Pierre-Aurélien Georges[1]

Les chaînes de clitiques : l'outil informatique au service de l'analyse comparative

I. Le Module MorphoSyntaxique du THESOC

1. Présentation

Le THESAURUS OCCITAN (ou THESOC en abrégé) est une base de données dialectales recouvrant tout le domaine occitan. Pour pouvoir figurer dans le THESOC, les faits linguistiques doivent remplir les deux conditions suivantes :

- les faits doivent être de nature orale et sont saisis dans la base avec leur transcription phonétique API.

- les faits doivent être précisément localisés, ce qui permet par la suite d'étudier la variation diatopique et de comparer les faits d'une localité à l'autre.

Actuellement, la base contient plus d'un million d'entrées lexicales, correspondant à 831 points d'enquête répartis sur toute l'Occitanie. Une partie de ces entrées lexicales est consultable en ligne sur Internet [2] . Le THESOC contient également des documents multimédia (plus de 1000 enregistrements sonores et plus de 500 documents visuels), des données bibliographiques, et des outils d'analyse tels qu'un module cartographique interactif permettant de générer des cartes linguistiques à la volée[3].

A côté de la base lexicale qui constitue le « cœur historique » du THESOC, signalons également la présence d'un volet de microtoponymie, ainsi que d'un volet plus particulièrement dédié à l'étude de la syntaxe et de la morphosyntaxe, qui a été mis en place depuis quelques années.

Au départ, ce volet était intitulé la « base TEXTES » [4] du THESOC, car il contenait essentiellement des textes et ethnotextes. Puis, au fur et à mesure des évolutions et améliorations apportées, il a été rebaptisé « Module MorphoSyntaxique » afin de lui donner un nom plus évocateur quand à sa finalité d'utilisation et ses possibilités d'exploitation des données, mais aussi pour tenir compte également de la présence de phrases isolées dans la base, qui viennent compléter les textes. Il peut s'agir par exemple de phrases issues de questionnaires d'enquête morphosyntaxique ou de phrases provenant de carnets d'enquête ou de certains atlas linguistiques tel que l'ALMC[5]. Comme pour le reste du THESOC, il est en outre possible d'attacher des documents multimédia à une phrase ou un texte de la base.

---

[1] Ingénieur de Recherche au laboratoire UMR 6039 - Bases, Corpus, Langage, Université de Nice Sophia-Antipolis, CNRS ; MSH de Nice, 98 bd E. Herriot, 06200 NICE. *email* : pageorge AT unice POINT fr

[2] *Cf.* Dalbera (1992-)

[3] Pour une présentation générale du THESOC, Cf. Oliviéri (2006)

[4] *Cf.* Georges (à paraître, a)

[5] Atlas Linguistique du Massif Central, Nauton (1957-1963)

2. Aperçu des fonctionnalités de la base MMS

Ce Module MorphoSyntaxique (MMS) est doté d'un certain nombre d'outils linguistiques. Signalons notamment la présence des éléments suivants :

- un transcripteur, qui permet de générer automatiquement une graphie phonologisante à partir de la transcription phonétique. Cette graphie, proche de la graphie mistralienne, permet un premier niveau de « lissage » qui gomme en quelque sorte la variation phonétique, et servira de point de départ à tous les autres traitements linguistiques effectués[6].

- un lemmatiseur, qui identifie chaque élément lexical d'une phrase ou d'un texte en se basant sur un dictionnaire intégré à la base. Afin de pouvoir gérer convenablement la variation (graphique, dialectale ou flexionnelle), ce dictionnaire est structuré sur deux niveaux hiérarchiques : les variantes sont regroupées sous différents lemmes, ce qui permet tour à tour d'effectuer une recherche ou un traitement, ou bien sur une variante particulière (telle flexion, avec telle graphie, dans tel dialecte), ou bien sur toutes les variantes associées à un lemme donné.

- un analyseur syntaxique, qui se base sur les données issues de la lemmatisation et sur le dictionnaire de la base pour proposer une ou plusieurs structures syntaxiques pour chaque phrase de la base[7]. Les arborescences ainsi générées se situent dans un cadre générativiste, mais les règles syntaxiques sur lesquelles l'analyseur se base pour effectuer son analyse sont modifiables par l'utilisateur.

Ces outils permettent d'automatiser en grande partie le travail d'étiquetage et d'annotation des données de la base. Une fois que les phrases et les textes ont ainsi été annotés, l'on peut utiliser les différentes fonctionnalités de recherches de la base pour sélectionner des données selon divers critères : par lemmes, par variantes, par catégorie syntaxique, par localité, par séquence de catégories syntaxiques ; recherche des phrases contenant tel ou tel fragment de structure syntaxique, etc. On peut ensuite générer et exporter un corpus de travail à partir des résultats de cette recherche. La base dispose d'ailleurs de fonctionnalités d'importation et d'exportation des données aux formats XML et TEI [8].

Dernièrement, une nouvelle fonctionnalité a fait son apparition dans MMS : il s'agit de la possibilité d'attribuer à certaines phrases des « étiquettes » définies par l'utilisateur (plus communément appelées *« tags »* en anglais), puis de faire des recherches croisées en fonction de ces étiquettes. C'est cette fonctionnalité que nous allons maintenant mettre en avant dans les exemples suivants pour illustrer en quoi cet outil informatique qu'est le Module MorphoSyntaxique pourrait être d'une grande utilité pour les linguistes syntacticiens ou morphosyntacticiens désirant travailler sur la variation dialectale.

