THE ACADIAN NOOJ MODULE: AUTOMATIC PROCESSING OF A REGIONAL ORAL FRENCH*

Sylvia Kasparian
Université de Moncton

ABSTRACT

Automated analysis of oral corpora is still in its infancy. Interest is growing, but tools are still scarce. This article 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, INTEX/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 morphological characteristics of nominal and verbal inflections; and (3) a disambiguation graph for *a*, which is the 3SG pronoun in Acadian French as well as the 3SG.PRES of the auxiliary *avoir*.

Key words: automatic language processing, INTEX/NooJ, oral corpus, regional varieties, Acadian French, *chiac*, mixed language, languages in contact

RÉSUMÉ

L'analyse informatisée de corpus oraux est encore à ses débuts. Malgré l'intérêt grandissant à l'étude de corpus oraux, les outils informatisés qui le permette sont encore rares. Notre article présente le module acadien NooJ que nous avons développé pour le traitement automatique de parlers régionaux acadiens. Cette variété de français parlée dans les provinces maritimes du Canada se caractérise par des traits d'oralité, de régionalisme et de contact de langues.

*This article presents the results of a research project, the *Dictionnaire électronique de l'acadien* ['*Electronic Acadian Dictionary*'], carried out jointly by Sylvia Kasparian, Laboratoire d'analyse de données textuelles (LADT) [Textual Data Analysis Laboratory], Université de Moncton; Gisèle Chevalier, Centre de recherche en linguistique appliquée (CRLA) [Applied Linguistics Research Centre], Université de Moncton; and Max Silberztein, INTEX/Nooj software designer, Université de Franche-Comté, France.

Notre défi consistait à adapter un outil existant, INTEX/NooJ afin qu'il puisse traiter les particularités de notre corpus. Nous vous présentons donc trois solutions développées avec NooJ : (1) La configuration des entrées du dictionnaire qui permettent de mettre en relation les variantes orthographiques et lexicales d'un mot, variantes provenant du français standard ou acadien, de l'anglais ; (2) Les grammaires développées pour la reconnaissance des caractéristiques morphologiques, flexions des noms et verbes ; enfin (3) un exemple de graphe de désambiguïsation, celui de *a*, qui peut être le pronom personnel 3sg en acadien ainsi que la la 3 sg présent de l'auxiliaire avoir.

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

1. INTRODUCTION

Linguistic studies that base their descriptions on large electronic corpora are gaining ground, and the number of increasingly sophisticated automated analysis tools continues to grow. Despite that, the automated analysis of oral corpora remains marginal. There are several reasons for this. First, the idea that the written word, which represents the standard, is the only corpus worthy of study is still firmly fixed in Western thought, even though various authors have denounced the prejudices to which the oral corpus has been subjected (cf. Blanche-Benveniste 1997 and recent work in corpus linguistics and conversational analysis). The phenomena unique to the oral corpus are often considered aberrations and dysfunctions when compared to the standardized grammar of the written corpus. Second, the costs associated with creating the oral corpus are undeniably a factor that restricts researchers. The creation of oral corpora is fastidious, time-consuming, and expensive. Finally, the greatest challenge facing the automated analysis of the oral corpus is still that of the specific character of the spontaneous word: the difficulty in delimiting the oral phrase, strong variability, non-canonical syntax, redundancy, incomplete sentences, and so on.

Another significant technical aspect that slowed down the development of automated tools for describing the oral corpora is the difficulty in standardizing the transcription of these corpora. There is a lack of homogeneity in the transcription of oral corpora, which, depending on the theoretical framework of which it forms a part, can take very different forms.

Since the capture and automation of significant amounts of text on the web as well as their ever-increasing dissemination, the tools for the automatic processing of written corpora have been strongly developed. However, equivalent tools for the oral corpus are still far from being readily available. Compared to the written corpora, there are still only a limited number of oral corpora, the largest of which is the oral part of the BNC (British National Corpus; see Burnard 1995), which includes 10 million occurrences in English. Corpus-related projects of this extent are rare for other languages. In fact, for French, the only large corpora are the Valibel (Francard, Geron, and Wilmet 2002). Cf. `valibel.fltr.ucl.ac.be`) database for

French in Belgium and the GARS/DELIC (Blanche-Bénvéniste 1990) for continental French, on the order of 3.5 to 4 million occurrences. There are also a number of important corpora of French spoken in Quebec: the *Corpus Sankoff-Cedergren, Montréal, 1971* (Sankoff, Sankoff, Laberge and Topham, 1976, for example), *Le corpus Montréal 1984* (Thibault and Vincent, 1990), and *Le corpus Montréal 1995* (Vincent, Laforest and Martel, 1995). Consequently, it has only been for the last 30 years, with the evolution of the branches of corpus linguistics and conversational analysis, that the study of the spoken language has generated a growing interest that has, on the one hand, spurred on the creation of oral corpora and, on the other, given a new impetus in recent years to the natural language processing community to develop software programs for the automatic processing of oral corpora. On the same subject, it would be worthwhile to consult Veronis (2004).

Although there is now a variety of computerized tools for the automatic processing of texts (HyperBase, Lexico, Alceste, Cordial, INTEX/NooJ, etc.),[1] the development or adaptation of existing tools to facilitate the creation, annotation, and description of corpora remains a major challenge. Several of these programs produce concordances and support linguists in their qualitative and quantitative lexical analyses, content analysis, or lexical statistics, but few of them operate on a morphosyntactical level, and none has yet been designed to analyze the oral language, regional varieties, or bilingual and/or multilingual corpora. Therefore, we confronted the challenge of automating the description of a regional oral language, Acadian, by adapting the formalism of INTEX/NooJ, a TAL software program developed by Max Silberztein (1993, 2004),[2] LASELDI, Université de Franche Comté, France.

