Revue de l'Université de Moncton

Revue de l'Université de Moncton

Module *NooJ* du français. Traitement automatique de corpus de français parlé régional

Gisèle Chevalier and Sylvia Kasparian

Volume 45, Number 1-2, 2014

Les chiffres et les lettres peuvent-ils se marier ? Quinze ans de recherches au Laboratoire d'analyse de données textuelles

URI: https://id.erudit.org/iderudit/1038912ar
DOI: https://doi.org/10.7202/1038912ar

See table of contents

Publisher(s)

Revue de l'Université de Moncton

ISSN

1712-2139 (digital)

Explore this journal

Cite this article

Chevalier, G. & Kasparian, S. (2014). Module *NooJ* du français. Traitement automatique de corpus de français parlé régional. *Revue de l'Université de Moncton*, *45*(1-2), 273–290. https://doi.org/10.7202/1038912ar

Article abstract

Automated analysis of oral corpora is still in its infancy. Interest is growing, but tools are still scarce. This paper presents processing tools that we have developed to analyze corpora of spontaneous oral speech in Acadian French. This variety of French spoken in the Maritime Provinces of Canada has three levels of characteristics: oral, regional, and mixed language traits. The challenge was to adapt an existing processing tool, *NooJ*, to find solutions to the problems presented by our corpora. We will present three different solutions developed with *NooJ*: (1) the configuration of dictionary entries that allows users to relate the orthographic and lexical representations of a word coming from standard French, traditional Acadian, English, or the vernacular; (2) grammars developed to process the morphological characteristics of nominal and verbal inflections; and (3) a disambiguation graph for the ambiguous form *a*, which is the 3SG pronoun in Acadian French as well as the 3SG.PRES of the auxiliary *avoir*.

Revue de l'Université de Moncton, vol.45, n^os 1 et 2, 2014, p. 273-290.

MODULE *NOOJ* DU FRANÇAIS. TRAITEMENT AUTOMATIQUE DE CORPUS DE FRANÇAIS PARLÉ RÉGIONAL

Gisèle Chevalier
et
Sylvia Kasparian
Université de Moncton

Résumé

Le traitement automatique des corpus oraux est en plein essor. L'intérêt gagne du terrain, mais les outils restent rares. Dans notre article, nous présentons un outil que nous avons développé pour l'analyse de corpus oraux spontanés en français acadien. Ces variétés de français parlées dans les Provinces maritimes du Canada ont trois niveaux de traits caractéristiques : elles sont orales, régionales et mixtes. Notre défi fut celui d'adapter et de créer un module *NooJ* acadien qui permette le traitement d'un corpus présentant de telles spécificités. Nous présentons ici trois solutions développées avec *NooJ* : 1) la configuration d'un dictionnaire qui permette la reconnaissance orthographique et lexicale de mots présentant des traits à la fois de français standard, d'acadien traditionnel et de l'anglais ou du vernaculaire; 2) les grammaires développées pour l'analyse des traits morphologiques de la flexion nominale et verbale; 3) un graphe de désambiguïsation pour *a*, qui représente non seulement la 3^e^ personne du singulier du présent du verbe *avoir*, mais aussi la 3^e^ personne du pronom personnel féminin

Cet article est issu d'une recherche dirigée de 2002 à 2008 par G. Chevalier et S. Kasparian, en collaboration avec Max Silberztein (Université de Franche Comté, France). Intitulé « Description automatique de l'Acadien », ce projet a été subventionné par la FINB, la FESR et le CRSH. Il s'agit d'une version revue et augmentée de l'article paru en 2004, « Éléments de solution pour le traitement automatique d'un français oral régional », dans *Traitement automatique des langues. Le traitement automatique de corpus oraux*. 45:2.41-62.

singulier en français acadien.

Mots clés : traitement automatique du langage, *NooJ*, oral, corpus, variété régionale, français acadien, chiac, langue mixte, langues en contact.

Abstract

Automated analysis of oral corpora is still in its infancy. Interest is growing, but tools are still scarce. This paper presents processing tools that we have developed to analyze corpora of spontaneous oral speech in Acadian French. This variety of French spoken in the Maritime Provinces of Canada has three levels of characteristics: oral, regional, and mixed language traits. The challenge was to adapt an existing processing tool, *NooJ*, to find solutions to the problems presented by our corpora. We will present three different solutions developed with *NooJ*: (1) the configuration of dictionary entries that allows users to relate the orthographic and lexical representations of a word coming from standard French, traditional Acadian, English, or the vernacular; (2) grammars developed to process the morphological characteristics of nominal and verbal inflections; and (3) a disambiguation graph for the ambiguous form *a*, which is the 3SG pronoun in Acadian French as well as the 3SG.PRES of the auxiliary *avoir*.

Keywords: automatic language processing, *NooJ*, oral, corpus, regional varieties, Acadian French, chiac, mixed language, languages in contact.

Introduction : Traitement automatique de corpus oraux en français régional

Bien qu'il existe maintenant de nombreux outils informatisés pour le traitement automatique de textes (*Hyperbase, Lexico, Alceste, Cordial, NooJ*, etc.)[1], le développement ou l'adaptation d'outils existants pour faciliter la constitution, l'annotation et la description des corpus reste un enjeu de première importance. Plusieurs de ces logiciels exécutent les concordances et appuient les linguistes dans leurs analyses lexicales

qualitatives et quantitatives, leur analyse de contenu ou de statistique lexicale, mais peu d'entre eux traitent du niveau morphosyntaxique et aucun n'est encore conçu pour l'analyse de l'oral, des variétés régionales ou de corpus qui présentent des occurrences de mélange de langues.

