Actes de l'atelier " Diversité Linguistique et TAL " (DiLiTAL 2017)

Fadoua Atta-Allah, Fatima Agnaou, Khalid Ansar, Aicha Bouhjar, Siham Boulaknadel, Malika Chakiri, Hammou Fadili, Jamal Frain, Jovan Kostov, Alice Millour, et al.

▶ **To cite this version:**

Fadoua Atta-Allah, Fatima Agnaou, Khalid Ansar, Aicha Bouhjar, Siham Boulaknadel, et al.. Actes de l'atelier " Diversité Linguistique et TAL " (DiLiTAL 2017). 2017. halshs-01541153

HAL Id: halshs-01541153
https://shs.hal.science/halshs-01541153

Submitted on 17 Jun 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution - NonCommercial - ShareAlike 4.0 International License

24e Conférence sur le Traitement Automatique des Langues Naturelles (TALN)

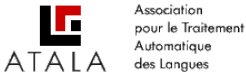
Orléans, France – 26-30 juin 2017

https://taln2017.cnrs.fr

Actes de l'atelier « Diversité Linguistique et TAL » (DiLiTAL 2017)

Fadoua Atta-Allah, Fatima Agnaou, Khalid Ansar, Aicha Bouhjar, Siham Boulaknadel, Malika Chakiri, Hammou Fadili, Jamal Frain, Jovan Kostov, Alice Millour, Satenik Mkhitaryan, Michael Zock (Eds.)

Sponsors :


ATALA
Association
pour le Traitement
Automatique
des Langues

Ortolang

cnrs
dépasser les frontières

Liberté · Égalité · Fraternité
RÉPUBLIQUE FRANÇAISE
MINISTÈRE
DE LA CULTURE

acatus
informatique

aktan

proxem

inbenta

Orléans
Mairie

ORLÉANS
MÉTROPOLE

LOIRE&ORLEANS
ECO

FRENCH TECH
LOIRE
VALLEY

UNIVERSITE D'ORLEANS

Laboratoire
Ligérien de
Linguistique

LABORATOIRE
D'INFORMATIQUE
LIFO
FONDAMENTALE
D'ORLEANS

LI
Laboratoire d'Informatique
EA 2101

Préface

Savoir communiquer est un des fondements de nos sociétés. Nos survies et notre coexistence sont intimement liées à notre faculté de nous faire comprendre, ce qui peut poser problème à cause de nos différences culturelles et linguistiques et à cause de nos expériences de vie, sensibilités, points de vue etc. Dans de très nombreux pays, on pratique plusieurs langues mais dont le statut est, généralement, différent (langue officielle, langue régionale). Cet état de fait engendre une grande disparité au niveau des outils et des ressources en traitement automatique des langues (TAL). Dans ce domaine les langues 'mineures/minorées' sont communément connues sous le nom de « langues peu dotées » (LPD). On note également que les décideurs s'y intéressent généralement peu, ce qui est un handicap incontestable à l'heure de la globalisation.

Conscients de cette situation, cet atelier vise à susciter une réflexion débouchant sur un élargissement des travaux en TAL notamment la prise en compte d'autres langues que celles habituellement traitées. Plus précisément, l'attention sera portée sur la nature des outils et ressources qu'il y a lieu de concevoir pour les LPD, comment les créer, comment former des spécialistes capables de les élaborer, voire comment assurer l'apprentissage de ces langues, afin de les sauvegarder et les pérenniser comme le préconisent les différentes initiatives mises en place par les grandes organisations internationales (Nations-Unies, Conseil de l'Europe, etc.), pour combler cette lacune.

La nécessité de traiter automatiquement les LPD découle, à la fois, des besoins scientifiques et humanitaires (santé, éducation, culture, littératie, etc.) mais également des enjeux d'ordre politique (accès à l'information et à l'enseignement). A cette fin, à travers la journée du 26 juin 2017, l'atelier DiLiTAL se donne pour objectif non seulement la sensibilisation de la communauté scientifique à ces enjeux et aux difficultés rencontrées dans le traitement des LPD, mais aussi à l'intérêt de créer des outils génériques permettant de traiter un très grand nombre de langues, tout en identifiant les besoins spécifiques en fonction des particularités des différentes langues (typologie). De ce fait, cette initiative vise à mettre en commun les différentes méthodes et techniques utilisées pour dynamiser la construction et la mutualisation des ressources, ainsi que le transfert des savoirs et des savoir-faire.

Comités

Comité de programme

Meftaha Ameur (IRCAM, Maroc)
Delphine Bernhard (LILPA, Université de Strasbourg)
Laurent Besacier (LIG, Université de Grenoble-Alpes)
Siham Boulaknadel (IRCAM, Maroc)
Ahmed Boukouss (IRCAM, Maroc)
Violetta Cavalli-Sforza (UAI, Maroc)
Malika Chakiri (Université Paris-Descartes)
Antoine Chalvin (CREE, INALCO)
Khalid Choukri (ELDA)
Rute Costa (UNL, Portugal)
Anaïd Donabedian (SeDyL, INALCO)
Hammou Fadili (CNAM)
Michel Francard (UCL, Belgique)
Kim Gerdes (LPP, Sorbonne Nouvelle)
Thibaut Grouas (DGLFLF)
Benaissa Ichou (IRCAM, Maroc)
Anne-Laure Ligozat (LIMSI, CNRS)
Mathieu Mangeot (GETALP, Université de Chambéry)
Joseph Mariani (LIMSI, CNRS)
Denis Maurel (LI, Université de Tours)
Azzedine Mazroui (FS-Oujda, Maroc)
Issouf Modi (MEN, Niger)
Abdelhak Mouradi (MESRSFC, Maroc)
Kamal Nait Zerrad (LACNAD, INALCO)
Patrice Pognan (PLIDAM, INALCO)
Sophie Rosset (LIMSI, CNRS)
Max Silberztein (ELLIADD, Université Franche-Comté)
Hamid Souifi (IRCAM, Maroc)
Dejan Stosič (ERSS, Université de Toulouse Jean-Jaurès)

Izabella Thomas (CRIT, Université de Franche-Comté)
Nora Tigziri (UMMTO, Algérie)
Mathieu Valette (ERTIM, INALCO)
Farida Yamouni (UMMTO, Algérie)
Michael Zock (LIF, CNRS & AMU)

Comité d'organisation

Fadoua Ataa Allah (IRCAM, Maroc)
Fatima Agnaou (IRCAM, Maroc)
Khalid Ansar (IRCAM, Maroc)
Aicha Bouhjar (IRCAM, Maroc)
Siham Boulaknadel (IRCAM, Maroc)
Malika Chakiri (Université Paris-Descartes)
Hammou Fadili (CNAM)
Jamal Frain (IRCAM, Maroc)
Jovan Kostov (PLIDAM, INALCO)
Alice Millour (STIH, Paris-Sorbonne)
Satenik Mkhitaryan (ERTIM, INALCO)
Michael Zock (LIF, CNRS & Aix-Marseille Université)

Table des matières

Session « oral »

DiLiTAL – Diversité Linguistique et TAL
Fadoua Atta-Allah, Fatima Agnaou, Khalid Ansar, Aicha Bouhjar, Siham Boulaknadel, Malika Chakiri, Hammou Fadili, Jamal Frain, Jovan Kostov, Alice Millour, Satenik Mkhitaryan, Michael Zock 1

L'amazighe dans les sciences du numérique : expérience de l'IRCAM
Fatima Agnaou, Khalid Ansar, Fadoua Ataa Allah, Aicha Bouhjar, Siham Boulaknadel 4

Problèmes de tokénisation pour deux langues régionales de France, l'alsacien et le picard
Delphine Bernhard, Amalia Todirascu, Fanny Martin, Pascale Erhart, Lucie Steiblé, Dominique Huck, Christophe Rey 14

Produire des ressources électroniques à partir de descriptions formelles : application aux langues peu dotées
Denys Duchier, Yannick Parmentier, Simon Petitjean, Emmanuel Schang 24

Retour d'expérience : l'utilisation de l'apprentissage profond (deep learning) dans le contexte de l'analyse sémantique des langues peu dotées
Hammou Fadili 33

Expériences d'étiquetage morphosyntaxique dans le cadre du projet RESTAURE
Pierre Magistry, Anne-Laure Ligozat, Sophie Rosset 45

Les chaînes coréférentielles en créole de la Guadeloupe
Emmanuel Schang, Jean-Yves Antoine, Anaïs Lefeuvre-Halftermeyer 54

Constitution d'un corpus d'arabe tunisien parlé à Orléans
Youssra Ben Ahmed 62

DiLiTAL – Diversité linguistique et TAL

Fadoua Atta-Allah[1], Fatima Agnaou[2], Khalid Ansar[3], Aicha Bouhjar[2], Siham Boulaknadel[1]
Malika Chakiri[4], Hammou Fadili[5], Jamal Frain[1], Jovan Kostov[6, 7], Alice Millour[8],
Satenik Mkhitaryan[9], Michael Zock[10]

(1) CEISIC, IRCAM, Rabat, Maroc
(2) CRDPP, IRCAM, Rabat, Maroc
(3) CAL, IRCAM, Rabat, Maroc
(4) Université Paris Descartes, 12, rue de l'Ecole de Médecine, 75006, Paris
(5) Laboratoire CEDRIC - CNAM, 2, rue Conté, 75003, Paris
(6) E.A. 4154 PLIDAM - INALCO, 2, rue de Lille, 75007, Paris
(7) PERL - USPC, 8, place Paul Ricœur, 75013, Paris
(8) E.A. 4509 STIH – Université Paris Sorbonne, 1, rue Victor Cousin, 75005, Paris
(9) E.A. 2520 ERTIM – INALCO, 2, rue de Lille, 75007, Paris
(10) LIF – UMR 7279, 163, avenue de Luminy, F-13288, Marseille

{agnaou,ansar,ataaallah,bouhjar,boulaknadel,frain}@ircam.ma,
chakirimalika@yahoo.fr, jovan.kostov@gmail.com, alice.millour@paris-
sorbonne.fr, satenik.mkhitaryan@inalco.fr, mikael.zock@lif.univ-mrs.fr

RESUME

L'atelier DiLiTAL se donne pour objectif de sensibiliser la communauté scientifique non seulemen aux enjeux et aux difficultés rencontrées dans le traitement des langues peu dotées (LPD), mai aussi à l'intérêt de créer des outils génériques permettant de traiter un très grand nombre de langues tout en identifiant les besoins spécifiques en fonction de leurs particularités. Cet atelier vise à mettr en commun les différentes méthodes et techniques utilisées pour dynamiser la construction et l mutualisation des ressources, ainsi que le transfert des savoirs et des savoir-faire.

ABSTRACT

Workshop Linguistic diversity and NLP - DiLiTAL
The goal of the DiLiTAL workshop is to raise awareness among the scientific community of th issues and difficulties encountered in the processing of under-resourced languages (ULL). We also want to point out the necessity of creating generic tools in order to process a large number o languages by identifying their particularities. This workshop's aim is to pool the different method and techniques used to boost the construction of resources, as well as the transfer of knowledge an experience in the processing of under-resourced languages.

MOTS-CLES : langues peu dotées, diversité, linguistique, TAL.
KEYWORDS: less resourced languages, diversity, linguistics, NLP.

1 Présentation générale de l'atelier

Savoir communiquer est l'un des fondements de nos sociétés. Nos survies et notre coexistence son intimement liées à notre faculté de nous faire comprendre, ce qui peut poser problème à cause d nos différences culturelles et linguistiques et de nos expériences de vie, sensibilités, points de vu etc. Dans de très nombreux pays, on pratique plusieurs langues dont le statut est différent : langu officielle, langue régionale etc. Cet état engendre une grande disparité au niveau des outils et de ressources en traitement automatique des langues (TAL). Dans ce domaine, les langue « mineures/minorées » sont communément connues sous le nom de « langues peu dotées » (LPD) On note également que les décideurs s'y intéressent généralement peu, ce qui est un handica incontestable à l'heure de la globalisation.

Cet atelier vise à susciter une réflexion débouchant sur un élargissement des travaux en TAL pou prendre en compte d'autres langues que celles habituellement traitées. L'attention de l'atelie DiLiTAL est portée sur la nature des outils et des ressources et sur la manière de les adapter au LPD, mais aussi sur la question de formation de spécialistes capables de les élaborer, de le pérenniser et de les réutiliser dans une visée d'enseignement et de diffusion de ces langues, selon le initiatives mises en place par les grandes organisations internationales (Nations-Unies, Conseil d l'Europe, etc.).

La nécessité de traiter automatiquement les LPD découle des besoins à la fois scientifiques e humanitaires (santé, éducation, culture, littératie, etc.), mais également des enjeux d'ordre politiqu (accès à l'information et à l'enseignement). L'atelier DiLiTAL se donne pour objectif d sensibiliser la communauté scientifique à ces enjeux et aux difficultés rencontrées dans le traitemen des LPD, mais aussi à l'intérêt de créer des outils génériques permettant de traiter un très gran nombre de langues, tout en identifiant les besoins spécifiques en fonction des particularités de différentes langues (typologie).

2 Axes thématiques de l'atelier DiLiTAL

L'atelier DiLiTAL a réuni des interventions qui présentent des travaux théoriques et/ou de applications concrètes qui s'articulent autour des thématiques suivantes :

2.1 Ressources et corpus : production, standardisation et archivage

L'ambition de cet axe est d'explorer les initiatives actuelles et futures qui ont pour but de collecter e de structurer des données langagières (spécialisées ou généralistes) des LPD. Ces données (lexiques corpus etc.) sont souvent utilisées pour l'entrainement des étiqueteurs morphosyntaxiques qu représentent, à leur tour, une étape préalable à des tâchas plus complexes comme l'analys syntaxique ou la traduction. Nous nous interrogeons également sur l'accessibilité » et la portabilit des données linguistiques, qui s'avèrent être des problèmes majeurs dans les travaux consacrés au LPD et c'est pour cette raisons que les contributions concernent principalement la création de ressources *open source*.

2.2 Outils pour le traitement des LPD

Les interrogations de ce second axe portent sur la pertinence de l'utilisation d'outils existants pou les LPD et sur la manière dont ils gèrent le multilinguisme. En effet, depuis l'arrivée d'UTF-8 (Unicode), le TAL s'est doté de la possibilité de diversifier son terrain d'action-recherche permettant de traiter d'autres langues que celles dites « majoritaires » (et bien dotées), comme l'anglais, l'espagnol, le français, le chinois etc. Dans une telle perspective, il nous a semble nécessaire de réfléchir à une amélioration de la gestion du multilinguisme par l'inclusion de nouvelles graphies et de nouveaux standards, tout en identifiant les contraintes méthodologiques auxquelles se heurtent les chercheurs dans le cadre d'un travail de recherche sur une LPD avec des outils existants.

2.3 Questions sociales, culturelles et éthiques

DiLiTAL a été également une occasion d'examiner les possibilités de coopération transfrontalière pour encourager l'échange d'expériences et pour dynamiser la recherche en TAL dans les pays où les LPD sont en usage. Il nous semble important d'entamer une réflexion concernant des aspects éthiques de la collecte, de l'archivage et du traitement des données linguistiques qui, eux, se trouvent ancrés dans des contextes sociétaux où les codes et les mœurs sont méconnus ou sensiblement différents de ceux du chercheur. Le but principal de cet axe est de réfléchir à une méthodologie de terrain qui ne vise pas uniquement à ménager la science, mais qui prend aussi en compte le contexte global, à savoir, le sujet parlant qui fait partie d'une culture, voire d'une catégorie sociale. Cet atelier a été l'occasion de discuter du TAL en termes de moteur des changements linguistiques. En effet, ce problème se pose dans le cadre des commissions terminologiques qui jouent un rôle prépondérant dans le processus de codification des langues dans différents pays du monde.

2.4 Retours d'expérience

Certaines contributions ont mis en lumière des expériences de traitement d'une LPD ou d'un groupe de LPD: la synthèse des expertises acquises par des chercheurs ayant participé à ce genre de travaux permet de nourrir une réflexion épistémologique nécessaire au développement des outils pour dépasser le cadre d'un seul type de langues. Ces expériences s'avèrent utiles pour identifier les domaines auxquels la recherche en TAL pourra contribuer. Quelle que soit la langue, un retour d'expérience est toujours précieux pour permettre d'identifier les convergences et les divergences dans les approches, constituant, de ce fait, un apport considérable pour cet atelier et pour la recherche-action qui en découlera.

L'atelier comprend sept interventions reflétant les recherches menées sur les LPD dans un contexte francophone et amazigh. Une conférence invitée (donnée par Joseph Mariani, LIMSI - CNRS) et une table ronde ont été également organisées dans le but de rendre compte des recherches sur les LPD dans le contexte scientifique actuel.

L'amazighe dans les sciences du numérique : expérience de l'IRCAM

Fatima Agnaou[1], Khalid Ansar[2], Fadoua Ataa Allah[3]
Aicha Bouhjar[4], Siham Boulaknadel[3]
(1) CRDPP, IRCAM, Rabat, Maroc
(2) CAL, IRCAM, Rabat, Maroc
(3) CEISIC, IRCAM, Rabat, Maroc
(4) CRDPP, IRCAM, Rabat, Maroc
{agnaou,ansar,ataaallah,bouhjar,boulaknadel}@ircam.ma

RESUME

Selon les études menées par l'UNESCO, les langues du monde courent, de plus en plus, le risque d'extinction. Ce qui amenuise la diversité culturelle qui fait la richesse de l'humanité. Ainsi, elle a entrepris, dans ses programmes de sauvegarde des langues en danger, des actions de préservation et de valorisation de ces langues en mettant en exergue les initiatives qui puissent les doter des technologies de support.

Avec la même vision, l'Institut Royal de la Culture Amazighe a mis en place une stratégie pour la promotion et le développement de l'amazighe. Cette stratégie s'est traduite à travers une démarche progressive en matière d'aménagement linguistique, d'implantation dans le système éducatif et de traitement de l'amazighe par le biais des sciences du numérique. Néanmoins, la voie de la concrétisation de cette stratégie est pleine de défis et d'entraves : elle nécessite, en effet, la coordination des acteurs concernés par le développement et la transmission de l'amazighe dans les domaines de la politique linguistique, la recherche scientifique et la communauté du logiciel libre.

ABSTRACT

Amazigh in the digital sciences: experience of the IRCAM

On the basis of the studies undertaken by UNESCO, it has been observed that world languages are, increasingly, exposed to the risk of extinction. This risk has influentially contributed in reducing cultural diversity which may well be viewed as one of the most prominent aspects of richness for humanity. To contend with this situation and save endangered languages, UNESCO has entertained a whole range of measures with an eye to preserving and valuing less endowed languages by carrying out a number of initiatives meant to endow these languages with the necessary technologies.

With the same line of thinking as background, the Royal Institute of Amazigh Culture has piloted and set up a strategy for the promotion and development of Amazigh. This strategy has been translated into action through steps involving the progressive planning of Amazigh, its implantation in the educational system and its introduction in the digital world. Nonetheless, an evaluation of the experience carried out by IRCAM, thus far, evinces that this strategy is beset with a whole range of obstacles and constraints, and that better results will ensue if a coordination holds between the actors interested in the development and transmission of Amazigh in the domains of linguistic policy, scientific research and language digitalization.

MOTS-CLES : amazighe, science du numérique, promotion, langue peu dotée.
KEYWORDS: Amazigh language, digital science, promotion, less resourced language.

1 Introduction

Dans un contexte global où la transformation digitale est l'un des défis majeurs dans tous les domaines (économique, social, culturel, …), toutes les langues se doivent de pouvoir répondre aux attentes de leurs locuteurs pour pouvoir être utilisées dans les sciences du numérique. C'est dans cet environnement global que les langues peu dotées, dont la langue amazighe, sont amenées à se développer. Se pose dès lors la question de savoir comment faire de l'amazighe une langue *suffisamment dotée* pour répondre aux défis de son implantation dans la sphère publique et ce d'autant qu'elle est à présent officielle, au Maroc, aux côtés de l'arabe depuis le 1er juillet 2011, soit une décennie après la création de l'Institut Royal de la Culture Amazighe (IRCAM).

En effet, le statut de la langue amazighe a changé depuis le Discours Royal d'Ajdir du 17 octobre 2001 et le Dahir (Décret Royal) qui porte création de l'IRCAM. Jusqu'alors absente de la scène publique, essentiellement orale, limitée à l'usage informel et présente sous sa forme dialectale, la langue amazighe fait désormais partie du « patrimoine culturel commun » et devient une « affaire nationale » puisqu'elle concerne tous les citoyens, soit plus de trente millions de personnes. L'identité culturelle du citoyen marocain est ainsi définie dans la pluralité de ses affluents et constitue un patrimoine à promouvoir, à valoriser et à préserver. Ce processus de patrimonialisation et plus récemment d'officialisation de la langue et de la culture amazighes induit des actions spécifiquement liées à l'aménagement du statut et du corpus de la langue. Conformément aux missions assignées à l'IRCAM, des travaux sur l'aménagement du corpus (graphie, orthographe, lexique et grammaire) ont été initiés afin de permettre à la langue de remplir les nouvelles fonctions qui lui sont à présent dévolues dans le domaine public ; prioritairement dans l'enseignement et les médias ainsi que dans les sciences du numérique.

Cette contribution expose trois aspects essentiels de la problématique liée à la stratégie de l'IRCAM en matière de développement de l'amazighe, à savoir : son aménagement linguistique, son implantation dans le système éducatif et son traitement par le biais des sciences du numérique.

2 Stratégie du développement de l'amazighe

D'une manière générale, la stratégie adoptée par l'IRCAM, en matière d'aménagement linguistique, inspirée de l'expérience corse (Marcellesi, 2003), consiste en une gestion matérielle de la variation géolectale, ce qui engendre un amazighe enrichi de l'apport des différents dialectes régionaux du nord, du centre et du sud du Maroc. Concrètement, il s'agit de neutraliser ou de contenir la variation dialectale lorsque c'est possible et préférable pour une communication élargie (sur le plan phonique essentiellement). Par contre, la variation est intégrée lorsqu'elle est source d'enrichissement de la langue : la variation lexicale ou morphosyntaxique, actualisée dans la langue par des termes ou structures concurrent(e)s, est notamment prise en charge par le biais de la synonymie. Les lacunes terminologiques font quant à elles l'objet d'une création néologique commune à l'ensemble des variantes nationales ; la néologie reste cependant l'ultime recours après que toutes les ressources de la langue aient été épuisées. Autrement dit, la stratégie polynomique permet de préserver aux différentes variétés leur vitalité et, par là même, d'obtenir l'adhésion des locuteurs et de garantir la

cohésion sociale. Mais cette approche n'exclut pas une visée standardisante sur le long terme à dessein d'atténuer les variations de surface et d'outiller la langue en enrichissant son lexique. Ainsi, au fonds commun sous-jacent aux différentes variétés, vient s'ajouter un pan terminologique que les différents géolectes auront en partage. Une norme standard nationale s'impose donc pour des raisons d'intercompréhension évidentes, mais également par souci didactico-pédagogique comme c'est le cas pour toutes les langues (Ameur et Boumalk, 2004 ; Boukous, 2007 ; Agnaou, 2009).

Par ailleurs, dans une perspective d'intégration de la langue amazighe dans les sciences du numérique, l'IRCAM a développé une feuille de route (Ataa Allah et Boulaknadel, 2012 ; Ataa Allah et Boulaknadel, 2014B), structurée selon le court, le moyen et le long terme, pour le développement de ressources et outils assurant son fonctionnement à l'instar des langues conformément équipées. Elle est représentée par une chaîne partant des traitements élémentaires, passant par la constitution de briques de ressources linguistiques, et allant vers des applications génériques. Cette chaîne va de l'adaptation et l'amélioration d'outils en fonction des nouvelles technologies jusqu'au développement d'applications, en respectant les stratégies de développement inspirées des travaux de Muhirwe (Muhirwe, 2007), à savoir la standardisation, la notion d'extensibilité et de documentation ainsi que l'adoption des technologies libres.

Plus concrètement, la première action, à court terme, à laquelle l'IRCAM s'est attelé afin de permettre le passage à l'écrit de la langue amazighe, a trait au choix au système d'écriture (le tifinaghe) et des graphèmes qui devaient constituer l'alphabet.

2.1 Graphie et encodage

Fixer une norme graphique de l'amazighe passe nécessairement par le choix d'un alphabet qui doit répondre à un double objectif :

- le maintien d'un lien solidaire avec les différentes variantes de l'alphabet tifinaghe actuel, d'où la nécessité de puiser dans le fonds des graphèmes disponibles dans les différentes variantes et de considérer la création de nouveaux symboles comme un dernier recours ;
- l'adaptation du nouvel alphabet aux structures de l'amazighe standard, requérant parfois l'introduction de quelques modifications.

Pour répondre à cet objectif, il est tenu compte de quatre principes : l'historicité, la simplicité, l'univocité du signe et l'économie. Dans une seconde étape, d'autres paramètres ont été pris en compte dans le choix des caractères. Il s'agit de la fréquence des graphèmes dans les différentes variantes du libyque - tifinaghe, de leur simplicité au niveau de l'écriture manuelle (facilité psycho-motrice), de l'esthétique des symboles et de la cohérence d'ensemble du système d'écriture proposé (Bouhjar, 2004). Après une étude approfondie du système phonético-phonologique des différentes variantes de l'amazighe présentes au Maroc, la sélection de 33 graphèmes a finalement été opérée. L'adoption officielle de la graphie tifinaghe en tant que système d'écriture pour la langue amazighe au Maroc a été annoncée le 10 février 2003.