[1]These types of tools have mainly been developed under the auspices of the European JADT (Journée d'analyse de données textuelles) ["Textual data analysis day"]; see the proceedings from these meetings (`www.cavi.univ-paris3.fr/JADT`) as well as the journal *Lexicometrica* (`www.cavi.univ-paris3.fr/lexicometrica`).

[2]Briefly, NooJ is a linguistic development environment that includes large-coverage dictionaries and grammars, and parses corpora in real time. NooJ includes tools to create and maintain large-coverage lexical resources as well as morphological and syntactic grammars. Dictionaries and grammars are applied to texts in order to locate morphological, lexical and syntactic patterns and tag simple and compound words. NooJ can build complex concordances, with respect to all types of finite-state and context-free patterns. NooJ users can easily develop extractors to identify semantic units in large texts, such as names of persons, locations, dates, technical expressions etc. NooJ dictionaries are associated with inflectional and derivational morphological descriptions for simple and compound words. Inflectional and derivational paradigms are formalized as structured libraries of graphs or text-based rules. Nooj's set of morphological operators can be adapted to each language. NooJ's morphological and syntactic grammars are structured libraries of graphs. NooJ's morphological and syntactic engines are unified, which allows syntactic grammars to include morphological operators.

NooJ can currently process a dozen languages including some Romance, Germanic, Slavic, Semitic, and Asian Languages, as well as Hungarian. NooJ dictionaries and gram-

Before presenting the components of the solution proposed for the automatic processing of Acadian using NooJ, we must clarify what we mean by Acadian French.

2. ACADIAN FRENCH

"Acadian French" consists of a group of varieties of French, spoken by speakers of French origin who settled in the area presently known as the Maritime provinces (Canada) 400 years ago. The colonists came mainly from Poitou, France.

The linguistic variation among the dialects of the different communities in New Brunswick, Nova Scotia, and Prince Edward Island remains significant despite efforts to standardize public education in each province. In the southeast region of New Brunswick, a variety of mixed language known as the *chiac* of Moncton has developed. The *chiac* language has a French matrix and a lexicon that has been generously enriched by English. English has permeated the Acadian language at every level: lexical, phonological, morphological, and, to some extent, syntactical. The degree of anglicization of the language varies and often depends on the subjects and the circumstances of the communication situation (Kasparian 2003).

The characteristics of the corpus of Acadian dialects can be grouped into three strata: oral, regional, and language contact.

2.1. The stratum of oral characteristics

This stratum is applicable to the numerous varieties of oral French. The goal of transcriptions of oral production is to reproduce as faithfully as possible the word pronounced (1).

(1)

hesitations	//euhm//
repetitions	tout c' que/ que t'as vu
skipped words	(ça-)fait-que
elisions	qu(i)est en juin; t(u)as vu; not(r)e
misfires	mon/ma licence

2.2. The stratum of regional characteristics

These include forms or usages of limited geographical distribution, which affect all levels of the language, as found in (2).

(2) a. *Phonomorphological level*:

dans rue	< dans la rue
icitte	< ici
awère	< avoir
cte	< ce, as in *cte point là*

mars are extremely simple objects to build and this tools can be shared among NooJ community members (cf. Silberztein `www.nooj4nlp.net/`).

b. *Morphosyntactical level*:

fait que	< ça fait que
il est un quart de trois	< il est trois heures moins le quart
je voulions tout	< je voulais tout

c. *Lexical level*:

zire	< dégoût
asteur	< maintenant
hardes	< haillons
yinque	< rien que

2.3. The stratum of language contact

Phenomena in this stratum include both borrowings (nouns, adjectives, verbs, adverbs, discursive markers, English expletives) and restructuring (reoganization of the morphosyntax of both languages to create a unique one called *chiac*). Consider, for example, the morphologically integrated English verbs, *watch*, in *watch*er (regarder), *drive* in *driv*er (conduire), and *freak*ant derived from freak ('épeurant' ou 'effrayant'), or English verbs with particles integrated into Acadian, as in "Alle est tu *pissé off?*" Here are two examples of utterances in *chiac*: "Je ne veux pas que mes enfants *turn*ont *out* comme des *bums* " and "*c'est des cool movies* intéressantes que j'ai *watché* hier."

3. THE CORPORA

Our research addresses the morphosyntactic labelling of the Acadian corpora that have already been transcribed and digitized by various researchers at different times and for different research purposes; thus, our research is based on the corpora, detailed in (3).

(3) a. *Chiac Kasparian H99* Corpus (84,600 words): some 30 spontaneous conversations between young adults aged 18–24 or between young people and their parents (Kasparian 1999);

b. *Parkton* Corpus (177,900 words): 29 sociological interviews collected from the residents of a poor socioeconomic neighbourhood (Poissant et al. 1995);

c. *Anna Malenfant* Corpus (20,000 words): six conversations in dyads between preteens aged 11 to 12 (Gauvin et Chevalier 1994);

d. *Péronnet-Kasparian* Corpus (35,000 words): 18 formal interviews with groups of university-educated young people working in francophone companies in three areas of New Brunswick (Péronnet-Kasparian 1992).

An example of Acadian French extracted from the *chiac Kasparian H99* Corpus (the English sections are in bold, the varieties of regional French in italics) is presented in (4).

(4) **1-9F1:** La **girlfriend** *à* Roger était dans le **car** *espèrait (attendait)* que Roger arrive/**I guess** qu'*a* laisse le **car run**er des quinze vingt minutes

1-10 F1: As-tu *entendu* le monde *qu'ont* **campé**/il y a du monde *qu'a* 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* encore en vie

1-11 H1: Il y a **one thing about it**/*al* est **canadian**

1-12 F1: *Al* est **show-off**

1-14 H1: Oui/*al* est **show-off 'cause she's good**

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

1-16 H1: *Ben* oui *vous autres*

1-17 F1: She's not impressing nobody

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* **busté** ...