---

[6] Rappelons que l'objectif de la base MMS est d'étudier la syntaxe et la morphosyntaxe. C'est pourquoi nous pouvons nous permettre ici de faire abstraction de la variation des réalisations phonétiques pour simplifier la suite des traitements linguistiques opérés. Toutefois, l'information n'est pas perdue puisque, à tout moment, l'utilisateur de la base a toujours accès à la transcription phonétique d'une phrase ou d'un texte, affichée juste à côté de sa graphie.

[7] A partir de maintenant, lorsque nous parlerons de « phrases » au sein de la base MMS, cela désignera toutes les phrases de la base, que ce soit des phrases isolées ou des phrases appartenant à un texte.

[8] Pour une présentation plus détaillée de la base MMS, *Cf.* Georges (à paraître, b).

II. Traitement des chaînes de clitiques dans MMS

1. Préambule

Il s'agit ici de montrer quelques possibilités offertes par l'outil informatique à travers un cas concret que nous avons choisi pour l'occasion : les chaînes de clitiques.

Nous ne reviendrons pas ici sur ce qui a déjà été dit à propos des chaînes de clitiques dans le domaine occitan, sujet abordé dans ce même ouvrage par un article de Guylaine Brun-Trigaud [mettre ici une référence bibliographique à un autre article présent dans les actes du colloque] auquel le lecteur pourra se référer pour plus d'informations sur le sujet.

Pour les besoins de la démonstration, nous nous limiterons aux données présentes dans l'ALMC sur les cartes n°1827 à 1838 et plus particulièrement, les réponses aux questions « donne-le-moi », « donne-le-lui », « donne-le-nous », « donne-le-leur », « dis-le-moi », « dis-le-lui », « dis-le-nous », et « dis-le-leur ». Cela nous permet en effet de disposer d'un ensemble homogène puisque toutes ces formes sont des impératifs[9].

2. Démarche suivie

Tout d'abord, les réponses des 55 points d'enquête de l'ALMC pour ces cartes-là ont été saisies dans la base MMS. Parfois, il y a deux réponses distinctes à une même question dans une même localité. Parfois aussi, il peut arriver qu'il n'y ait pas de réponses à la question demandée dans une certaine localité. Cela ne pose pas de problèmes particuliers pour la saisie dans la base MMS, mais cela explique pourquoi le nombre de réponses présentes dans la base est différent pour chaque question : il varie entre 38 et 60 réponses par question, avec une moyenne qui se situe malgré tout autour de 55 réponses par question.

Nous avons associé à toutes ces réponses des étiquettes en fonction des critères suivants :

- toutes les phrases dans lesquelles on observe l'ordre « clitique datif ; clitique accusatif » se sont vu attribuer l'étiquette DAT_ACC

- toutes les phrases dans lesquelles on observe l'ordre inverse, « clitique accusatif ; clitique datif » se sont vu attribuer l'étiquette ACC_DAT

Notons au passage que la lemmatisation préalable des phrases, bien que facultative, permet ici de gagner un temps précieux puisqu'elle permet alors d'effectuer une recherche de toutes les séquences « pronom clitique datif suivi de pronom clitique accusatif », puis d'affecter d'un seul coup une étiquette donnée à toutes les phrases listées dans les résultats de cette recherche.

---

[9] Il est évident que les chaînes de clitiques ne se comportent pas de la même manière à la forme impérative qu'à la forme affirmative. En témoigne le français « il me le donne », mais « donne-le-moi ! » où l'on voit bien que l'ordre et la nature des clitiques s'en trouve modifiés. Dès lors, dans une approche comparative, il faut bien prendre garde de ne comparer que des choses qui sont comparables, c'est-à-dire de ne comparer que des impératifs avec des impératifs ou des affirmatives avec des affirmatives.

Parallèlement à ce premier jeu d'étiquettes DAT_ACC *versus* ACC_DAT, nous avons introduit d'autres jeux d'étiquettes :

- toutes les phrases contenant le verbe « donner » se sont vu attribuer l'étiquette « DONNE… »

- toutes les phrases contenant le verbe « dire » se sont vu attribuer l'étiquette « DIS… »

De la même manière, nous avons créé des étiquettes « …MOI… », « …LUI… », « …NOUS… », « …LEUR… » pour repérer les phrases qui contiennent ces différents pronoms.

A partir de là, il est tout à fait possible de combiner ces jeux d'étiquettes pour former de nouvelles étiquettes si on le souhaite.

Signalons au passage qu'un petit nombre de phrases n'ont pas pu être étiquetées pour diverses raisons : ou bien elles ne correspondaient pas au schéma « Verbe + clitique + clitique » auquel nous avons choisi de nous intéresser, ou bien leur transcription phonétique laissait planer un doute ou une ambiguïté si bien que nous n'avons pas pu analyser leurs constituants avec certitude.

Après avoir étiqueté les différentes phrases, on peut ensuite jongler avec ces jeux d'étiquettes pour effectuer des recherches dans la base ou générer des statistiques. Par exemple, l'utilisateur peut très facilement demander au moteur de recherche de la base la requête suivante : « afficher la liste des localités dans lesquelles il existe au moins une phrase étiquetée DIS… + ACC_DAT + …MOI… et au moins une phrase étiquetée DONNE… + ACC_DAT + …MOI… »[10]. Ce que nous avons fait, et la base nous a alors indiqué qu'il y a 18 localités qui correspondent à ces critères de recherche. Il est possible ensuite de consulter la liste des localités en question et d'accéder à toutes les réponses qui ont été récoltées dans ces points d'enquête, notamment pour consulter la transcription phonétique de ces réponses.

