Pierre-Aurélien Georges
Université de Nice Sophia-Antipolis
pa.georges AT laposte.net

La base TEXTES du THESOC

Introduction

Le THESAURUS OCCITAN, ou THESOC, est une base de données informatique destinée à collecter un ensemble de faits linguistiques oraux, attestés dans les différents parlers de la zone occitane, et précisément localisés géographiquement[1].

La partie lexicale de la base de données [2] permet de consulter les réponses aux questionnaires des différents atlas linguistiques, de diverses manières :

- pour une question donnée dans une localité donnée
- pour l'ensemble des réponses fournies, soit à une question donnée, soit pour une localité donnée

On peut également rechercher des réponses par un système de mots-clés et de thèmes (nature, habitat, homme, etc.).

Comme on le voit dans l'écran ci-dessous, chaque réponse est composée entre autres des champs phonétique, graphie phonologisante, lemme et étymon[3].

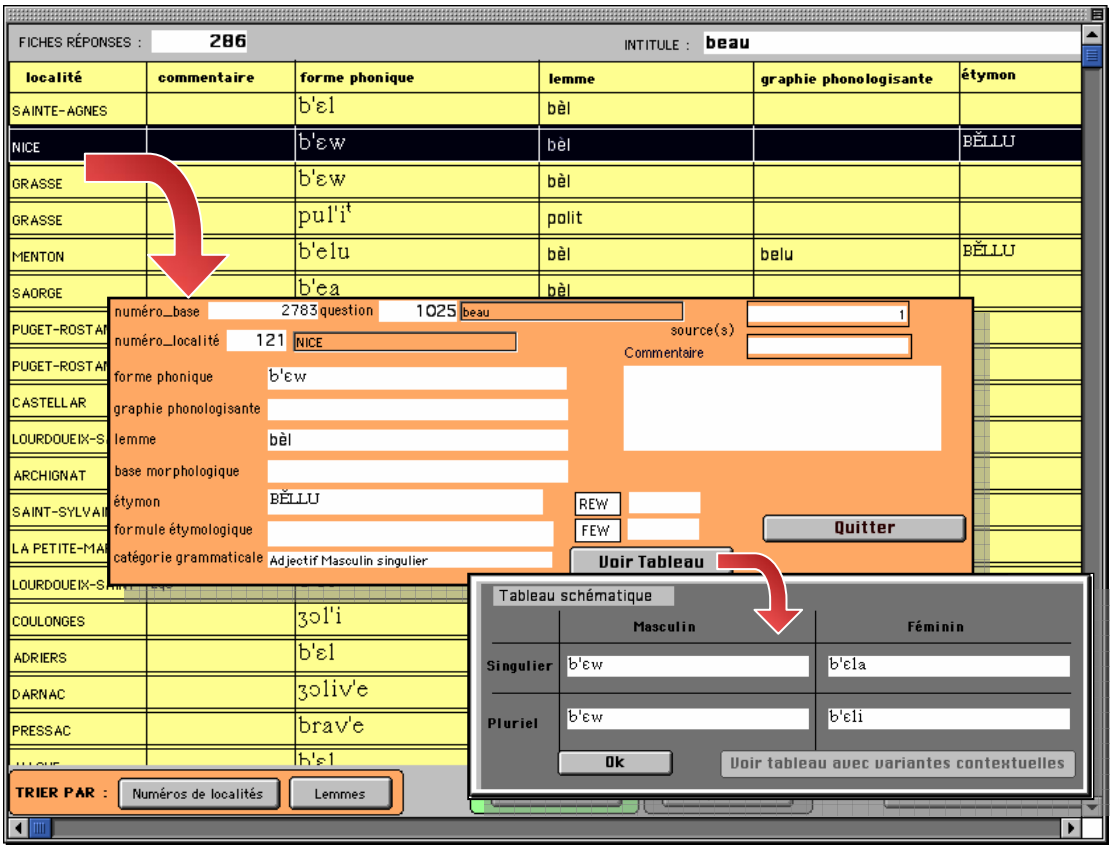
---

[1] Pour plus de précisions, *cf.* Dalbera (1997).

[2] *Cf.* la présentation de M. Oliviéri et G. Brun-Trigaud dans ce volume.