Our challenge is to automate the description of a mixed, regional, oral language. To do that, we had to adapt the INTEX/NooJ formalism (Max Silberztein 1993, 2004) to process our corpus, which entailed the additional problem of the lack of homogeneity in the transcription conventions. At the time this corpus was transcribed, there were no common transcription standards for special regional features or unique characteristics of register. Some examples of the many types of writing that we found in the transcription of our corpus are shown in (5).

(5)

avoir	→	aouèr, awèr, awér
ici	→	icitte, icite, iciT
chiac	→	shiaque, shiac, chiak
ces	→	ctés, ctes, ctès
tout ce que tu as vu	→	tout c(e) que/ que t(u) as vu; touT c'que t'as vu
notre	→	not, note, noT, not(r)e

4. ACADIEN NOOJ MODULE

Several significant modifications to NooJ with regard to INTEX responded to the challenges of processing our corpus: (i) the inclusion of a single dictionary of atomic linguitic unities (ALU) of different sizes: simple words, complex words, frozen or semi-frozen expressions; (ii) the classification of units in a hierarchy of variants, lemmas, and super-lemmas; and (iii) the recording in the dictionary entry of the semantic and syntactic properties of the predicates (V N1).

In this article, we present three solutions prepared using NooJ for automating the processing of certain characteristics of our corpus (6):

(6) a. Super-lemmas and dictionaries: solution for orthographic, regional, and English variants.

b. Morphology: developed grammars
- inflectional markers of the gender/number of Acadian nouns
- inflectional markers of verbs: Acadian and English verbs that are mopholog-ically integrated into French; English verbs with particles

c. Disambiguation graphs: the example of the auxiliary *a*, which can be both the third-person singular present of the verb *avoir* and the Acadian pronoun *elle*.

4.1. *Dicionatary of Acadian French*: Super-lemmas

The ranking by NooJ of units into variants, lemmas, and super-lemmas, as well as the inclusion of a single dictionary of atomic linguistic units of different sizes, allowed us to regulate the orthographic, regional, and English variants at the same time. In NooJ the super lemma is "a word that acts as a canonical form for the lexical entries as well as all their inflected and derived forms" (Silberztein, 2002–2008:80). When constructing the dictionary, we also used the canonical entry for orthographic variants as well as for lexical variation.

We will briefly show how the entries in the dictionary are configured to link together graphic, lexical, and morphological variants of the same word, and to establish semantic or notional links of equivalency between these variants in the *chiac* vocabulary originating from different sources: standard French, the vernacular, "traditional" Acadian, and English.

Linking all of the variants of the same word to the same super-lemma (second word in the dictionary entry) makes it possible to search by means of a simple query for all the occurrences of this super-lemma as well as the associated lemmas in the text (occurrences of all the orthographic and regional variants of this word). An excerpt from the Acadian dictionary prepared by NooJ is provided in (7).

(7) *Excerpt of entries from the dictionary* `Acadico.dic`

```
à, PREP
abandonner,V+FLX=Aimer
a,elle,PRO+CLS+3+f+s
abandouner,abandonner,V+FLX=Aimer
abat,N+FLX=Crayon
abatis,N+FLX=Agrès
abattant,A+FLX=Lacté
abattoir,N+FLX=Crayon
abattouer,abattoir,N+FLX=Crayon
aberouer,abreuvoir,N+FLX=Crayon
aberver,abreuver,V+FLX=Aimer
abeurver,abreuver,V+FLX=Aimer
abîmer,V+FLX=Aimer
aboiteau,N+FLX=Agneau
abominer,V+FLX=Aimer
abondant,A+FLX=Lacté
abord,N+FLX=Crayon
aborder,V+FLX=Aimer
abouette,N+FLX=Table
abouetter,V+FLX=Aimer
abouler,V+FLX=Aimer
```

...

```
about,PART+En
aboutissement,N+FLX=Crayon
âbre,arbre,N+FLX=Crayon
assoyont,asseoir,V+FLX=Asseoir+PR+3+p+acad
asteur,asteur,ADV+LG=ac+FC=maintenant
asteure,asteur,ADV+LG=ac+FC=maintenant
astheure,asteur,ADV+LG=ac+FC=maintenant
asthme,asthme,N+FLX=Crayon+m+s
bad,A+FLX=Nice+LG=en
bâdrer,badrer,V+FLX=Aimer+LG=ac+FC=déranger
badrer,V+FLX=Aimer+LG=ac+FC=déranger
bag,N+FLX=Crayon+LG=en
blusher,V+FLX=Aimer+LG=en
bodrer,badrer,V+FLX=Aimer+LG=ac
boire,V+FLX=Boire
```

It should be noted that the entries in (7) that represent regional Acadian or English variants, such as *a* for the 3SG.FEM pronoun *elle*, the pronunciation *abattouer* for *abattoir* and *abandouner* for *abandonner*, or the case of *asteur* (which appears in the following three spellings — *asteur*, *asteure*, and *astheure*) are given in (8). Each spelling has a separate entry in the dictionary, but all three spellings link back to the same super-lemma adverb, *asteur*, for which `LG=ac` is indicated, i.e., the Acadian language, followed by the annotation `FC=maintenant`, which indicates the equivalent of the term in standard French:

(8)
```
asteur,asteur,ADV+LG=ac+FC=maintenant
asteure,asteur,ADV+LG=ac+FC=maintenant
astheure,asteur,ADV+LG=ac+FC=maintenant
```

Another example is the case of *badrer* (the English verb 'to bother', morphologically integrated into French), for which we have three spellings too and three dictionary entries, and *badrer* as the super-lemma under which the other two forms are grouped. These three forms correspond to the standard French, `FC = déranger`, indicated at the end of the dictionary entry (9):

(9)
```
bâdrer, badrer,V+FLX=Aimer+LG=ac+FC=déranger
badrer,V+FLX=Aimer+LG=ac+FC=déranger
bodrer,badrer,V+FLX=Aimer+LG=ac+FC=déranger
```

A search using super-lemmas allows us to retrieve with a single command the concordances of all the spellings or variants of this same lemma in the corpus. Searching on the field `FC=déranger` allows us to retrieve all the concordances for *déranger*, as well as the Acadian variants, *badrer*, *bâdrer*, and *bodrer*.