Le grand défi de l'analyse informatisée de l'oral reste celui des spécificités de la parole spontanée : difficulté de délimitation de la phrase orale, forte variabilité, syntaxe non canonique, redondance, phrases inachevées, etc. Un autre aspect technique non négligeable qui ralentit l'évolution des outils automatiques de description des corpus oraux est le défi de la normalisation de la transcription des corpus. Il y a un manque d'homogénéité dans la transcription de ces corpus qui, selon le cadre théorique dans lequel ils s'inscrivent, se présentent sous des formes très variées[2].

Nous avons donc relevé le défi d'automatiser la description d'une langue orale régionale, le français acadien et sa variante anglicisée, le chiac, en adaptant le formalisme de *NooJ*, logiciel de traitement automatique du langage (TAL) développé par Max Silberztein (1993, 2003, 2004).

Tel que décrit par son auteur (www.nooj4nlp.net), *NooJ* est un environnement de développement linguistique qui permet de formaliser des phénomènes linguistiques aux niveaux orthographique, lexical, morphologique, syntaxique et sémantique, sous forme de grammaires et de dictionnaires. *NooJ* comprend des outils permettant de créer et de gérer des ressources lexicales importantes, ainsi que des grammaires morphologiques et syntaxiques. Les dictionnaires et grammaires sont appliqués aux textes afin d'identifier les structures morphologiques, lexicales et syntaxiques et de marquer des mots simples et composés[3].

Les outils de *NooJ* permettent même à des chercheurs peu versés en informatique de produire les ressources linguistiques requises pour l'analyse des états de langue qui les intéressent. Toutefois, il ne sera pas possible dans l'espace qui nous est imparti de rendre compte en détail des travaux accomplis pour la réalisation du module acadien de *NooJ*. Nous nous limiterons ici à décrire trois outils fondamentaux qui ont été développés pour appuyer nos recherches, soit la structure des entrées du dictionnaire de la langue, qui permet de mettre en relation les variantes de toutes sortes (phonétiques, orthographiques, lexicales), la grammaire flexionnelle qui rend compte des formes morphologiques non standard et,

pour finir, les graphes de désambiguïsation. Nous allons d'abord donner un bref aperçu de ce qu'on appelle « les français acadiens » et, ce faisant, des défis que pose le traitement d'un parler spontané, régional, et mixte de surcroît.

1. Le français parlé en Acadie

Le français parlé en Acadie est issu d'un ensemble de parlers issus du français implanté dans l'actuelle région des Provinces maritimes (Canada), il y a plus de 400 ans, par des colons venant majoritairement du Poitou (France).

La variation linguistique entre les parlers des différentes communautés du Nouveau-Brunswick, de la Nouvelle-Écosse et de l'Île-du-Prince-Édouard reste significative malgré les efforts de normalisation de l'instruction publique dans chaque province. Dans la région du Sud-Est du Nouveau-Brunswick s'est développée une variété de langue mixte, appelée le chiac. Fondamentalement, le chiac présente une matrice française auquel se mêlent des emprunts lexicaux à l'anglais, qui ont une couverture plus large que dans les français majoritaires (québécois ou hexagonal, par exemple) de même que des traces d'influence de l'anglais sur les plans phonologique, morphologique et syntaxique. L'importance qualitative et quantitative de l'anglais sur le français acadien se fait sentir à des degrés divers, selon le locuteur et la situation de communication (Kasparian, 2003)[4].

On peut décrire les spécificités relevées dans les corpus de parler acadien que nous avons analysés[5] en les regroupant en trois strates :

1. La strate des traits d'oralité (applicables à de nombreuses variétés de français oral). Les transcriptions de productions orales visent à reproduire le plus fidèlement possible les paroles énoncées, y inclus les hésitations (*// euhm //*), les répétitions (*tout c' que/ que t'as vu*), les mots omis (*(ça-)fait-que*), les élisions (*qu(i) est en juin*) *t(u)'as vu; not(r)e*)…
2. La strate des traits régionaux (formes ou usages connaissant une distribution géographique restreinte), qui touche tous les niveaux de la langue : le niveau phono-morphologique, ex. *dans rue* « dans la rue »*, icitte* « ici »*-, awère* « avoir »*-, cte* « ce » *(cte point-là)*; le

niveau morphosyntaxique, ex. *fait que* « ça fait que »-*, il est un quart de trois* « il est trois heures moins le quart », *je voulions touT* « je voulais tout »; le niveau lexical, *zire* « vomir », *asteur* « maintenant », *hardes* « vêtements ».

3. La strate des phénomènes de contact de langue où l'on retrouve à la fois des emprunts lexicaux, exceptionnellement, de toutes les catégories grammaticales : noms, adjectifs, verbes, adverbes, marqueurs discursifs et jurons en anglais et des restructurations (réorganisation de la structure morphosyntaxique française pour accommoder des marques de l'anglais, ce qui donne naissance au chiac proprement dit. Citons les verbes anglais intégrés morphologiquement, comme *watcher* (regarder), *driver* (conduire), *freakant* (dérivé de *freak* au sens de « épeurant ») ou les verbes anglais à particules, comme *alle est tu pissée off?*[6]

Les extraits ci-dessous font ressortir le type de phénomènes qu'il nous incombe de décrire pour traiter automatiquement les corpus acadiens et chiac. On y trouve en police italique des particularismes régionaux et en police grasse des emprunts à l'anglais.