Evidemment, si l'on se contente de ne faire qu'une seule requête dans le moteur de recherche, l'investissement, en termes de temps passé, qu'il a fallu pour saisir les réponses dans la base et les étiqueter est supérieur au bénéfice obtenu. En revanche, dès lors que l'on commence à effectuer plusieurs requêtes différentes, que l'on manipule les jeux d'étiquettes pour effectuer différents traitements ou différentes recherches, le bénéfice que l'on en retire devient vite largement supérieur à l'investissement de départ.

[10] Pour être précis, il faut également ajouter comme critère : « et ne contenant aucune phrase étiquetée DIS… + DAT_ACC + … MOI… ni aucune phrase étiquetée DONNE... + DAT_ACC + …MOI… » afin d'éliminer les localités dans lesquelles les deux formes concurrentes (DAT_ACC et ACC_DAT) ont été recueillies, et pour lesquelles il existe donc vraisemblablement une variation libre dans l'ordre de ces clitiques.

3. Résultats obtenus

Ainsi, en procédant de la sorte avec différentes combinaisons d'étiquettes, nous avons pu dresser le tableau suivant en seulement quelques clics de souris :

	"Dis-le-moi"	*"Dis-moi-le"*
"Donne-le-moi"	18 localités	1 seule localité : **MARCHASTEL**
"Donne-moi-le"	0	34 localités

	"Dis-le-lui"	*"Dis-lui-le"*
"Donne-le-lui"	15 localités	1 seule localité : **MARCHASTEL**
"Donne-lui-le"	1 seule localité : **PIERREFORT-PAULHENC**	6 localités

	"Dis-le-nous"	*"Dis-nous-le"*
"Donne-le-nous"	17 localités	2 localités : **MARCHASTEL** et SAINT-GERMAIN-DU-TEIL
"Donne-nous-le"	0	34 localités

	"Dis-le-leur"	*"Dis-leur-le"*
"Donne-le-leur"	18 localités	6 localités : **MARCHASTEL**, MENET, MILLAU, ROCHEPAULE, SAINT-GERMAIN-DU-TEIL, SAINT-BONNET-DE-SALERS
"Donne-leur-le"	1 seule localité : **PIERREFORT-PAULHENC**	16 localités

Figure 1. Ordre des clitiques suivant le verbe utilisé

Comme on le voit sur la figure 1, en règle générale, quel que soit le clitique datif utilisé, il semblerait que le choix du verbe « donner » ou « dire » n'influence pas vraiment l'ordre des clitiques datif et accusatif, à quelques exceptions près.

Si l'on regarde de plus près les quelques cas qui semblent faire exception à cette tendance générale, on voit que deux localités reviennent de manière récurrente (en gras sur la figure 1) :

- Marchastel (15), point d'enquête n°32 de l'ALMC
- et Pierrefort-Paulhenc (15), point d'enquête n°43.

Le paradigme de Marchastel (15) constitue ainsi une exception généralisée à cette tendance :

intitulé	réponse (phonétique)
donne-le-moi	b'ajlɔ lu mi
donne-le-lui	b'ajlɔ lu i
donne-le-nous	b'ajlɔ lu nu
donne-le-leur	b'ajlɔ lu lur
dis-le	diʒɔ zu
dis-le-moi	diʒɔ m u
dis-le-moi	diʒɔ m zu
dis-le-nous	diʒɔ nuz u
dis-le-leur	diʒɔ ju zu
dis-le-lui	diʒɔ ju zu
on ne le sait pas	lɔ̃d u sa pa
on ne peut pas le savoir	lɔ̃ pɔ pa u s'awre
tu veux le savoir ?	bɔ zu s'awre
tu veux le savoir ?	u bɔs s'awre

Figure 2. Séquences de clitiques recueillies à Marchastel

Malheureusement, l'ALMC ne nous fournit ces chaînes de clitiques que pour les deux verbes « donne » et « dis ». Aussi serait-il donc intéressant, à la vue de ces données, de faire une enquête complémentaire sur la commune de Marchastel pour vérifier avec d'autres verbes si l'on constate également cette différence dans l'ordre des clitiques datif et accusatif. Si c'est le cas, cela signifiera alors que le choix du verbe peut influencer l'ordre des clitiques dans ces dialectes-là, et qu'il serait alors bon de pouvoir recueillir aussi de nouvelles données pour les 54 autres points d'enquête afin de compléter les réponses déjà attestées et de pouvoir vérifier avec d'autres verbes que « donne » ou « dis » si l'ordre des clitiques y reste identique.