Parallèlement, le tifinaghe a franchi une étape technologique initiée par le codage, un processus qui a connu deux phases contrastées :

- Une première phase transitoire a consisté en une adaptation de la norme ISO 8859-1 pour coder les caractères tifinaghes en codage ANSI afin de répondre à l'urgence de l'introduction de l'amazighe dans le système éducatif marocain en 2003. Cependant, la portée de ce codage privé est limitée et la gestion des textes comportant plusieurs systèmes

d'écriture est difficile. D'ailleurs, le traitement des textes multilingues devait jongler à la fois avec les différentes normes de codage et avec les polices associées.

- La deuxième phase a concerné l'intégration de l'écriture du tifinaghe dans le plan multilingue de base. Pour ce faire, une proposition de norme pour le codage informatique de l'alphabet tifinaghe a été soumise au consortium Unicode. Ce dernier a réservé aux caractères tifinaghes quatre sous-ensembles dont l'espace hexadécimal est de 2D30 à 2D7F. Le fait que les caractères tifinaghes sont codés de manière consécutive a permis de simplifier des traitements automatiques potentiels, en particulier la reconnaissance automatique de la langue amazighe. Le premier sous-ensemble représente les 33 lettres alphabétiques de base préconisées par l'IRCAM. Les labiovélaires (ⴳⵯ, ⴽⵯ) ont été aménagées de sorte que la diacrité soit indépendante et considérée comme un caractère autonome. Le deuxième sous-ensemble contient les 8 caractères de la liste étendue, qui a été définie par l'IRCAM pour l'intérêt historique, scientifique et stylistique. En effet, sur le plan historique, l'intégration de certaines lettres ancestrales répond au souci de la préservation du patrimoine culturel. Sur le plan scientifique, leur utilisation permettra des études comparatives inter-dialectales dans la mesure où il est apparu nécessaire de conserver certains graphèmes afin de rendre les spécificités régionales, à des fins stylistiques, dans la production écrite, notamment dans le domaine de la poésie connu pour être réfractaire à tout processus de normalisation. Le troisième sous-ensemble est formé de 4 lettres néo-tifinaghes utilisées fréquemment dans le reste du Maghreb. Et le quatrième sous-ensemble contient 11 lettres touarègues modernes dont l'usage est attesté (Andries, 2008).

Face à ces deux systèmes différents d'encodage, il a été nécessaire de procéder à la réalisation de convertisseurs (Ataa Allah *et al.*, 2013) pour l'exploitation du fond documentaire amazighe en codage ANSI et la transition vers l'Unicode.

Le travail effectué sur la standardisation de la graphie et son encodage a permis de produire le support numérique ⴰⴷ ⵏⵍⵎⴷ ⵜⵉⴼⵉⵏⴰⵖ [Ad nlmd tifinaghe] « Apprenons les tifinaghes » (Agnaou et Afoulki, 2006) en vue de contribuer au développement de la compétence de la lecture chez les apprenants. Les objectifs opérationnels de ce support sont : (i) familiariser l'apprenant avec les signes graphiques traduisant les phonèmes de la langue amazighe, (ii) soutenir l'expression orale, (iii) préparer et faciliter le passage à l'écrit et (iv) développer le goût de la lecture dans cette langue nouvellement introduite dans le système d'éducation et de formation. Outre ces objectifs d'ordre technique et didactique, ces supports ont une finalité éducative et citoyenne, notamment la découverte et le partage du patrimoine culturel commun, le vivre ensemble et l'acceptation de la différence. En parallèle, ils visent la formation de lecteurs et de lectrices efficients capables de continuer leur propre éducation et de pérenniser l'usage de la langue amazighe et de la protéger contre sa déperdition.

Sur le plan didactique, ce support a été élaboré de sorte à ce qu'il développe les composantes fondamentales de la lecture (Smith, 2004), notamment la conscience phonémique, le principe alphabétique, la fluidité, le vocabulaire et la compréhension à partir d'un corpus adapté au niveau et aux besoins des apprenants. D'autres ressources ont été développées dans le domaine lexical.

2.2 Ressources lexicales

Le lexique est un domaine assez pourvu mais les premières recherches couvrent généralement des parlers locaux, rarement toute une aire géolectale donnée. L'insertion de la langue amazighe dans la vie publique au Maroc a eu pour conséquence l'extension de ses domaines d'usage ce qui a suscité des besoins sur le plan terminologique afin de pouvoir dénommer des réalités nouvelles. La

démarche globale de l'IRCAM s'appuie sur un certain nombre de principes et critères hiérarchisés et suit les principales étapes de tout travail terminologique. Ainsi, trois principes fondamentaux ont dicté la démarche à adopter lors du travail terminologie : la pertinence, la motivation (transparence des propositions dans la mesure où seul un haut degré de lisibilité garantit un haut degré d'appropriation des termes) et le respect de la morphologie et de la syntaxe de l'amazighe. Ces principes et critères ont permis de produire les lexiques sectoriels suivants : le lexique grammatical (Ameur *et al.*, 2009 ; Ameur *et al.*, 2011), le lexique des médias (Ameur *et al.*, 2009 ; Ameur *et al.*, 2013) et le lexique administratif (Ameur *et al.*, 2015).

Parallèlement, l'IRCAM a entrepris un projet de construction de bases de données lexicales et terminologiques numériques. La base de données terminologiques contient des entrées terminologiques relevant de plusieurs thématiques (El Azrak et El Hamdaoui, 2011). Cette base de données a été exploitée dans la réalisation de l'application mobile « LEXAM » (Frain *et al.*, 2014). Par ailleurs, une base de données lexicales, qui regroupe le lexique usuel appartenant aux différentes variantes de l'amazighe, a été élaborée. Cette dernière fait l'objet d'une exploitation Web et libre d'accès afin de répondre aux besoins des acteurs travaillant dans les domaines de la linguistique, l'enseignement, la traduction et la communication.

En vue de développer la compétence lexicale chez les apprenants de l'amazighe, l'IRCAM a réalisé cinq ressources numériques, en l'occurrence ⵜⴰⵎⴰⵡⴰⵜ ⵏ ⵜⴰⵍⴰⴼⴰⵏⵜ [Tamawalt inu tawlafant] « Mon vocabulaire illustré » (Agnaou et Karoum, 2009), ⵜⴰⵎⴰⵡⴰⵜ ⵏ ⵉⵎⵣⵣⵢⴰⵏ [Tamawalt n imzzyann] « Dictionnaire pour enfants » (Ataa Allah, 2011), ⵜⵉⵏⵎⵍ ⵏ ⵜⴰⵎⴰⵣⵉⵖⵜ [Tinml n tmazight] « Ecole amazighe » (Zenkouar *et al.*, 2007), ⵉⵣⵡⵉⵍ ⵙ ⵜⴰⵎⴰⵣⵉⵖⵜ [Izwiln s tmazight][1] « Chiffres en amazighe » et ⴰⵡⴰⵍ ⵏⵓ ⴰⵎⵣⵡⴰⵔⵓ [Awal inu amzwaru][2] « Mes premiers mots ». Ces ressources ont été conçues dans le but d'outiller l'apprenant de l'amazighe du vocabulaire de base en langue qui soit en harmonie avec le processus de l'habilitation de la langue et avec les thèmes fixés dans les programmes scolaires.

D'autres contenus numériques ont été conçus pour le développement de la compétence communicative. Il s'agit de deux ressources, à savoir Imudar iramyarn d inamyarn « Animaux sauvages et domestiques » ⵉⵖⴷⴰⴷ ⵙ ⵉⴱⵓⵅⴰ [IgDaD d ibukha] « Oiseaux et insectes » (Zenkouar *et al*., 2008). Elles visent, en plus du développement du vocabulaire de la faune, la maîtrise de la compréhension et la production orale chez les apprenants.

2.3 Ressources grammaticales

On peut penser que les faits morphosyntaxiques soient moins sujets à la variation dans la mesure où il s'agit du niveau le plus stable. Or, une observation approfondie de certains faits morphosyntaxiques permet de relever que ce n'est pas le cas. En effet, certains phénomènes tels que l'état construit, la morphologie verbale et les déterminants, entre autres, résistent au processus d'unification. Disposant de peu de recherches et d'ouvrages de référence élaborés dans une perspective comparative et normative, l'IRCAM, conscient de telles contraintes, a concentré ses efforts sur la publication d'ouvrages de référence à visée standardisante sur le plan morphosyntaxique à partir d'une approche comparative inter-dialectale (Boukhris *et al*. ,2008 ; Laabdelaoui *et al*. ,2012). A partir de ces premiers ouvrages, il s'agissait également de doter la

[1]https://play.google.com/store/apps/details?id=ma.ircam.chiffres&hl=fr
[2] https://play.google.com/store/apps/details?id=com.ircam.vocabetlettres&hl=fr

langue amazighe des premiers outils de base de grammaire en TAL. Le premier outil élaboré concerne l'étiqueteur morphosyntaxique qui a pour objectif la catégorisation grammaticale des unités lexicales d'un corpus de textes. Dans cette optique, il y a eu lieu d'élaborer un premier corpus numérique de référence à visée exhaustive (Boulaknadel et Ataa Allah, 2012 ; Boulaknadel et Ataa Allah, 2013). Ce corpus est constitué de textes bruts, représentatif des variantes de l'amazighe marocain et des différents genres (conte, poésie et articles de presse). Bien que le projet a progressé, il faut néanmoins noter qu'il a fallu surmonter un certain nombre de difficultés liées essentiellement à l'homogénéisation de la graphie et de l'orthographe du corpus. Par ailleurs, un jeu d'étiquettes, inspiré du modèle EAGLES (EAG, 1996), a été proposé (Ataa Allah *et al.*, 2014) afin d'assurer l'exploitation dans différentes applications du TAL, notamment dans le contexte multilingue. Le second outil concerne le conjugueur qui prend appui sur le *Manuel de conjugaison* (Laabdelaoui *et al.*, 2012). Il est conçu selon une méthode à base de règles fondée sur une extension, par les arbres discriminatoires, de l'approche à deux niveaux (Koskenniemi, 1984). Il permet l'accès à distance à la conjugaison des verbes de la langue amazighe (Ataa Allah et Boulaknadel, 2014A).

D'autres outils sont présentement à l'étude mais, vu la pluralité des plateformes existantes et l'absence d'études en traitement automatique de l'amazighe, ils requièrent en amont des recherches approfondies menées en parallèle. Ces outils concernent l'analyse morphologique, les entités nommées et la traduction automatique.

Les travaux de recherche sur la morphologie de la langue amazighe ont été initiés par deux études (Nejme *et al.*, 2016 ; Ataa Allah, 2014), reposant respectivement sur les systèmes Nooj (Silberztein, 2007) et Xerox (Beesley et Karttunenauri, 2003). Ces études se basent sur le principe des transducteurs bidirectionnels utilisés à la fois pour l'analyse et la génération. Du point de vue de la génération, le plus haut niveau représente une description abstraite des règles morphotactiques qui définissent les conditions d'ordre et de combinaison entre les différents morphèmes. Ce niveau est associé à un niveau plus concret, dans lequel les règles morpho-phonologiques s'appliquent et lient chaque lexie à ses formes de surface. Les règles modélisées par des transducteurs séparés à chaque niveau sont combinées avec un lexique qui contient les formes radicales de l'ensemble des lexies. Concernant l'analyse, le transducteur traite les formes entrantes pour reconnaître les segmentations éventuelles en identifiant le radical et les différents affixes qui lui sont collés.

Parallèlement à ces travaux, des études d'analyses syntactico-sémantiques ont été menées. Elles consistent à réaliser un extracteur automatique des entités nommées amazighes (Talha *et al.*, 2015 ; Boulaknadel *et al.*, 2014) en exploitant le principe des transducteurs à états finis pour définir les contextes d'apparition des unités à extraire.

En outre, dans la perspective de réaliser un système de traduction automatique de la langue amazighe, deux études ont été entamées. La première porte sur l'alignement de corpus parallèles pour l'amazighe (Miftah *et al*., 2017). Tandis que la deuxième (Taghbalout *et al.*, 2017) concerne l'intégration de la langue amazighe dans le projet UNL (Universal Networking Language) (Boguslavsky *et al.*, 2005). Ce projet tend à favoriser l'éclosion du multilinguisme dans la société d'information et de permettre à toute personne d'accéder à l'information sur Internet par le biais de sa langue native sans limitations des barrières linguistiques. Il permet de coder une représentation du sens d'un texte pour servir de pivot inter-langue dans les systèmes UNL de la traduction automatique.

3 Points forts et limites

Dans le cadre de sa stratégie relative à la promotion de la langue amazighe, l'IRCAM accorde une grande importance à l'aménagement de la langue et à son intégration dans les sciences du numérique. Ainsi, des efforts ont été fournis en vue de doter la langue amazighe de fondements essentiels et de ressources indispensables pour son développement. Désormais, l'amazighe est passée de la situation de langue vernaculaire orale avec des transcriptions différentes assez disparates à celle d'une langue dotée d'une écriture normée et encodée selon les standards du plan multilingue de base ISO-Unicode (norme 10646). Elle dispose à présent d'une orthographe stabilisée, d'une grammaire, de ressources et d'outils numériques.

En dépit de ces efforts pour développer une stratégie, qui favorise l'intégration de l'amazighe dans les sciences du numérique, celle-ci est confrontée à différentes contraintes.

Ainsi, la contrainte d'ordre linguistique est à mettre en relation avec le système d'écriture de l'amazighe qui pose les difficultés majeures suivantes :

- Variation dialectale : le processus de standardisation étant encore tout récent, les ressources disponibles nécessitent d'être redressées afin de répondre aux normes orthographiques préconisées par l'IRCAM. Cette étape reste incontournable pour pouvoir exploiter de manière optimale ces ressources dans les sciences du numérique.
- Pratiques scripturaires : l'usage du tifinaghe n'étant pas généralisé, le spécialiste du traitement automatique de l'amazighe se retrouve face à une multitude de formes d'écriture et de modes de transcription. Ce fait limite considérablement la portée de l'action vu l'ampleur des tâches préliminaires à effectuer avant toute exploitation numérique.
- Disponibilité de ressources : le nombre restreint de ressources enfreint fortement le développement numérique de l'amazighe pour des raisons évidentes de fiabilité des résultats. De même, sur le plan qualitatif, il faut relever que la nature des ressources disponibles ne couvre généralement que le domaine littéraire ; les médias émergeant tout doucement. Ce constat met en péril la représentativité des corpus exploités.
- Morphologie amazighe : certaines catégories grammaticales n'étant pas encore stabilisées dans la mesure où elles nécessitent une étude plus poussée, la réalisation d'outils de traitement s'en trouve entravée.
- Sur le plan éducatif, les limites relevées concernent principalement les points suivants :

    - Le niveau de compétences des enseignants exige une mise à niveau constante par la formation à l'utilisation technique des outils élaborés.
    - La non généralisation de l'équipement, de salles de cours, en ordinateurs et Internet.

4 Conclusion

En définitive, il est à noter que la langue amazighe a relevé le défi de son introduction dans le monde du numérique puisqu'à présent une panoplie d'outils existent et sont mis à la disposition des utilisateurs. Cependant, beaucoup reste encore à réaliser et à entreprendre compte tenu des contraintes soulevées pour suffisamment doter la langue amazighe. Dans cette perspective, des travaux de recherche sont en cours et d'autres sont envisagés.

Références

ANDRIES P. (2008). UNICODE 5.0 EN PRATIQUE, CODAGE DES CARACTERES ET INTERNATIONALISATION DES LOGICIELS ET DES DOCUMENTS. DUNOD, FRANCE.

AGNAOU F., AFOULKI, M. (2006). ⴰⴷ ⵏⵍⵎⴷ ⵜⵉⴼⵉⵏⴰⵖ (AD NLMD TIFINAGHE). PRODUCTIONS DE L'INSTITUT ROYAL DE LA CULTURE AMAZIGHE. RABAT.

AGNAOU F. (2009). « CURRICULA ET MANUELS SCOLAIRES : POUR QUEL AMENAGEMENT LINGUISTIQUE DE L'AMAZIGHE MAROCAIN ? », ASINAG N°3, RABAT, PUBLICATIONS DE L'IRCAM, PP. 109-126.

AGNAOU F., KARROUM, M. (2009). ⵜⴰⵎⴰⵡⴰⵍⵜ ⵉⵏⵓ ⵜⴰⵡⵍⴰⴼⴰⵏⵜ (TAMAWALT INU TAWLAFANT). PRODUCTIONS DE L'IRCAM, RABAT.

AMEUR M., BOUMALK A. (2004). STANDARDISATION DE L'AMAZIGHE : ACTES DU SEMINAIRE ORGANISE PAR LE CENTRE DE L'AMENAGEMENT LINGUISTIQUE, RABAT, 8-9 DECEMBRE 2003. PUBLICATIONS DE L'IRCAM.

AMEUR M., BOUMALK A., IAZZI E.M, SOUIFI H., ANSAR K. (2009). VOCABULAIRE DES MEDIAS, FRANÇAIS – AMAZIGHE – ANGLAIS - ARABE. PUBLICATIONS DE L'IRCAM. IMPRIMERIE EL MAARIF AL JADIDA. RABAT.

AMEUR M., BOUHJAR A., BOUMALK A., EL AZRAK N., LAABDELAOUI R. (2009). VOCABULAIRE GRAMMATICAL. PUBLICATIONS DE L'IRCAM. IMPRIMERIE EL MAARIF AL JADIDA. RABAT.

AMEUR M., BOUHJAR A., BOUMALK A., EL AZRAK N., LAABDELAOUI R. (2011). VOCABULAIRE GRAMMATICAL DE L'AMAZIGHE : APPLICATION PHRASEOLOGIQUE. PUBLICATIONS DE L'IRCAM. IMPRIMERIE EL MAARIF AL JADIDA. RABAT.

AMEUR M., ANSAR K., BOUHJAR A., EL AZRAK N. (2013). TERMINOLOGIE AMAZIGHE DE L'AUDIOVISUEL. PUBLICATIONS DE L'IRCAM. IMPRIMERIE EL MAARIF AL JADIDA. RABAT.

AMEUR M., ANSAR K., BOUHJAR A., EL AZRAK N. (2015). TERMINOLOGIE ADMINISTRATIVE. PUBLICATIONS DE L'IRCAM. IMPRIMERIE EL MAARIF AL JADIDA. RABAT.

ATAA ALLAH F. (2011). CONCEPTION D'UN DICTIONNAIRE IMAGIER SONORE EN LIGNE DE LA LANGUE AMAZIGHE. ACTES DE THE 8TH MULTIDISCIPLINARY SYMPOSIUM ON DESIGN AND EVALUATION OF DIGITAL CONTENT FOR EDUCATION (SPDECE 2011). CIUDAD REAL, ESPAGNE, 15-17 JUIN 2011. 158-164.

ATAA ALLAH F., BOULAKNADEL S. (2012). TOWARD COMPUTATIONAL PROCESSING OF LESS RESOURCED LANGUAGES: PRIMARILY EXPERIMENTS FOR MOROCCAN AMAZIGH LANGUAGE. THEORY AND APPLICATIONS FOR ADVANCED TEXT MINING. RIJEKA: INTECH. NOVEMBRE 2012.

ATAA ALLAH F. (2014). FINITE-STATE TRANSDUCER FOR AMAZIGH VERBAL MORPHOLOGY. LITERARY & LINGUISTIC COMPUTING. OXFORD UNIVERSITY PRESS, DOI:10.1093/LLC/FQU045.

ATAA ALLAH F., BOULAKNADEL S. (2014A). AMAZIGH VERB CONJUGATOR. ACTES DE THE 9TH INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION (LREC 2014). REYKJAVIK. ICELAND. 26-31 MAI 2014. 1051-1055.

ATAA ALLAH F., BOULAKNADEL S. (2014). LA PROMOTION DE L'AMAZIGHE A LA LUMIERE DES TECHNOLOGIES DE L'INFORMATION ET DE COMMUNICATION. ASINAG N°9, PUBLICATIONS DE L'IRCAM, PP. 33-48.

ATAA ALLAH F., BOULAKNADEL S., SOUIFI H. (2014B). « JEU D'ETIQUETTES MORPHOSYNTAXIQUES DE LA LANGUE AMAZIGHE », ASINAG N°9, PUBLICATIONS DE L'IRCAM, PP. 171-184.

BEESLEY K. R., KARTTUNENAURI L. (2003). FINITE STATE MORPHOLOGY. CSLI PUBLICATIONS, STANFORD. CA.

BOGUSLAVSKY I., CARDEÑOSA J., GALLARDO G., IRAOLA L. (2005). THE UNL INITIATIVE. AN OVERVIEW. COMPUTATIONAL LINGUISTICS AND INTELLIGENT TEXT PROCESSING. 6TH INTERNATIONAL CONFERENCE, CICLING 2005. MEXICO CITY. MEXIQUE. FEVRIER 13-19.

BOUHJAR A. (2004). LE SYSTEME GRAPHIQUE TIFINAGHE-IRCAM. ACTES DU SEMINAIRE ORGANISE PAR LE CENTRE DE L'AMENAGEMENT LINGUISTIQUE, STANDARDISATION DE L'AMAZIGHE. PUBLICATIONS DE L'IRCAM. IMPRIMERIE EL MAARIF AL JADIDA. RABAT.

BOUKHRIS F., BOUMALK A., ELMOUJAHID E., SOUIFI H. (2008). LA NOUVELLE GRAMMAIRE DE L'AMAZIGHE. PUBLICATIONS DE L'IRCAM. IMPRIMERIE EL MAARIF AL JADIDA. RABAT.

BOUKOUS A. (2007). L'ENSEIGNEMENT DE L'AMAZIGHE (BERBERE) AU MAROC : ASPECTS SOCIOLINGUISTIQUES. REVUE DE L'UNIVERSITE DE MONCTON. NUMERO HORS SERIE. 81-89.DOI : 10.7202/017709AR

BOULAKNADEL S., ATAA ALLAH F. (2012). INITIATIVE POUR LE DEVELOPPEMENT D'UN CORPUS DE LA LANGUE AMAZIGHE », ACTES DE LA 5EME CONFERENCE INTERNATIONALE SUR LES TECHNOLOGIES D'INFORMATION ET DE COMMUNICATION POUR L'AMAZIGHE (TICAM 2012). RABAT. MAROC. 26-27 NOVEMBRE 2012.

BOULAKNADEL S., ATAA ALLAH F. (2013). BUILDING A STANDARD AMAZIGH CORPUS. ADVANCES IN INTELLIGENT SYSTEMS AND COMPUTING: PROCEEDING OF THE INTERNATIONAL CONFERENCE ON INTELLIGENT HUMAN COMPUTER INTERACTION (IHCI 2011). PRAGUE. CZECH REPUBLIC. AUGUST 29-31, 2011. 179: 91-98. SPRINGER BERLIN HEIDELBERG. ISBN: 978-3-642-31602-9.

BOULAKNADEL S., TALHA M., ABOUTAJDINE D. (2014). AMAZIGHE NAMED ENTITY RECOGNITION USING A RULE BASED APPROACH. ACTES DE LA 11TH ACS/IEEE INTERNATIONAL CONFERENCE ON COMPUTER SYSTEMS AND APPLICATIONS (AICCSA'2014). DOHA. QATAR. 10-12 NOVEMBRE. 478-484.

EAGLES. (1996). RECOMMENDATION FOR THE MORPHOSYNTACTIC ANNOTATION OF CORPORA. EAGLES DOCUMENT EAG-TCWG-MAC/R. HTTP://WWW/ILC.CNR.IT/EAGLES96/HOME.HTML.

EL AZRAK N., ELHAMDAOUI A. (2011). REFERENTIEL DE LA TERMINOLOGIE AMAZIGHE : OUTIL D'AIDE A L'AMENAGEMENT LINGUISTIQUE. ACTES DU 4EME ATELIER INTERNATIONAL SUR L'AMAZIGHE ET LES

NOUVELLES TECHNOLOGIES DE L'INFORMATION ET DE COMMUNICATION, SOUS LE THEME : LES RESSOURCES LANGAGIERES: CONSTRUCTION ET EXPLOITATION (NTIC 2011). RABAT. MAROC. 24-25 FEVRIER 2011.

FRAIN J., ATAA ALLAH F., AIT OUGUENGAY Y. (2014). LEXIQUE AMAZIGHE POUR MOBILE, ACTES DE LA 6EME CONFERENCE INTERNATIONALE SUR LES TECHNOLOGIES D'INFORMATION ET DE COMMUNICATION POUR L'AMAZIGHE (TICAM 2014). RABAT. MAROC. 24-25 NOVEMBRE 2014.

KOSKENNIEMI K. (1984). TWO-LEVEL MORPHOLOGY: A GENERAL COMPUTATIONAL MODEL FOR WORD-FORM RECOGNITION AND PRODUCTION, THÈSE DE DOCTORAT, UNIVERSITÉ DE HELSINKI, FINLANDE.

LAABDELAOUI R., BOUMALK A., IAZZI E.M, SOUIFI H., ANSAR K. (2012).MANUEL DE CONJUGAISON AMAZIGHE. PUBLICATIONS DE L'IRCAM. IMPRIMERIE EL MAARIF AL JADIDA. RABAT.