1. CK : 1-9F1[7] : La **girlfriend** *à* Roger était dans le **car** *espèrait* (attendait) que Roger arrive / **I guess** qu'*a* laisse le **car** r**un**er des quinze vingt minutes

2. CK : 1-10F1 : As-tu *entendu* le monde *qu'ont* **campé** / il y a du monde *qu'a* (est) resté dans leurs **cars** / *i* ont dit *sur le* radio *à m*atin / il y a du monde *qu'a* resté dans leurs **cars** toute la *souèrée* (soirée) avec le **motor** qui **run**ait / les **RCMP check***iont* pour *ouère* (voir) *si qu'étiont* (étaient) encore en vie

3. CK : 1-15 F1 : *Al* (elle) est *après de* (en train de) **turn**er **off** le monde / **everybody** en parle à l'*ouvrage* (travail)

4. CK : 1-18 H1 : Oui *ben* / les jeunes sont **impressed_**/ *pis* tu sais *comment c'est que* Roger est **by the time** qu'*i* sort / j'ai *rouvré* (rouvert) la porte / je voulais *i* parler / *pis* t'*arrais* (aurais) dû *entende le train* (entendre le bruit) / j'ai dit "Ton **muffler** est-*tu* **bust**é

2. L'élaboration du module acadien de *NooJ*

Concrètement, la tâche d'élaborer un module acadien peut se résumer à décrire les unités linguistiques que ne reconnait pas *NooJ* après l'application des ressources linguistiques du français commun, essentiellement, le DELAF.nod[8]. Le développement des ressources lexicales acadiennes s'est fait principalement à partir de quatre corpus, dont trois du Sud-Est, soit les corpus Anna-Malenfant (1994), Chiac Kasparian (1999) et Parkton (1999), un corpus du Nord-Est (Beaulieu, 1996), et un corpus recueilli dans trois régions du Nouveau-Brunswick, le corpus Péronnet-Kasparian (1992). En appliquant les ressources lexicales intégrées à *NooJ,* la majorité des éléments linguistiques qui constituent nos corpus sont reconnus comme formes françaises[9]. Les formes non reconnues, les '*UNKNOWNs*', sont mises à part. Elles forment notre matière première.

Dans un premier temps, il s'agissait de dresser l'inventaire des *UNKNOWN*s sous la forme d'un dictionnaire de vocables (nous l'avons intitulé ACADICO.dic). Il fallait ensuite décrire les particularismes morphologiques dans une grammaire flexionnelle (flexions nominales et verbales, ACADICO.flx). Il restait enfin à générer le dictionnaire des formes fléchies : ACADICO.nod, qui sera appliqué aux corpus, conjointement avec le DELAF.nod. Le système étant incrémentiel, il y a toujours possibilité de bonifier le dictionnaire par l'ajout successif de mots rencontrés dans de nouveaux corpus, à la seule condition que l'on génère une nouvelle version d'ACADICO.nod périodiquement.

La démarche est simple en soi, mais l'entreprise demeure compliquée vu la diversité des faits à décrire et les exigences du traitement que l'on veut faire des données d'analyse. Ayant affaire à la variation linguistique intra et interlangue, il faut être en mesure de mettre en correspondance les variantes acadiennes et anglaises avec les formes du français commun. Rappelons en outre la difficulté que présente l'hétérogénéité des conventions de transcription des formes non normalisées.

Nous nous pencherons ici sur trois solutions apportées par *NooJ* pour automatiser le traitement de certaines spécificités du corpus acadien[10] :

1. La configuration des entrées du dictionnaire pour la mise en relation des variantes orthographiques, régionales et anglaises;

2. La description des flexions nominales et verbales des formes non standard et anglaises et, en dernier lieu;
3. La désambiguïsation des formes hyper fréquentes ambiguës, comme la graphie *a* qui représente à la fois la forme populaire *a* du pronom *elle* et la forme de l'auxiliaire *avoir* au présent.

2.1. *La configuration des entrées du dictionnaire ACADICO.dic*

En plus des informations morphosyntaxiques conventionnelles (catégorie grammaticale et paradigme flexionnel), les entrées du dictionnaire ACADICO.dic sont conçues de façon à fournir des informations quant à la langue d'origine du mot vedette, quand ce n'est pas le français (LG=ac /en) et à la « glose » des mots acadiens et anglais en français courant (FC=xyz).

Le dispositif développé dans *NooJ* pour le traitement de la variation donne la possibilité de hiérarchiser les éléments de l'entrée en lemme et super lemme, ce qui permet de régler d'un seul coup la question des variantes orthographiques, régionales et anglaises. Par exemple, les différentes orthographes du mot *asteur,* contraction de « à cette heure »[11] seront reconnues comme des variantes du « super lemme » *asteur* et mises en relation avec la forme standard.

asteur, asteur, ADV+LG=ac+FC=maintenant
asteure, asteur, ADV+LG=ac+FC=maintenant
astheure, asteur, ADV+LG=ac+FC=maintenant
à cette heure, asteur, ADV+LG=ac+FC=maintenant+UNAMB
à c't'heure, asteur, ADV+LG=ac+FC=maintenant+UNAMB

Cette forme d'entrée sert à extraire les occurrences du vocable de divers corpus, qui suivent différents protocoles de transcription. En tapant l'expression naturelle <asteur> dans la boite « locate pattern », NooJ générera une table de concordances contenant n'importe quelle variante orthographique du terme dans les corpus auxquels on l'appliquera. NooJ pourra également les repérer, au moyen de la requête <FC=maintenant>.