En combinant autrement les différents jeux d'étiquettes, nous avons également obtenu les chiffres suivants :

	"Dis-le-moi"	*"Dis-moi-le"*
"Dis-le-lui"	15 localités	1 seule localité : PIERREFORT-PAULHENC
"Dis-lui-le"	0	7 localités

	"Donne-le-moi"	*"Donne-moi-le"*
"Donne-le-lui"	19 localités	0
"Donne-lui-le"	0	36 localités

	"Dis-le-leur"	*"Dis-leur-le"*
"Dis-le-lui"	14 localités	2 localités : MILLAU, SAINT-GERMAIN-DU-TEIL
"Dis-lui-le"	0	5 localités

	"Donne-le-leur"	*"Donne-leur-le"*
"Donne-le-lui"	19 localités	0
"Donne-lui-le"	9 localités	25 localités

	"Dis-le-nous"	*"Dis-nous-le"*
"Dis-le-lui"	14 localités	2 localités : MILLAU, SAINT-GERMAIN-DU-TEIL
"Dis-lui-le"	0	5 localités

	"Donne-le-nous"	*"Donne-nous-le"*
"Donne-le-lui"	19 localités	0
"Donne-lui-le"	0	36 localités

Figure 3. Ordre des clitiques suivant le paradigme des personnes

On peut y observer qu'en règle générale, le paradigme est uniforme : l'ordre des clitiques datif et accusatif reste le même à toutes les personnes, contrairement à ce qui se passe en français par exemple avec la troisième personne qui se comporte différemment des deux autres personnes. Seules 3 localités sur les 55 contreviennent à cette tendance : on retrouve la localité Pierrefort-Paulhenc (15) qui se distinguait déjà dans la figure 1, ainsi que les localités de Millau (12) et St Germain du Teil (12), respectivement points d'enquête n°51 et 38 de l'ALMC.

On aura remarqué que dans les tableaux de la figure 1 et de la figure 3, le total des 4 cases blanches n'est pas toujours égal au nombre de points d'enquête de l'ALMC (à savoir 55 localités). Cela résulte de trois facteurs :

- Comme indiqué précédemment, certaines phrases n'ont pas pu être étiquetées et n'ont donc pas été prises en compte dans le calcul de ces chiffres.

- Nous avons filtré dans nos critères de recherche les localités pour lesquelles deux formes concurrentes sont attestées, selon les modalités précisées plus haut dans une note de bas de page.

- Lorsque l'on croise deux questions différentes comme c'est le cas ici, les localités qui ne possèdent aucunes réponses pour l'une ou l'autre de ces deux questions ne peuvent pas être prises en compte.

Comme on peut le voir à travers ces deux exemples, cette procédure de comparaison assistée par l'outil informatique permet, en quelques clics de souris, non seulement de dégager des tendances sur une aire dialectale et d'évaluer la corrélation ou non entre deux phénomènes linguistiques observés, mais également de faire ressortir les points d'enquête qui se comportent singulièrement et ce de manière récurrente. Cela permet ensuite au chercheur de se focaliser plus précisément sur ces points particulièrement intéressants du point de vue du phénomène linguistique considéré.

On pourra alors étudier en quoi et pourquoi ces points d'enquête se distinguent de la tendance générale : s'agit-il d'une particularité du dialecte local, et si oui, quelle est cette particularité et quel est son fonctionnement ? Ou bien s'agit-il simplement d'un artefact lié aux conditions particulières de l'enquête (utilisation de plusieurs témoins différents sur un même point d'enquête, variation libre qui était passée inaperçue faute d'avoir récolté suffisamment de matériau linguistique, erreur de transcription phonétique, etc.). Dans ce dernier cas, le « contre-exemple » peut alors être « éliminé » et vient alors renforcer la tendance générale constatée.

Cela est tout particulièrement intéressant pour les syntacticiens qui travaillent dans un cadre théorique générativiste : l'approche « Principes et Paramètres »[11] de la grammaire générative dissocie d'une part des Principes généraux universels communs à toutes les langues, et d'autre part des Paramètres qui sont positionnés différemment d'une langue à l'autre. Cette « conception paramétrique de la variation linguistique »[12] permet de comparer les dialectes et les langues du monde du point de vue de la syntaxe. Dans ce cadre-là, cette procédure de comparaison et de recherche de corrélations éventuelles pourrait aider les chercheurs à dégager de nouveaux principes et à valider ou invalider leurs hypothèses de travail.

[11] *Cf.* Chomsky (1981).

[12] Chomsky (1982: 349).

4. Commentaires critiques

Il nous faut à présent répondre à un certain nombre de critiques pertinentes qui pourraient être soulevées suite à la démarche que nous venons de présenter.

Tout d'abord, il faut signaler que dans un souci de rigueur scientifique, il faudrait en principe faire la distinction entre pronom neutre et pronom masculin de la troisième personne du singulier, car certains dialectes produisent deux réalisations différentes pour le pronom clitique qui se traduit en français par « le », suivant qu'il s'agisse du pronom neutre, comme dans « je le lui ai dit », ou du pronom masculin de la troisième personne du singulier, comme dans « je le lui ai donné (le livre) »[13]. Il faudrait donc s'interdire de comparer sur un même plan des formes comme « donne-le-lui » et « dis-le-lui », car l'ordre des clitiques pourrait peut-être y être influencé par la nature même des clitiques (suivant qu'il s'agit d'un pronom neutre ou non). Cependant, il s'agissait ici uniquement de démontrer l'utilité de l'outil informatique et de présenter simplement son utilisation et son fonctionnement. Les chercheurs désireux d'utiliser cet outil auront tout le loisir de peaufiner la méthode présentée ici et de prendre en compte ces considérations plus en détail[14].