4.2. Morphology: Nominal and verbal inflection

In INTEX/NooJ, grammars can be prepared either in the form of grammatical rules (models of inflections that consist of a codified phrase) assembled in the dictionar-

ies of inflections (for example, the inflection for *cheval*), as in example (10), or in the structured set of graphs (for example, of Acadian verbs), as in Figure 6.

4.2.1. Gender/number of Acadian nouns

In Acadian, certain words have a different gender than they do in standard French; for example, Acadians would say "*une* autobus-SG.F" instead of "*un* autobus-SG.M". Gender and number of certain nouns and adjectives may also differ from that of standard French, particularly for words that end in "al". As a result, we have "des *chevals*" instead of *chevaux* and "des élus *provincials*" instead of *provinciaux*. For example, the first line, in bold, in (10) shows the grammar developed for the inflections of *cheval*. This grammar gives three rules:

(i) If nothing comes after *cheval*, it is the masc + sing form (**<E>/m+s**)

(ii) If *s* is added after *cheval*, it is the masc + plur form (**+ s/m+p**)

(iii) If one letter is removed from the end and *ux* is added, it is the masc + plur form (**<B1>ux/m+p**)

The application of this grammar to the lemma *cheval* automatically generates the corresponding inflected forms in the NooJ inflection dictionary.

(10) *Grammar and flexions of cheval in Acadian*:

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

A search on the super-lemma *cheval* will then give us all the graphic and morphological variants of the word *cheval* as shown in the concordance table (Figure 1).

4.2.2. Inflections of Acadian verbs

The conjugation of verbs involves the same inflectional forms as in standard French for the formal Acadian register. It is mainly in the register of vernacular Acadian French or in the networks of intimate conversations that we find the regional forms. Below we present one of the unique features of Acadian conjugation, that of inflections of the third-person plural. In Acadian we find the conjugation in *-ont* (pronounced *-ant* in certain regions of Acadie) in all tenses and moods in the third-person plural: *ils allont, ils alliont, ils iriont, qu'ils alliont*. This is a vestige of the sixteenth-century French that is still very much alive in contemporary Acadie.

This inflection also applies to English verbs integrated into Acadian: "ils *mind*ont pas" = "ils ne s'en formalisent pas". The English verbs are integrated into Acadian in the form of verbs belonging to the first conjugation (regular) class. Depending

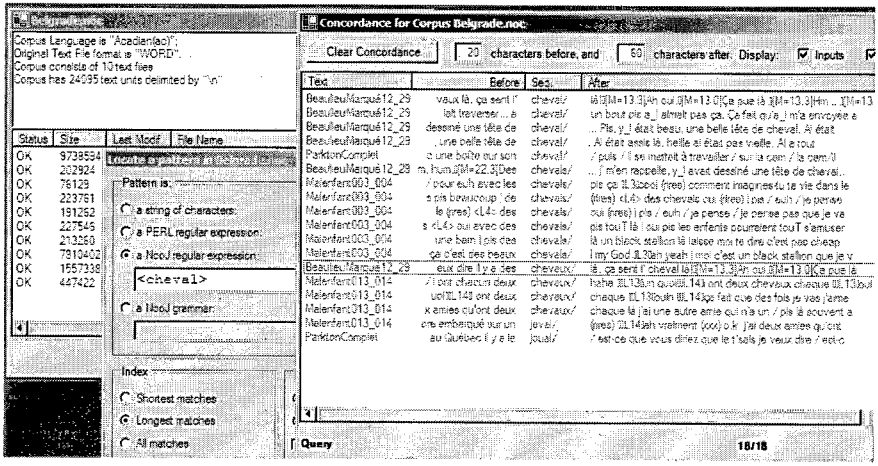

FIGURE 1

Concordances of the super-lemma *cheval*

on the final consonant of the English verb, different types of conjugations have evolved. Below are several examples of inflections of Acadian verbs.

The automatic inflection of the verb *dire* in NooJ (by applying the developed graphs of Acadian verb grammar) will give two forms of the third-person plural: *ils disent* and *ils disont* (Figure 2).