La structure lemme-super lemme ne s'applique pas uniquement aux mots invariables. En reliant les variantes orthographiques du verbe badrer au

lemme <badrer>, NooJ pourra extraire du corpus toutes les variantes morphologiques du verbe pour toutes les variantes orthographiques.

badrer, V+FLX=Aimer+LG=ac+FC=déranger
bâdrer, badrer, V+FLX=Aimer+LG=ac+FC=déranger
bodrer, badrer, V+FLX=Aimer+LG=ac+FC=déranger

Cette même structure permet d'associer les variantes de <cheval> en français courant et en français acadien.

cheval, cheval, N+FLX=Cheval
jeval, cheval, N+FLX=Cheval+LG=ac
joual, cheval, N+FLX=Cheval+LG=ac
Les formules FLX=Aimer et FLX=Cheval dans les entrées sont des indications pertinentes pour la grammaire flexionnelle dont il sera question dans la prochaine section.

2.2. *La grammaire flexionnelle*

2.2.1. La flexion nominale

La morphologie nominale du français acadien est similaire à celle du français courant, avec, comme seule exception, la tendance à « régulariser » les paradigmes exceptionnels du français courant. La formule FLX=Cheval, intégrée à chaque entrée des variantes de <cheval> dans ACADICO.dic (ci-haut) renvoie à la règle flexionnelle suivante pour les noms se terminant en – al :

Cheval = <E>/m+s + s/m+p+ac + <B1>ux/m+p;

En vertu de cette règle, on ajoute un *s* à cheval pour former le pluriel acadien, et on remplace le *l* par *–ux* pour obtenir le masculin pluriel courant en *–aux*. Lors de la génération du dictionnaire ACADICO.nod, *NooJ* appliquera la règle flexionnelle à tous les noms répertoriés qui finissent par *–al* dans ACADICO.dic., dont *jeval* et *joual*. La figure 1 présente les entrées du dictionnaire des formes fléchies pour les variantes du super-lemme <cheval>, incluant donc *jeval* et *joual*.

```
cheval, cheval, N+FLX=Cheval+m+s
chevals, cheval, N+FLX=Cheval+m+p
chevaux, cheval, N+FLX=Cheval+m+p
jeval, cheval, N+FLX=Cheval+FC=Cheval+m+s
jevals, cheval, N+FLX=Cheval+FC=Cheval+m+p
jevaux, cheval, N+FLX=Cheval+FC=Cheval+m+p
joual, cheval, N+FLX=Cheval+FC=Cheval+m+s
jouaux, cheval, N+FLX=Cheval+FC=Cheval+m+p
```

Figure 1 : Variantes de <cheval> dans ACADICO.nod

De cette façon, la requête de localiser le super lemme <cheval> nous donne alors toutes les variantes graphiques et morphologiques de « cheval » attestées dans les corpus étudiés (Figure 2).

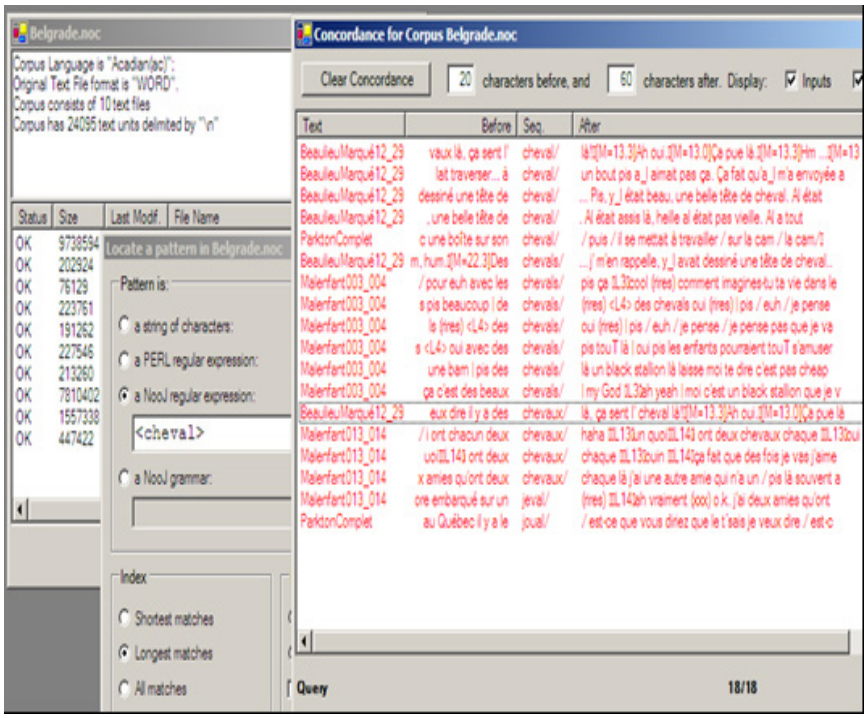

Figure 2 : Concordances du super lemme <cheval>

2.2.2. La flexion des verbes acadiens

La flexion verbale typique du français acadien est la forme *–ont* à la 3e personne du pluriel : *ils allont, ils alliont, ils iriont, qu'ils alliont*. C'est un vestige du 16e siècle, mais elle est encore très productive dans les variétés acadiennes les plus conservatrices. Il a suffi d'ajouter une ligne à la description morphologique des verbes (les quelque 94 paradigmes du Bescherelle), spécifiant que la 3e personne se construit par l'ajout de la variante *-ont* en plus de la variante *–ent* à tous les temps grammaticaux, et que la première est marquée du trait (+ac). La règle s'applique même aux verbes empruntés à anglais qui se conjuguent sur le modèle des verbes du premier groupe. La figure 3 présente les formes du verbe *minder*, générées par la règle flexionnelle FLEX=AIMER, parmi lesquelles se retrouve *(ils) mindont (pas)* (ils ne s'en formalisent pas), ils mandiont pas, ils manderiont pas…