En outre, d'aucuns diront que les questionnaires de phrases, technique très souvent utilisée lors des enquêtes linguistiques, entraînent de la part des informateurs des réponses « moins naturelles » qu'un discours libre par exemple, et augmentent le risque de calques sur le français. On peut dès lors se poser la question de la pertinence d'utiliser ce type de méthodes d'enquêtes lorsqu'il s'agit d'étudier la syntaxe ou la morphosyntaxe dialectale. Par ailleurs, l'étude des chaînes de clitiques rencontre un problème supplémentaire à l'utilisation des questionnaires de phrases : d'une part, les chaînes de clitiques s'avèrent finalement assez rares à l'oral, et d'autre part, lorsque l'on demande à un informateur de traduire une phrase contenant une chaîne de clitiques du type « il le lui donne », il n'est pas rare que ce dernier « élimine » un des clitiques lors de la traduction, de sorte que l'on perd un élément de la chaîne de clitiques et que l'on se retrouve au final avec quelque chose du type « i b'ajlɔ »[15].

Ainsi, comme Séguy (1973) l'a fait remarquer :

> « la syntaxe proprement dire, en dehors de quelques traits épais, échappe à l'enquête par questionnaire. Si on demande aux informateurs de traduire dans leur dialecte des énoncés en français standard, on n'obtient aucune différence dans les régions centrales, et, dans les régions périphériques, on provoque souvent des calques. Les discours sollicités, comme ceux qui ont servi de tests à X. Ravier, bien qu'ils soient libres, présentent une syntaxe monotone et appauvrie. Le seul moyen est d'écouter du langage spontané, des conversations entre patoisants. »

Pour prendre en compte ces différentes remarques et y apporter une solution, plutôt que de se baser sur des phrases isolées issues de questionnaires de phrases, il est tout à fait

---

[13] *Cf.* Tuaillon (1969). Cette distinction n'est d'ailleurs pas cantonnée à la région lyonnaise, mais se retrouve également dans certains parlers des Alpes Maritimes, comme l'indique Dalbera (1991: 604), ainsi que dans la région limousine, où j'ai d'ailleurs eu l'occasion de le constater personnellement sur le terrain lors d'une enquête linguistique réalisée à 6 km au sud de la commune d'Eymoutiers (87), au lieu-dit « Meilhaguet ».

[14] L'ALMC ne nous fournit malheureusement pas suffisamment de données pour pouvoir étudier séparément le pronom neutre et le pronom masculin de la troisième personne du singulier au sein des chaînes de clitiques. Pour bien faire, il faudrait donc envisager d'avoir recours à des données provenant d'autres sources.

[15] Mot à mot : « (il) lui donne »

possible à la place d'utiliser des ethnotextes, discours libres ou enregistrements de conversations spontanées, en appliquant exactement la même méthode que celle décrite précédemment : la base MMS contient en effet à la fois des textes et des phrases isolées, et le système d' « étiquettes » qui vient d'être présenté ici s'applique tout aussi bien aux phrases isolées qu'aux phrases contenues dans un texte.

Si l'on s'intéresse plus particulièrement aux chaînes de clitiques, il est vrai que pour arriver à récolter le même nombre d'occurrences de telles chaînes de clitiques, il faudra une quantité de discours libres beaucoup plus conséquente que si l'on utilise des questionnaires de phrases qui orientent la réponse de l'informateur et provoquent l'apparition de chaînes de clitiques avec une plus grande probabilité. On pourrait cependant également envisager une approche mixte, en utilisant à la fois les phrases isolées issues des questionnaires, et les textes issus de discours libres.

D'une manière générale, ces critiques s'appliquent donc surtout à la nature même des données utilisées et à la façon dont elles ont été recueillies en enquête et comment elles sont exploitées, mais ne remettent pas en cause le procédé de traitement de ces données en lui-même, puisque ce même traitement peut être envisagé avec différent types de données.

III. Perspectives

La méthode que nous venons de présenter pourrait être utilisée pour effectuer des recherches sur différents aspects de la morphosyntaxe et de la syntaxe des dialectes, quel que soit le cadre théorique utilisé. Un cas a cependant retenu notre attention : l'utilisation de cette méthode dans un cadre générativiste soulève quelques réflexions intéressantes et nécessite de prendre quelques précautions que nous allons maintenant aborder.

1. Utilisation dans un cadre générativiste

Si l'on se situe dans un des cadres théoriques de la grammaire générative, les fonctionnalités de la base MMS qui ont été présentées ci-dessus pourraient être utilisées autour de l'approche « Principes et Paramètres » de deux manières complémentaires : soit de façon heuristique, pour tenter de rechercher la présence éventuelle de corrélations entre deux paramètres sans avoir *a priori* de position sur le sujet, soit *a posteriori*, pour confirmer ou infirmer une hypothèse sur la question.

Dans ce dernier cas, je pense notamment à la corrélation établie par Rizzi (1982) entre « sujet nul / *pro drop* » (ou disons plutôt, « pronom sujet non réalisé », de manière plus neutre) et inversion libre du sujet et du verbe[16].

Cette corrélation est parfois remise en question par certains linguistes, et l'on peut à juste titre se demander si ce principe d'une corrélation entre paramètre du sujet nul et inversion libre du sujet pourrait toujours tenir si l'on étudiait désormais les faits linguistiques à l'échelle de la micro-variation dialectale ?