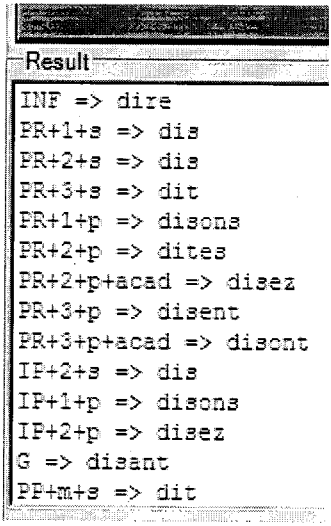

FIGURE 2

Inflections of the verb *dire* with the Acadian grammar

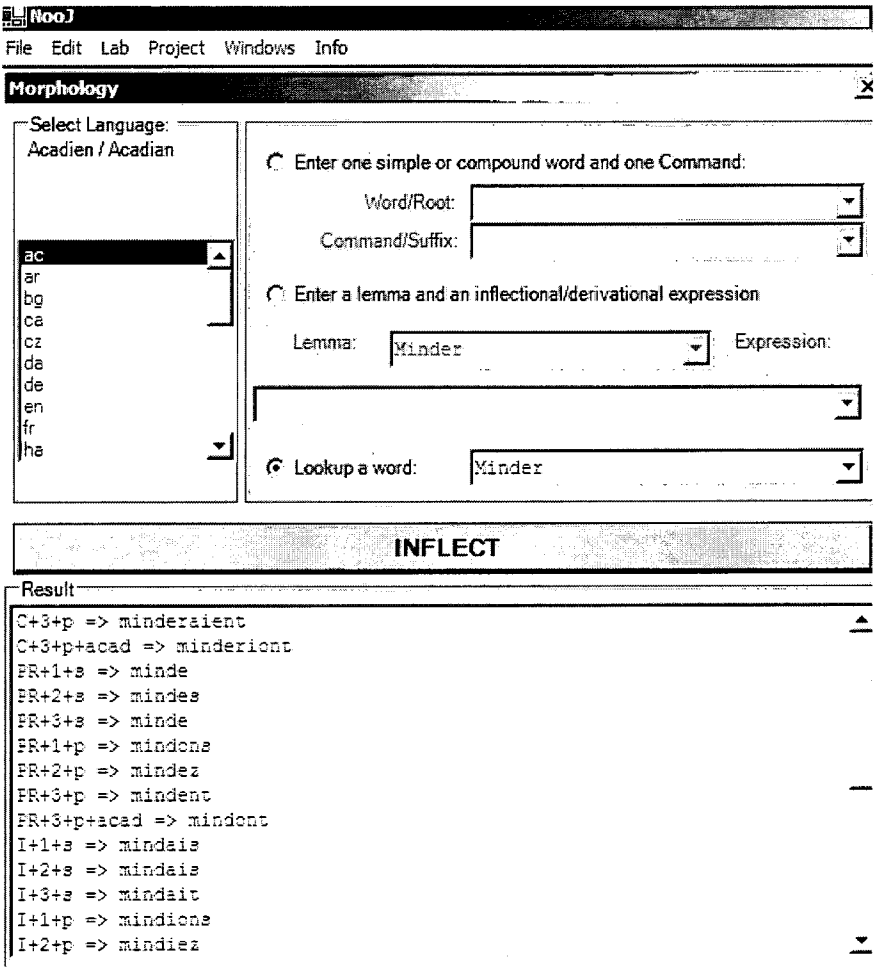

FIGURE 3

Acadian inflections of *minder*, from the English verb *to mind*

The application of the inflections to English verbs will produce the Acadian conjugation of those verbs. Figure 3 shows the window of the verb '*mind*er' being automatically inflected by NooJ.

For verbs with several spellings, NooJ will inflect all the forms. Thus, for example, for the three spellings of *badrer* 'to bother' (*déranger*), we have the three forms *bâdrer*, *badrer*, and *bodrer* inflected in all tenses and persons.

So, when we enter the search query <V+3+p>, verbs in the third-person plural, we get at the same time all the verbs in common French (*elles sont*, *ils ont*, *mes parents comprennent*), the verbs in the Acadian inflection (*i écoutont*, *i m'écoutont*,

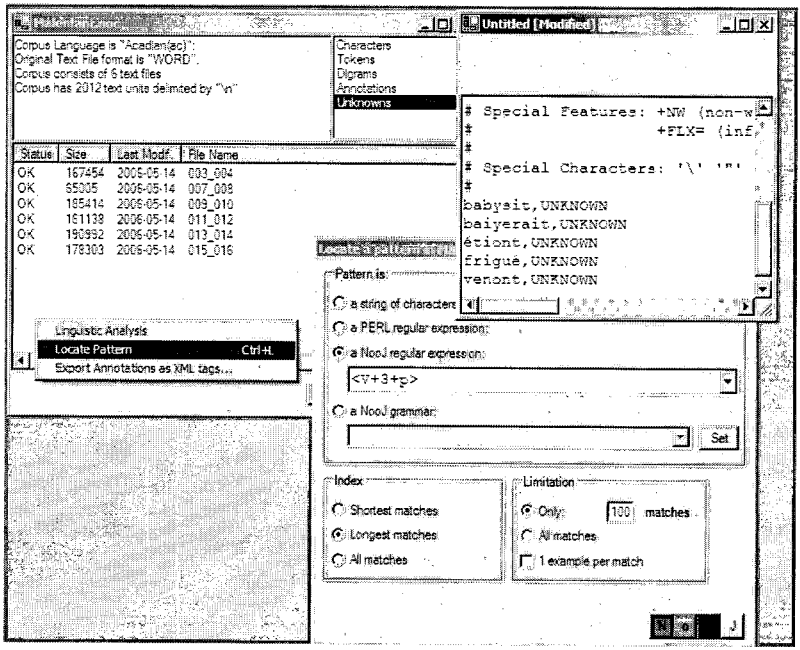

FIGURE 4

Search for third-person plural verbs with NooJ

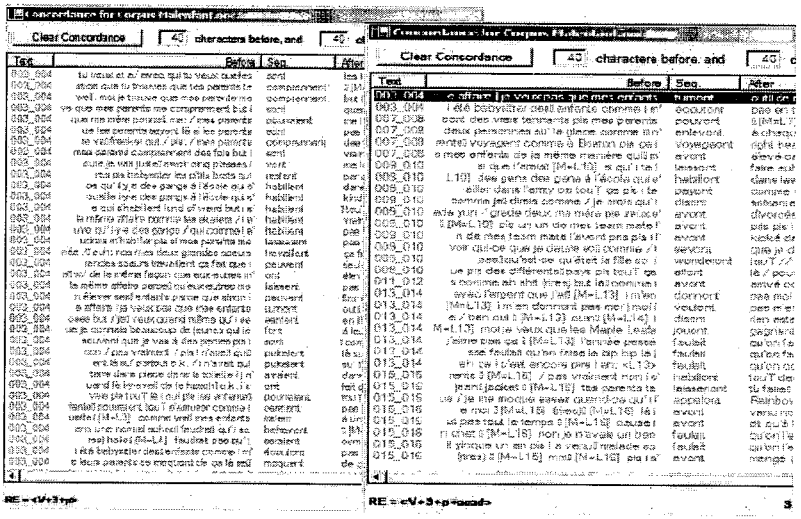

FIGURE 5

Concordances of third-person plural verbs and concordances of Acadian verbs in the third person plural

i avont, *i appelont*, *tes parents te laisseriont*), and the integrated English verbs (*i* turn*ont*, *i* wonder*ont*, *i* puk*aient*, *ils se* behav*ent*). If, on the other hand, only the Acadian forms are of interest to us, the search query <V+3+p+acadien> allows us to isolate the concordances of these Acadian forms (cf. Figure 4 for search screen V+3+p and Figure 5, Windows–NooJ screen for the query and the concordances of the third-person plural verbs and the Acadian verbs in the third person plural).