[3] Pour plus d'explications sur ces types de transcription, *cf.* http://thesaurus.unice.fr/.

Capture d'écran du THESOC : consultation des réponses



La base TEXTES

Au sein du THESOC, la base TEXTES est destinée à recueillir un ensemble de textes oraux en vue de leur analyse morphosyntaxique et syntaxique et de la comparaison des structures entre les différents dialectes. Cette étude de la variation syntaxique à la fois géographique et diachronique permettra également de dresser une cartographie des mécanismes syntaxiques identifiés.[4]

Les conditions d'intégration dans la base

Par rapport à la base lexicale du THESOC, les conditions d'oralité et de localisation ont été quelque peu assouplies dans la base TEXTES : du point de vue de la localisation, ce sont des aires géographiques qui remplacent les localités précises, tandis que la notion d'oralité est étendue à l'« oral-

4. *Cf.* Oliviéri (2003) et dans ce numéro.

écrit », c'est-à-dire des textes écrits dans une langue « orale » : presse populaire, théâtre, etc.

Les données

Les textes

Les différents textes respectant ces conditions d'oralité et de localisation qui sont présents dans la base, ou qui auraient vocation à y être, peuvent être d'origines diverses :

- des ethnotextes, sous forme d'enregistrements libres recueillis au cours d'enquêtes lexicales,
- des émissions de radio,
- des pièces de théâtre,
- des articles de la presse populaire,
- des chansons, comptines, poésie populaire…

Bien entendu, d'un point de vue strict, on peut se poser la question de savoir s'il est judicieux de comparer la syntaxe d'un ethnotexte, tout à fait oral, avec une pièce de théâtre, qui relève de l'oral-écrit, ou d'une poésie, dont l'agencement en rimes et en strophes doit répondre à certaines contraintes supplémentaires. C'est pourquoi chaque texte de la base est clairement identifié et étiqueté avec son « genre », et lors de l'utilisation des fonctionnalités de recherche de la base, que nous évoquerons un peu plus loin, il est possible de ne sélectionner que certains genres de textes, ou d'en éliminer certains des résultats de la recherche, afin d'obtenir un corpus de travail homogène et cohérent.

Les phrases

A côté de ces textes, la base contient également un ensemble de phrases isolées, qui sont principalement de deux natures :

- des données provenant des Atlas linguistiques : les cartes morpho-syntaxiques, les locutions et phrases isolées situées dans les marges des cartes,
- des réponses à des questionnaires morpho-syntaxiques, notamment celui utilisé pour les Parlers des Alpes-Maritimes (PAM).

Ces phrases isolées sont soumises aux mêmes traitements informatiques que les textes, et les fonctionnalités de recherche de la base

font apparaître dans leurs résultats, à la fois les phrases et les textes correspondants aux critères demandés[5], et ce sur un même plan.

A propos de la graphie

En ce qui concerne la graphie des textes et des phrases, notre politique d'intégration dans la base TEXTES est la suivante : lorsqu'il existe déjà une version écrite du texte, nous conservons systématiquement la graphie originale du texte tel qu'il a été rédigé par son auteur, qu'elle soit de type alibertin, mistralien, italianisant, ou autre. Lorsqu'il s'agit d'un ethnotexte, de nature orale et donc n'ayant pas de version écrite, la base est dotée d'un transcripteur automatisé qui génère une graphie « phonologisante », proche de la graphie mistralienne, à partir de la transcription API. Mais il reste également possible de saisir une autre transcription graphique pour ce dernier type de textes.

Couverture géographique

La situation géographique de notre laboratoire (Nice) fait que, pour l'instant, la base Textes comporte essentiellement des données concernant les dialectes des Alpes Maritimes, mais elle contient également des textes d'autres régions, notamment le Languedoc et l'Auvergne. A terme, l'objectif est de couvrir tout le domaine occitan et nous serons donc amenés à l'enrichir progressivement avec d'autres textes occitans.

Multimédia

Les textes et les phrases peuvent être accompagnés des enregistrements sonores correspondants, notamment en ce qui concerne les ethnotextes, les pièces de théâtre et les émissions de radio.

Aux textes peuvent également être associés des documents iconographiques, notamment pour conserver une représentation de la mise en page originale, par exemple lorsqu'il s'agit de poésie populaire, ou bien pour illustrer le thème du texte.