MARCELLESI J. B. (2003). SOCIOLINGUISTIQUE. EPISTEMOLOGIE, LANGUES REGIONALES, POLYNOMIE. PARIS, L'HARMATTAN.

MIFTAH N., ATAA ALLAH F., TAGHBALOUT I. (2017). SENTENCE-ALIGNED PARALLEL CORPUS AMAZIGH-ENGLISH. ACTES DE L'INTERNATIONAL CONFERENCE ON INFORMATION AND COMMUNICATION SYSTEMS. IRBID. JORDANIE. 4-6 AVRIL 2017.

MUHIRWE, J. (2007), TOWARDS HUMAN LANGUAGE TECHNOLOGIES FOR UNDER RESOURCED LANGUAGES. COMPUTING AND ICT RESEARCH. (ÉD) JOSEPH KIZZA ET AL. KAMPALA.

NEJME F., BOULAKNADEL S., ABOUTAJDINE D. (2016). AMAMORPH: FINITE STATE MORPHOLOGICAL ANALYZER FOR AMAZIGHE. JOURNAL OF COMPUTING AND INFORMATION TECHNOLOGY. 91-110.

SMITH, F. (2004). UNDERSTANDING READING: A PSYCHOLINGUISTIC ANALYSIS OF READING AND LEARNING TO READ. MAHWAH. NEW JERSEY. LAWRENCE ERLBAUM ASSOCIATES PUBLISHERS.

SILBERZTEIN M. (2007). AN ALTERNATIVE APPROACH TO TAGGING. NLDB 2007: 1-11

TALHA M., BOULAKNADEL S., ABOUTAJDINE D. (2015). DEVELOPMENT OF AMAZIGHE NAMED ENTITY RECOGNITION SYSTEM USING HYBRID APPROACH. ACTES DE LA 16TH INTERNATIONAL CONFERENCE ON INTELLIGENT TEXT PROCESSING AND COMPUTATIONAL LINGUISTICS. CAIRE. EGYPTE.14-20 AVRIL 2015.

TAGHBALOUT I., ATAA ALLAH F., EL MARRAKI M. (2017). TOWARDS UNL BASED MACHINE TRANSLATION FOR MOROCCAN AMAZIGH LANGUAGE. INTERNATIONAL JOURNAL OF COMPUTATIONAL SCIENCE AND ENGINEERING.

ZENKOUAR L., AIT OUGUENGAY Y. (2007). ⵜⵉⵏⵎⵍ ⵏ ⵜⵎⴰⵣⵉⵖⵜ (TINML N TMAZIGHT). PRODUCTIONS DE DE L'INSTITUT ROYAL DE LA CULTURE AMAZIGHE, RABAT.

ZENKOUAR L., AIT OUGUENGAY Y. (2008). ⵉⵎⵓⴷⴰⵔ ⵉⵔⴰⵎⵢⴰⵔⵏ ⴷ ⵉⵏⴰⵎⵢⴰⵔⵏ (ANIMAUX SAUVAGES ET DOMESTIQUES). PRODUCTIONS DE DE L'INSTITUT ROYAL DE LA CULTURE AMAZIGHE. RABAT.

Problèmes de tokénisation pour deux langues régionales de France, l'alsacien et le picard

Delphine Bernhard[1] Amalia Todirascu[1] Fanny Martin[2] Pascale Erhart[1]
Lucie Steiblé[1] Dominique Huck[1] Christophe Rey[2]
(1) LiLPa - EA 1339, Université de Strasbourg
(2) Laboratoire Habiter le Monde - HM - EA 4287, Université de Picardie Jules Verne, Amiens

{dbernhard,todiras,pascale.erhart,lucie.steible,dominique.huck}@unistra.fr
{fanny.martin,christophe.rey}@u-picardie.fr

RÉSUMÉ

La tokénisation est une étape essentielle dans tout système de traitement automatique des langues, d'autant plus que de nombreux outils dépendent du découpage obtenu. La tâche est particulièrement ardue pour les textes qui ne respectent pas les conventions orthotypographiques ou les langues pour lesquelles ces conventions ne sont pas stabilisées. Nous nous intéressons ici aux cas de deux langues régionales de France, l'alsacien et le picard. Nous présentons les défis posés par ces deux langues, et proposons des critères de découpage implémentés dans des tokéniseurs.

ABSTRACT

Tokenization Issues for Two Regional Languages of France, Alsatian and Picard

Tokenization is an essential step in any language processing system, especially as many tools rely on the obtained segmentation. The task is particularly difficult for texts that do not meet typographical syntax or languages for which the usage of typographic signs is not stabilized. Here we focus on the case of two regional languages of France, Alsatian and Picard. We present the challenges of these two languages, and propose segmentation criteria implemented in tokenizers.

MOTS-CLÉS : Tokénisation, Alsacien, Picard, conventions orthotypographiques.

KEYWORDS: Tokenization, Alsatian, Picard, typographical syntax.

1 Introduction

La tokénisation est une des premières étapes de tout système de traitement automatique des langues (Webster & Kit, 1992). Cette tâche de découpage d'un texte en mots et phrases est donc d'une importance primordiale. Si la tokénisation pose encore rarement des problèmes pour les langues bien dotées, du moins pour les documents respectant les conventions orthotypographiques, la situation est toute autre pour de nombreuses langues moins dotées. Dans de nombreux cas il n'existe pas d'acte officiel de standardisation de la langue, y compris pour ce qui est de l'utilisation de l'espacement et des signes de ponctuation. C'est le cas notamment des deux langues régionales de France que nous considérons dans cet article : l'alsacien et le picard. Ces deux langues appartiennent à des familles différentes, mais posent des problèmes similaires, notamment les marques d'oralité à l'écrit (épenthèses - insertions de sons pour faciliter l'articulation, contractions signalées ou non par une

apostrophe). La tokénisation se base généralement sur des délimiteurs, qui marquent les frontières de mots et de phrases. Certains de ces délimiteurs sont non ambigus (comme le point d'exclamation, les double-points), d'autres sont ambigus (comme l'apostrophe, l'espace, le tiret ou le point) et nécessitent donc des traitements plus fins. Par exemple, en picard, l'apostrophe ne doit pas être considérée comme une frontière de mot dans la forme *k'min* (chemin) ou *batt'meints* (battements). On trouve le même phénomène en alsacien pour le déterminant *d'r* (le) ou le participe passé *g'hàlte* (arrêté). Il s'agit dans notre cas de développer des tokéniseurs afin de (i) produire des corpus annotés manuellement avec des informations morphosyntaxiques et (ii) développer des outils d'étiquetage morphosyntaxique pour ces deux langues régionales.

Nous passons tout d'abord en revue les travaux existants sur le découpage de textes en mots. Nous présentons ensuite les problèmes spécifiques qui se posent pour l'alsacien et le picard et les solutions proposées. Nous faisons ensuite une évaluation des systèmes de tokénisation développés.

2 État de l'art

La mise en œuvre d'un système de tokénisation passe dans un premier temps par la définition de ce qui constitue un token dans la langue considérée. Webster & Kit (1992) considèrent le token comme un "nœud terminal" (*terminal node*), qui, du point de vue des traitements ultérieurs, ne sera pas découpé en unités plus petites. Ainsi, un token pourra être constitué de plusieurs mots graphiques, comme c'est le cas par exemple des expressions idiomatiques, des mots composés (*pomme de terre*) ou de certains nombres (10 000). A l'inverse, un mot graphique peut être décomposé en deux unités (par exemple *au* décomposé en *à le*). Webster & Kit (1992) soulignent également l'importance de la tâche pour laquelle la tokénisation est effectuée, qui contraint la décomposition ou non des unités linguistiques en unités plus petites.

Les outils de TAL (outils de traitement de corpus, étiqueteurs, analyseurs syntaxiques) mettent en œuvre des stratégies de tokénisation plus ou moins évoluées même pour la langue standardisée. On distingue les approches à base de règles et les approches par apprentissage. Les premières définissent des règles de découpage ou d'identification des tokens qui sont, au moins en partie, dépendantes de la langue cible et qui peuvent prendre la forme d'expressions régulières (Grefenstette & Tapanainen, 1994). Les approches par apprentissage peuvent procéder par apprentissage supervisé, à l'aide d'un corpus pré-tokénisé, ou de manière non-supervisée. Par exemple, Jurish & Würzner (2013) utilisent un modèle de Markov caché (*Hidden Markov Model*) pour classifier les frontières de segments. L'approche non supervisée de Wrenn *et al.* (2007) repose quant à elle sur des arbres préfixes (*trie*) afin d'analyser les propriétés statistiques des frontières de tokens.

Enfin, les stratégies développées pour un type de textes à l'écrit peuvent ne pas fonctionner aussi bien pour d'autres types de textes pourtant dans la même langue. La problématique de la tokénisation resurgit ainsi pour les textes de spécialité, par exemple dans le domaine médical (Rabary *et al.*, 2015). Dans ce cas, le nombre très important de sigles et abréviations (comme *chir.* pour *chirurgie*) pose problème à un tokéniseur développé pour des textes journalistiques, de même que les mesures (comme *3x/j*). D'autres cas problématiques apparaissent dans les textes parus sur les médias sociaux, liés aux phénomènes de faute de frappe (espaces manquants, superflus ou mal placés), aux marques d'oralité (contractions) et aux répétitions de signes de ponctuation ou graphèmes, ainsi que tous les tokens spécifiques à ce type de communication comme les émoticônes ou mots-dièse (Laarmann-Quante & Dipper, 2016).

3 Tokénisation pour l'alsacien

3.1 Recommandations orthotypographiques existantes

Même s'il existe des propositions récentes de conventions orthographiques (par exemple ORTHAL (Zeidler & Crévenat-Werner, 2008)), l'écriture des dialectes alsaciens n'est pas strictement normée et les usages sont donc très diversifiés. Il est d'ailleurs très difficile des trouver des recommandations explicites pour l'usage des apostrophes ou des graphies séparées ou synthétiques, à part quelques exemples :

Grammaire de l'alsacien, d'Edmond Jung La grammaire formule des préconisations pour les adverbes pronominaux composés (*wo/da* + préposition), les articles et les pronoms personnels. Il est notamment conseillé de détacher les pronoms personnels sujets et objets dans tous les cas : *gìm mer s* (donne-le moi) et non *gìmmers* (Jung, 1983, p. 106-108).

Orthographe alsacienne, de Edgar Zeidler et Danielle Crévenat-Werner La méthode ORTHAL donne des recommandations très souples pour "l'écriture agglomérée". Ainsi, les graphies suivantes sont toutes considérées comme acceptables (Zeidler & Crévenat-Werner, 2008, p. 62) : *ìnere Stund* / *ìn ere Stund* / *ìn're Stund* (dans une heure). Ces diverses possibilités ne sont pas sans poser des problèmes à la tokénisation, car si l'on souhaitait avoir un découpage cohérent, il faudrait alors procéder à un découpage au sein de groupes de caractères alphanumériques contigus (comme par exemple découper *ìnere* en *ìn ere*). Nous avons dans ce cas fait le choix de ne procéder au découpage que si la séparation est marquée par une espace ou un signe spécifique (tiret et apostrophe). Par ailleurs, les recommandations données ci-dessus ne reflètent pas nécessairement les usages qui se rencontrent dans les données, pour une période qui s'étend sur 200 ans (1816 étant l'année de parution de la première pièce de théâtre en alsacien, le *Pfingstmontag* de Georges-Daniel Arnold).

3.2 Critères de découpage tokénisation en alsacien

Nous proposons de différencier les deux types de tokens suivants :

1. les signes graphiques qui relèvent du champ lexical ou grammatical ;
2. les signes graphiques qui relèvent de l'épenthèse, essentiellement le <n> et le <w> euphoniques. L'alsacien tend à représenter graphiquement des phénomènes phonétiques, ce pourquoi il a été choisi de découper les épenthèses, comme c'est souvent le cas dans le cadre de l'étiquetage de corpus oraux (Benzitoun *et al.*, 2012).

La Table 1 donne quelques exemples des deux types de phénomènes, relevés dans des textes variés et d'auteurs différents.

3.3 Développement d'un tokéniseur pour l'alsacien

Le tokéniseur développé repose sur des expressions régulières inspirées de (Grefenstette, 1998; Pointal, 2004). Les expressions régulières permettent de distinguer différents types d'unités :

– Les caractères et chaînes de caractères qui doivent être séparés de la suite lorsqu'ils sont situés après une espace : `d', s', z', n', üf'` etc. Cette règle permet de considérer les articles, prépositions et conjonctions élidés avant un mot comme des unités indépendantes.

Graphie initiale	Tokénisation proposée	Phénomène	Genre et source
zitter'm Ààfàng (depuis le début)	zitter␣'m Ààfàng	préposition + article au datif	encyclopédie : (Wikipedia, 2015)
in'ra Volkssproch (dans un dialecte)	in␣'ra Volkssproch	préposition + article au datif	encyclopédie : (Wikipedia, 2017)
uf'e'me Schàrebbà (sur un char à bancs)	uf␣'e'me Schàrebbà	préposition + article au datif	récit : (Sonnendrücker & Kauss, 1998)
hàt fànga-n-à drucka (a commencé à imprimer)	hàt fànga␣-n-␣à drucka	épenthèse	encyclopédie : (Wikipedia, 2017)
mine-n-Anforderunge (mes exigences)	mine␣-n-␣Anforderunge	épenthèse	théâtre : (Stoskopf, 1906)
Heere-n-Er (entendez-vous)	Heere␣-n-␣Er	épenthèse	théâtre : (Redslob, 1907)
geh-w-i (je vais)	geh␣-w-␣i	épenthèse	guide : (Keck & Daul, 2010)

TABLE 1 – Exemples de découpages en mots proposés pour l'alsacien.

– Les caractères et chaînes de caractères qui doivent être séparés de ce qui précède lorsqu'ils sont situés avant une espace : `'m'r, 'ma, 'me` etc. Cette règle permet notamment de considérer les pronoms situés à droite d'un verbe comme des unités lexicales indépendantes.
– Les suites de caractères qui doivent être séparés de ce qui suit lorsqu'ils sont situés en milieu de mot : `-n-`.
– Les nombres, abréviations, URLs, adresses mail.

La tokénisation a notamment pour objectif d'obtenir des unités cohérentes pour l'annotation morpho-syntaxique. Il est donc nécessaire de séparer les mots grammaticaux des mots relevant des classes ouvertes du lexique.

3.4 Évaluation du tokéniseur pour l'alsacien

Nous avons constitué un corpus de test à partir d'extraits appartenant à des genres variés, qui correspondent aux types d'écrits qu'il est possible de trouver en alsacien : théâtre (426 tokens), poésie (368 tokens), récit en prose (732 tokens), Facebook (333 tokens) et Wikipédia alémanique (778 tokens). Le corpus de test comprend 2 633 tokens en tout (signes de ponctuation inclus) et la tokénisation a été effectuée manuellement à partir des critères de découpage détaillés dans la section précédente. Nous avons également comparé les résultats de notre tokéniseur spécifique aux dialectes alsaciens avec le tokéniseur fourni avec le TreeTagger (Schmid, 1994), en prenant en compte le lexique des abréviations fournies pour l'allemand.

Les résultats de la tokénisation ont été évalués à l'aide de la commande `diff` et figurent dans les tables 2 et 3. Nous avons mesuré le nombre de vrais positifs (VP), faux positifs (FP, insertions par rapport à la tokénisation de référence) et faux négatifs (FN, absence de découpage par rapport à la référence), nous permettant ensuite de calculer la précision, le rappel et la F-mesure. Notre tokéniseur adapté à l'alsacien obtient des niveaux de performance supérieurs au tokéniseur générique fourni par le TreeTagger, même si les performances de ce dernier restent relativement bonnes. De manière globale, 2 621 tokens ont été correctement reconnus par notre tokéniseur, et 2 528 par le tokéniseur

du TreeTagger.

Genre	VP	FP	FN	Précision	Rappel	F-mesure
Facebook	332	1	1	0,997	0,997	0,997
Poésie	366	0	2	1,000	0,995	0,997
Récit	731	0	1	1,000	0,999	0,999
Théâtre	419	1	7	0,998	0,984	0,991
Wikipédia	773	1	5	0,999	0,994	0,996

TABLE 2 – Résultats de l'évaluation du tokéniseur spécifique aux dialectes alsaciens.

Genre	VP	FP	FN	Précision	Rappel	F-mesure
Facebook	325	1	8	0,997	0,976	0,986
Poésie	345	7	23	0,980	0,938	0,958
Récit	711	14	21	0,981	0,971	0,976
Théâtre	403	5	23	0,988	0,946	0,966
Wikipédia	744	6	34	0,992	0,956	0,974

TABLE 3 – Résultats de l'évaluation du tokéniseur du TreeTagger.

Nous avons étudié les erreurs de tokénisation les plus fréquentes. Elles concernent essentiellement des caractères ambigus (*r*, *d*, *z*, *s*) qui doivent ou ne doivent pas être détachés selon le contexte : ainsi *'r* doit être détaché dans *dass'r's* mais pas dans *widd'r* (à noter que dans *dass'r's* il y a deux segmentations à opérer). Ces cas sont difficiles à gérer, compte-tenu des nombreuses variantes graphiques qui peuvent être trouvées (par exemple *us-em* et *us'm*) et de l'apostrophe qui peut être utilisée à la fois pour marquer des élisions en frontière de token et au milieu de mots qui ne doivent pas être découpés (ex : *Z'erscht*, équivalent à *zuerst* en allemand standard). Les deux outils sont de ce point de vue plutôt conservateurs car le nombre de faux négatifs est supérieur au nombre de faux positifs.

4 Tokénisation pour le picard

4.1 Recommandations existantes

Le picard n'existe pas en tant que langue standardisée, ni normée sur le plan de la graphie, cependant, on l'écrit depuis plusieurs siècles déjà et sa présence sur la scène littéraire, sous ses différentes variétés, est importante, ainsi qu'en témoignent de récents succès de librairie et notamment celui des volumes de bandes dessinées en picard d'Astérix et Tintin [1]. Malgré les nombreux débats dans les années 1960-1970 autour de la standardisation et de l'orthographe en picard (centralisation de la langue, question de la variation : uniformisation de la graphie et du lexique), à ce jour, aucune standardisation n'est engagée sur l'ensemble du domaine picard. Plus encore, cette situation est vécue

1. Il n'y a donc pas de contradiction formelle entre la non-standardisation du picard à l'échelle du domaine linguistique picard, la présence de variation et la question de la vitalité notamment (mais pas exclusivement) littéraire du picard. En effet, il n'y a pas ici de vérité ni de lien *stricto sensu* entre standardisation et vitalité sur l'ensemble du domaine picard. Par ailleurs, c'est peut-être ce particularisme de la non-standardisation qui évite de poser un « carcan » trop rigide revendiquant ainsi par la variation une forme de liberté.

aujourd'hui comme une « liberté assumée » contre la standardisation. À ce titre, le picard peut être défini comme une « langue de la liberté » (Martin, 2015). Comme l'écrit Jean-Michel Éloy :

> « [...] en domaine picard, la question de la standardisation n'est pas posée du tout en ce qui concerne les formes de langue. Même si l'on ne parle que de standardisation des procédés graphiques, son importance est diversement appréciée. Les débats sur "l'orthographe du picard", qui furent très vifs dans les années 60 et 70, sont aujourd'hui curieusement éteints – un colloque sur ce thème en 2010 avait d'ailleurs conclu au *statu quo*, et avait eu peu d'écho. » (Eloy, 2014, 10-11)

Nombreux sont les ouvrages en picard, qui mentionnent la non existence de standard et la cooptation d'une « liberté assumée » concernant la graphie. Certains auteurs mentionnent leurs références pour la graphie (Debrie, 1996, 1972; Carton, 1963, 1964, 2001).

4.2 Critères de découpage de mots et tokénisation automatique en picard

Cette variation graphique du picard pose des problèmes pour la tokénisation automatique. Nous nous sommes appuyés sur l'ouvrage *Éche pikar bèl é rade* (Debrie, 1983a), pour proposer des critères du découpage de mots en picard. Les cas les plus problématiques sont les suivants :

- le tiret peut être un séparateur (*Est-ce-què*) mais il peut faire partie de certains mots (par exemple dans certains verbes) : *quandis n'mariye-té* [picard de Belgique] / *kan k i s'marite* [picard de l'Amiénois] (quand elles se marient) ;
- le point peut être utilisé comme séparateur de phrase, entrer dans la composition d'un sigle ou signaler un allongement de la consonne qui conduit à une nasalisation) *I se proumon.ne* [picard du Cambrésis] (il se promène) ;
- l'apostrophe peut avoir plusieurs interprétations possibles : (a) une marque qui oriente par rapport à la prononciation comme dans les exemples *té mérit'roès (tu mérites), f'rais (ferais)*, où une voyelle est supprimée ; (b) une marque de l'élision d'une voyelle *L'aute* (l'autre) ; (c) en fin de mot *Dis-l'*. Il est fréquent d'avoir plusieurs apostrophes dans le même mot avec utilisations différentes *Qu't'os (que tu as), Coreed'l'histoire (encore de l'histoire)* ;
- l'espace. On peut avoir des séquences espace + lettre + espace. Il s'agit du phénomène d'épenthèse, la lettre marque une liaison entre les deux mots : par exemple la lettre z dans *lé z éfans* (les enfants).

La graphie non-standardisée soulève des nombreux problèmes pour le découpage automatique du picard. Outre l'ambiguïté des séparateurs, tel l'apostrophe ou le tiret, l'espace peut apparaître après l'apostrophe et certains pronoms peuvent être agglutinés au mot précédent ou suivant. Ce dernier cas a été traité différemment de l'alsacien. Pour le découpage automatique des mots en picard, nous utilisons, comme pour l'alsacien, des expressions régulières adaptées. D'abord, nous annotons les mots composés incluant le tiret ou l'espace, à l'aide d'un lexique de mots composés construit à partir de plusieurs lexiques picards (Debrie, 1987, 1986, 1985, 1983b, 1981, 1975). Ce lexique contient des locutions prépositionnelles (*a travér dech'*) ou adverbiales (*Tout ein heùt*), des expressions figées (*pi vlaù qu'*). Les mots composés annotés à l'aide du dictionnaire ne seront pas soumis à la procédure de découpage. Ensuite, nous utilisons les séparateurs non-ambigus (!, ;) et nous proposons des règles de découpage spécifiques pour les séparateurs ambigus :

- l'apostrophe est identifiée comme séparateur à l'aide d'une liste de mots outils (déterminants, pronoms, démonstratifs). La présence d'un de ces mots outils avant l'apostrophe, en début ou à la fin du mot, indique qu'on doit découper le mot au niveau de l'apostrophe ;
- le point est considéré comme séparateur de phrases, sauf pour les sigles, les nombres, les marques de nasalisation ;

Graphie initiale	Tokénisation proposée	Phénomène	Source
quand is n'mariye-té (quand elles se marient)	quand␣is n'␣mariye-té	conjonction + pronom, négation + verbe	(Debrie, 1983a)
Est-ce-què	Est␣ce␣què	verbe + pronom + conjonction	(Debrie, 1983a)
I se proumon.ne (il se promène)	I se proumon.ne	pronom + pronom + verbe	(Debrie, 1983a)
O z avon (nous avons)	O z avon	pronom + consonne d'appui + auxiliaire	(Debrie, 1983a)
Té mérit'roès qu'j'el diche à tin père, quand qu'il arvarro (Tu mériterais que je le dise à ton père, quand il arrivera)	Té mérit'roès qu'␣j'␣el diche à tin père␣, quand qu'␣il arvarro	pronom relatif + pronom personnel + pronom, conjonction + pronom	(Debrie, 1983a)
Eze z'éfan i s'abiye (les enfants s'habillent)	Eze z'␣éfan i s'␣abiye	consonne d'appui + nom, pronom personnel + verbe	(Debrie, 1983a)

TABLE 4 – Exemples de découpages en mots proposés pour le picard.

– Le tiret est reconnu comme séparateur dans certains cas particuliers (est-ce-que), sauf pour les mots composés trouvés dans le dictionnaire ou dans certaines formes de verbes ;
– Certains pronoms (*is, i, il*) ou prépositions (*ed ˆ*) agglutinés au mot précédent ou suivant sont découpés en unités indépendantes.

4.3 Évaluation du tokéniseur pour le picard

Nous avons évalué le tokéniseur sur un corpus regroupant des extraits de genres divers (4 191 tokens) : un extrait d'un roman (506 tokens), deux extraits de deux nouvelles (824 tokens), un extrait d'une collection de lettres (452 tokens), un extrait de théâtre (532 tokens), 2 extraits de poésie narrative (635 tokens), 3 extraits de poésie (1 242 tokens). Le genre le plus représenté est la poésie, suivi des nouvelles et de la poésie narrative. En général, ces genres posent des problèmes aux tokéniseurs, à cause de la structure du texte spécifique. Comme pour l'alsacien, nous avons pris en compte le nombre de vrais positifs (VP), de faux positifs (FP) et de faux négatifs (FN) pour calculer la précision, le rappel et la F-mesure. L'évaluation a été faite avec *diff*. Les résultats sont de bonne qualité (tableau 5), en particulier pour la poésie et pour le roman. Les résultats obtenus pour le théâtre et pour la poésie narrative sont les moins précis, ce qui confirme les attentes :

Les erreurs les plus fréquentes relevées dans les résultats du tokéniseur picard sont :
– découpage excessif, dû à la confusion entre un mot outil (réfléchi, conjonction) et le début d'un verbe : *s'roit* sera découpé en *s'␣roit* alors qu'il s'agit d'un seul mot ; *qu'meinchi* (commencer) sera découpé alors qu'il s'agit d'un seul mot. Dans certains cas, le point peut également être considéré comme séparateur et le mot se trouve découpé, alors que le découpage ne doit pas se faire ;
– découpage erroné, quand plusieurs découpages sont possibles : *ch'l'* pourra se découper *ch'␣l'* ou non selon le contexte. Le mot outil le plus long sera prioritaire.