4.2.3. English verb-particle construction, integrated into Acadian French

One thing that characterizes the most anglicized Acadian dialects is their integration of English verbs with particles. This has led to a restructuring of the verbal structure: English verb + French flexion + English verbal particle (11).

(11)	j'ai *freaké out*	< freak out 'avoir peur'
	du *stuff* qui va *on* dans la vie	< go on 'se passer'
	il a *timbé off* la *cliff*	< fall off 'tomber (en bas) de la falaise'
	mes enfants *turn*ont *out* comme des *bums*	< turn out 'devenir'

The graph in Figure 6 was prepared to describe English verbs with particles: thus, the application of this graph allows us to locate all the verbs with particles in the corpus, as well as the direct paths, verbs + particles — as in the example "on *burn out*" or "ma mère va *freaker out*," and these are intercalated with an adverb, as in "ça *work* pas *out*" (cf. the concordances in (12).

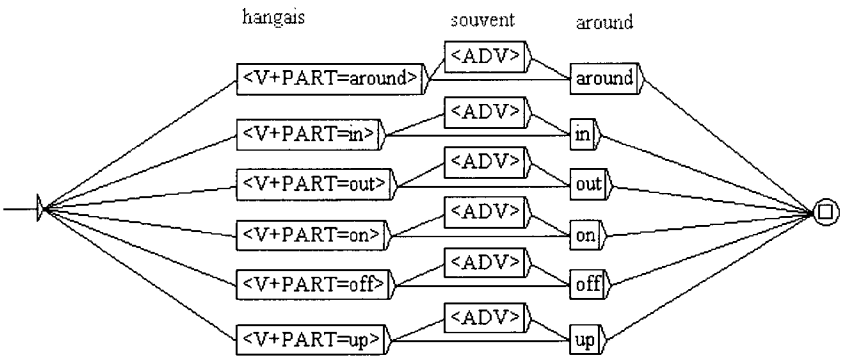

FIGURE 6

Graph of the Acadian conjugation of verbs with particles

(12) *Concordances of verbs with particles*:

tcheque fille ou ben tcheque fille la *beat up* pis a' braillait dans la bus l L16

t dans la bus l L16 tcheque fille la *beat up* (rires) L15 (rires) L16 c'tait pas

asse l tait tcheque looser l alle a *beat up* tcheque fille ou ben tcheque fille

9 o.k. c'est fini L10 on burn out on *burn out*/je vas aller la woure L 9 hein

mariage L 9 o.k. c'est fini L10 on *burn out* on burn out/je vas aller la woure

s) L15 je m'ai endormie presque/je *dozais off* l ou / chu assise en avant
l tcheque autre fille pis comme t'as *find out* t'as le droit L15 mm L16 I
ein je la drive right bad/pis l a' *fly up* pis a' fly de mme monte pis l a' start
arrire dfense quand que zeux en avant *freakant out* i ont besoin de tchques-
L15 un peu woin L16 quand-ce que as-tu *freak out* quand que quand Kirk
si les parents sont pas l ma mre va *freaker out*/pis:/mes parents
passe jusqu'/on s'a on s'a on s'a *frigu up* aux Jeux rgionaux cause que
es on a juste deux t.v. but mon frre a *mov out* so on est yinque trois/so mes
o 'steure a' parle pu L15 alle est tu *piss off* L16 a fait deux jours qu' a' me
i pis si que chu puni je sors pareil je *sneak out* pis (rires) L7 right L8 euh
affaire || je veux pas que mes enfants *turnont out* de mme so je crois qu' i est
ue je joues au ball hockey pis a so je *use pas up* touT mon argent/|| o.k.
L10 yeah ben quosse qu'arrive si a a *work pas out* L 9 hum/la NFL L10

4.3. Disambiguation graph: The example of "a"

As mentioned earlier, the third-person subject pronouns in Acadian are listed in (13) for the feminine and the masculine.

(13) a,elle,PRO+CLS
elle,PRO+CLS
alle,elle,PRO+CLS
al,elle,PRO+CLS
i,elles,PRO+CLS+3+f+p
i,il,PRO+CLS+3+m+s
i,ils,PRO+CLS+3+m+p

It should be noted that, by the same token, the form *i* is used for *elles*, *ils* and *il*, the forms *a*, *al*, or *alle* for *elle*. On the other hand, *a* can also represent the third-person singular present tense of the auxiliary *avoir*.

To resolve these ambiguities, we constructed a disambiguization graph, a grammar to disambiguate the different meanings contained in a lemma. Figure 7 makes possible the disambiguation of the Acadian pronoun *a* and third-person singular of the auxiliary *avoir*.[3]

Figure 7 indicates the two possible routes for *a*:

(i) the line heading upwards indicates that *a is a pronoun* if it is followed by a verb in the third person, and that it can be followed by other pronouns such as *le*, *la*, *les/leur*, *lui/en y*, placed before the verb.

(ii) the line heading downwards indicates that *a is a verb* if it is preceded by the personal pronouns contained in the first box (*je*, *j'*, *tu*, *t'*) or by *i-y-* in the second box; it is also a verb if it is followed by a past participle, which may be preceded by an adverb.

[3]This graph was prepared by Gisèle Chevalier.