Outre l'aspect didactique, voire ludique, de ces fonctionnalités multimédias, la présence des enregistrements sonores au sein même de la base se révèle d'un grand intérêt car elle permet aux chercheurs d'avoir un

5. Il reste cependant tout à fait possible de ne sélectionner que les textes ou que les phrases dans les résultats de la recherche.

accès direct aux sources des données, afin de pouvoir contrôler toutes les étapes du processus de leur exploitation. Ainsi, s'il subsiste un doute sur une particularité de la transcription phonétique ou graphique d'un texte, par exemple, on pourra se référer à l'enregistrement sonore original pour contrôler ou corriger la transcription.

Méthodologie

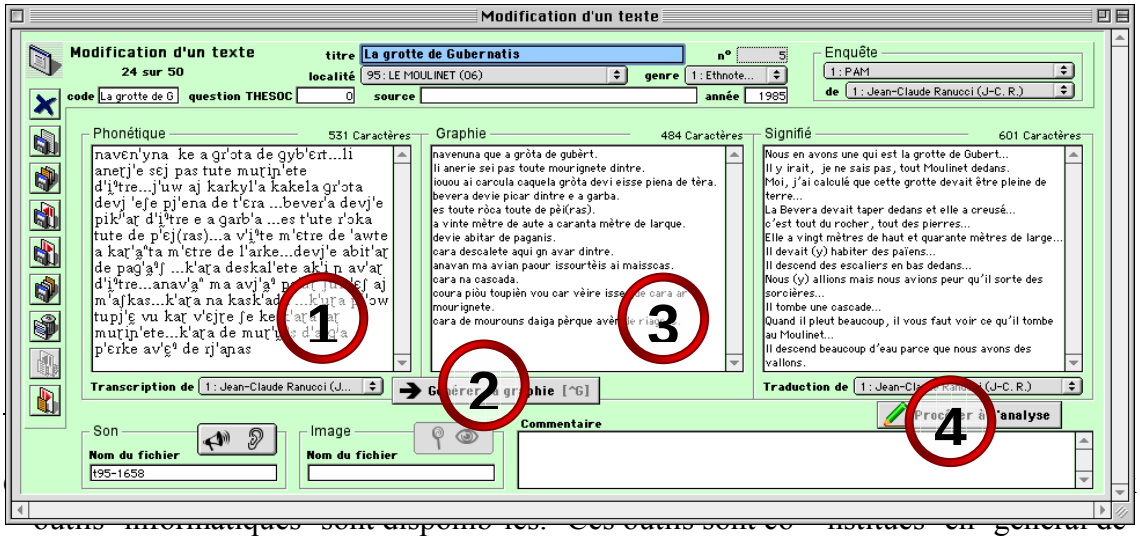
La base TEXTES est dotée d'un certain nombre d'outils informatiques destinés à faciliter et à automatiser, dans la mesure du possible, le traitement linguistique des textes et des phrases.

Préparatifs

Ces traitements linguistiques automatisés se basent sur la transcription graphique. Leur application nécessite donc au préalable de faire quelques aménagements :

- En ce qui concerne les textes de nature orale (typiquement, les ethnotextes) dont on ne dispose pas d'une version écrite, il est nécessaire de préparer une transcription graphique. On peut pour cela faire appel au transcripteur graphique de la base (2) qui génère automatiquement une transcription graphique de type « phonologisante » (3) à partir de la transcription en phonétique API (1) du texte oral[6].