Genre	VP	FP	FN	Précision	Rappel	F-mesure
Poésie	1 199	19	24	0,984	0,980	0,982
Poésie narrative	563	36	36	0,939	0,939	0,939
Nouvelle	759	36	29	0,955	0,963	0,959
Roman	480	8	18	0,984	0,964	0,974
Lettre	424	12	16	0,972	0,964	0,968
Théâtre	503	13	16	0,974	0,969	0,972

TABLE 5 – Résultats de l'évaluation du tokéniseur pour le picard

Afin de comparer avec un tokéniseur disponible pour le français, notre choix a été l'étiqueteur TreeTagger (Schmid, 1994), le même outil choisi pour la comparaison avec le tokéniseur alsacien. Nous avons comparé notre tokeniseur avec deux configurations différentes de TreeTagger : TreeTagger avec la configuration standard pour le français (TreeTaggerBase) ; TreeTagger adapté pour le picard, utilisant une liste des mots outils en picard finissant par une apostrophe (prépositions, déterminants) et le dictionnaire picard de mots composés (TreeTaggerPicard). Ces deux dernières ressources sont utilisées également par le tokeniseur picard.

Configuration	Genre	VP	FP	FN	Précision	Rappel	F-mesure
TreeTaggerBase	Poésie	1 014	217	202	0,824	0,834	0,829
	Poésie narrative	486	62	119	0,887	0,803	0,843
	Nouvelle	685	68	80	0,910	0,895	0,903
	Roman	426	24	66	0,947	0,866	0,904
	Lettre	415	27	33	0,939	0,926	0,933
	Théâtre	510	15	11	0,971	0,979	0,975
TreeTaggerPicard	Poésie	1 031	196	190	0,840	0,844	0,842
	Poésie narrative	545	58	62	0,904	0,898	0,901
	Nouvelle	714	54	64	0,930	0,918	0,924
	Roman	455	5	48	0,989	0,905	0,945
	Lettre	412	28	35	0,936	0,922	0,929
	Théâtre	515	12	9	0,977	0,983	0,980

TABLE 6 – Résultats de l'évaluation du tokéniseur du TreeTagger français (TreeTaggerBase) et du TreeTagger utilisant les ressources spécifiques au picard (TreeTaggerPicard)

Le tokéniseur picard obtient des meilleurs résultats par rapport à TreeTagger (sauf pour le théâtre). Ces résultats s'expliquent par la mise en place des règles et des ressources adaptées au picard. TreeTagger se distingue du tokéniseur picard par la différence de traitement de l'apostrophe (qui est souvent considérée comme token séparé). TreeTaggerBase a systématiquement obtenu des performances plus faibles que TreeTaggerPicard. Malgré les ressources ajoutées, les résultats du TreeTaggerPicard restent inférieurs aux performances du tokéniseur du picard, à l'exception du théâtre où les tendances sont inversées (F-mesure de 0,980 contre 0,972 pour le tokéniseur picard) (voir tableau 6).

5 Conclusion et perspectives

Nous avons présenté les problèmes liés à la mise en place des outils de tokénisation pour l'alsacien et le picard, deux langues régionales dont l'orthographe est peu standardisée. Nous avons traité certains séparateurs ambigus (tels le point ou l'apostrophe), le cas des mots agglutinés et de l'épenthèse. Le découpage automatique rencontre des difficultés similaires pour les deux langues : plusieurs découpages possibles, plusieurs interprétations pour les séparateurs, ou les phénomènes d'épenthèse. Les erreurs les plus fréquentes sont liées aux séparateurs ambigus. A l'avenir, les outils de tokénisation seront évalués sur des corpus de plus grande taille, incluant d'autres genres (conte, dialogue, etc.) et les règles de découpage seront améliorées. Ils seront utilisés pour le développement des outils de traitement automatique de l'alsacien et du picard (étiqueteur, lemmatiseur).

Remerciements

Les travaux décrits dans cet article ont bénéficié du soutien de l'ANR (projet RESTAURE - référence ANR-14-CE24-0003).

Références

BENZITOUN C., FORT K. & SAGOT B. (2012). TCOF-POS : un corpus libre de français parlé annoté en morphosyntaxe. In *JEP-TALN 2012 - Journées d'Études sur la Parole et conférence annuelle du Traitement Automatique des Langues Naturelles*, p. 99–112, Grenoble, France.

CARTON F. (1963). Essai d'adaptation de l'orthographe Feller au picard moderne. *Nos patois du Nord*, **8**(supplément), 134–139.

CARTON F. (1964). L'adaptation de l'orthographe Feller au picard moderne. *Nos patois du Nord*.

CARTON F. (2001). Orthographe picarde Feller-Carton. *Linguistique picarde*.

DEBRIE R. (1972). *Propos sur l'orthographe*. Amiens : Archives départementales de la Somme.

DEBRIE R. (1975). *Lexique picard des parlers ouest-amiénois*. Centre d'Études Picardes.

DEBRIE R. (1981). *Lexique picard du Vimeu*. Centre d'Études Picardes.

DEBRIE R. (1983a). *Eche pikar bèl é rade*. Ed.-disques Omnivox.

DEBRIE R. (1983b). *Lexique picard des parlers est-amiénois*. Centre d'Études Picardes.

DEBRIE R. (1985). *Lexique picard du Ponthieu*. Centre d'Études Picardes.

DEBRIE R. (1986). *Lexique picard des parlers du Santerre*. Centre d'Études Picardes.

DEBRIE R. (1987). *Lexique picard du Vermandois*. Centre d'Études Picardes.

DEBRIE R. (1996). Essai d'orthographe picarde. *Le Courrier Picard*.

J.-M. ELOY, Ed. (2014). *Standardisation et vitalité des langues de France*, volume 9 of *Carnets d'Ateliers de Sociolinguistique*. Paris : L'Harmattan.

GREFENSTETTE G. (1998). *Re : Corpora : Sentence splitting*. Corpora List http ://torvald.aksis.uib.no/corpora/1998-4/0035.html.

GREFENSTETTE G. & TAPANAINEN P. (1994). What is a word, What is a sentence ? Problems of Tokenization. In *3rd International Conference on Computational Lexicography (COMPLEX'94)*, Budapest, Hungary.

JUNG E. (1983). *Grammaire de l'alsacien, dialecte de Strasbourg avec indications historiques*. Strasbourg, France : Oberlin.

JURISH B. & WÜRZNER K.-M. (2013). Word and Sentence Tokenization with Hidden Markov Models. *JLCL*, **28**(2), 61–83.

KECK B. & DAUL L. (2010). *L'alsacien pour les nuls*. Pour les nuls (Éd. de poche), ISSN 1625-0486. Paris, France : First éd.

LAARMANN-QUANTE R. & DIPPER S. (2016). An Annotation Scheme for the Comparison of Different Genres of Social Media with a Focus on Normalization. In *Proceedings of the LREC Workshop on Normalisation and Analysis of Social Media Texts (NormSoMe)*.

MARTIN F. (2015). *Espaces et lieux de la langue au XXIe siècle en Picardie. Approche complexe de la structuration des répertoires linguistiques en situations ordinaires – enquête en Picardie*. Thèse de doctorat, Université de Picardie Jules Verne, Amiens.

POINTAL L. (2004). Tree Tagger Wrapper. [en ligne ; accédé le 4 avril 2016] `https://perso.limsi.fr/pointal/dev:treetaggerwrapper`.

RABARY C. T., LAVERGNE T. & NÉVÉOL A. (2015). Etiquetage morpho-syntaxique en domaine de spécialité : le domaine médical. In *Actes de la 22e conférence sur le Traitement Automatique des Langues Naturelles*, Caen, France.

REDSLOB R. (1907). *D'r Schlitterhannes. Elsaessisches Bauerndrama in zwei Akten*. Strassburg, 1907.

SCHMID H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of the International Conference on New Methods in Language Processing*, p. 44–49.

SONNENDRÜCKER P. & KAUSS A. (1998). *Kochersberg : récits en dialecte avec version française*. Strasbourg, France : Bf.

STOSKOPF G. (1906). *D'r Hoflieferant. Elsaessische Komœdie in 3 Aufzuegen*. Strassburg, 1906.

WEBSTER J. J. & KIT C. (1992). Tokenization as the initial phase in NLP. In *Proceedings of the 14th conference on Computational linguistics-Volume 4*, p. 1106–1110.

WIKIPEDIA (2015). Elsässisches Museum (Straßburg) — Alemannische Wikipedia. [En ligne, accédé le 01/06/2017, Page Version ID : 654412].

WIKIPEDIA (2017). Johannes Mentelin — Alemannische Wikipedia. [En ligne, accédé le 01/06/2017, Page Version ID : 754490].

WRENN J. O., STETSON P. D. & JOHNSON S. B. (2007). An unsupervised machine learning approach to segmentation of clinician-entered free text. In *AMIA Annual Symposium Proceedings*, volume 2007 : American Medical Informatics Association.

ZEIDLER E. & CRÉVENAT-WERNER D. (2008). *Orthographe alsacienne : bien écrire l'alsacien de Wissembourg à Ferrette*. Colmar : J. Do Bentzinger.

Produire des ressources électroniques à partir de descriptions formelles : application aux langues peu dotées

Denys Duchier[1] Yannick Parmentier[1] Simon Petitjean[2] Emmanuel Schang[3]
(1) LIFO, Université d'Orléans, 45067 Orléans, France
(2) CRC991, Heinrich Heine Universität, D-40225 Düsseldorf, Allemagne
(3) LLL, Université d'Orléans, 45067 Orléans, France
denys.duchier@univ-orleans.fr, yannick.parmentier@univ-orleans.fr,
simon.petitjean@phil.uni-duesseldorf.de, emmanuel.schang@univ-orleans.fr

RÉSUMÉ
Dans cet article, nous montrons comment les langages de description, et le système XMG2 en particulier, peuvent être utilisés pour produire à moindre coût (en termes d'hommes.années) une ressource électronique linguistiquement précise et relativement couvrante, ouvrant ainsi la voie à la création de ressources pour les langues peu dotées. Nous insistons également sur le fait que l'utilisation d'un langage de description permet (sans aucun sur-coût) de documenter une langue et d'évaluer la capacité d'une théorie linguistique à décrire un certain jeu de données.

ABSTRACT
Creating e-resources from formal descriptions : application to under-resourced languages

In this paper, we show how description languages, and the XMG2 framework in particular, can be used to design at a lesser cost (in terms of man-year) a linguistically precise and relatively large electronic resource. This paves the way to the creation of resources for under-resourced languages. We furthermore emphasize the fact that the use of description languages makes it possible (at no additional cost) to document a natural language and evaluate the capacity of a linguistic theory to describe a given data set.

MOTS-CLÉS : langage formel, langage de description, ressources électroniques, métagrammaire.

KEYWORDS: formal language, description language, electronic resources, metagrammar.

1 Introduction

Le fait de pouvoir disposer de ressources linguistiques (telles que des lexiques, des grammaires ou encore des corpus annotés) dans un format électronique *ouvert* est non seulement important pour permettre une sauvegarde de la connaissance que l'on a d'une langue, mais aussi pour permettre d'adopter une vue introspective sur celle-ci (en appliquant des traitements sur ces ressources pour en vérifier la cohérence ou encore en extraire des informations statistiques plus ou moins fiables selon leur taille et leur représentativité), ou encore bien sûr pour permettre le développement d'applications de traitement automatique des langues nécessitant une modélisation relativement précise de la langue (cas de la traduction automatique ou du dialogue homme-machine par exemple).

Malheureusement, le coût de développement de telles ressources est habituellement très élevé (de l'ordre de plusieurs homme.années pour des grammaires électroniques noyaux comme par exemple la

grammaire d'arbres adjoints du français FTAG (Abeillé *et al.*, 1999)).

Alors que les premières ressources linguistiques électroniques étaient produites à la main, il est devenu courant de recourir à des méthodes (totalement ou en partie) automatiques pour les extraire à partir de données de base. Ainsi, il est devenu possible d'extraire un lexique à partir d'un corpus brut (Sagot *et al.*, 2006), ou encore de produire une grammaire noyau par transformation de règles canoniques (Prolo, 2002). [1]

Parmi les approches semi-automatiques de création de ressources linguistiques, on peut distinguer les approches basées sur des *langages de description* des autres approches, par le fait qu'elles ne nécessitent aucune ressource externe préalable, et sont de ce fait adaptées à la création rapide de prototypes de ressources linguistiques *from scratch* [2]. Les langages de description reposent sur des mécanismes d'abstraction permettant de définir de manière déclarative les régularités (voire dans certains cas les irrégularités) d'une langue. [3] Cette description déclarative est ensuite traitée automatiquement (*compilée*) pour produire les unités décrites (par exemple les entrées d'un lexique ou d'une grammaire électronique). Comme nous le verrons, ces descriptions offrent divers avantages dont principalement (i) le fait de constituer en elle-même, de par leur structure modulaire et hiérarchique, une documentation des unités de la langue (mots ou règles syntaxiques dans notre cas), et (ii) le fait de permettre de vérifier la portée d'une théorie linguistique (en observant l'adéquation entre structures décrites et structures observées chez les locuteurs de la langue considérée).

Dans ce qui suit, nous utiliserons le système libre et ouvert XMG2 (Petitjean *et al.*, 2016) pour illustrer l'utilisation pratique des langages de description pour produire des ressources linguistiques. XMG2 est un système permettant notamment de définir des langages de description de manière modulaire, et de générer à la volée un compilateur pour chacun de ces langages. [4]

L'article est structuré comme suit. En Section 2, nous présentons le concept de description formelle de ressource linguistique (métagrammaire) et l'état de l'art dans ce domaine. En Section 3, nous présentons brièvement le système XMG2. En Section 4, nous montrons deux cas d'utilisation des langages de description (et du système XMG2) pour décrire des langues peu dotées. Concrètement, nous montrons comment un langage de description permet de décrire un lexique nominal en définissant des combinaisons d'affixes en fonction d'une racine nominale donnée pour produire un lexique électronique d'une langue bantoue (sous-section 4.1), ou encore de décrire de manière concise un ensemble de règles grammaticales arborescentes en capturant des généralisations entre celles-ci pour produire une grammaire du são-tomense (sous-section 4.2). Enfin (Section 5) nous concluons et présentons des pistes de recherche dans ce domaine.

2 Décrire des ressources linguistiques : les métagrammaires

Les langages de description sont un outil utilisé depuis un certain temps déjà en informatique, comme en témoignent les exemples du langage HTML pour les interfaces graphiques (ou pages web) sur

1. À noter que l'annotation automatique de corpus repose souvent sur la disponibilité préalable de ressources externes, par exemple des corpus d'entraînement pour la création de corpus arborés par analyse syntaxique automatique.

2. « À partir de rien ».

3. Cette description déclarative est souvent appelée *métagrammaire*, car son utilisation originelle résidait dans la description de grammaires.

4. XMG2 peut être qualifié de *méta-compilateur*, puisqu'il permet de compiler le compilateur d'un langage de description en fonction d'une définition de ce langage.

internet ou encore du langage LaTeX pour la description de documents. Leur utilisation en traitement automatique des langues est elle aussi relativement ancienne puisqu'elle remonte aux années 80 avec notamment les langages PATRII (Shieber, 1984) et DATR (Evans & Gazdar, 1996) pour la représentation de ressources lexicales. Dans ces contextes, un langage de description n'est autre qu'un langage formel dont la syntaxe est définie au moyen d'une grammaire hors-contexte dans la hiérarchie de Chomsky (1957), et dont la sémantique est définie en termes d'interprétation des expressions de ce langage, pour produire par exemple un affichage (cas du langage HTML), un document PDF (cas du langage LaTeX) ou encore une ressource linguistique particulière (cas des langages PATRII/DATR). Il convient de noter que les langages PATRII/DATR étaient assez limités en termes d'expressivité, n'offrant que peu de moyens de définir des abstractions (seules des macros ou des règles de transformation étaient possibles).

Le concept de métagrammaire est apparu quant à lui à la fin des années 90, dans les travaux de Candito (1999). La métagrammaire correspondait alors à la description d'une grammaire d'arbres adjoints (Joshi *et al.*, 1975) des verbes en français et italien. Cette description reposait sur une analyse tridimensionnelle de la langue. Dans une première dimension était décrite une hiérarchie de cadres de sous-catégorisation pour les prédicats verbaux (représentation de la valence, l'ordre et la catégorie des arguments du verbe), dans une deuxième dimension des règles de transformation (redistributions des fonctions grammaticales comme par exemple lors du passage de l'actif au passif) et enfin, en troisième dimension, une hiérarchie de réalisations syntaxiques pour les différentes fonctions grammaticales présentes dans chacun des cadres de sous-catégorisation considérés. La motivation à ces travaux était de contrôler la richesse syntaxique des formalismes grammaticaux fortement lexicalisés (comme le sont les grammaires d'arbres adjoints) au moyen d'un outil descriptif à la fois formel et linguistiquement motivé (ici par la structure tridimensionnelle).

Ces travaux ont marqué le début d'une lignée de recherches sur les langages de description pour les ressources linguistiques, notamment les grammaires d'arbres. Les limitations des travaux de Candito résidaient dans (i) la structure tridimensionnelle rigide de la métagrammaire, et (ii) l'utilisation de formules logiques de description d'arbres dont les variables avaient une portée globale à la métagrammaire. [5] Les principales alternatives à l'approche de Candito correspondent aux systèmes LexOrg (Xia, 2001) (qui propose d'utiliser des variables locales dans les formules logiques de description d'arbres), et FRMG (Villemonte De La Clergerie, 2010) et XMG (Crabbé *et al.*, 2013) (permettant tous deux de définir finement la portée des variables utilisées dans les formules de description). C'est ce dernier système XMG qui a servi d'inspiration au développement du système XMG2, que nous allons présenter dans la section suivante. À la différence du système XMG, qui *ne* permet *que* de compiler des grammaires décrites au moyen du langage de description XMG, le système XMG2 permet (1) de définir modulairement un langage de description, et (2) de compiler un compilateur pour ce langage (qui sera ensuite utilisé par le linguiste pour décrire et compiler sa ressource linguistique).

3 Le système XMG2

Définition modulaire d'un langage de description. Comme nous l'avons mentionné précédemment, le système XMG2 [6] permet de compiler un compilateur pour divers langages de description. Ces

5. Un identifiant de variable dénotant une information linguistique donnée ne pouvait plus être réutilisé ailleurs dans la description pour référer à une autre information, ce qui revient à gérer un ensemble d'identifiants de variables très grand.

6. `http://dokufarm.phil.hhu.de/xmg/?animal=xmg`

langages doivent au préalable avoir été définis formellement. Pour cela, XMG2 offre un ensemble de langages élémentaires (appelés *briques de langage* dans la terminologie XMG2) qui peuvent être assemblés déclarativement (Petitjean *et al.*, 2016). Les briques de langages actuellement disponibles nativement dans XMG2 incluent notamment :

– B1 un langage de description de structures de traits,
– B2 un langage de description d'arbres à base de dominance/précédence entre nœuds (Rogers & Vijay-Shanker, 1994),
– B3 un langage de description de formules sémantiques « plates » (Bos, 1995),
– B4 un langage de description de *frames* sémantiques (Lichte & Petitjean, 2015).

Un assemblage de ces briques prend la forme d'un fichier texte au format YAML (liste de clés-valeurs) comme décrit dans (Petitjean, 2014; Petitjean *et al.*, 2016). Il est ainsi possible par exemple d'assembler les briques de langage B1 et B2 ci-dessus, et à partir de cet assemblage, de générer un compilateur pour un langage de description permettant de décrire des arbres syntaxiques dont les nœuds utiliseraient des structures de traits comme étiquettes (cf exemple de Petitjean *et al.* (2016)).

Utilisation d'un langage de description. À partir de la définition formelle d'un langage de description L par assemblage de briques de langage, le système XMG2 génère (méta-compile) un compilateur pour ce langage. Ce dernier peut alors être utilisé par un expert linguiste pour sa tâche de description des unités d'une langue donnée. En d'autres termes, cet expert va écrire une métagrammaire au moyen du langage L, et cette métagrammaire sera ensuite compilée par le compilateur précédemment généré pour produire les unités de la langue (c'est-à-dire la ressource linguistique), comme illustré sur la Figure 1.

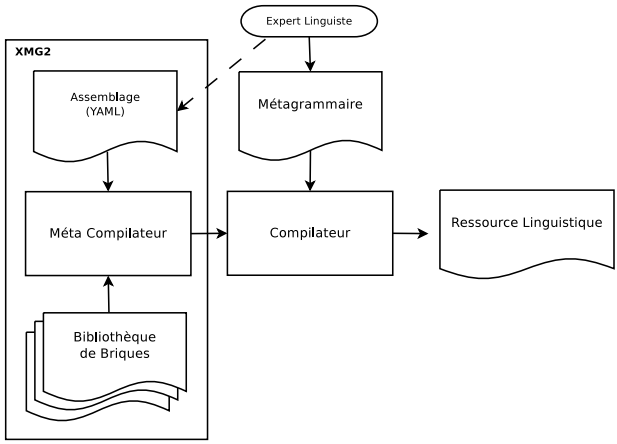


FIGURE 1 – Utilisation du système XMG2

Dans l'exemple de l'assemblage B1-B2 ci-dessus, une fois le compilateur pour le langage L_{B1-B2} généré, il peut être utilisé par un expert linguiste pour compiler une métagrammaire écrite dans le langage L_{B1-B2} et ainsi produire une ressource électronique (au format XML). Les éléments de cette ressource sont des arbres dont les nœuds sont équipés de structures de traits (règles d'une grammaire lexicale fonctionnelle ou d'une grammaire d'arbres adjoints par exemple).

À propos du logiciel XMG2. Il convient de noter que le logiciel XMG2 est distribué sous licence GPL, ce qui signifie entre autres que (1) son code source est accessible librement, et (2) qu'il est possible de le télécharger et de l'utiliser gratuitement.[7] En outre, les formats manipulés par le système XMG2 (format utilisé pour la définition d'un assemblage de briques, format utilisé pour la métagrammaire, format XML utilisé pour représenter la ressource linguistique électronique produite) sont tous ouverts (leur spécification est disponible librement), ce qui facilite la réutilisation de ces ressources. Ces formats pourraient constituer une norme de répresentation de ressources linguistiques *de facto*, si par exemple une communauté d'experts se mettait à les utiliser pour représenter leurs propres ressources (que celles-ci aient été produites par le système XMG2 ou un autre).

4 Application aux langues peu dotées

Nous allons ici illustrer l'utilisation de langages de description (et du système XMG2) pour produire deux types de ressources électroniques pour des langues peu dotées. Dans un premier temps, nous allons montrer comment décrire un lexique morphologique de l'ikota (langue bantoue), et ensuite comment décrire une grammaire d'arbres du são-tomense.

4.1 Décrire un lexique nominal de l'ikota

L'ikota est une langue bantoue parlée principalement au Gabon (où le nombre de locuteurs est estimé à 25000) et en République Démocratique du Congo. C'est une langue tonale à deux tons qui compte dix classes nominales et possède un accord généralisé dans le syntagme nominal. Elle compte en outre 3 classes verbales, et un infinitif où l'élément verbal est préfixé par une classe nominale.

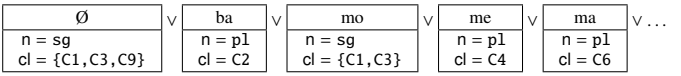
Notre modélisation linguistique se base sur les travaux de Duchier *et al.* (2012), qui utilisaient déjà le concept de métagrammaire (notamment le langage XMG), pour décrire la morphologie verbale de l'ikota. Ici, nous décrivons les formes nominales en ikota comme composées d'une racine nominale (RN) précédée d'un préfixe défini en fonction de la classe nominale de la racine considérée, comme indiqué ci-dessous :

classe nominale	préfixe
CL 1	mò-, Ø-
CL 2	bà-
CL 3	mò-, Ø-
CL 4	mè-
CL 5	ì-, ʤ-
CL 6	mà-
CL 7	è-
CL 8	bè-
CL 9	Ø-
CL 14	ò-, bò-

Nous allons utiliser cette modélisation pour créer un lexique noyaux de formes nominales de l'ikota (lexique qui permettra entre autre de confronter cette modélisation linguistique aux données de terrain).

7. Plusieurs compilateurs pour langages de description (dont L_{B1-B2}) ont été pré-méta-compilés et sont disponibles en ligne à l'adresse `http://xmg.phil.hhu.de/index.php/upload/compile_grammar`.