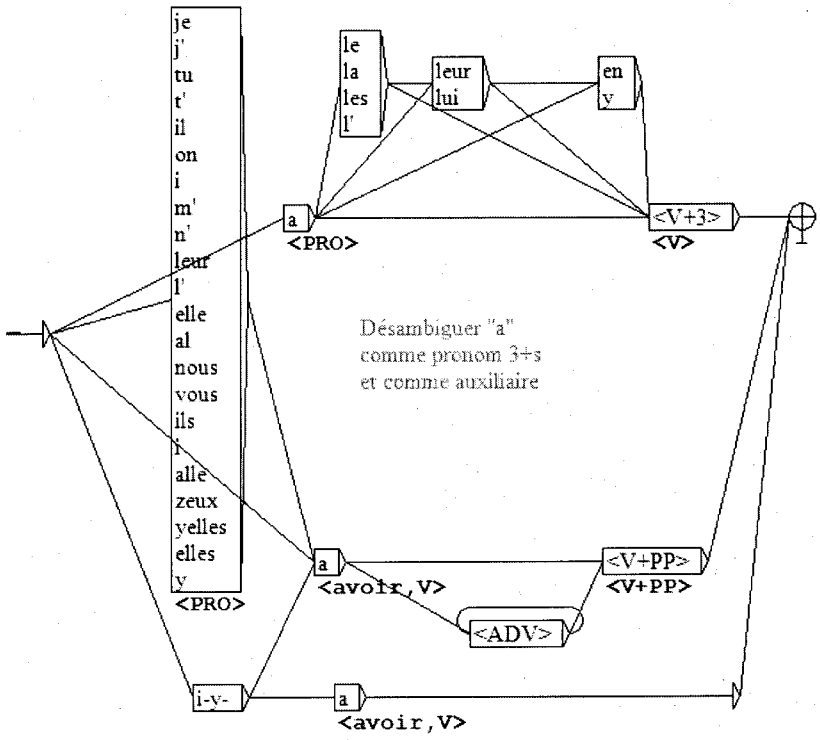

FIGURE 7

Graph of the disambiguation of *a*

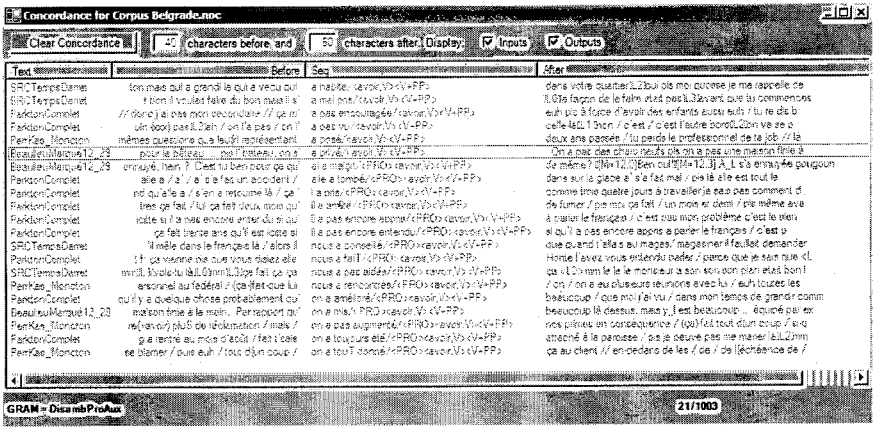

FIGURE 8

Concordances of the verb *avoir* in the third-person singular

The application of the graph in Figure 7 makes it possible to precisely and uniquely locate the *a* verbs or pronouns. Figure 8 is the screen display of concordances obtained by the search on the verb *avoir* in the third-person singular.

5. CONCLUSION

The preceding examples describe the evolution of the Acadian NooJ module and provide a good illustration of the possibilities NooJ offer for solving problems related to the automatic processing of oral corpora, regardless of whether these problems are orthographic, phonological, morphological, or syntactical. The Acadian NooJ module is now well advanced and continues to expand: the dictionaries are finished, and the grammars of flexions for nouns and adjectives, the morphology of verbs, and English verbs (with and without particles) integrated into Acadian have also been completed.

The use of super-lemmas and lemmas that take into account a hierarchy of the units in variants, as well as the inclusion of a single dictionary of the atomic linguistic units of different sizes in the lemmas, have made it possible to deal with many of the questions related to the processing of regional variants and corpora that are heterogeneous in terms of transcription.

We have now built the foundation of the Acadian NooJ module, but the research avenues for following up the automatic description of the Acadian dialects are infinite at every level, whether lexical (e.g., frozen expressions, words used in discourse), morphosyntactical (e.g., structures with prepositions), or semantic (e.g., the aspect properties of English verbs with particles).

The NooJ modules available free on line include Arabic, Bulgarian, English, Hebrew, Italian, Spanish, Armenian, Chinese, French, Hungarian, and Latin. However, the Acadian NooJ module is the first module of oral and regional French. The Acadian NooJ module, now available on-line in NooJ resources (cf. `www.nooj4nlp.net/pages/resources.html`), should make possible significant advances in the study and understanding of Acadian French in all its richness.

ACKNOWLEDGMENTS

This research study is the fruit of a long-term team effort. I would like to thank the entire research team, particularly my colleague Gisèle Chevalier for her devotion to the project, as well as our research assistants, Mike Long, Mathieu Lanteigne, Philippe Desjardins, Gemaël Melanson, and Michelle Mongeau for their hard work. The research project was supported by the New Brunswick Innovation Foundation (2003–2005) and a SSHRCC grant to small universities, administered by the FESR, Université de Moncton.

REFERENCES

Beauchemin, N., P. Martel, and M. Théoret. 1992. *Dictionnaire de fréquence des mots du français parlé au Québec*. New York, Peter Lang.

Beaulieu, L. 1996. *Corpus du Nord-Est*. Université de Moncton, New Brunswick.

Blanche-Benveniste, C. 1997. *Approches de la langue parlée en français*. Paris: Orphys.

——. 2003. La langue parlée. In *Le grand livre de la langue française*, ed. Marina Yaguello, 317–342. Paris, Seuil.