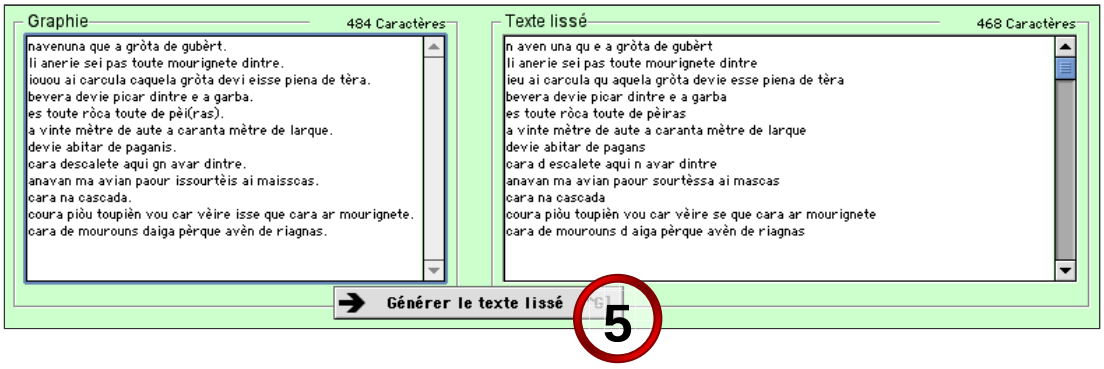
Capture d'écran du formulaire de saisie de textes :



fonctions de lecture en boucle d'une partie de l'enregistrement audio, avec une incrémentation automatique. *Cf.* le chapitre 2 de Delais-Roussarie (2002) pour plus de précisions.

- Ensuite, que le texte comporte une version écrite originale ou non, il faut effectuer ce que l'on appelle un « lissage » de cette transcription graphique. Cette étape consiste à retirer la ponctuation, effectuer des opérations telles que la conversion des chiffres en toutes lettres (lorsque le texte a été saisi en graphie) et vérifier et corriger au besoin le découpage de la chaîne en mots (surtout dans le cas où la graphie est générée automatiquement à partir de la phonétique par le transcripteur), notamment en ce qui concerne les amalgames, les mots composés et les mots valises. On dispose pour cela d'une fonction (5) qui retire automatiqueme nt la ponctuation et si gnale à l'utilisateur la présence éventuelle de chiffres ou d' autres caractères spéciaux (symbole monétaire, symbole « & », etc.) dans le texte. L'utilisateur peut a insi gagner du temps sur ces étapes et se concentrer sur le découpage de la chaîne en mots. Dans cet exemple, on voit par exemple dans la première phrase de la graphie générée, « *navenuna que a gròta de gubèrt* », que les trois premiers mots ne sont pas séparés, faute de présence de blancs dans la transcription API. Il faut donc les séparer en « *n aven una* ». De même, dans la suite de cette phrase, il faut découper « *que* » en « *qu e* », car il s'agit ici de deux unités syntaxiques.