Un autre phénomène rencontré est la production de flexions verbales non standard attribuables, comme pour les flexions nominales, au processus de régularisation des verbes irréguliers, comme *disez*,*faisez* à la 2e personne du pluriel ou *il faulait qu'ils faisent* au subjonctif. Elles sont également marquées (+ac) dans les dictionnaires ACADICO.nod, même si elles ont cours en français populaire dans toute la francophonie (Figure 4).

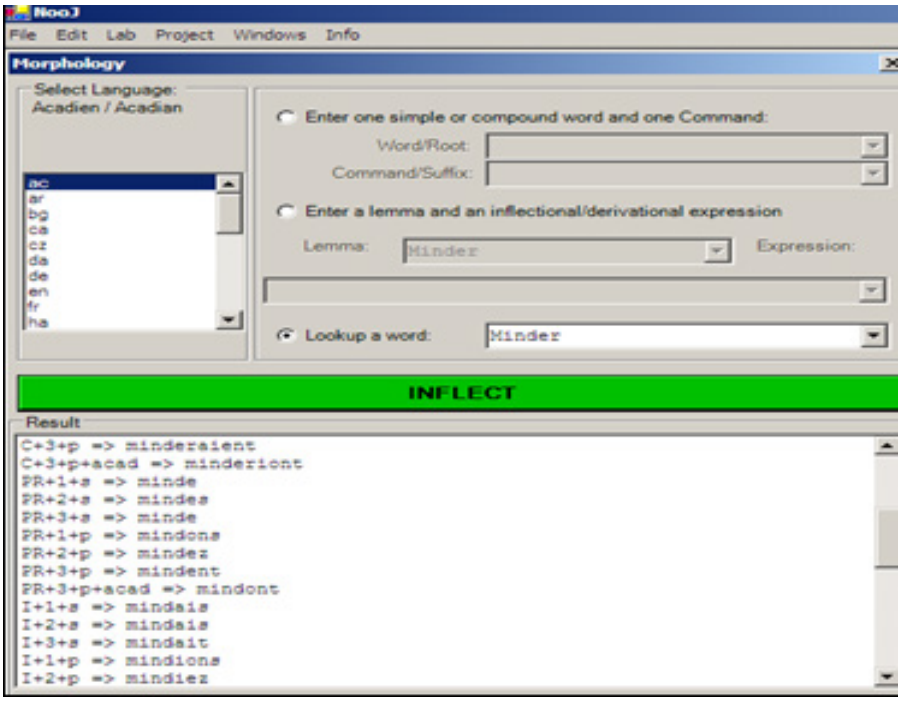

Figure 3 : Application de la grammaire flexionnelle acadienne au verbe mixte *minder*

Un autre phénomène rencontré est la production de flexions verbales non standard attribuables, comme pour les flexions nominales, au processus de régularisation des verbes irréguliers, comme *disez*, *faisez* à la 2e personne du pluriel ou *il faulait qu'ils faisent* au subjonctif. Elles sont également marquées (+ac) dans les dictionnaires ACADICO.nod, même si elles ont cours en français populaire dans toute la francophonie (Figure 4).

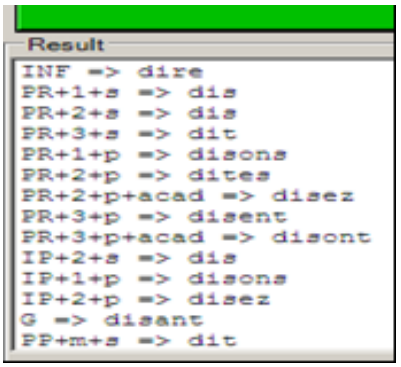

Figure 4 : Flexions du verbe *dire* avec la grammaire acadienne

Ainsi quand nous posons la requête <V+3+p>, soit « verbe à la 3e personne du pluriel », on obtient à la fois les verbes de français commun (*elles sont, ils ont, mes parents comprennent*), les verbes à flexion acadienne (*i écoutont, i m'écoutont, i avont, i appelont, tes parents te laisseriont*) et les verbes anglais conjugués selon la forme acadienne ou standard (*i turnont, i wonderont, i pukaient, ils se behavent*). Si par contre, uniquement les formes acadiennes nous intéressent, la requête <V+3+p+ac> permet d'isoler les concordances de ces formes acadiennes.

2.3. *Graphes de désambiguïsation : l'exemple de* a*, soit auxiliaire, soit pronom*

La transcription du pronom *elle* par la graphie *a* pour représenter sa prononciation effective, entraîne une fâcheuse ambiguïté avec l'auxiliaire *avoir* au présent de la 3e personne du singulier. Cette ambiguïté vient brouiller toute étude sur les pronoms sujets en acadien. Pour lever ce type d'ambiguïtés, on a recours à la construction de graphes qui décrivent la distribution syntaxique propre à chaque catégorie.