Oliviéri (2004) a montré que le comportement des sujets post-verbaux de l'occitan n'est pas tout à fait le même qu'en italien. L'ordre des constituants en occitan est certes plus libre

[16] Pour résumer, en italien on peut dire « *Maria dorme.* » tout aussi bien que « *Dorme.* » tandis qu'en français le sujet doit être obligatoirement réalisé : « *Marie dort.* » mais « **dort.* ». Et parallèlement à cela, l'inversion libre du sujet est possible en italien mais pas en français : « *Dorme Maria.* » mais « **Dort Marie.* ».

qu'en français, mais ne l'est pas autant qu'en italien, malgré une même absence de clitique sujet[17] et une morphologie verbale riche. Cette corrélation potentielle pourrait donc être remise en question s'il se révélait, au cours d'une étude plus approfondie, que l'inversion du sujet en occitan n'est pas aussi libre qu'on pourrait le supposer ; et ce malgré le fait qu'il s'agisse bien, dans sa grande majorité, de dialectes "à sujet nul".

Pour confronter à la réalité des faits dialectaux cette hypothèse d'une corrélation entre paramètre du sujet nul et inversion libre du sujet, l'idée serait donc de pouvoir comparer différentes grammaires pour voir s'il existe des grammaires dans lesquelles on observe l'un de ces deux paramètres sans pour autant observer l'autre. Autrement dit, existe-t-il des grammaires dans lesquelles on observerait par exemple une inversion libre du sujet sans pour autant avoir la présence de sujet nuls, ou bien l'inverse ?

Notez que l'on parle ici de « grammaires » et non pas de « dialectes » : dans une première approximation, on pourrait être tenté de penser que les deux termes sont ici interchangeables, tout du moins dans ce contexte bien précis, et que si l'on compare plusieurs dialectes entre eux, cela revient à comparer plusieurs grammaires entre elles. Ainsi, la tentation serait grande de vouloir rechercher s'il existe des localités dans lesquelles on observe des phrases avec sujets nuls mais dans lesquelles on n'observe pas de phrases contenant des inversions du sujet.

Mais il faut faire attention : il ne faut pas perdre de vue que les données linguistiques recueillies sur le terrain ne sont que des productions individuelles ; or il se pourrait, d'après ce qui est apparu dans les recherches récentes (Séminaire de recherche, 2008-2009) que chaque locuteur possède en fait sa propre grammaire, son propre idiolecte. En prenant en compte cette hypothèse, si l'on compare des données provenant de plusieurs dialectes sans s'être assuré au préalable qu'il n'y a bien eu dans ces données qu'un et un seul informateur pour chacun de ces dialectes, le risque de mélanger des phrases provenant de différents locuteurs d'un même dialecte, c'est-à-dire de mélanger différents idiolectes, et donc différentes grammaires, nous guette. L'idéal serait-il donc de s'assurer de n'avoir qu'un seul informateur par point d'enquête ? Cela va à l'encontre de l'idée communément admise qu'il est préférable, dans l'absolu, d'avoir plusieurs informateurs par point d'enquête, de manière à lisser les particularités de chaque locuteur et à avoir une meilleure vision du dialecte de cette localité. Par ailleurs, cela compliquerait énormément la tâche, puisque cela signifierait alors qu'on ne pourrait pas compléter les données recueillies auprès d'un premier informateur par des données recueillies auprès d'un second informateur lors d'une enquête ultérieure sur la même localité.

De toute façon, en allant encore plus loin dans la réflexion, même si l'on s'assurait de n'avoir utilisé qu'un seul informateur par point d'enquête, cela ne règlerait pas totalement le problème, car non seulement chaque locuteur possède son propre idiolecte, mais en réalité il possède même plusieurs grammaires : on n'écrit pas de la même façon que l'on parle, et l'on ne parle pas de la même façon avec ses amis ou sa famille que lorsqu'on s'adresse en public à un auditoire. De même que l'on n'écrit pas non plus de la même façon suivant que l'on rédige un article de linguistique ou que l'on envoie une carte postale à sa grand-mère. Dès lors, comment s'assurer que toutes les phrases qui ont été recueillies dans une même localité ont

[17] Dans la grande majorité des dialectes occitans, exception faite des dialectes du nord de l'Occitanie, notamment dans certaines zones du Limousin et de l'Auvergne.

bien été produites non seulement par le même informateur, mais par une seule et même grammaire de cet informateur ?

Il n'est pas impossible de penser, en effet, que certains paramètres comme la présence de sujets nuls ou l'inversion libre du sujet puissent être paramétrés différemment d'une grammaire à l'autre au sein d'un même individu. Prenons un exemple qui se situe dans le même ordre d'idée : à l'oral, dans un registre familier, la majorité des gens diront plus facilement « Marie elle dort » que « Marie dort » et pourront tout aussi bien dire « elle dort Marie » alors que dans un écrit avec un registre soutenu, ils se seraient contentés de la forme standard « Marie dort ».

Si l'on utilise des phrases isolées provenant de questionnaires d'enquête, la situation paraît donc insoluble : difficile de s'assurer que toutes les questions ont été posées dans une même situation, un même cadre, un même contexte, avec la même expression, la même intonation, et le même registre de langue. Et il est encore plus difficile de s'assurer que l'informateur y a systématiquement répondu avec le même registre de langue, le même contexte, bref, la même grammaire.