Pour ce faire, nous allons définir des blocs élémentaires (appelées *classes* dans la terminologie XMG2), et contenant des contributions pour les préfixes définis ci-dessus et pour les racines nominales. Les valeurs possibles pour les préfixes (et les traits associés) sont définis dans une classe *Prefixe* :



Notons que cette description des préfixes peut s'écrire directement au moyen de la brique de langage `tf_morph` disponible dans le logiciel XMG2, comme suit (le symbole `|` représente la disjonction) :

```
class Prefixe
{
        <morph>{
        {
        { n=sg; cl=@{C1,C3,C9}; prefix <- ""}
        |
        { n=pl; cl=C2; prefix <- "ba"}
        |
        { n=sg; cl=@{C1,C3}; prefix <- "mo"}
        |
        { n=pl; cl=C4; prefix <- "me"}
        |
        { n=pl; cl=C6; prefix <- "ma"}
        |
        { n=sg; cl=C5; prefix <- "dz"}
        |
        { n=sg; cl=C7; prefix <- "e"}
        |
        { n=pl; cl=C8; prefix <- "be"}
        |
        { n=sg; cl=C14; prefix <- "bo"}
        }
    }
}
```

De manière similaire, nous pouvons définir des classes pour décrire un nom comme appartenant à une certaine classe nominale au singulier *ou* à une autre au pluriel.

Nous pouvons ensuite définir les racines nominales (RN) considérées en indiquant à chaque fois leur classe nominale, comme par exemple « mbòka » (village, classe 14) :

$$RN \quad \rightarrow \quad (mbòka \wedge \mathsf{cl{=}C14}) \vee \ldots$$

Enfin, un nom sera décrit comme correspondant à la conjonction d'un préfixe et d'un racine nominale (avec la contrainte additionnelle que les traits associés à chacune de ces contributions doivent pouvoir s'unifier) :

$$Nom \quad \rightarrow \quad Prefixe \wedge RN$$

Toutes les combinaisons préfixe-RN sont calculées par le compilateur et seules celles étant valides (traits unifiables) sont conservées. Le lexique ainsi produit est en adéquation avec les observations de terrain (Magnana Ekoukou, 2015).

On remarque que la structure de la métagrammaire est très proche de la modélisation linguistique. Cela est rendu possible par l'utilisation d'un langage de description permettant de définir des alternatives de valeurs pour des champs ordonnés linéairement. En d'autres termes, l'utilisation d'un langage de description configurable (ici, par assemblage modulaire) permet d'avoir une métagrammaire reflétant la modélisation linguistique et constituant de ce fait en elle-même une documentation de la ressource (dans notre exemple, elle contient une motivation à la structure interne des noms).

Le lexique décrit ici permet de générer quelques formes fléchies par racine nominale. La description peut paraître verbeuse, mais l'ajout de nouvelles racines étant limité à la classe *RN*, l'extension de la ressource est grandement facilitée. On peut ainsi passer rapidement à une échelle supérieure et vérifier empiriquement la capacité du modèle linguistique à prédire la structure des noms en ikota. De plus, le lexique ainsi généré, dont les formes fléchies sont étiquettées avec des traits morpho-syntaxiques, pourrait ensuite servir au développement d'outils comme par exemple un analyseur morphologique.

4.2 Décrire une grammaire du são-tomense

Le são-tomense (appelé aussi forro) est un créole du portugais parlé sur l'île de São-Tomé. Comme de nombreux créoles, le são-tomense contient des marqueurs pré-verbaux exprimant le temps, l'aspect et le mode (marqueurs TMA). Ces marqueurs ne sont présents que dans certains ordres. Notre modélisation linguistique correspond aux travaux de Schang *et al.* (2012), et traite ces marqueurs comme des projections verbales.

Nous allons ainsi décrire des règles grammaticales pour les verbes en são-tomense prenant la forme d'arbres élémentaires, et contenant des nœuds pour les différentes combinaisons de marqueurs autorisées. Un exemple de combinaison autorisée est donné en Figure 2.

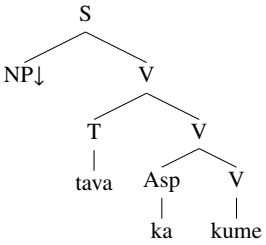


FIGURE 2 – Arbre élémentaire pour le verbe *manger* au passé progressif

Pour représenter les différents marqueurs, nous définissons des classes métagrammaticales (descriptions de fragments d'arbres) pour chacun d'eux (cf fragments (b) et (c) en Figure 3). Afin de contrôler les combinaisons entre ces fragments, nous étiquettons les nœuds V d'un trait proj(ection) restreignant les valeurs autorisées. On ajoute de plus un fragment pour le sujet canonique (cf fragment (a)) et pour la racine verbale (cf fragment (d)). Le compilateur va ensuite chercher les combinaisons autorisées de ces fragments (c'est-à-dire chercher tous les modèles d'arbres contenant les fragments en question) et ne conserver que l'arbre souhaité.

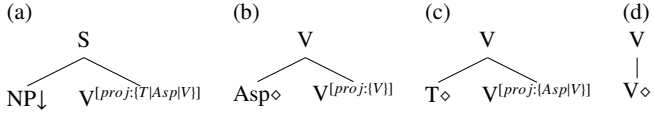


FIGURE 3 – Classes métagrammaticales pour les marqueurs TMA en são-tomense.

Pour vérifier cette modélisation, et produire une grammaire exemple du são-tomense, nous pouvons utiliser le langage de description `synsem` inclus dans XMG2 (assemblage des briques B1 et B2 vues précédemment). Ce langage permet de décrire des arbres syntaxiques dont les nœuds contiennent des structures de traits. Par exemple, le fragment d'arbre (a) s'écrit comme suit dans le langage `synsem` :

```
class SujetCan
export ?X ?Y ?Z
declare ?X ?Y ?Z
{ <syn>{
        node ?X [cat = s, proj = v]{
                node ?Y (mark=subst)[cat = np]     node ?Z [cat = v, proj = @{t,asp,v}]
                }
        }
}
```

L'expression `node ?X { node ?Y }` représente une formule de logique de description d'arbre contenant deux variables de nœuds identifiées par `?X` et `?Y`, et telles que le nœud dénoté par la variable `?X` domine directement celui dénoté par la variable `?Y`. L'expression `node ?X node ?Y` représente deux noeuds frères ordonnés (précédence). Les nœuds ainsi dénotés contiennent en outre une catégorie syntaxique et un trait proj dont la valeur peut être une disjonction (symbole `@`) de valeurs constantes. Les autres fragments sont définis de manière similaire. Une classe `Intransitif` est finalement définie comme la conjonction des classes `SujetCan`, `Aspect`, `Temps` et `RV`. On indique alors au compilateur qu'on souhaite évaluer la classe `Intransitif` et ce dernier produira le (ou les) arbre(s) correspondant(s) au format XML.

Cette illustration de l'utilisation d'un langage de description pour décrire une grammaire d'arbres est limitée. On peut là aussi se demander si écrire 4 classes pour produire un arbre est un gain. L'idée est que la description peut être facilement étendue et que les fragments peuvent être réutilisés dans divers contextes. On a ainsi un gain qui augmente au cours du développement de la ressource (et la maintenance est facilitée puisque l'information y est factorisée). Dans le cas du français, la métagrammaire de Crabbé *et al.* (2013) décrit 6000 schémas d'arbre non lexicalisés à partir de 293 classes. Comme l'ont de plus montré Crabbé *et al.* (2013), le développement d'une grammaire couvrante s'accompagne d'une méthodologie de conception. En suivant cette méthodologie, la métagrammaire est structurée hiérarchiquement, et cette hiérarchie (qui capture des généralisations linguistiques) peut être vue comme une documentation de la syntaxe de la langue décrite.

5 Conclusion et travaux futurs

Dans cet article, nous avons présenté comment les langages de description peuvent être utilisés pour développer des ressources linguistiques à faible coût, s'avérant ainsi particulièrement utiles pour tester des modélisation linguistiques et construire des ressources noyaux pour des langues peu dotées. Nous avons également vu comment le système XMG2 permet une définition modulaire de langages par assemblage de briques élémentaires. Les compilateurs pour ces langages sont alors générés automatiquement et utilisables directement pour des tâches de descriptions linguistiques. Enfin, nous avons illustré concrètement l'utilisation de langages de description pour produire deux ressources exemples pour deux langues sous-dotées, l'ikota et le são-tomense.

Les travaux autour des métagrammaires ne sont pas nouveaux (des métagrammaires pour grammaires d'arbres du français, de l'anglais et de l'allemand existent depuis près d'une dizaine d'années), cependant les avancées récentes autour de la définition modulaire de langages de description ouvre la voie à des utilisations personnalisables en fonction de la langue et de la modélisation linguistique choisies. Des travaux sont en cours pour décrire d'autres langues (telles que l'arabe ou le guadeloupéen), et d'autres niveaux linguistiques (sémantique notamment). Cela passe par la définition de nouvelles briques, et aussi de procédés de résolution des descriptions pour produire les unités voulues.

Références

Abeillé A., Candito M. & Kinyon A. (1999). FTAG : current status and parsing scheme. In *Proceedings of Vextal '99*, p. 283–292, Venice, Italy.

Bos J. (1995). Predicate Logic Unplugged. In *Proceedings of the tenth Amsterdam Colloquium, Amsterdam*.

Candito M. (1999). *Organisation Modulaire et Paramétrable de Grammaires Electroniques Lexicalisées*. PhD thesis, Université Paris 7.

Chomsky N. (1957). *Syntactic Structures*. The Hague : Mouton.

Crabbé B., Duchier D., Gardent C., Le Roux J. & Parmentier Y. (2013). XMG : eXtensible MetaGrammar. *Computational Linguistics*, **39**(3), 591–629.

Duchier D., Magnana Ekoukou B., Parmentier Y., Petitjean S. & Schang E. (2012). Décrire la morphologie des verbes en ikota au moyen d'une métagrammaire. In *19e conférence sur le Traitement Automatique des Langues Naturelles (TALN 2012) - Atelier sur le traitement automatique des langues africaines (TALAf 2012)*, p. 97–106, Grenoble, France.

Evans R. & Gazdar G. (1996). DATR : a language for lexical knowledge representation. *Computational Linguistics*, **22**(2), 167–216.

Joshi A. K., Levy L. S. & Takahashi M. (1975). Tree Adjunct Grammars. *Journal of the Computer and System Sciences*, **10**, 136–163.

Lichte T. & Petitjean S. (2015). Implementing semantic frames as typed feature structures with XMG. *Journal of Language Modelling*, **3**(1), 185–228.

Magnana Ekoukou B. (2015). *Description de l'Ikota (B25), langue bantu du Gabon. Implémentation de la morphosyntaxe et de la syntaxe*. PhD thesis, Université d'Orléans, France.

Petitjean S. (2014). *Génération Modulaire de Grammaires Formelles*. PhD thesis, Université d'Orléans, France.

Petitjean S., Duchier D. & Parmentier Y. (2016). XMG2 : Describing Description Languages. In *Logical Aspects of Computational Linguistics (LACL 2016)*, volume 10054 of *Lecture Notes in Computer Science*, p. 255–272, Nancy, France : Springer-Verlag.

Prolo C. A. (2002). Generating the XTAG English Grammar Using Metarules. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING'2002)*, p. 814–820, Taipei, Taiwan.

Rogers J. & Vijay-Shanker K. (1994). Obtaining trees from their descriptions : An application to tree-adjoining grammars. *Computational Intelligence, 10 :401–421*.

Sagot B., Clément L., De La Clergerie É. & Boullier P. (2006). The Lefff 2 syntactic lexicon for French : architecture, acquisition, use. In *LREC 06*, p. 1–4, Gênes, Italy.

Schang E., Duchier D., Magnana Ekoukou B., Parmentier Y. & Petitjean S. (2012). Describing São Tomense Using a Tree-Adjoining Meta-Grammar. In *11th International Workshop on Tree Adjoining Grammars and Related Formalisms (TAG+11)*, p. 82–89, Paris, France.

Shieber S. M. (1984). The design of a computer language for linguistic information. *COLING-84*, p. 362–366.

Villemonte De La Clergerie É. (2010). Building factorized TAGs with meta-grammars. In *TAG+10*, p. 111–118, New Haven, CO, United States.

Xia F. (2001). *Automatic Grammar Generation from two Different Perspectives*. PhD thesis, University of Pennsylvania.

Retour d'expérience : l'utilisation de l'apprentissage profond (deep learning) dans le contexte de l'analyse sémantique des langues peu dotées

Hammou Fadili[1, 2]
(1) Laboratoire CEDRIC du Conservatoire National des Arts et Métiers de Paris
192, rue Saint Martin, 75141, Paris cedex 3, France
(2) Pôle Systèmes d'Information et du Numérique, Programme Maghreb de la FMSH Paris
190, avenue de France 75013, Paris, France
Hammou.fadili@cnam.fr / fadili@msh-paris.fr

Résumé

On estime à plusieurs milliers le nombre de langues parlées dans le monde et seulement quelques dizaines disposent de ressources (informatiques, textuelles, etc.) permettant leur traitement automatique. Celles ne disposant pas ou disposant de peu de ressources sont appelées langues peu dotées (LPD). Plusieurs rapports de l'UNESCO a rment que la plupart des langues peu dotées sont en voie de disparition. De plus, plusieurs spécialistes des langues, estiment que leur disparition est accélérée par les phénomènes informatiques (internet, réseaux sociaux, etc.) qui les marginalisent encore plus. Cependant, d'après les mêmes spécialistes, l'intégration des langues peu dotées dans le monde des nouvelles technologies pourrait constituer une opportunité pour leur développement, leur sauvegarde et donc pour leur survie. En e et, mettre à disposition des utilisateurs des outils les incitant à la découverte et à la création dans les LPD, aidées par des passerelles avec d'autres langues mieux dotées (LMD), telles que les fonctionnalités : de liens, d'alignement, de traduction, d'analyse et de synthèse, etc. pourrait avoir un impact positif sur la popularité de leur utilisation et par conséquent sur leur développement. Dans cet article, nous présentons une expérience exploitant les nouvelles technologies d'apprentissage profond dans le contexte de l'analyse sémantique des langues peu dotées. Le but est de montrer à travers un exemple d'approche qu'on peut exploiter certaines technologies facilement adaptables aux langues soufrant du manque de ressources en termes de contenus et d'outils informatiques ; espérant que cela pourra, en plus, aider à sensibiliser et à inciter les chercheurs du domaine à proposer des solutions génériques intégrant dans leur conception le support des LPD.

Abstract

Feedback : use of deep learning in the context of poorly endowed languages.

It is estimated that there are several thousand languages spoken in the world and only a few dozen have resources (tools, corpuses, annotations, etc.) for automatic processing. Those with little or no resources are called poorly endowed languages (PEL). Several UNESCO reports state that most

poorly endowed languages are endangered. In addition, several language specialists believe that their disappearance is accelerated by the phenomena of new technologies (Internet, social networks, etc.) which further marginalize them. However, according to the same specialists, the integration of poorly endowed languages into the world of new technologies could constitute an opportunity for their development, their preservation and therefore for their survival. Indeed, make available to the users, tools encouraging them to discover and to create in the PEL, helped by bridges with better endowed languages (BEL), such as functionalities of: links, alignment, Translation, analysis, etc. Could have a positive impact on the popularity of their use and consequently on their development. In this article, we present an experiment exploiting the new technologies of deep learning in the context of the semantic analysis of PEL. The aim is to show through an example of approach that we can exploit certain technologies that are easily adaptable to languages su ering from the lack of resources in terms of content and computer tools; Hoping that it will also help to raise awareness and encourage researchers in the eld to propose generic solutions integrating in their design the support of LPD.

MOTS-CLÉS : Langues peu dotées, apprentissage profond, modèles de langue, sémantique, représentations vectorielles des mots.

KEYWORDS: Poorly endowed languages, deep learning, language model, semantics, word embeddings.

1 Introduction

On estime à plusieurs milliers le nombre de langues parlées dans le monde et seulement quelques dizaines disposent de ressources (informatiques, textuelles, etc.) permettant leur traitement automatique. Celles ne disposant pas ou disposant de peu de ressources sont appelées langues peu dotées (LPD). Plusieurs rapports de l'UNESCO a rment que la plupart des langues peu dotées sont en voie de disparition. De plus, plusieurs spécialistes des langues, estiment que leur disparition est accélérée par les phénomènes informatiques (internet, réseaux sociaux, etc.) qui les marginalisent encore plus. Cependant, d'après les mêmes spécialistes, l'intégration des langues peu dotées dans le monde des nouvelles technologies pourrait constituer une opportunité pour leur développement, leur sauvegarde et donc pour leur survie. En e et, mettre à disposition des utilisateurs des outils les incitant à la découverte et à la création dans les LPD, aidées par des passerelles avec d'autres langues mieux dotées (LMD), telles que les fonctionnalités : de liens, d'alignement, de traduction, d'analyse et de synthèse, etc. pourrait avoir un impact positif sur la popularité de leur utilisation et par conséquent sur leur développement.

C'est dans ce sens que nous avons mené une expérience exploitant les nouvelles technologies d'apprentissage profond pour le traitement automatique d'un cas d'une langue peu dotée. Nous avons essayé à travers ce travail de privilégier des méthodes basées sur des apprentissages non/peu supervisées et des méthodes statistiques capables d'e ectuer des traitements sur des données brutes n'ayant subi aucun traitement au préalable. Le but est de montrer la faisabilité de contourner les

problèmes liés aux manques de données structurées, annotées, etc., d'outils et de règles de traitements, dont sou rent les LPD. Nous espérons, en plus, à travers ces expériences, aider à sensibiliser la communauté travaillant dans le domaine du TALN en général, et les inciter à proposer des approches et outils génériques, réutilisables et applicables dans le contexte de n'importe quelle langue y compris celles peu dotées. Cet article est organisé comme suivant, la première partie est consacrée à un bref rappel sur les langues peu dotée. La deuxième partie rappelle quelques éléments technologiques retenus en termes de modèle de langue, de modèles de données et de traitements. La partie suivante est consacrée à la description de l'ensemble des éléments au sein d'une architecture adaptée, à leur fonctionnement et aux résultats des tests menés. La dernière partie conclue le présent article.

2 Contexte des langues peu dotées

D'une manière générale, les langues peu dotées, sont des langues qui sou rent de plusieurs problèmes : problèmes liés à la graphie, au manque d'un système d'écriture stable, au manque de ressources informatiques et linguistiques. Le manque de ressources langagières concerne les dictionnaires, thésaurus, corpus traités, etc. ; le manque d'outils numériques concerne les outils du traitement automatique de la langue naturelle : analyseurs morphologiques, syntaxiques, sémantique, etc. Tous ces éléments rendant di cile, voire impossible l'analyse sémantique des langues peu dotées, peuvent être classés suivant les 3 axes ci-après :

Modélisation & modèle de langue : la langue naturelle est très complexe en général. Ceci est dû au nombre important (presque in ni) de cas possibles d'utilisation, d'exceptions et de règles, etc. La modélisation des caractéristiques de la langue naturelle est une grande problématique, d'une manière générale ; problématique encore plus accentuée dans le cas des LPD.

Connaissances & données : un autre problème dont sou rent les LPD concerne les données prétraitées qui sont d'une grande nécessité dans les systèmes de traitements des données. Elles sont utilisées pour instancier les modèles des données et pour exprimer les règles de raisonnement et d'apprentissage. Le manque de ce type de données, dans le cas des LPD constitue un frein majeur pour leur exploitation automatique.

Outils : les outils informatiques nécessaires à l'automatisation des tâches des traitements, bien adaptés au LPD sont également rares.

A n de contourner ces problèmes, nous avons expérimenté et réutilisé une solution ayant fait ses preuves dans le cas d'autres langues mieux dotées (l'Anglais et le Français,). Elle a l'avantage d'être indépendante ou plutôt peu dépendante de la langue traitée.

3 Démarche expérimentale

Cette partie a pour but de décrire l'expérience menée, en étudiant le comportement sur une LPD, en l'occurrence le Berbère, de l'approche mise en place pour le Français et l'Anglais. Cette approche étant générique, i.e. pas ou peu dépendante de la langue traitée, nous a permis de contourner certains problèmes liés aux LPD, décrits précédemment. On y exploite des apprentissages non supervisés ou peu supervisés des modèles de représentation et de traitements pour le traitement automatique de la sémantique du texte comme la désambiguïsation.

3.1 Modèle de langue

Pour le choix et la représentation du modèle de langue, nous avons besoin de résoudre et de contourner certaines di cultés : celles relatives à la modélisation et à la formalisation de la langue naturelle, puis celles relatives aux choix technologiques pour les représenter et aux réalisations informatiques pouvant les supporter. Dans le premier cas, nous avons à modéliser deux notions importantes du domaine de l'analyse des données non structurées, à savoir la notion de contexte et la notion des relations sémantiques a n de mieux caractériser la sémantique. Ces notions déjà modélisées pour d'autres LMD ont été validées et adaptées pour le cas du Berbère ; grâce au concourt des linguistiques du domaine qui nous ont aidé à adapter le modèle et créer un modèle générique de langue.

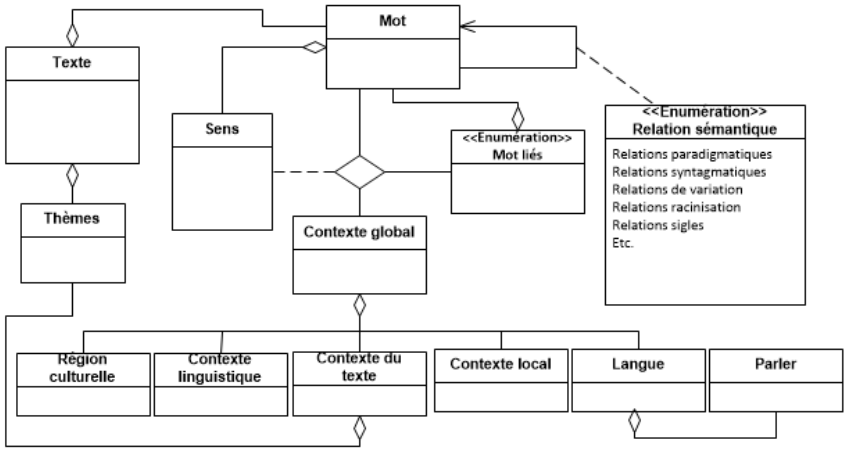


FIGURE 1: Modèle pour l'apprentissage

Sur le plan technologique, plusieurs problèmes devaient être également pris en compte et résolus. Cela concerne principalement les choix technologiques et applicatifs pouvant d'une part supporter le modèle de langue et d'autre part optimiser et simpli er les traitements. Nous avons évité

d'utiliser les modèles technologiques de langue basés sur des représentations vectorielles « sparses » de type sac de mots classique, représentant les mots comme des vecteurs dans des espaces vectoriels de très grandes dimensions (taille de tout le vocabulaire) ; di cile à mettre en place. Pour cela, nous avons fait le choix d'adopter les nouvelles représentations vectorielles « denses » de type « Word embeddings » et son implémentation « word2vec ». Pour cela, nous avons utilisé, les 2 implémentations de Word2vec : Skip-gram (Mikolov et al. 2013a) et CBOW (Mikolov et al. 2013b). Le but est d'entrainer un réseau de neurones profond pour obtenir une représentation vectorielle sémantique réduite de chaque mot à partir de sa représentation initiale et ses contextes locaux.

Dans le cas de Skip-Gram :

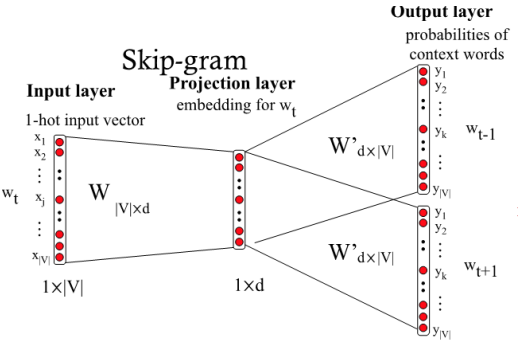


FIGURE 2: Skip-gram

L'apprentissage se fait en maximisant la fonction objective des Log probabilités.

La normalisation des probabilités se fait par un Softmax.

$$\Sigma_{t=1}^{T}\Sigma_{j=-c,j\neq 0}^{c}\log P(w^{t+j}\mid w^{t})$$

$$P(w^{t+j}\mid w^{t})=\frac{\exp(\mathbf{v}_{w^{t+j}}\cdot\mathbf{v}_{w^{t}})}{\Sigma_{i=1}^{V}\exp(\mathbf{v}_{w_i}\cdot\mathbf{v}_{w^{t}})}$$

k est la taille du contexte local

Dans le cas de CBOW (Continuous Bag Of Words) :

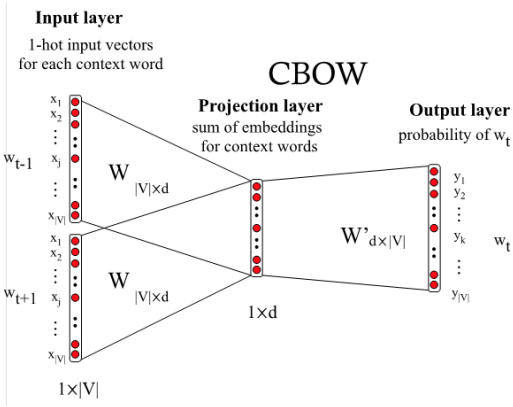


FIGURE 3: CBOW

L'apprentissage se fait en maximisant la fonction objective des Log probabilités.