Le graphe présenté dans la figure 5 propose les deux chemins possibles de *a* de façon à décider si la graphie représente le pronom ou le verbe dans chacune de ses occurrences dans le texte :

1. la ligne qui va vers la partie supérieure du graphe indique que *a* est un pronom lorsqu'il est suivi d'un verbe à la 3e personne, ou encore,

dans les cases superposées, lorsqu'il peut être suivi d'autres pronoms comme *le, la, les / leur, lui / en / y*, placés avant le verbe.

2. la ligne qui va vers le bas du graphe indique que *a* est un verbe s'il est précédé des pronoms personnels contenus dans la longue boîte à gauche ou du pronom impersonnel transcrit *i-y-a* « il y a »; enfin *a* est un verbe auxiliaire s'il est suivi d'un participe passé ou un adverbe pouvant se placer avant ce participe passé.

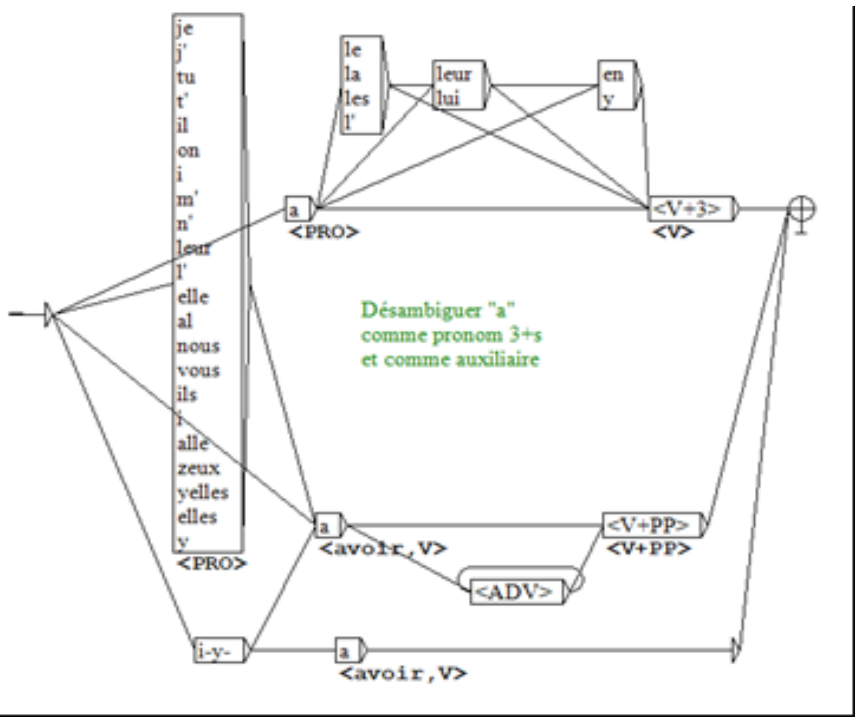

Figure 5 : Graphe de désambiguïsation de *a*

L'application de ce graphe aux corpus acadiens permet de localiser précisément et uniquement les occurrences de *a* verbes ou pronoms et de regrouper les occurrences dans des tables de concordances séparées. La figure 6 reproduit un extrait des concordances obtenues par l'application de la grammaire de désambiguïsation du pronom et de l'auxiliaire *a*.