Mais la situation n'est pas désespérée pour autant. En réalité, la solution permettant de s'affranchir de tous ces problèmes est même relativement simple : lors des enquêtes linguistiques de terrain, il faut utiliser une unité de collecte plus grande que la phrase : le « texte », ou plutôt l'ethnotexte. En effet, au sein d'un ethnotexte, on peut s'attendre à une homogénéité beaucoup plus importante, qu'il serait en pratique difficile d'atteindre avec des phrases isolées.

Ainsi, toutes les phrases d'un ethnotexte :

- ont été prononcées à la même époque (unité temporelle)
- dans un même lieu (unité spatiale)
- sont généralement issues du même locuteur, donc du même idiolecte (unité du locuteur), si l'on met de côté les textes qui contiennent des conversations entre plusieurs personnes[18].
- sont généralement prononcées dans un même contexte, un même genre, voire un même style, un même registre (unité de la grammaire utilisée par le locuteur), sauf lorsque le locuteur change délibérément de registre pour chercher à imiter quelqu'un, à jouer un rôle, ou à donner une tournure spéciale à certaines phrases de son discours. Dans ce cas, les phrases en question se repèrent en général assez facilement puisque l'intention du locuteur était justement d'introduire une différence qui soit facilement perceptible par son interlocuteur. Ces quelques phrases-là peuvent donc facilement être écartées le cas échéant.

[18] Cela étant, on pourrait très bien imaginer de traiter les conversations entre plusieurs personnes de la manière suivante : chaque locuteur se voit attribuer une étiquette X, Y, Z. Chaque phrase du texte est étiquetée en fonction de son locuteur : X, Y, ou Z. On peut ensuite utiliser ce jeu d'étiquettes pour virtuellement faire comme s'il s'agissait de trois textes différents, dans lesquels il n'y a qu'un seul et même locuteur du début jusqu'à la fin du texte. On peut alors traiter ces trois textes virtuels comme les autres textes de la base.

En d'autres termes, pour s'affranchir de tous les problèmes mentionnés ci-dessus, il suffit de revoir quelque peu la méthode utilisée : pour évaluer s'il existe une corrélation entre les paramètres X et Y, au lieu de rechercher les localités dans lesquelles on a recueilli des phrases contenant la caractéristique X et des phrases contenant la caractéristique Y[19], on préfèrera plutôt rechercher les textes qui contiennent des phrases contenant la caractéristique X et des phrases contenant la caractéristique Y en leur sein. Ainsi, on peut s'assurer que l'on est bien en train de chercher à déterminer s'il existe une corrélation entre la présence d'un paramètre X et la présence d'un paramètre Y au sein d'une seule et même grammaire.

Pour reprendre notre exemple concret, si l'on veut valider ou invalider l'hypothèse d'une corrélation entre présence de sujets nuls et inversion libre du sujet, il faudra plutôt rechercher si l'on peut trouver des ethnotextes de longueurs assez conséquentes qui contiennent des phrases possédant une inversion du sujet mais qui ne contiennent aucune phrase avec un sujet nul (ou bien l'inverse) au lieu de rechercher s'il existe des localités dans lesquelles on observe des phrases possédant une inversion du sujet mais dans lesquelles on n'observe aucunes phrases avec un sujet nul.

Notons qu'en procédant de la sorte, les ethnotextes n'ont alors en théorie pas nécessairement besoin de provenir de localités différentes, puisque l'on compare non plus des dialectes mais des grammaires individuelles. Cependant, on peut s'attendre à ce que la diversité linguistique soit plus importante entre deux locuteurs géographiquement distants qu'entre deux locuteurs du même village. Si les ethnotextes proviennent d'un grand nombre de point d'enquête variés, cela n'en sera donc que plus bénéfique.

2. Etudes diachroniques

Jusqu'ici nous avons essentiellement traité :

- de la variation diatopique, puisque nous avons comparé des faits linguistiques entre différentes localités.

- de la variation individuelle[20], puisque nous avons abordé la problématique de l'utilisation de faits linguistiques provenant de différents locuteurs, afin de comparer différentes grammaires individuelles, dans un cadre génératiViste.

Sur un tout autre point de vue, cette méthode comparative pourrait également être utilisée dans un cadre diachronique. L'hypothèse formulée par Dalbera (1992) suggère que la diversité dialectale (en synchronie) est le reflet de l'évolution diachronique de la langue :

> « Mais l'une des directions d'exploitation les plus claires de la description des faits dans l'espace, à notre sens, est le domaine de la reconstruction. L'idée a été maintes fois avancée que, au plan linguistique, l'espace pouvait être exploité comme projection du temps. En d'autres termes, que la variation et la répartition des faits dans l'espace constituaient des indices de première importance pour l'analyse diachronique. A travers la variation, ce sont des stades d'évolution que l'on appréhende; à travers la répartition, ce sont des indications sur la succession de ces stades que l'on peut se procurer. »[21]

---

[19] ou bien : « des phrases contenant la caractéristique X mais aucune phrase contenant la caractéristique Y », selon la nature de ce que l'on cherche à étudier.

[20] Pourrait-on dire socio-linguistique ?

[21] Dalbera (1992: 138).

On pourrait alors utiliser cette méthode comme base pour comparer tel ou tel phénomène ou caractéristique linguistique entre les différents parlers de la zone occitane dans le but de rechercher les différents stades d'évolution de la langue.