La normalisation des probabilités se fait par un Softmax.

$$\Sigma_{t=1}^{T} \log P(w^t \mid w^{t-c},...,w^{t-1},w^{t+1},...,w^{t+c})$$

$$P(w^t \mid w^{t-c},...,w^{t-1},w^{t+1},...,w^{t+c}) = \frac{\exp(\mathbf{v}_{\overline{w}^t} \cdot \mathbf{v}_{w^t})}{\Sigma_{i=1}^{V} \exp(\mathbf{v}_{\overline{w}^t} \cdot \mathbf{v}_{w_i})}$$

$\mathbf{v}_{\overline{w}^t}$ la moyenne (pondérée) des représentations vectorielles des mots contextuels de w^t

Cette technologie permet de coder l'historique des utilisations d'un mot dans un vecteur dense d'environ 300 dimensions. Il s'agit d'apprendre une projection d'un espace vectoriel initial « sparce » dans un espace vectoriel « sémantique » et « dense » représentant toutes utilisations passées a n de prédire les utilisations futures. Cette technologie a été testée avec succès, bien adaptée pour toutes les langues. Dans notre cas, la technologie « Word embeddings » a été exploitée pour instancier et représenter notre modèle sémantique d'apprentissage. Cela correspond à une extension et un enrichissement de la technologie « Word embeddings » initiale. Ces travaux nous ont permis, d'une part, une formalisation approfondie et générique et une prise en compte de certaines caractéristiques importantes de la langue naturelle, pouvant in uencer le sens ; et d'autre part, l'exploitation de certaines technologies de pointe sur lesquelles nous avons apporté des améliorations, a n de faire progresser certains résultats dans ce domaine en général et dans le contexte des LPD en particulier. En l'appliquant à un cas concret d'une langue, nous avons pu, montrer l'apport considérable de ces technologies pour le traitement automatique des LPD.

3.2 Génération d'instances pour l'apprentissage

Cette partie est consacrée à la présentation des approches statistiques basées sur des apprentissages non supervisés pour l'extraction des caractéristiques sémantiques et des thèmes latents du texte. Ceci a n de représenter convenablement les mots du texte et aussi pour compléter l'instanciation du modèle sémantique étendu de langue pour l'apprentissage.

Pour cela, nous avons exploité 3 types d'outils et de méthodes :

- Modèles neuronaux
    - Représenter vectoriellement les mots
    - Word2vec - Skip-grams, CBOW (Mikolov, 2013)
- Outils de TALN
    - Extraction du contexte local
    - Extraction du contexte linguistique
- Modèles de groupements thématiques
    - Extraire les thèmes latents
    - Latent Dirichlet allocation - LDA (Blei, 2009)

Des éléments du contexte tels que la langue, le parler, etc., sont instanciés à la main une seule fois avant l'application du processus. Automatiser cette tâche est possible, car des travaux sur la détection automatique de la langue et des parlers berbères existent, on peut citer par exemple ceux de (W. Adouane et al., 2016).

Dans le cas des LPD ne disposant pas de Wordnet ou équivalent, nous avons adopté, contrairement à l'approche initiale, une méthode de « groupement » ou « Clustering » non supervisée pour le calcul des classes des sens. Nous avons testé et exploité la méthode de (Schütze, 1998), qui est une méthode de désambiguïsation automatique et non-supervisée basée sur les résultats d'une « clustérisation » des données et qui prend en compte les cooccurrences du second ordre dans la représentation du contexte des mots.

La combinaison de toutes les caractéristiques permet la représentation vectorielle du modèle sémantique pour l'apprentissage (contexte local, contexte linguistique, thématiques et domaines traités (contexte global)).

3.3 Outils

Sur le plan des outils, plusieurs plateformes (Weka, Rapidminer, Orange, ...) et classi eurs (Bayes, Arbres de décision, SVM, Réseaux de neurones…) ont été testés avant de faire le choix d'exploiter un réseau de neurones profond pour le Traitement Automatique de la Langue Naturelle (TALN). Il su t que l'UNICODE soit supporté, pour pouvoir exploiter ces plateformes et ces technologies, ce qui est le cas du berbère.

4 Architecture

Nous avons encapsulé un certain nombre d'éléments, décrits dans ce document, dans un processus d'analyse sophistiqué permettant l'extraction automatique des caractéristiques sémantiques du modèle sémantique retenu, comme source d'entrée pour un réseau de neurone profond. Ceci a n de faciliter le processus d'analyse, d'interprétation et d'exploitation sémantiques automatiques des données non structurées.

4.1 Apprentissage profond (Deep learning) pour le TALN des LPD

Notre contribution initiale, en plus d'être basée sur des recherches améliorant les modèles de langues existantes, permettant de caractériser au maximum les textes, leurs mots, leurs utilisations et leurs sens, et par conséquent permettant d'améliorer leurs analyses et leurs interprétations ; est basée sur l'apprentissage profond des modèles étendus de représentation et de traitement de la sémantique des textes. Ces architectures profondes sont des réseaux comportant plusieurs couches, pouvant modéliser avec un haut niveau d'abstraction des modèles de données, articulés autour de transformations non linéaires. Dans cette partie, nous avons exploité la modularité des architectures profondes pour mieux les adapter au contexte des LPD. Par exemple, l'emplacement des modèles symboliques dans le processus général, était problématique ; car ils sont di ciles à mettre en place, surtout dans le cas des langues peu dotées. Un texte doit y être représenté par les mots le constituant ainsi que par les propriétés issues des di érents traitements linguistiques ou des formalisations spéci ques telles les ontologies, faisant défaut dans le cas des LPD (cf. paragraphe ci-dessus).

Pour contourner ces di cultés, nous avons privilégié l'intégration des modèles numériques statistiques en amont du processus. Ces modèles sont basés sur des calculs mathématiques dans des espaces sémantiques optimisés, bien adaptés. L'avantage est qu'on n'a pas besoin d'informations (pré)-traitées au préalable pour modéliser et décrire ce type de modèles ; on y exploite directement les données à analyser et les modèles mathématiques pour déduire les modèles de représentations.

4.2 Architecture

Concrètement, les tests ont été réalisés grâce à un réseau de neurones multicouches, permettant des apprentissages mixtes comme suivant :

- les premières couches ou couches basses ont été consacrées à l'extraction des informations latentes (features). Elles utilisent des apprentissages non supervisés, exploitant des approches numériques basées sur les fréquences et les agencements des mots, et sur les modèles mathématiques statistiques et probabilistes, pour le codage de la sémantique

- les dernières couches ont été consacrées à la prise de décision. Elles utilisent des apprentissages supervisées, exploitant des approches symboliques basées sur les structures des données, leur description ainsi que sur des systèmes de règles, générées dans les couches précédentes, pour e ectuer les traitements

Cette démarche a l'avantage de combiner les deux approches, à des degrés di érents, et avec des ordonnancements spéci ques. Son originalité réside dans le fait que, d'une part, elle intègre des notions importantes dans le modèle de langue : la notion de contexte global et la notion des relations sémantiques ; d'autre part, elle exploite ces notions pour générer un modèle sémantique riche de données sur lequel on applique des algorithmes d'apprentissage.

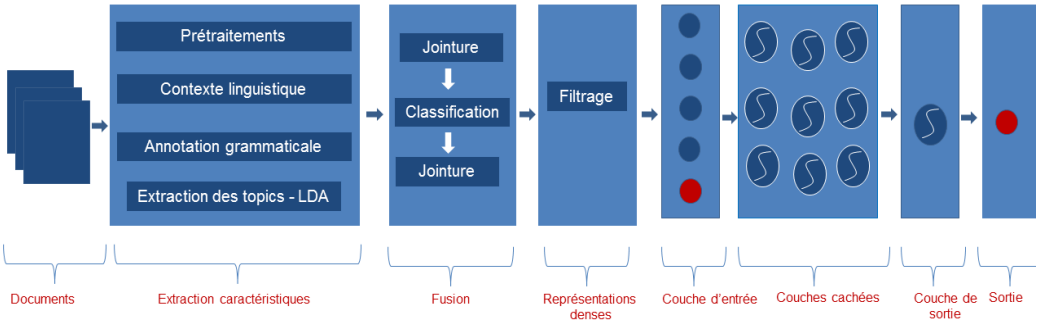


FIGURE 4: Réseaux de neurones multicouches

5 Tests

Nous avons développé et testé des modules permettant d'extraire les caractéristiques sémantiques et les instances dans le format « .ar » supporté par la plateforme d'apprentissage Weka. Nous avons séparé les données générées en trois parties comme suivant :

- 20% des données pour la validation, ceci a n d'optimiser les hyper-paramètres du système : le pas d'apprentissage, le type de la fonction d'activation et le nombre de couches

- 60% pour l'entrainement, ceci a n d'estimer les meilleurs coe cients (wi) de la fonction du réseau de neurones, minimisant l'erreur entre les sorties réelles et les sorties désirées

- et 20% pour les tests, ceci a n d'évaluer les performances du système.

La génération du chier .ar , s'est fait grâce, dans un premier, aux modules développés, qui nous ont permis d'instancier le modèle (POS, Contexte local, Topics, classe, etc.) via un chier .csv. La plateforme Weka a ensuite été utilisée pour l'implémentation d'un réseau de neurones de type « perceptron multicouche » appliqué aux instances pour l'évaluation.

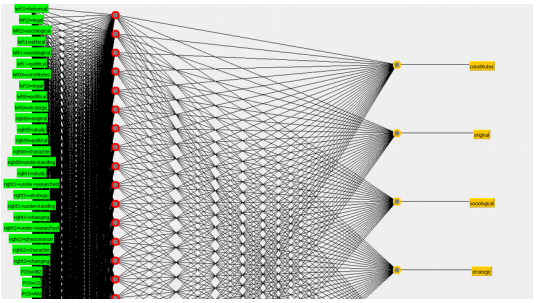


FIGURE 5: Application du perceptron multicouche

Ci-après quelques résultats des tests :

Correctly Classified Instances	71.4286%
Incorrectly Classified Instances	28.5714%

TABLE 6 : Vue d'ensemble

	Precision	Recall	F-Measure	ROC Area	Class
	0	0	0	0.917	iyya
	1	1	1	1	anezli
	1	1	1	1	tasnametti
	1	1	1	1	astrategic
	0.5	1	0.667	0.917	tisfrass
	0.5	1	0.667	0.917	yann
	0	0	0	0.917	azamul
Weighted Avg.	0.571	0.714	0.619	0.952	

TABLE 7 : Mesures

a	b	c	d	e	f	g	<-- classified as
0	0	0	0	0	1	0	| a=iyya
0	1	0	0	0	0	0	| b= anezli
0	0	1	0	0	0	0	| c= tasnametti
0	0	0	1	0	0	0	| d=astrategic
0	0	0	0	1	0	0	| e=tisfras
0	0	0	0	0	1	0	| f=yann
0	0	0	0	1	0	0	| g= azamul

TABLE 8 : Matrice de confusion

Une fois le modèle d'apprentissage correctement instancié, les résultats des tests montrent des performances similaires que sur les langues mieux dotées. Le système à réussit grâce à l'apprentissage sur une partie des instances à déduire le sens réel de tous les mots qui étaient mal classés au départ.

6 Conclusion

Cet article décrit une expérience basée sur une approche d'apprentissage indépendante de la langue traitée, pour instancier le modèle de données d'apprentissage et lui appliquer des méthodes d'apprentissages non ou peu supervisés. Elle a l'avantage de contourner les problèmes majeurs rencontrés dans l'analyse des données non structurées dans le contexte des langues peu dotées, à savoir le manque d'outils et de données annotées, sémantiquement et automatiquement exploitables. Les expériences menées sur des exemples con rment d'une part l'apport considérable du modèle proposé pour la détection du sens réel des mots dans le texte et d'autre part, l'application de l'approche sur les langues peu dotées. Nous espérant que cette expérience peut aider à sensibiliser et à inciter les chercheurs du domaine à concevoir et à développer des solutions génériques applicables à plusieurs langues dont les LPD. Ce qui pourrait contribuer à développer et peut-être même à sauvegarder ce type de langue en voie de disparition, faute de moyens et soutiens.

Références

ADOUANE W., SEMMAR N., JOHANSSON R. (2016). Romanized Berber and Romanized Arabic Automatic Language Identi cation Using Machine Learning. In *Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects,* 53–61, Osaka, Japan.

BLEI D., LAFFERTY J. (2009). Topic Models. *Text Mining, Classi cation, Clustering, and Applications* A. Srivastava and M. Sahami, editors. Chapman & Hall/CRC Data Mining and Knowledge Discovery Series.

DEERWESTER S. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6), 391–407.

FADILI H. (2013). Towards a new approach of an automatic and contextual detection of meaning in text, Based on lexico-semantic relations and the concept of the context. *IEEE-AICCSA,* Ifrane, (Morroco), May.

MIKOLOV T., CHEN K.,CORRADO G., DEAN J. (2013a). E cient estimation of word representations in vector space. *arXiv preprint,* arXiv:1301.3781.

MIKOLOV T., SUTSKEVER I., CHEN K.,CORRADO G., DEAN J. (2013b).Distributed representations of phrases and their compositionality. In *NIPS*.

SCHÜTZE H. (1998). Automatic word sense discrimination. *Computational Linguistics*, 24(1), 97–124.

TURNEY D. (2006). Similarity of semantic relations. *Computational Linguistics*, 32(3):379–416.

Expériences d'étiquetage morphosyntaxique dans le cadre du projet RESTAURE

Pierre Magistry[1] Anne-Laure Ligozat[2] Sophie Rosset[1]
(1) LIMSI, CNRS, Université Paris-Saclay, Bât 508, Campus Universitaire F-91405 Orsay, France
(2) LIMSI, CNRS, ENSIIE, Université Paris-Saclay, Bât 508, Campus Universitaire F-91405 Orsay, France
prenom.nom@limsi.fr.fr

RÉSUMÉ

Le projet RESTAURE vise à outiller en outils TAL trois langues régionales de France : l'alsacien, l'occitan et le picard. Dans cet article, nous abordons la question de l'étiquetage morphosyntaxique et rapportons les performances de différents systèmes proposés dans la littérature. Notre objectif est d'aborder les trois langues de manière homogène afin de pouvoir comparer les méthodes dans la variété de situations que présentent nos données. Ces expériences doivent guider notre réflexion pour le développement d'outils semisupervisés.

ABSTRACT

POS Tagging Experiments in the RESTAURE Project

The RESTAURE project aims at developing resources and NLP tools for three low-resource regional languages of France : Alsacien, Occitan and Picard. In this paper we report preliminary experiments in POS tagging. We benefit from this variety of cases to test methods in realistic situations. These experiments are conducted in order to drive the design of semisupervised methods.

MOTS-CLÉS : étiquetage morphosyntaxique, langues peu dotées.

KEYWORDS: Pos tagging, low-resource languages.

1 Introduction

Nos travaux s'inscrivent dans le projet RESTAURE, qui vise à outiller trois langues régionales de France : l'alsacien, l'occitan et le picard. Avec ces trois langues, nous disposons d'un éventail de cas d'étude pour l'outillage de langues peu dotées. Le projet s'étend de la compilation des corpus et de ressources lexicales à la réalisation d'outils de TAL. Dans cet article nous nous focalisons sur la tâche d'étiquetage morphosyntaxique.

2 Description des corpus

Les ressources mises à notre disposition par les différents partenaires du projet varient grandement d'une langue à l'autre, en fonction de l'état d'avancement de l'outillage de la langue concernée. Nous présentons ici les corpus que nous utilisons dans nos expériences.

2.1 Alsacien

Pour travailler sur l'alsacien, nous disposons d'un corpus annoté par deux expertes qui nous sert de corpus de référence pour l'évaluation de nos systèmes. Ce corpus est composé de quatre textes : deux articles tirés de Wikipedia, des recettes de cuisine et un court extrait d'une pièce de théâtre. Notons que les articles de Wikipedia sont rédigés en haut-rhinois tandis que les deux autres textes sont en bas-rhinois.

Nous disposons aussi de textes non annotés issus de la Wikipedia que nous utilisons comme données brutes.

2.2 Picard

Dans le cadre du projet, les partenaires de l'Université de Picardie travaillent à la constitution d'un corpus de référence pour le picard. Ce travail de construction du corpus est encore en cours ; les chiffres donnés à la table 1 sont donc les comptes de la version du corpus que nous utilisons au moment de la rédaction de cet article, mais ils seront amenés à évoluer rapidement. Ce corpus se compose essentiellement d'œuvres littéraires, et sera prochainement complété par des textes de presse régionale.

Une partie de ce corpus a été annotée en partie du discours. Là aussi, nous utilisons ce sous-corpus comme corpus d'évaluation de nos systèmes, tandis que la partie non annotée constitue l'ensemble de nos données brutes.

2.3 Occitan

Le corpus annoté dont nous disposons pour l'occitan est moins varié que pour les deux autres langues puisqu'il est issu d'une seule œuvre (Escorregudas en Albigés de Sergi Viaule). Nous prévoyons d'enrichir notre corpus avec des textes de presse, mais dans l'immédiat nous limitons nos expériences sur l'occitan à la partie supervisée (Section 4). D'autres expériences sur cette langue sont décrites dans (Vergez-Couret & Urieli, 2015).

2.4 Vue d'ensemble

Des informations quantitatives sur nos corpus sont détaillées dans le tableau 1. Dans toutes les expériences qui suivent, nous utilisons la tokenisation de référence pour les évaluations. Lorsque nous utilisons les corpus bruts, nous utilisons le tokeniseur à base de règles qui a servi à la pré annotation du corpus de référence pour l'alsacien, et une tokenisation « naïve » pour le picard.

Les jeux d'étiquettes varient d'une langue à l'autre. Pour l'alsacien, la liste des étiquettes a été affinée par rapport aux travaux précédents. Ces différences expliquent en partie les variations dans les résultats obtenus. Une conversion vers un jeu d'étiquettes commun aux trois langues est prévue, mais n'est pas encore disponible. La taille des différents jeux est aussi indiquée au tableau 1.

Langue	Texte	Tokens	Types	Jeu d'étiquettes
alsacien	Wikipedia 1	400	210	
	Wikipedia 2	503	252	
	cuisine	364	203	
	théatre	232	141	
alsacien	total annoté	1499	719	16
alsacien	données brutes	70 536	15 362	
occitan	un texte annoté	31 207	5 901	33
picard	lesi ziepe	366	191	
	Philéas Lebesgue	188	124	
	Simons L'Gampe	348	213	
picard	total annoté	11 814	3 843	15
picard	données brutes	1 769 018	140 072	

TABLE 1 – Corpus de référence

3 Étiquetage morphosyntaxique des langues peu dotées

L'étiquetage morphosyntaxique de langues peu dotées est une thématique qui bénéficie déjà d'une littérature abondante. Mais la diversité des situations fait que les méthodes proposées sont rarement applicables directement à un cas de figure particulier.

Ainsi beaucoup ont recours à des corpus alignés (à la suite de (Yarowsky *et al.*, 2001)), et une grande partie de ces expériences utilisent le corpus Europarl (Koehn, 2005). Si les résultats obtenus par ces méthodes sont assez encourageants, elles ne sont applicables qu'à un sous-ensemble assez restreint des langues peu dotées. Cet ensemble exclut les langues auxquelles nous nous intéressons, puisqu'Europarl couvre les langues européennes officielles. Les langues sur lesquelles nous travaillons sont beaucoup moins standardisées et peu ou pas enseignées, ce qui provoque une diversité des pratiques graphiques à laquelle ne sont pas sujettes les langues d'Europarl. De plus, pour utiliser les algorithmes proposés par cette lignée de travaux, il est nécessaire de disposer de corpus alignés. De tels corpus ne sont pas disponibles pour les langues que nous devons traiter.

Une autre pratique que nous laissons de côté dans un premier temps est le recours à des représentations vectorielles des mots par analyse de sémantique distributionnelle (telle que *word2vec*). Pour les travaux qui utilisent de telles méthodes, des corpus de données brutes de plusieurs dizaines de millions de tokens sont généralement requis. Notre corpus le plus grand, celui du picard, n'atteint pas deux millions de tokens.

Dans notre situation, nous sommes poussés à nous intéresser à l'adaptation des ressources ou des outils disponibles pour les langues proches des langues que nous traitons et qui sont mieux dotées. On utilisera ainsi comme langues « sources » : l'allemand pour l'alsacien, le catalan pour l'occitan et le français pour le picard. Nous avons recours à des modèles pré entraînés pour *Treetagger* (Schmid, 1995) ou le *Stanford Tagger* (Toutanova *et al.*, 2003) et nous comparons ceux-ci avec un étiqueteur que nous entraînons en utilisant les corpus Tiger (Brants *et al.*, 2004) pour l'allemand et Sequoia

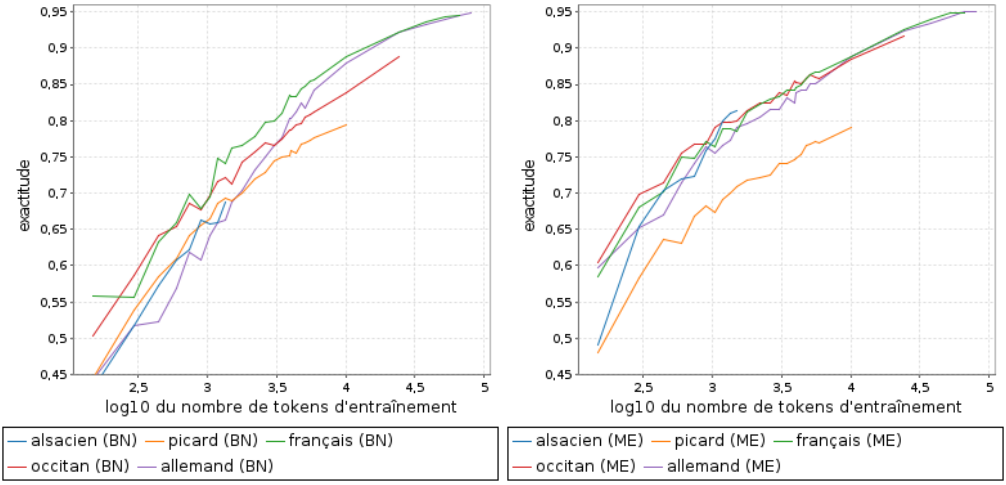


FIGURE 1 – Courbes d'apprentissage sur les différents corpus, en Bayes Naif (BN) ou MaxEnt(ME)

(Candito & Seddah, 2012) pour le français.

Dans un premier temps, nous commençons par observer les résultats que l'on peut obtenir avec le peu de données annotées dont nous disposons et des méthodes supervisées classiques (Section 4). Puis nous testons les méthodes de transfert (Bernhard & Ligozat, 2013) (Section 5), enfin nous essayons l'auto-entraînement (*self-training*) en utilisant une première annotation des corpus par transfert.

4 Apprentissage supervisé

Pour nos expériences, nous utilisons notre propre implémentation d'un étiqueteur très inspiré de MElt (Denis & Sagot, 2009), qui permet d'intégrer facilement de l'information lexicale issue de ressources externes et d'utiliser un classifieur bayésien naïf (BN) ou un Maximum d'Entropie (ME) de façon interchangeable. Dans le cas idéal d'une langue standardisée pour laquelle un large corpus d'entraînement et des lexiques à large couverture sont disponibles (par exemple sur le français ou l'allemand), MElt est au niveau de l'état de l'art. Dans les expériences présentées ici, nous n'intégrons pas de lexiques externes supplémentaires (nous extrayons simplement celui du corpus d'entraînement). Les scores rapportés sont donc légèrement en deçà des performances rapportées pour MElt. La possibilité d'intégrer des lexiques externes nous sera utile dans la suite de nos travaux.

Afin d'observer l'influence de la taille du corpus d'entraînement indépendamment des types de textes considérés, nous procédons par échantillonnages aléatoires sur la totalité du corpus. Pour différentes tailles d'entraînement, nous extrayons cette quantité de phrases pour l'entraînement de notre étiqueteur et nous le testons sur le reste du corpus. Cette opération est réalisée n fois et nous calculons les moyennes des scores obtenus, qui sont reportées à la Figure 1.

Pour l'alsacien et le picard, nous disposons de corpus d'évaluation qui sont composés de textes

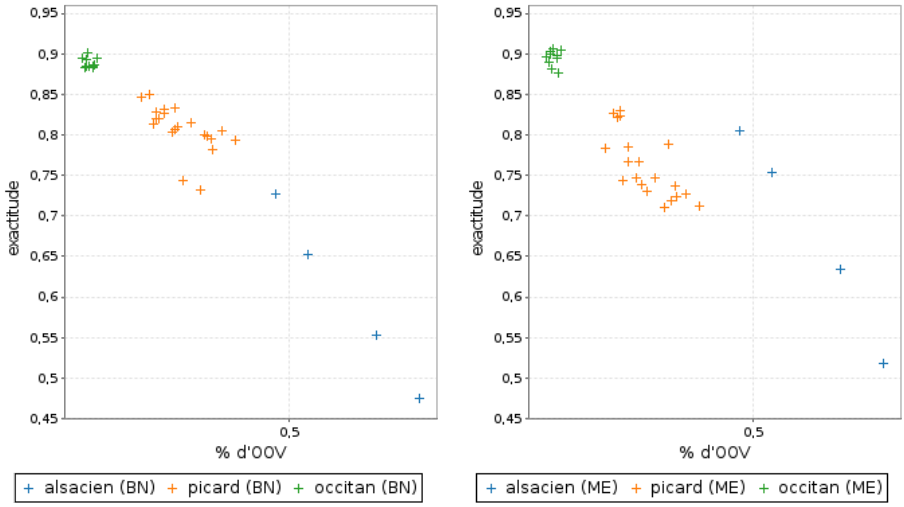


FIGURE 2 – Scores obtenus en fonction de différentes proportions de tokens inconnus. Sur les différents corpus, en Bayes Naif (BN) ou MaxEnt(ME)

d'origines variées. Il nous semble donc intéressant d'observer les différents cas de figure qui se présentent à nous. Nos corpus étant d'une taille très restreinte, nous évaluons indépendamment l'étiquetage de chaque texte avec un étiqueteur entraîné sur l'ensemble des autres textes (de la langue concernée).