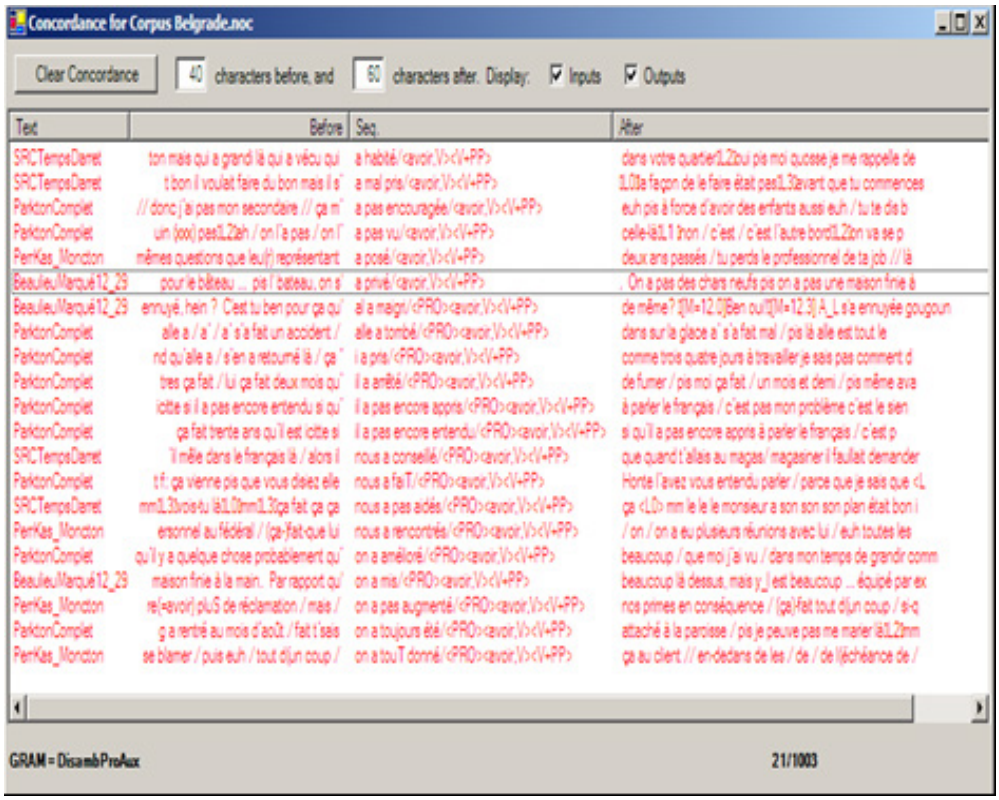

Figure 6 : Concordances du verbe avoir à la 3^e^ personne du singulier

Conclusion

L'environnement *NooJ* nous a permis de relever le défi de traiter automatiquement des corpus de productions dans une variété de français vernaculaire régionale et mixte. Le module *NooJ* du français acadien et du chiac fonctionne au moyen d'un dictionnaire français de référence de large couverture, le DELAF.nod, conjointement avec le dictionnaire du français acadien, ACADICO.nod, qui a priorité sur le dictionnaire français à l'étape de l'analyse lexicale du corpus. Le dictionnaire du français acadien « reconnait » les variantes orthographiques et morphologiques des unités

lexicales non standards du français, y compris les unités lexicales empruntées à l'anglais, qu'elles soient intégrées morphologiquement (les noms et les verbes) ou invariables (marqueurs discursifs, conjonction, adverbes, particules verbales).

Vu la définition à la fois large et étroite que nous avons donnée du français acadien, soit, d'une part, l'ensemble des formes du français standard plus des régionalismes et des emprunts à l'anglais et, d'autre part, l'accent mis sur les formes qui se trouvent dans les productions spontanées de registre familier ou populaire, le module acadien de *NooJ* rend possible les recherches sur la variation touchant les particularismes acadiens (le lexique du vieil acadien, la troisième personne du pluriel en *–ont*) et ceux du chiac (les mots anglais fléchis en français, les verbes à particules). Il peut également servir à l'analyse des formes vernaculaires communes aux français du Canada et, pour finir, à l'extraction des formes de la langue parlée en général, comme l'étude des variantes morphologiques des pronoms personnels, par exemple.

ACADICO.nod ayant été élaboré à partir de corpus oraux incluant une des variétés de français les plus conservatrices, celle du Sud-Est du Nouveau-Brunswick, il peut prétendre couvrir, en principe, les variantes morphosyntaxiques du français oral populaire de la plupart des français parlés au Canada[12]. À ce titre, la prochaine étape logique de notre entreprise consisterait à appliquer le module acadien de *NooJ* à des corpus représentatifs d'autres variétés de français canadiens, incluant le français québécois, et à examiner la nature des formes non reconnues générées par le logiciel. Nous faisons l'hypothèse que la plupart des formes non reconnues, les *UNKNOWNs*, relèveraient des conventions de transcription du corpus et son degré de granularité, d'une part, et, d'autre part, des traces d'oralité (hésitations, reprises, troncations, ruptures syntaxiques), des emprunts nominaux et verbaux à l'anglais de faible fréquence, à l'exclusion des morphèmes grammaticaux et de phénomènes dérivationnels ou flexionnels[13]. De par sa nature incrémentielle, le dictionnaire pourrait être bonifié par l'ajout d'items lexicaux venus de différentes régions francophones du Canada.

Nous avons construit maintenant le socle du module *NooJ* acadien, qui pourrait être appelé à devenir un module des français parlés au Canada. Nous nous sommes limitées ici aux fonctions lexicales et

morphosyntaxiques de *NooJ*. Le module ouvre des pistes de recherches innombrables pour le traitement automatisé de corpus en français, que ce soit au niveau lexical (par exemple, les expressions figées, les mots du discours), morphosyntaxique (les structures avec les prépositions, l'usage des pronoms) ou sémantique (réseaux sémantiques, identification des entités nommées, etc.).

Bibliographie

Chevalier, G. (2008a). L'interrogation de corpus oraux en français périphérique. Communication présentée à la *Journée d'étude du labo MoDyCo (Paris X, Nanterre) : Français périphérique et traitement des corpus oraux non standard* tenue le jeudi 17 avril 2008, sous la direction de Colette Noyau.

----- (2008b). Les français du Canada : faits linguistiques, faits de langue. *Alternative francophone.* 1:1.80-97.

Chevalier, G., Kasparian. S., et Silberztein, M. (2004). Éléments de solution pour le traitement automatique d'un français oral régional. *In* Véronis, J. (dir.). Le traitement automatique des corpus oraux. *Revue française en traitement automatique du langage (TAL).* 2:45.41-62.

Kasparian, S. (2003). Parler bilingue et actes identitaires : le cas des Acadiens du N.-B. *In* Stebbins, R.A., Romney, C., et Ouellet, M. (dir.). *Francophonies et langue dans un monde divers en évolution : contacts interlinguistiques et socioculturels*. Winnipeg : Presses universitaires de St Boniface. 159-177.

Péronnet, L. (1988). *Le parler acadien du sud-est du Nouveau-Brunswick. Eléments grammaticaux et lexicaux*. New York : Peter Lang.

Perrot, M.-È. (1995). *Aspects fondamentaux du métissage français / anglais dans le chiac de Moncton (Nouveau-Brunswick, Canada)*. Thèse de doctorat, Université de la Sorbonne Nouvelle Paris III.

----- (2000). Ordre des mots et restructurations dans le chiac de Moncton : l'exemple du syntagme nominal. *Cahiers de linguistique de l'Inalco*. 1-3.