Par ailleurs, le Module MorphoSyntaxique contient également un champ « date » qui permet de renseigner, pour chaque texte et phrase de la base, à quelle date ils ont été recueillis sur le terrain. En supposant que l'on puisse arriver à avoir suffisamment de données réparties sur une période de temps relativement étendue, on pourrait alors envisager d'utiliser la méthode décrite ici non plus pour comparer des données linguistiques en synchronie entre différentes localités ou entre différents locuteurs, mais cette fois-ci en diachronie.

IV. Conclusion

La base MMS dispose d'outils et de fonctionnalités souples, qui peuvent être utilisés de différentes manières et selon différents cadres théoriques, ce qui lui permet d'assurer pleinement son rôle d'outil informatique d'aide à la recherche en syntaxe et morphosyntaxe sur les dialectes occitans.

En particulier, ce système d'« étiquettes » que nous venons de présenter ici ainsi que les différentes fonctionnalités du Module MorphoSyntaxique qui y sont associées permettent d'évaluer rapidement, sur un très grand nombre de données et de localités, la présence ou l'absence de corrélations entre deux (ou plusieurs) paramètres linguistiques qui sont soupçonnés d'entrer en compte dans l'explication de la variation des faits observés.

On dispose donc là d'un outil qui permet non seulement de valider ou d'invalider des hypothèses linguistiques en les confrontant à la réalité des faits linguistiques, mais aussi de faire ressortir de manière heuristique d'éventuelles corrélations qui auraient pu passer inaperçues jusque-là.

Suivant le type des données disponibles et suivant la nature des recherches envisagées, on pourra appliquer la méthode que nous venons de décrire pour comparer des phrases isolées entre différentes localités, des phrases provenant de textes entre différentes localités, des phrases provenant de différents textes que l'on compare entre eux, ou une quelconque combinaison des trois, selon les besoins et les avantages et inconvénients en jeu.

Au-delà des quelques exemples simples et d'une taille raisonnable que nous venons de voir, pour dégager des corrélations d'ordre général plus importantes et plus intéressantes, il faudra bien sûr exploiter un volume plus important de données linguistiques et sur un plus grand nombre de localités. C'est là que nous pourrons alors vraiment mesurer toute la puissance de l'outil informatique, qui permet de réaliser des travaux qui n'auraient tout simplement pas été humainement envisageables auparavant.

V. Bibliographie

Chomsky, Noam, 1981, *Théorie du Gouvernement et du Liage - les Conférences de Pise*, Editions du Seuil, 1991, Paris.

Chomsky, Noam, 1982, *La Nouvelle Syntaxe*, Editions du Seuil, 1987, Paris.

Dalbera, Jean-Philippe, 1991, *Les pronoms personnels atones dans les parlers des Alpes-Maritimes. Champ et mécanismes de variation*, Actes du XVIIIe congrès international de Linguistique et Philologie Romanes, t. III, Université de Trèves (Trier), 1986, Max Niemeyer Verlag, Tübingen p. 599-613.

Dalbera, Jean-Philippe, 1992, *Dialectologie et morphonologie*, Actes du Congrès International de Dialectologie, Bilbao, IKER-7 p. p. 135-149.

Dalbera, Jean-Philippe *et al.*, 1992-, *Thesaurus Occitan*, CNRS - UMR 6039 "BCL". Université de Nice - Sophia Antipolis, http://thesaurus.unice.fr.

Georges, Pierre-Aurélien, à paraître, a, *Présentation de la base Textes associée au THESOC*, actes du colloque La dialectologie hier et aujourd'hui (1906-2006), Lyon, 2006, Brigitte Horiot (Ed.). p. 75-87.

Georges, Pierre-Aurélien, à paraître, b, *Le THESOC: bases de données et outils d'analyse consacrés à l'étude des dialectes occitans*, actes du colloque Bases de données, Méthodes, Modèles de description: de nouvelles perspectives pour la recherche sur les langues régionales et minoritaires? Tübingen, 2008, Stauffenburg Verlag (DeLingulis).

Nauton, Pierre, 1957-1963, *Atlas Linguistique et ethnographique du Massif Central*, Editions du CNRS, Paris.

Oliviéri, Michèle, 2003, *Y-a-t-il des frontières dialectales en syntaxe?* actes du 128e Congrès des sociétés historiques et scientifiques (CTHS): Relations, échanges et coopération en Méditerranée, Bastia, 14-21 avril 2003, à paraître.

Oliviéri, Michèle, 2004, "Paramètre du sujet nul et inversion du sujet dans les dialectes italiens et occitans", *Cahiers de grammaire n°29, Questions de linguistique et de dialectologie romane, 2004*, p. p. 105-120.

Oliviéri, Michèle; Brun-Trigaud, Guylaine, 2006, *Présentation du logiciel Thesaurus Occitan*, actes du colloque La dialectologie hier et aujourd'hui (1906-2006), Lyon, à paraître, Brigitte Horiot (Ed.).

Rizzi, Luigi, 1982, *Issues in Italian Syntax*, Dordrecht, Foris.

Séguy, Jean, 1973, "Les atlas linguistiques de la France par régions", *Langue Française*, Volume 18, Numéro 1, p. 65-90.

Tuaillon, Gaston, 1969, "Substrat et structure: à propos d'un solécisme du français populaire de Lyon et de sa région", *Travaux de Linguistique et de Littérature Romanes (TraLiLi), t VII,1*, p. 169-176.

Séminaire de recherche. 2008-2009. CNRS - UMR 6039 "BCL". Université de Nice - Sophia Antipolis.