Il est ainsi possible d'observer l'influence de différents paramètres, notamment celle de la couverture lexicale du corpus d'entraînement. Les scores obtenus indépendamment sur chaque document sont reportés sur la figure 2 où chaque point représente le score obtenu sur un document en fonction du pourcentage de formes inconnues qu'il contient.

Si ces résultats semblent encourageants au regard de la faible quantité de données dont nous disposons, nos corpus annotés sont coûteux et difficiles à étendre. Soulignons que l'axe du nombre de tokens de la figure 1 est au logarithme. L'effort d'annotation nécessaire pour rejoindre les scores des langues mieux dotées n'est donc pas négligeable. Étant donnée la taille actuelle de nos corpus annotés, nous préférons limiter l'usage de ceux-ci comme corpus de *développement*, pour évaluer et affiner d'autres méthodes.

5 Étiquetage par transfert

Dans cette section nous proposons une évaluation de l'approche par transposition de corpus. On appelle ici « langue source » la langue proche pour laquelle nous disposons de ressources annotées ou de modèles d'étiquetage préentraînés, et « langue cible » la langue que nous cherchons à analyser. Nous procédons en plusieurs étapes :

— utilisation d'un étiqueteur en langue source sans modification ;

— traduction de mots outils fréquents sur la base de listes ne dépassant pas les centaines d'entrées.

Pour ces expériences, nous comparons Treetagger, le Stanford Tagger avec des modèles préentraînés ainsi que notre étiqueteur présenté à la Section 4, que nous entraînons cette fois sur le corpus en langue source. Pour notre étiqueteur, nous testons le classifieur bayésien naïf et la régression logistique. Dans les configurations classiques, nous nous attendons à ce que cette dernière donne les meilleurs résultats, mais ce n'est pas ce que nous observons ci-dessous.

5.1 Étiqueteurs de langue source

Le point de départ de cette approche est d'utiliser un étiqueteur en langue source. Dans cette configuration, les corpus ne sont pas modifiés et nous utilisons les modèles pour la *langue cible* tels qu'ils sont distribués avec l'étiqueteur ou en entraînant notre étiqueteur sur Sequoia (français) ou Tiger (allemand).

langue	Texte	TreeTagger	Stanford	Bayes Naïf	MaxEnt
alsacien	Wikipedia1	0.49	0,59	0,48	0,50
	Wikipedia2	0,49	0,51	0,49	0,50
	recette	0,50	0,53	0,49	0,49
	théâtre	0,62	0,65	0,59	0,61
	tous	0,51	0,55	0,50	0,51
picard	lesi ziepe	0,38	0,33	0,43	0,40
	Philéas Lebesgue	0,56	0,58	0,60	65
	Simons L'Gampe	0,83	0,84	0,81	0,83
	tous	0,55	0,53	0,55	0,58

TABLE 2 – Scores d'exactitude de différents étiqueteurs entraînés pour la langue source, testés sur la langue cible sans modification.

En plus d'un score global par langue, nous donnons le résultat pour chaque texte de notre corpus alsacien. Pour le picard, nous ne mentionnons les résultats par texte que pour deux des plus mauvais et le meilleur. Les résultats varient fortement d'un texte à l'autre, particulièrement en picard (plus de 50 points d'écart entre les deux extrêmes). Cela reflète la variété des distances prises par les auteurs des textes avec la langue voisine.

5.2 Transposition des mots outils

Nous reproduisons ici une expérience comparable à (Bernhard & Ligozat, 2013) sur nos nouvelles données. Pour cela nous utilisons les listes de 169 traductions issues de ces travaux pour l'alsacien vers l'allemand. Pour le picard, nous générons cette liste automatiquement à partir d'une partie de nos corpus et d'une annotation avec la traduction mot à mot en français. En partant des 200 formes les plus fréquentes, nous conservons celles qui correspondent à une catégorie grammaticale fermée et qui est différente du français. Nous aboutissons ainsi à une liste de 77 paires de traductions.

Chaque occurrence des formes contenues dans ces mini lexiques est remplacée par sa traduction en langue source dans le texte en langue cible avant qu'il soit soumis à l'étiqueteur. En sortie d'étiquetage,

langue	Texte	TreeTagger	Stanford	Bayes Naïf	MaxEnt
alsacien	Wikipedia1	0,78	0,83	0,78	0,78
	Wikipedia2	0,74	0,77	0,73	0,75
	recette	0,65	0,76	0,67	0,67
	théâtre	0,74	0,78	0,71	0,76
	tous	0,73	0,78	0,73	0,74
picard	lesi ziepe	0,63	0,66	0,68	0,66
	Philéas Lebesgue	0,63	0,61	0,63	0,62
	Simons L'Gampe	0,84	0,85	0,82	0,83
	tous	0,70	0,69	0,70	0,71

TABLE 3 – Scores d'exactitude de différents étiqueteurs entraînés pour la langue source, testés sur la langue cible avec traduction des mots outils.

les formes d'origine sont rétablies, afin d'obtenir le corpus étiqueté en langue cible. Par exemple, les formes *ànder ànschtàtt àwer ìwer ùff ùn* de l'alsacien seront remplacées par les formes de l'allemand *ander, anstaat, aber, über, auf, und* que connait l'étiqueteur.

Les résultats des différents étiqueteurs sont donnés dans le tableau 3.

Cette méthode améliore les résultats sur tous les textes, mais elle profite bien plus aux textes les plus éloignés de la langue source, pour aboutir à des scores plus homogènes.

6 Réapprentissage endogène

Pour finir, nous essayons de combiner les différents étiqueteurs afin de tirer partie des données brutes dont nous disposons.

En utilisant l'étiqueteur de Stanford avec transposition comme dans la configuration précédente, nous annotons nos corpus bruts. Ces données annotées automatiquement servent ensuite à entraîner notre étiqueteur que nous évaluons comme précédemment. Les résultats sont donnés dans le tableau 4.

Cette opération n'améliore pas vraiment les scores globaux, mais améliore ou dégrade les scores de différents textes. Il faut noter que les meilleurs scores obtenus sont encore loin d'être satisfaisants et sont comparables à ceux obtenus avec de très petits corpus d'entraînement (cf. Figure 1).

7 Discussion

Nous avons présenté une série d'expériences sur un ensemble de données variées en utilisant différentes approches classiques.

Nous ne parvenons pas à apporter d'amélioration nette à l'état de l'art. Si les approches par corpus parallèles et sémantique distributionnelle sont matériellement difficiles voir impossibles à mettre en œuvre dans les cas concrets de langues peu outillées qui se présentent à nous, nous observons ici que la méthode par transposition des ressources en langue proche amène son lot de difficultés et qu'il n'est pas trivial de l'améliorer. D'un autre côté, l'approche supervisée obtient rapidement de

langue	Texte	Bayes Naïf	MaxEnt
alsacien	Wikipedia1	0,79	0,81
	Wikipedia2	0,80	0,81
	recette	0,75	0,74
	théâtre	0,75	0,71
	tous	0,78	0,78
picard	lesi ziepe	0,64	0,69
	Philéas Lebesgue	0,61	0,68
	Simons L'Gampe	0,81	0,84
	tous	0,70	0,71

TABLE 4 – Scores d'exactitude de différents étiqueteurs entraînés sur des corpus étiquetés automatiquement avec la méthode par transposition.

meilleurs résultats, mais ils sont eux aussi loin d'être satisfaisants et la constitution de corpus annotés de taille suffisante n'est pas envisageable.

En effet, pour les approches supervisées, nous nous attendons à avoir besoin de corpus d'autant plus grands que la variation dialectale, ainsi que les diversités de choix graphiques en l'absence de norme imposée, auxquelles vient s'ajouter la pratique fréquente d'alternance de code entre langues régionales et langues institutionnalisées voisines, accentuent la dispersion des données. La richesse et la variété des phénomènes observables dans nos corpus nous emmènent bien loin des corpus homogènes disponibles pour les « grandes » langues standardisées. (et les moyens engagés pour y travailler sont eux aussi sans comparaison).

En testant la transposition des mots grammaticaux, nous avons observé une grande disparité de résultats d'un texte à l'autre lorsque la différence de distance à la langue source est plus ou moins marquée. Dans certains cas, un algorithme généralement plus performant dans une configuration d'apprentissage supervisé classique (le MaxEnt) peut se révéler moins performant qu'un algorithme plus basique (Bayes Naïf), les features plus lexicalisées de notre étiqueteur « maison » le pénalisant par rapport à celui de Stanford pour cette approche par transposition. À la vue de cette disparité des résultats, il semble difficile, voir impossible de définir une unique bonne stratégie pour l'ensemble d'un corpus. Il sera peut-être nécessaire de se doter de moyens d'identifier des sous-parties du corpus pour entraîner plusieurs modèles et définir une méthode de sélection du modèle le plus pertinent.

Pour la suite de nos travaux, il nous semble nécessaire de combiner les deux approches en nous intéressant aux méthodes semi-supervisées pouvant utiliser la transposition en amorce.

Remerciements

Ces travaux ont bénéficié du soutien de l'ANR (projet RESTAURE - référence ANR-14-CE24-0003).

Références

BERNHARD D. & LIGOZAT A.-L. (2013). Es esch fàscht wie Ditsch, oder net ? Étiquetage

morphosyntaxique de l'alsacien en passant par l'allemand. In *TALARE 2013*, p. 209–220, Les Sables d'Olonne, France.

BRANTS S., DIPPER S., EISENBERG P., HANSEN-SCHIRRA S., KÖNIG E., LEZIUS W., ROHRER C., SMITH G. & USZKOREIT H. (2004). TIGER : Linguistic Interpretation of a German Corpus. *Research on Language and Computation*, **2**(4), 597–620.

BRAS M. & THOMAS J. (2008). Batelòc : cap a una basa informatisada de tèxtes occitans. In A. R. . D. SUMIEN, Ed., *IXème Congrès International de l'Association Internationale d'Études Occitanes*, p. 661–670, Aachen, Germany : Shaker.

CANDITO M. & SEDDAH D. (2012). Le corpus Sequoia : annotation syntaxique et exploitation pour l'adaptation d'analyseur par pont lexical. In *TALN 2012 - 19e conférence sur le Traitement Automatique des Langues Naturelles*, Grenoble, France.

DENIS P. & SAGOT B. (2009). Coupling an annotated corpus and a morphosyntactic lexicon for state-of-the-art POS tagging with less human effort.

KOEHN P. (2005). Europarl : A parallel corpus for statistical machine translation. In *MT summit*, volume 5, p. 79–86 : Citeseer.

SCHMID H. (1995). Improvements In Part-of-Speech Tagging With an Application To German. In *In Proceedings of the ACL SIGDAT-Workshop*, p. 47–50.

TOUTANOVA K., KLEIN D., MANNING C. D. & SINGER Y. (2003). Feature-rich Part-of-speech Tagging with a Cyclic Dependency Network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, p. 173–180, Stroudsburg, PA, USA : Association for Computational Linguistics.

VERGEZ-COURET M. & URIELI A. (2015). Analyse morphosyntaxique de l'occitan languedocien : l'amitié entre un petit languedocien et un gros catalan. In *TALARE 2015*, Caen, France.

YAROWSKY D., NGAI G. & WICENTOWSKI R. (2001). Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of the first international conference on Human language technology research*, p. 1–8 : Association for Computational Linguistics.

Les chaînes coréférentielles en créole de la Guadeloupe

Emmanuel Schang [1] Jean-Yves Antoine [2] Anaïs Lefeuvre-Halftermeyer [3]
(1) LLL (UMR 7270), Université d'Orléans/CNRS - 10 Rue de Tours 46527 Orléans cedex 2, France
(2) LI, Université François Rabelais Tours, 3 place Jean Jaurès, 41000 Blois, France
(3) LIFO, Université d'Orléans - 6, rue Léonard de Vinci 45067 Orléans Cedex 2, France
emmanuel.schang@univ-orleans.fr,
jean-yves.antoine@univ-tours.fr,anais.halftermeyer@univ-orleans.fr

RÉSUMÉ

Cet article présente une étude des chaînes coréférentielles en créole de la Guadeloupe à partir d'un corpus oral annoté en relations coréférentielles (1096 relations coréférentielles). Nous contrastons les résultats de notre expérimentation avec les données observées pour le français sur le corpus ANCOR (Muzerelle *et al.*, 2013). Bien que ces deux langues partagent largement leur lexique, elles diffèrent significativement sur le plan grammatical (absence de genre, SN sans déterminants en créole), ce qui donne toute sa valeur à leur comparaison.

ABSTRACT

Coreference Chains in Guadeloupean Creole

This paper presents a study on coreference chains in Guadeloupean Creole. A corpus of spoken Guadeloupean Creole was annotated in coreference relations (1096 relations) and we have compared the results with those found in the ANCOR corpus (French). While French and Gwadloupéyen share most of their lexicon, they differ greatly in their grammars (no gender, bare NPs in Creole *inter alia*), which makes their comparison valuable.

MOTS-CLÉS : Coréférence, créole, gwadloupéyen, corpus, annotation.

KEYWORDS: Coreference, Guadeloupean Creole, annotation, corpus.

1 Introduction

Le créole français de Guadeloupe (gwadloupéyen, ou CG) tire l'essentiel de son lexique du français [1] mais sa syntaxe diffère du français, notamment sur les points suivants :

— l'absence de genre grammatical,
— la forte utilisation des SN sans déterminants.

Le genre et le nombre de l'antécédent ainsi que le type d'article sont habituellement considérés comme des traits importants pour les systèmes de résolution des coréférences par aprentissage automatique développés par le TAL (Recasens *et al.*, 2011; Recasens Potau, 2010; Désoyer *et al.*, 2015). La description des stratégies coréférentielles en créole est intéressante car nous disposons d'un point de comparaison avec le français qui ne met pas en question les relations lexicales, mais uniquement la grammaire. Dans ce papier, nous aborderons ces questions au travers de deux points :

1. Par des apports consécutifs depuis le 17ème siècle et sous l'influence d'un important bilinguisme, v. (Bernabé, 1983) entre autres.

— la mise en évidence de chaînes coréférentielles prévalentes en créole,
— l'utilisation des syntagmes nominaux sans déterminants (SNSD).

Dans un premier temps, nous présenterons succinctement les principaux points de la grammaire du gwadloupéyen qui seront pertinents pour la suite de l'exposé. Puis nous présenterons la méthode suivie pour l'annotation du corpus et enfin, nous détaillerons quelques résultats.

2 Quelques éléments de la grammaire du gwadloupéyen

Le gwadloupéyen se distingue du français par l'utilisation de marqueurs de temps, mode et aspects (TMA) antéposés au verbe (Damoiseau, 2012; Bernabé, 1983; Vaillant, 2008) et qui se combinent avec la négation et certains adverbiaux à la place de la flexion verbale du français. Le verbe ne s'accorde pas avec son sujet. Il n'y a pas de passif.

Dans le domaine nominal, qui touche directement à la question des expressions référentielles, on trouve principalement les différences suivantes :

Absence de genre : Il n'y a pas de genre grammatical en CG, comme l'attestent les exemples en (1).

(1)
a. on nonm "un homme"
b. on fanm "une femme"
c. ti gason-la bèl "le petit garçon est beau"
d. tifi-la bèl "la petite-fille est belle"

Défini et démonstratif postposés : Contrairement au français, l'article défini et le démonstratif (2-a) se trouvent en marge droite du SN [2] et le nombre, lorsqu'il est marqué, est toujours accompagné du défini (2-b) :

(2) a. vwati wouj -la/lasa
voiture rouge DEF/DEM
'La/cette voiture rouge'
b. sé timoun-la (*sé timoun)
PL personne-DEF
'les enfants'

Un pluriel marqué renvoie donc à des entités 'identifiables' dans le contexte (non génériques) car l'article défini *-la* véhicule la notion de 'spécificité'. Nous continuerons cependant à nommer celui-ci 'défini', conformément à l'usage établi.

SN sans déterminants : On trouve dans la grande majorité des langues créoles des syntagmes nominaux sans déterminants (SNSD) avec un usage référentiel (Baptista & Guéron, 2007). Les SNSD véhiculent plusieurs valeurs sémantiques différentes, décrites, pour le GC dans (Gadelii, 2007).

2. Le défini et le démonstratif sont habituellement graphiés précédés d'un tiret.

(3) a. fo ou achté épis a kolombo
il.faut 2SG acheter épice à Colombo
'il faut que tu achètes des épices à Colombo.'
b. i ka fè kolombo ban nou
3SG IPFV faire Colombo pour 1PL
'il fait le/un colombo pour nous'

Nous ne pouvons pas détailler, faute de place, les exemples sans entrer dans une longue discussion pour chacun d'eux en exposant quels contextes rendent leur usage possible. On retiendra que les SNSD ont un usage beaucoup plus étendu que les SN sans déterminants en français (pour lesquels on pourra lire (Bouchard, 2003)). L'usage des SNSD pour des expressions singulier ou pluriel (identifiables par des pronoms de reprise au singulier ou au pluriel : *i* ou *li* vs *yo*) est un trait majeur à prendre en compte.

On retiendra de cette section que le gwadloupéyen ne permet pas d'utiliser l'accord en nombre [3] et en genre comme trait pour la résolution des coréférences, bien que le lexique soit semblable à celui du français. Nous avons conduit une première étude comparative entre le français et le créole en corpus pour appréhender l'impact de ces particularités.

3 Méthodologie

Le corpus ANCOR-Centre (Antoine *et al.*, 2016; Lefeuvre *et al.*, 2014; Muzerelle *et al.*, 2013) est un corpus annoté en relations anaphoriques et coréférentielles sur du français oral. Il est composé principalement d'entretiens et comprend environ 115 000 mentions et 51 500 relations anaphoriques. La méthodologie qui a présidé à la constitution de ce corpus est exposée en détails dans (Antoine *et al.*, 2013).

Le corpus ANCOR-971 est un corpus constitué à partir de la même méthodologie que le corpus ANCOR, mais sur des enregistrements de créole guadeloupéen (entretiens libres) disponibles en ligne (Glaude, 2013). Les points de divergence sont les suivants :

— les catégories annotées pour les mentions (SNSD (noté Bare), Indéfini, Défini, Démonstratif),
— annotation des principales fonctions syntaxiques (sujet, objet, attribut notamment).

Les mentions (ainsi que les relations) ont été annotées par un locuteur natif du gwadloupéyen et vérifiées par un linguiste spécialiste de cette langue. La plateforme d'annotation utilisée était GLOZZ (Widlöcher & Mathet, 2012) et les résultats ont été exploités avec les outils d'interrogation GLOZZ-QL, ANCOR-QI (Lefeuvre *et al.*, 2014) notamment. Concernant les liens coréférentiels uniquement [4], les relations annotées étaient les suivantes :

— directe : la reprise et l'antécédent ont la même tête lexicale,
— indirecte [5] : la reprise et l'antécédent ont la même tête lexicale,
— anaphore : reprise par un pronom.

3. Tout au moins de façon similaire à ce qui se fait en français.
4. Les autres relations, qui correspondent à des anaphores associatives, ne seront pas prises en compte dans cet article car elles ne portent pas sur la coréférence.
5. Aussi appelée 'anaphore infidèle'.

4 Résultats

Le corpus 971 contient 2731 entités et 1225 relations (dont 1096 relations coréférentielles), ce qui représente une base de travail considérablement plus large par rapport aux études antérieures sur le créole. A titre d'exemple, (Gadelii, 2007) a travaillé sur un corpus contenant cent exemples.

4.1 Relations

(Antoine *et al.*, 2016) ont étudié les chaînes coréférentielles en français oral dans les situations propres au corpus ANCOR (interaction orale spontanée en situation de dialogue finalisé). Ils ont observé que les chaînes répondaient à un patron prototypique (c'est-à-dire au sens de la chaîne la plus probable distributionnellement) N-N-N-P-P [6]. Cela signifie que, typiquement, une chaîne s'ancre sur un nom, suivi de deux reprises nominales puis de reprises pronominales.

Comparons ces données avec celles du gwadloupéyen.

4.1.1 Ancrage des chaînes

La proportion de SN par rapport aux pronoms est plus importante en créole qu'en français, comme le montre le Tableau 1.

	N	Pr
Français	51.2%	48.8%
Créole	70% (1901)	30% (819)

TABLE 1 – Proportion de SN vs Pronoms dans les deux langues

Les chaînes coréférentielles sont principalement ancrées par des N même si la proportion de chaînes ancrées par des pronoms (20%) est plus forte qu'en français (7% dans le corpus ANCOR).

Les 59 cas d'usage de pronoms comme ancre proviennent du pronom pluriel *yo* 'ils, eux' dans un usage indéfini, ou du pronom singulier *sa* ayant une référence mal délimitée (abstraite au sens de (Dipper & Zinsmeister, 2012)) et un seul cas de cataphore.

4.1.2 Structure des chaînes

Les chaînes coréférentielles du corpus créole sont en moyenne plus courtes que celles trouvées dans le corpus français. En créole, elles sont en moyenne légèrement en dessous de 3 maillons (2,87) alors que dans le corpus français, elles sont en moyenne constituées de 4 maillons (Antoine *et al.*, 2016).

L'analyse avec ANCORQI de la distribution des relations partant d'une ancre, ou internes à la chaîne, nous fournit des graphes de transitions comme la figure 1, qui concerne les chaînes à ancre nominale.

6. De façon à simplifier l'argumentation, nous parlerons de N à place de SN.

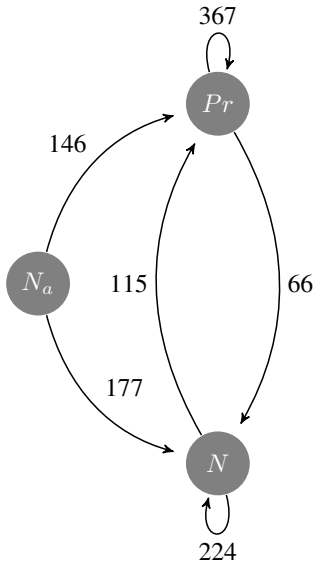


FIGURE 1 – Chaînes coréférentielles ancrées par un N (N_a) par type et en nombre de relations.

L'analyse des transitions de la figure 1, de même que la connaissance du nombre moyen de relations dans une chaîne, montre que les chaînes nominales prévalentes en créole sont du type N-P-P-P ou N-N-N-N. Quant aux chaînes ancrées par un pronom (30% des cas), elles sont constituées très majoritairement d'une séquence P-P-P-P de pronoms. Comme en français, lorsque dans une chaîne on utilise un pronom, il est rare de revenir au N par la suite.

4.2 Pronoms et SNSD

On notera (Table 2) que les SN sont en majorité des Syntagmes Nominaux Sans Déterminants (SNSD). Ceux-ci ont fait l'objet de nombreuses études dans le cadre des langues créoles (Baptista & Guéron, 2007). Dans (Gadelii, 2007), on trouve l'idée que les SNSD du CG sont principalement des sujets reprenant un topique (continued topics).

Cette hypothèse peut se vérifier facilement sur le corpus en croisant les traits annotés : nouvelle entité du discours (non), fonction (sujet) et type (SNSD). Il apparaît cependant que les SNSD apparaissent assez peu en sujet (Table 3), et en proportion, beaucoup moins que les pronoms personnels (3SG et 3PL) dont c'est la fonction privilégiée. Par ailleurs, il est à noter que contrairement aux pronoms personnels, la majorité des SNSD apparaissent comme nouvelle entité du discours (646 ancres contre 448 reprises) et que lorsqu'ils sont en reprise, ils n'apparaissent pas majoritairement en sujet (Table 4). L'hypothèse avancée dans (Gadelii, 2007) n'est donc pas confirmée par ce corpus.

Pour aller un peu plus loin dans l'analyse des SNSD par rapport aux pronoms, on pourra remarquer que le nombre d'entités du discours (mentions) entre deux membres d'une chaîne dont le second terme est un pronom (N-P ou P-P) est bien inférieur à celui des chaînes dont le second terme est un SNSD (N-SNSD ou P-SNSD). La moyenne du nombre d'entités entre un pronom et son antécédent

SNSD	définis	indéfinis	démonstratifs
56,5% (1074)	21,2% (403)	20,78% (395)	1,52% (29)

TABLE 2 – Distribution des SN par type.

Type	sujet	objet	autre
SNSD	9,5% (102)	28,5% (306)	62% (666)
defini	19,35% (78)	23,82% (96)	56,82% (229)
pronom pers.	56,93% (115)	7,92% (16)	35,15% (71)

TABLE 3 – Fonctions grammaticales des SNSD, SN définis et pronoms

Type	sujet	objet	autre
SNSD reprise	14,02% (60)	21,96% (94)	64,02% (274)
SNSD ancre	6,58% (42)	33,23% (212)	60,19% (384)

TABLE 4 – Fonctions grammaticales des SNSD (reprise ou ancre)

est de 3.07 (médiane = 1) tandis que pour les SNSD, elle est de 8,83 (médiane = 2). Ce qui indique que la résolution des pronoms (toutes catégories confondues) est locale par rapport aux SNSD.