Blanche-Benveniste, C., ed. 1990. *Le français parlé: études grammaticales*. Paris: Éditions CNRS.

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

Bilger, M., F. Gadet, and K. Van Den Eynde, ed. 1998. *Analyse linguistique et approches de l'oral*. Louvain/Paris: Peeters.

Burnard, L., ed. 1995. *The British National Corpus users reference guide*. Oxford: Oxford University Computing Services.

Chapados, A., G. Chevalier, and S. Kasparian. 2004. Description du verbe *aller* en français acadien du N.-B., intégrée à INTEX. *Lexicometrica*, Numéro spécial: *Des enquêtes aux corpus littéraires: l'analyse de données textuelles*. `www.cavi.univ-paris3.fr/lexicometrica/`. Accessed 10 February, 2009.

Chevalier, G., and M. Long. 2005. *Finder out*, pour qu'on les *frig* pas *up*, comment *c'qui workont out*, les verbes à particules en chiac. In *Français d'Amérique: approches morphosyntaxiques* (Actes du Colloque international Grammaire comparée des variétés de français d'Amérique), ed. P. Brasseur and A. Falkert, 201–212. Paris: l'Harmattan.

Chevalier, G., S. Kasparian, and M. Silberztein. 2004. Éléments de solution pour le traitement automatique d'un français oral régional, *TAL* [*Traitement automatique des langues*] 45(2):41–62.

Francard, M., G. Geron, and R. Wilmet. 2002. La banque de données VALIBEL: des ressources textuelles orales pour l'étude du français en Wallonie et à Bruxelles. In *Romanische Korpuslinguistik-Korpora und gesprochene Sprache/Romance Corpus Linguistics-Corpora and Spoken Language*, ed. C. Pusch et W. Raible, 71–80. Tübingen: Gunter Narr.

Gadet, F. 1992. *Le français populaire*. Paris: PUF.

Gauvin, K, and G. Chevalier. 1994. Corpus Anna-Malenfant: du parler acadien de Dieppe (N.-B., Canada). Université de Moncton, Moncton.

Gross, M. 1975. *Méthodes en syntaxe: régime des constructions complétives*. Paris: Hermann.

Habert, B., A. Naxarenko, and A. Salem. 1997. *Les linguistiques de corpus*. Paris: Armand Colin.

Kasparian, S. 2003. Parler bilingue et actes identitaires: le cas des Acadiens du Nouveau-Brunswick. *Francophonies et langue dans un monde divers en évolu-*

tion: contacts interlinguistiques et socioculturels, ed. R.A. Stebbins, C. Romney et M. Ouellet, 159–177. Winnipeg: Presses Universitaires de St. Boniface.

Kasparian, S., and J. De Finney, ed. 2004. *L'analyse de données textuelles: De l'enquête aux corpus littéraires. Lexicometrica*, Numéro spécial: *Des enquêtes aux corpus littéraires: l'analyse de données textuelles.* www.cavi.univ-paris3.fr/lexicometrica/. Accessed 10 February, 2009.

Long, M., and G. Chevalier. 2004. Construction d'un dictionnaire informatisé de type DELAF des verbes à particules du français acadien chiac. Paper presented at the 7th Journées INTEX-NooJ, Belgrade.

Morel, M-A. 1985. *Analyse linguistique d'un corpus d'oral finalisé.* GRESCO, "Communication parlée", Caen.

Sankoff, D., G. Sankoff, S. Laberge and M. Topham. 1976. Méthodes d'échantillonnage et utilisation de l'ordinateur dans l'étude de la variation linguistique. *Cahiers de linguistique de l'Université du Québec* 6: 85–125.

Silberztein, M. 1990. Dictionnaires électroniques du français, *Langue française* 87:71–83.

——. 1996. Analyse automatique de corpus avec INTEX. *LINX* 34:269–276.

Silberztein, M. 2002–2008. *NooJ V2 Manual.* www.nooj4nlp.net/pages/references.html.

——. 2004. NOOJ: A Cooperative, Object-oriented Architecture for NLP. In *Cahiers de la MSH Ledoux*, vol. 1: *INTEX pour la linguistique et le traitement automatique des langues*, ed. C. Muller, J. Royeauté et M. Silberztein, 351–361. Besancon: Presses Universitaires de Franche-Comté.

Silberztein, M., and M. Long. 1993. The Intex manual: INTEX software documentation. Available online at: intex.univ-fcomte.fr/. Accessed 10 February, 2009.

Thibault, P. and D. Vincent. 1990. *Un corpus de français parlé*, Québec: CIRAL, Université Laval.

Veronis, J., ed. 2004. Le traitement automatique des corpus oraux. *TAL* [*Traitement automatique des langues*] 45(2):7–14.

Vincent, D., M. Laforest, and G. Martel. 1995. Le corpus de Montréal 1995: adaptation de la méthode d'enquête sociolinguistique pour l'analyse conversationnelle. *Dialangue* 6:29–46.

INTERNET LINKS

(Accessed 10 February, 2009)

www.tlfq.ulaval.ca/bdlp/ — Pan-francophone lexigraphic database designed by Claude Poirier, CIRAL, Université Laval.

valibel.fltr.ucl.ac.be/ — Valibel centre for research on linguistic varieties in French, located in Belgium

intex.univ-fcomte.fr/ — INTEX software.

www.cavi.univ-paris3.fr/lexicometrica/ — *Revue Lexicometrica*

www.image-zafar.com/english/alceste.htm — Alceste software

ancilla.unice.fr/~brunet/pub/commande.html — HyperBase software

www.cavi.univ-paris3.fr/ilpga/tal/lexicowww/ — Lexico software

intex.univ-fcomte.fr/ — Sphinx software

www.grimmersoft.com/ — WordMapper software

www.cavi.univ-paris3.fr/JADT — Actes des Journées d'analyse de données textuelles

www.nooj4nlp.net/ — NooJ software

www.nooj4nlp.net/pages/resources.html — Acadian NooJ module