Poirier, P. (1993). *Le glossaire acadien; Édition critique établie par Pierre M. Gérin.* Moncton : Éditions d'Acadie et Centre d'études acadiennes

Silberztein, M. (2003). *NooJ Manual*. http://www.nooj4nlp.net.

----- (2004). *NOOJ*: an Object-Oriented Approach. *In* Muller, C., Royauté, J., et Silberztein, M. (dir.), *INTEX pour la linguistique et le traitement automatique des langues*. Bezançon : Presses Universitaires de Franche-Comté. 359-369.

Corpus dépouillés pour l'élaboration du module *NooJ*

Gauvin, K., et Chevalier, G. (1994). *Corpus Anna-Malenfant (N.-B., Canada)*, Université de Moncton, Moncton (Canada).

Six conversations en dyades entre jeunes de 11 à 12 ans (20 000 mots)

Kasparian, S. (1999). *Corpus chiac Kasparian*. Université de Moncton, Moncton (Canada).

Une trentaine de conversations spontanées entre jeunes de 18 à 24 ans ou entre les jeunes et leurs parents (84 600 mots)

Poissant, G. (1995). *Corpus Parkton*. CRLA, Université de Moncton, Moncton (Canada).

29 entrevues sociologiques recueillies auprès de résidents d'un quartier de niveau socio-économique faible (177 900 mots)

Péronnet, L., et Kasparian S. (1992). *Corpus Péronnet-Kasparian*, Université de Moncton, Moncton (Canada).

18 entrevues formelles auprès de jeunes cadres ayant une formation universitaire et œuvrant dans des entreprises francophones dans trois régions du N-B. (35 000 mots)

Beaulieu, L. (1996). *Corpus sociolinguistique du français acadien du Nord-Est du Nouveau-Brunswick*, Université de Moncton, Campus de Shippagan, Shippagan, Nouveau-Brunswick (Canada).

16 entrevues semi-dirigées avec des locuteurs représentatifs de différentes couches sociales, les deux sexes, niveau d'éducation et types de réseaux sociaux (210 000 mots)

Corpus témoins

Étienne, G. (2008). *Corpus Gérard Étienne du français de la péninsule acadienne des années 1970.* Transcription réalisée par les bons soins du

CRLA (Centre de recherche en linguistique appliquée), Université de Moncton.

Hallion, S. (2000). *Étude du français parlé au Manitoba*. Thèse de doctorat, Université de Provence, Aix-en-Provence, 3 vol., 464 f. + 859 f. (corpus).

---

1 Ce type d'outils ont été notamment développés dans le courant des JADT européennes (Journée d'Analyse de Données textuelles), voir les actes de ces journées (www.cavi.univ-paris3.fr/JADT) ainsi que la revue *Lexicometrica*. (www.cavi.univ-paris3.fr/lexicometrica).

2 Outre l'exemple de *asteur* (point 2.3), mentionnons le cas des formules interrogatives complexes telle que *où est-ce que* rendue par l'une ou l'autre des graphies suivantes : *où ce que, où'c'que, ouesque, ousque, ouske, yousque, ayousque*…

3 *NooJ* est utilisé par des chercheurs travaillant sur plus d'une douzaine de langues appartenant aux langues romanes, germaniques, slaves, sémitiques, aux langues d'Asie et au hongrois.

4 Pour les aspects micro-linguistiques, voir Péronnet (1988) et Chevalier (2008b) en ce qui a trait aux particularismes acadiens. Concernant les emprunts à l'anglais dans le chiac de Moncton, voir l'étude exhaustive de Perrot (1995).

5 La liste des corpus dépouillés se trouve dans une section de la bibliographie.

6 Voir également la description de la restructuration du groupe nominal dans le chiac de Moncton dans Perrot (2000).

7 Le code CK : 1-9F1 placé devant les citations identifie la source et le sujet de l'énonciation. En l'occurrence, CK renvoie au corpus Chiac-Kasparian, 1 — identifie le premier des 30 entretiens du corpus de polylogues dans lesquels plusieurs femmes ou hommes interviennent. 9F1 indique le numéro de l'intervention (9) et le fait qu'elle a été faite par la première femme qui prend la parole (F1). La même codification s'applique dans les autres citations. CK : 1-18 H1 à la citation (4) renvoie à l'intervention 18 du premier entretien du corpus Chiac-Kasparian produite par le premier locuteur homme à prendre la parole.

8 DELAF.nod tient pour Dictionnaire électronique des formes fléchies des mots du français.

9 Il peut se produire une confusion entre les graphies françaises et anglaises. Il y a ainsi confusion entre *but* « mais », emprunté à l'anglais et *but* au sens de « but ». La désambiguïsation des occurrences et leur étiquetage grammatical approprié est une étape ultérieure dans l'établissement des corpus.

10 On peut suivre l'évolution du projet en consultant les articles cités dans les références.

11 Voir les multiples graphies de cette expression au 16^e siècle dans le *Glossaire acadien* de Pascal Poirier (édition critique établie par Pierre Gérin (1993).

12 Les entrées du *Glossaire acadien* de Pascal Poirier (édition critique établie par Gérin, 1993), qui sont du vocabulaire propre au « vieil acadien » ont été ajoutées à la base lexicale.

13 C'est la conclusion que l'on a tirée du dépouillement du *Corpus Gérard Étienne du français de la péninsule acadienne* (2008) et du *Corpus du français franco-manitobain* de Sandrine Hallion (2000) (Chevalier, G., 2008, inédit).