On peut faire l'hypothèse raisonnable que, en l'absence de genre grammatical, les pronoms sont utilisés pour reprendre une entité du discours immédiatement disponible et, dès lors qu'une ambiguïté est possible, la reprise par un SNSD est préférée.

On remarque par ailleurs que les SN définis [7] trouvent leur antécédent à une distance (en nombre de mentions) moyenne de 14 (médiane = 4). On le voit donc, ils reprennent en moyenne des mentions situées plus loin que les SNSD.

Ceci est compatible avec les théories de l'accessibilité des référents (Gundel, 2010; Gundel *et al.*, 2003; Gundel, 2003). On peut proposer l'échelle d'accessibilité suivante (accessibilité du référent en fonction de la distance par rapport à l'antécédent) :

++ accessible : pronom
+ accessible : SNSD
- accessible : SN défini

5 Conclusion

Bien qu'il soit impossible de détailler tous les paramètres annotés dans le corpus de créole guadeloupéen dans cet article, cette étude préliminaire, qui demande à être complétée, nous a permis de mettre en avant quelques faits intéressants :

— bien que le créole ne fasse pas usage du genre grammatical, on retrouve deux points communs avec le français : lorsque dans une chaîne coréférentielle un pronom est utilisé, il est rare de poursuivre la chaîne avec un N,
— le créole fait usage de SN sans déterminants pour reprendre des entités moins proches (distance calculée en nombre de mentions entre l'antécédent et la reprise) que les pronoms. On peut faire

7. Les SN indéfinis étant principalement des ancres, ils ne sont pas pertinents pour notre analyse.

l'hypothèse que l'absence de genre en créole favorise la reprise par un SN sans déterminant (accessibilité du référent par la tête nominale).

Les résultats de cette étude nous conduisent à envisager de développer ces mêmes investigations pour comparer d'autres langues créoles avec leur langue lexificatrice. Par exemple, la comparaison des créoles portugais du Golfe de Guinée ou de Haute-Guinée (forro et kriyol respectivement) avec le portugais serait intéressante car ceux-ci diffèrent du portugais (la langue qui a fournit l'essentiel de leur lexique) par l'absence de genre grammatical. Ces études devraient conduire à une meilleure compréhension à la fois de la créolisation [8] (v. entre autres (Mufwene, 2005)) et des processus de résolution des coréférences en général. Enfin, ce travail contribue à la prise en compte de langues dites 'peu dotées' dans les réflexions linguistiques générales et dans le traitement automatique des langues naturelles.

Remerciements

Les auteurs tiennent à remercier : Laura Noreskal pour ses annotations, Flora Badin poour ses scripts, les membres du GDRI SEEPiCLa pour leurs retours sur des versions préliminaires du texte ainsi que les relecteurs anonymes de DILITAL.

Références

ANTOINE J.-Y., LEFEUVRE A. & SCHANG E. (2016). Codage en chaîne ou en première mention de la coréférence : Approcher la structure des chaînes de référence par comparaison des deux annotations. *SHS Web of Conferences*, **27**(nil), 02001.

ANTOINE J.-Y., SCHANG E., MUZERELLE J., LEFEUVRE A., PELLETIER A., ESHKOL I., MAUREL D. & VILLANEAU J. (2013). *Corpus ANCOR_Centre*. Rapport interne.

BAPTISTA M. & GUÉRON J. (2007). *Noun phrases in creole languages : a multi-faceted approach*, volume 31. John Benjamins Publishing.

BERNABÉ J. (1983). *Fondal-natal*. l'Harmattan Paris.

BOUCHARD D. (2003). Les sn sans déterminant en français et en anglais. *Essais sur la grammaire comparée du français et de l'anglais*, p. 55–95.

DAMOISEAU R. (2012). *Syntaxe créole comparée*. Karthala et CNDP-CRDP edition.

DÉPREZ V. (2005). Morphological number, semantic number and bare nouns. *Lingua*, **115**(6), 857–883.

DÉSOYER A., LANDRAGIN F., TELLIER I., LEFEUVRE A. & ANTOINE J.-Y. (2015). Coreference Resolution for Oral Corpus : a machine learning experiment with ANCOR corpus. *Traitement Automatique des Langues*, **55**(2), 97–121.

DIPPER S. & ZINSMEISTER H. (2012). Annotating abstract anaphora. *Language Resources and Evaluation*, **46**(1), 37–52.

GADELII K. (2007). The bare np in lesser antillean. *Creole Language Library*, **31**, 243.

GLAUDE H. (2013). *Corpus Créoloral*. oai :crdo.vjf.cnrs.fr :crdo-GCF, SFL Université Paris 8 - LLL Université Orléans.

8. On entend par *créolisation* les forces qui produisent une langue nouvelle née de contacts entre plusieurs langues.

GUNDEL J. K. (2003). Information structure and referential givenness/newness : How much belongs in the grammar. In *Proceedings of the HPSG'03 Conference*, p. 143–162.

GUNDEL J. K. (2010). Reference and accessibility from a Givenness Hierarchy perspective. *International Review of Pragmatics*, **2**(2), 148–168.

GUNDEL J. K., HEGARTY M. & BORTHEN K. (2003). Cognitive status, information structure, and pronominal reference to clausally introduced entities. *Journal of Logic, Language and Information*, **12**(3), 281–299.

LEFEUVRE A., ANTOINE J.-Y. & SCHANG E. (2014). Le corpus ANCOR_Centre et son outil de requêtage : application à l'étude de l'accord en genre et nombre dans les coréférences et anaphores en français parlé. In *SHS Web of Conferences*, volume 8, p. 2691–2706 : EDP Sciences.

MANUELIAN H. & FATTIER D. (2011). L'utilisation des déterminants en créole haïtien : Etude de quelques chaînes de référence.

MUFWENE S. S. (2005). *Créoles, écologie sociale, évolution linguistique*. Editions L'Harmattan.

MUZERELLE J., LEFEUVRE A., ANTOINE J.-Y., SCHANG E., MAUREL D., VILLANEAU J. & ESHKOL I. (2013). ANCOR, premier corpus de français parlé d'envergure annoté en coréférence et distribué librement. *Actes de TALN*, p. 555–563.

RECASENS M., HOVY E. & MARTÍ M. A. (2011). Identity, non-identity, and near-identity : Addressing the complexity of coreference. *Lingua*, **121**(6), 1138–1152.

RECASENS POTAU M. (2010). Coreferència : Teoria, anotació, resolució i avaluació.

VAILLANT P. (2008). Grammaires factorisées pour des dialectes apparentés. In *TALN 2008 : Actes de la 15ème conférence annuelle sur le Traitement Automatique des Langues Naturelles*.

WIDLÖCHER A. & MATHET Y. (2012). The Glozz platform : a corpus annotation and mining tool. In *Proceedings of the 2012 ACM symposium on Document engineering*, p. 171–180 : ACM.

ZRIBI-HERTZ A. & JEAN-LOUIS L. (2013). From noun to name : definiteness marking in modern martinikè. *Crosslinguistic studies on Noun Phrase structure and reference*, p. 269–315.

Constitution d'un corpus d'arabe tunisien parlé à Orléans

Ben Ahmed Youssra
Laboratoire Ligérien de Linguistique (UMR 7270), 10 Rue de Tours – BP 46527, 45065, France
ben.ahmed.yossra@gmail.com

RESUME

La constitution de corpus des parlers arabes se heurte à l'absence de ressources et le manque d'outils pour le traitement de ces derniers. Comme conséquence directe de cette situation, nous avons rencontré de maints problèmes lors de la construction d'un corpus de l'une des variétés de ces parlers, i.e. l'arabe tunisien. Nous proposons dans cet article d'exposer les principaux choix méthodologiques et techniques pour lesquels nous avons opté afin de répondre aux contraintes auxquelles nous avons été confrontée.

ABSTRACT

Composition and exploitation of a Tunisian Arabic corpus - Orleans
The composition of Arabic corpus dialect faces the absence of resources and the lack of tools in order to treat them. As a direct consequence of this condition, we encountered many problems during the construction of a corpus of one of the varieties of these dialects, i.e. Tunisian Arabic.
We propose in this article to present the main methodological and technical choices for which we chose to respond to the restrictions we confronted.

MOTS-CLES : arabe tunisien, corpus oral, transcription, annotation.
KEYWORDS : tunisian arabic, corpus oral, transcription, annotation.

1 Introduction

Cette étude s'aventure dans un terrain à peu près vierge, celui de l'analyse des expressions du futur en arabe tunisien.

C'est la multitude d'emplois de la forme du «futur» en français (futur historique, futur de vérité générale, d'injonction, futur de bilan, etc.) qui a attiré notre attention et nous a amenée à explorer la question dans une autre langue, notamment en arabe tunisien. Il nous a semblé utile, dans ce cadre, de s'interroger sur l'existence du futur et son expression dans les systèmes verbaux de ces deux langues, qui se différencient fondamentalement sur le plan aspectuel.

Dans la plupart des études sur le futur, les linguistes ont eu recours soit à la fabrication d'exemples soit à l'emprunt d'exemples écrits ; des choix susceptibles d'empêcher de voir son fonctionnement réel.

Cette considération nous a poussée à proposer une étude contrastive de l'emploi du futur en nous basant sur des données orales authentiques. Notre choix s'est porté, pour la partie française, sur des données puisées dans les Enquêtes Socio-Linguistiques à Orléans (désormais ESLO). Quant à celle de l'arabe, nous avons constitué un corpus auprès des locuteurs tunisiens résidant à Orléans.

Dans ce papier, après une première partie où il sera notamment question des principaux choix opérés lors de la constitution du corpus de l'arabe tunisien (désormais AT), nous nous attarderons dans la deuxième partie sur les problèmes rencontrés lors de la phase de transcription. En troisième partie, nous aborderons les difficultés affrontées lors du processus d'annotation.

2 Méthodologie

Pour la constitution de notre corpus d'arabe tunisien[1], nous avons suivi en grande partie la démarche adoptée dans la construction du corpus ESLO. C'est sur ce dernier que nous nous sommes basée dans la sélection des données en français.

D'une taille conséquente (à ce jour, environ 7 millions de mots) et d'une diversité de genres (conférences universitaires, entretiens, repas en famille ou entre amis…), ce corpus offre un avantage essentiel dans le domaine de la temporalité où le pertinence des analyses est tributaire d'une bonne prise en compte de la situation de communication : il contient des données situées, enrichies par des métadonnées renseignant sur la situation de communication et précisant pour chaque locuteur son profil en termes d'âge, de sexe et de catégorie socio-professionnelle.

S'agissant d'une recherche contrastive, les choix présidant à la constitution du corpus de l'AT ont été en grande partie dictés par la recherche de la plus grande comparabilité possible avec le corpus ESLO.

A commencer par le lieu de l'enquête : même si Orléans constitue sans doute un « non-choix » (Abouda & Baude, 2009 : 133), il nous a paru prudent de respecter cette unité de lieu, d'autant que notre corpus de l'AT a pu ainsi intégrer, pour être partagé, la base de données constituée par le programme « Langues en Contact à Orléans » (LCO) en bénéficiant des moyens et de l'expérience accumulée dans ce cadre. Il va de soi que ce choix, qui a écarté l'autre possibilité un temps envisagée de mener une enquête comparable dans une ville tunisienne, présentait quelques inconvénients : il était particulièrement long et difficile de constituer un corpus suffisamment grand et équilibré.

Le corpus, que nous avons constitué entre 2013 et 2014, représente un volume de 17h, fractionné en 37 enregistrements, ce qui nous semble suffisant pour les investigations envisagées. Il parvient à

[1] Il s'agit d'une langue vernaculaire non standardisée, parlée par douze millions de personnes vivant principalement en Tunisie et par de nombreuses familles résidant à l'Europe, notamment en France. Il s'emploie principalement dans tous les usages informels, i.e. à la maison, dans la rue, au travail ou à l'école, et même dans les médias audiovisuels.

capter une certaine diversité de locuteurs (en fonction des variables d'âge, de sexe, de CSP, de régions, etc.), ce qui permet d' « améliorer [sa] représentativité »[2].

Corpus de l'arabe tunisien (2013-2014)	
Nombre d'heures d'enregistrements	17h
Situation de parole	Entretien face à face
Âge	- 20- 30 ans : 54% - 30- 50 : 35% - 50 et plus : 11%
Sexe	- Hommes : 65% - Femmes : 35%
Niveau scolaire	- primaire collège 15% - secondaire 15% - bac 24% - supérieur 46%
Profession	- Chef de cuisine - Animateur - Commerçant - Assistant d'éducation - Etudiant - Gérant - Femme de ménage…

TABLEAU 1 : Caractéristiques quantitatives du corpus

Respectant les procédures suivies par ESLO, nécessaires pour rendre le corpus disponible, nous avons procédé à une documentation précise de nos données[3].

En ce qui concerne le mode de recueil des données, nous avons privilégié l'entretien en face-à-face, « situation certes très formelle, mais qui avait l'avantage d'être (…) contrôlable » (Abouda & Baude, 2009 : 134).

Aussi, dans le même souci de comparabilité, nous avons réalisé un questionnaire basé sur les six thèmes retenus par ESLO (logement/Orléans, travail, loisirs, questions évaluatives sur Orléans, langue, recette), afin de faire parler les locuteurs, en ciblant les contextes propices à l'émergence des formes verbales au futur.

[2] Cf. Habert 2000.

[3] Pour chaque locuteur, nous avons réalisé une fiche d'information récapitulant l'âge, le sexe, le niveau scolaire… complétée par des indications sur l'enregistrement (n°, type (situation de parole), participant(s), lieu, date et durée de l'enregistrement, situation d'enregistrement…)

3 Transcription

Après la collecte des données, s'est posée la question de leur transcription. Cette étape a soulevé plusieurs interrogations depuis le système graphique jusqu'aux outils de transcription, en passant par le mode de transcription et les conventions adoptées.

3.1 Système de notation

Les travaux sur l'arabe tunisien, peu nombreux, hésitent entre les deux systèmes graphiques, i.e. latin et arabe. Le choix de l'un ou l'autre système est dicté par de nombreux paramètres, allant de la tradition du champ, jusqu'aux préférences idéologiques, en passant par la facilités techniques.

C'est précisément pour cette dernière raison, que nous avons opté pour une transcription avec une graphie latine, qui aura également l'avantage de fournir un corpus partageable et facilement lisible par les non-natifs.

3.2 Mode de transcription

Ce sont non seulement les objectifs de recherche qui définissent le choix d'un mode de transcription, mais aussi les spécificités de la langue parlée. En bref, ainsi le note Gadet (2008 : 37) « la transcription ne peut être regardée comme une opération banale, car on transcrit pour donner à voir quelque chose.»

La transcription de notre corpus a impliqué un choix entre plusieurs types de notation (phonétique, phonologique, morphologique et usuelle). Nous avons opté finalement pour une notation orthographique (usuelle) *d'inspiration phonologique* qui tient compte de l'aspect morphosyntaxique des énoncés. Ce choix a été motivé par les raisons suivantes :

- l'objet d'étude ne nécessite ni une notation phonétique, ni une notation phonologique stricte ;
- la simplicité de ce mode de notation permet un décodage rapide et facile par le lecteur en écartant les ambiguïtés et les hésitations, principalement au niveau syntaxique ;
- l'ajout possible des deux autres modes de transcription (phonétique ou/et phonologique) selon les besoins des chercheurs.

Nonobstant, l'absence d'un standard stabilisé a exigé la reprise des pratiques orthographiques les plus usitées au sein de la communauté scientifique.

3.3 Les conventions de transcription

En ce qui concerne l'outil de transcription, nous avons choisi TRANSCRIBER[4], un logiciel d'aide à la transcription manuelle de fichiers audio qui permet de transcrire de nombreuses langues y compris non européennes.

[4] Téléchargeable sur : http://www.ldc.upenn.edu/mirror.Transcriber/

Lors de la transcription, un problème d'encodage s'est manifesté par le fait que quelques caractères spéciaux ne s'affichent pas correctement. Il nous a paru dés lors indispensable d'opter pour l'encodage UTF-8.

De différents facteurs sont entrés en jeu pour la détermination des conventions de transcription, allant des finalités de la recherche, jusqu'à la taille du corpus, en passant par le type des données primaires (audio ou vidéo).

Les conventions varient selon la nature de la langue, i.e. écrite (comme le cas du français) ou orale (comme pour l'arabe tunisien). Elles se divisent en deux types : des conventions « *spécifiques* » à chaque langue ; et des conventions « *communes* » à tout corpus oral quelle que soit la langue.

Privé d'une tradition orthographique solide, nous avons choisi de transcrire l'arabe tunisien en nous basant sur les propositions de l'INALCO (1996-1998). En ce qui concerne les phénomènes associés à l'oralité, ont été maintenues les conventions avancées par le LLL pour le corpus du français de l'ESLO.

4 Annotation

L'annotation, étape incontournable en ce qu'elle permet de croiser approches qualitative et quantitative, consiste dans l'apport d'informations de nature différente. On parle dans ce sens d'une «valeur ajoutée» (Leech 1997) aux données brutes.

S'il existe pour l'arabe standard des étiqueteurs morphosyntaxiques, relativement importants (Arabic Part-of-speech Tagger[5], Sakher[6], Sebawai[7], Aramoph[8], etc.), la situation concernant l'arabe tunisien est nettement différente ; car il n'y a pas à notre connaissance de systèmes complets et disponibles pour l'étiquetage de celui-ci.

Etant donné que « l'arabe dialectal se distingue de l'arabe classique par une syntaxe simplifiée, un lexique plus riche en vocables étrangers et une phonologie altérée » (Boukadida, 2008 : 37), nous n'avons pas pu se servir de ces étiqueteurs. Comme conséquence de cette situation, nous avons été amenée, pour exploiter le corpus constitué, à baliser manuellement les occurrences de futur. Dans le fichier Transcriber, nous avons ainsi ajouté une balise sur chaque portion du texte exprimant un futur : (<Event desc="FUT" type="lexical" extent="begin"/> … <Event desc="FUT" type="lexical" extent="end"/>). Ce balisage dans le corps de transcription rend possible un retour facile vers les occurrences dans leur contexte et un accès rapide au signal sonore, nécessaire pour l'analyse de données.

[5] L'étiqueteur APT 'Arabic Part-of-speech Tagger' de Khoja (Khoja 2001) se présente comme une adaptation à l'arabe du système du British National Corpus (BNC) qui combine des techniques statistiques et des règles linguistiques pour déterminer tous les traits morphologiques d'une unité lexicale.

[6] Analyseur morphologique Sakher est un système développé par Chalabi (Chalabi, 2004). Il traite aussi bien l'arabe classique que l'arabe moderne, et il permet de déterminer la racine possible d'un mot en supprimant tous les affixes et suffixes, et en décrivant la structure morphologique de celui-ci.

[7] Analyseur morphologique Sebawai est un système développé par Darwish en 2003. Il permet de trouver les racines de mots.

[8] Aramorph est un analyseur distribué par le LDC (Linguistic Data Consortium) qui permet de segmenter un mot en trois séquences (préfixe racine post-fixe).

Les 2731 occurrences de futur ainsi identifiées ont par la suite été extraites grâce au logiciel d'analyse textométrique TXM[9], et exportées dans un tableau CSV, afin d'y être annotées. Chacune des occurrences du futur a ainsi été sous-spécifiée pour un certain nombre de traits morphosyntaxiques et sémantiques.

La dernière étape de ce processus a consisté à réinjecter sous TXM les occurrences et leurs annotations affinées dans l'objectif d'obtenir une analyse linguistique fine croisant approches qualitative et quantitative.

La réinjection sous TXM après annotation des occurrences identifiées permettra une analyse linguistique fine croisant approches qualitative et quantitative.

5 Conclusion

Dans cet article, il a été question de la constitution et du traitement de corpus de l'arabe tunisien parlé à Orléans.

La rareté des travaux sur l'arabe tunisien et le manque d'outils de traitement automatique de cette langue nous ont posé beaucoup de problèmes lors de la constitution de notre corpus, depuis le recueil de données jusqu'à l'annotation, en passant par la transcription. Afin de construire un corpus partageable et exploitable, il a été nécessaire d'opter pour quelques choix méthodologiques jugés les plus adéquats.

Bien qu'il ne soit pas assez représentatif, notre corpus peut être perçu comme un échantillon diversifié permettant d'observer plusieurs phénomènes qui découlent du fonctionnement du parler tunisien. Il propose une catégorie bien spécifique de locuteurs, habituellement non intégrés dans les corpus oraux, i.e. les locuteurs « sans papiers ». Néanmoins, cet échantillon ne représente pas toute la communauté tunisienne d'Orléans.

Après avoir transcrit et enrichi le corpus d'AT, il est devenu dés lors possible d'examiner nos données sur les plans quantitatif et qualitatif. Travailler sur des corpus oraux authentiques de deux langues différentes nous a permis d'observer la multiplicité d'emplois du futur et de nous approcher des propriétés de son fonctionnement à l'oral.

Bien que ce travail ne soit qu'une ébauche, la nature des données constituées peut apporter des informations concrètes sur la validité des modèles généraux de la description grammaticale.

[9] http://textometrie.ens-lyon.fr/

Références

Abouda L., Baude O. (2005). Du français fondamental aux ESLO. Colloque international Français fondamental, *corpus oraux, contenus d'enseignement*. 50 ans de travaux et d'enjeux, SIHFLES - Laboratoire ICAR, Lyon, 8, 9 et 10 décembre 2005.

Abouda L., Baude O. (2006). Constituer et exploiter un grand corpus oral : choix et enjeux théoriques. Le cas des ESLO, *in* F. Rastier, M. Ballabriga (dir.), *Corpus en Lettres et Sciences sociales — Des documents numériques à l'interprétation,* actes du XXVII colloque d'Albi, *Langages et signification,* publiés par C. Duteil-Mougel et B. Foulquié.

Abouda L. (2015). *Syntaxe et Sémantique en corpus. Du temps et de la modalité en français oral,* mémoire HDR, Université d'Orléans.

Baude O. (coord.) (2006). *Corpus oraux, Guide des bonnes pratiques.* CNRS éditions et P.U.O.

Baude O. (2008). Le droit de la parole, dans Bilger, Mireille (éd.). *Données orales : les enjeux de la transcription*. Perpignan. PUP.

Benjelloun S. (2002). Une double graphie, latine et arabe, pour enseigner l'arabe marocain, in : D. Caubet, S. Chaker, J. Sibille (éds), *Codification des langues de France*, 331-340, L'Harmattan, Paris.

Bergounioux G. (dir.) (1992). Enquêtes, Corpus et Témoins, *Langue Française* 93.

Bergounioux G., et al. (1992). « L'Etude socio-linguistique sur Orléans (1966-1991), 25 ans d'histoire d'un corpus», *Langue française*, 93, 74-93.

Bilger M., Cappeau P. (2004). L'oral ou la multiplication des styles. *Langage et Société* 109, 13-30.

Bilger M. (2008). Les enjeux des choix orthographiques dans Bilger, Mireille (éd.) *Données orales – Les enjeux de la transcription*. Perpignan. PUP, 248-257.

Blanche-Benveniste C., Jeanjean C. (1987). *Le français parlé : transcription et édition*, Paris, Didier-Erudition.

Boukadida N. (2008). Connaissances phonologiques et morphologiques dérivationnelles et apprentissage de la lecture en arabe (Etude longitudinale).

Bourdieu P. (2003). (sous la direction de) *La misère du monde*, Paris, Seuil – Collection Point.

Caubet D. (1999). Arabe maghrébin : passage à l'écrit et institutions, In *Faits de Langues*, vol. 7, n° 13, 235-244.

Caubet D. (2002). Arabe maghrébin, langue de France : entre deux graphies, in : D. Caubet, S. Chaker, J. Sibille (éds), *Codification des langues de France,* p.331-340, L'Harmattan, Paris, 2002.

Cerquiglini, B. (1999), *Les langues de la France*, rapport aux ministres de l'Éducation nationale et de la Culture et de la Communication.

Gadet F. (2000). Derrière les problèmes méthodologiques du recueil des données, dans M. Bilger (dir.), *Linguistique sur corpus*, Presses Universitaires de Perpignan.

Gadet F. (2008). L'oreille et l'œil à l'écoute du social, dans Bilger, Mireille (éd.). *Données orales: les enjeux de la transcription*. Perpignan. PUP, 35-47.

Habert B., Nazarenko A., Salem A. (1997). *Les linguistiques de corpus*, Paris, A. Colin.

Habert B. (2000). Des corpus représentatifs : de quoi, pour quoi, comment ?, dans M. Bilger (dir.), *Linguistique sur corpus*, Presses universitaires de Perpignan.

Khoja S. (2001). Arabic part-of-speech tagger. In Proceedings of the Student Workshop at the Second Meeting of the North American Chapter of the Association for Computational Linguistics, Carnegie Mellon University, Pittsburgh, 81–86.

Leech G. (1997). Introduction corpus annotation. In R. Garside, G. Leech, A. McEnery (Eds.), Corpus annotation: Linguistic information from computer text corpora. London, Longman, 1-18.

Maurer B. (1999). Quelles méthodes d'enquête sont effectivement employées aujourd'hui en sociolinguistique, dans L.-J. Calvet et P. Dumont (dir.), *L'enquête sociolinguistique*, L'Harmattan.

Mondada L. (2008). La transcription dans la perspective de la linguistique interactionnelle, dans Bilger, Mireille (éd.). *Données orales : les enjeux de la transcription*. Perpignan. PUP, 78-109.