Capture d'écran du formulaire d'analyse de textes : lissage du texte

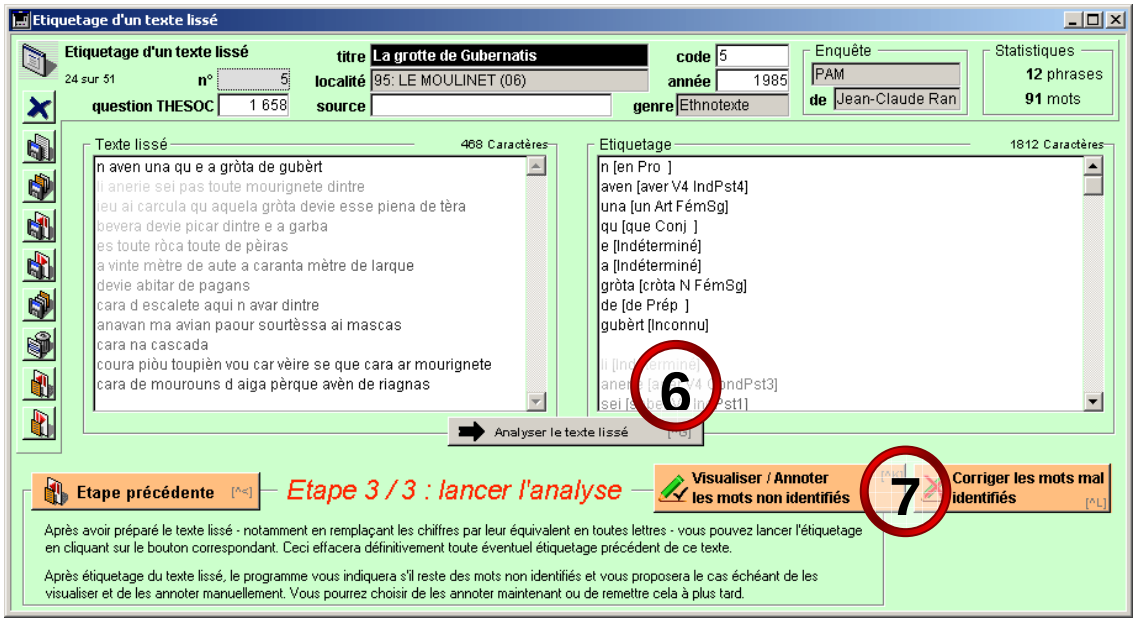


Etiquetage morphosyntaxique

Après avoir effectué ces différents préparatifs, l'étiqueteur morphosyntaxique de la base permet d'annoter automatiquement les mots d'un texte donné en les associant à une variante du dictionnaire. L'étiqueteur compare un par un les mots d'un texte donné avec le contenu de ce dictionnaire, et les mots se voient ainsi affectés d'une catégorie syntaxique et d'une flexion, et sont rattachés à un lemme.

Dans l'étape suivante, l'analyse du texte lissé (6) par l'étiqueteur morphosyntaxique permet d'affecter à chaque occurrence une étiquette comportant : le lemme auquel le mot est rattaché, sa catégorie grammaticale et sa flexion. Les résultats sont affichés avec un mot par ligne, chaque phrase étant délimitée par une ligne vide :

Capture d'écran du formulaire d'analyse de textes : étiquetage morphosyntaxique



Pour le premier mot de la phrase, *n*, l'étiqueteur a trouvé une et une seule variante correspondant dans le dictionnaire. Le mot est donc annoté avec les informations de cette variante. Il en est de même, pour *aven*, *una* et *qu*, qui sont respectivement identifiés comme un verbe, un article et une conjonction. En revanche, les mots *e* et *a* n'ont pas pu être identifiés automatiquement car il y a, pour le premier, deux variantes, et pour le deuxième, trois variantes, présentes dans le dictionnaire avec la même graphie. L'étiqueteur n'est donc pas en mesure de se prononcer sur l'identité de ces mots et les étiquette comme [indéterminé]. Les mots *gròta* et *de* sont bien étiquetés avec les informations de leurs variantes correspondantes mais le nom propre *Gubèrt* n'a aucune variante correspondante dans le dictionnaire, il est donc annoté comme [inconnu].

Signalons également que les annotations générées automatiquement par cet étiqueteur morphosyntaxique peuvent être complétées ou modifiées manuellement si besoin (7).

Structure du dictionnaire

L'étiquetage morphosyntaxique est basé sur un dictionnaire annoté qui contient pour chaque variante des informations associées, telle que sa catégorie grammaticale. Etant donné que nous avons choisi de conserver la graphie originale de chaque texte, le dictionnaire de la base doit contenir un grand nombre de variantes pour étiqueter correctement tous ces textes.

Afin de prendre en compte toutes ces variations, le dictionnaire de la base est structuré en deux niveaux : le niveau des variantes qui enregistre individuellement chaque forme avec la localité dans laquelle elle est attestée ainsi que sa flexion. A un niveau supérieur, on regroupe ensuite toutes les variantes correspondant à un même lemme. Cela permet de factoriser les informations concernant le signifié et la catégorie grammaticale, informations qui sont les mêmes pour toutes les variantes d'un lemme.

Structure du dictionnaire:[7]

Lemmes		
lu	article	"le/la"
li	pronom pers.	"lui"
fòl	adjectif	"fou"
bèl	adjectif	"beau"

Variantes d'un lemme		
fuala	fém. sing.	Nice
fuale	masc. plur.	Nice
fuali	fém. plur.	Nice
fòlha	fém. sing.	Languedoc[7]

Variantes d'un lemme		
bèl	masc. sing.	Nice
beu	masc. sing.	Nice
bela	fém. sing.	Nice
bella	fém. sing.	Nice

Dans cette représentation schématique, on peut voir les différents types de variation :

- phonologique entre *bèl* (devant voyelle) et *beu* (devant consonne) à Nice,
- dialectale entre *fòlha* dans le Languedoc et *fuala* à Nice,
- graphique entre *bela* et *bella* à Nice,
- et flexionnelle entre *fuala*, *fuali* et *fuale* à Nice.

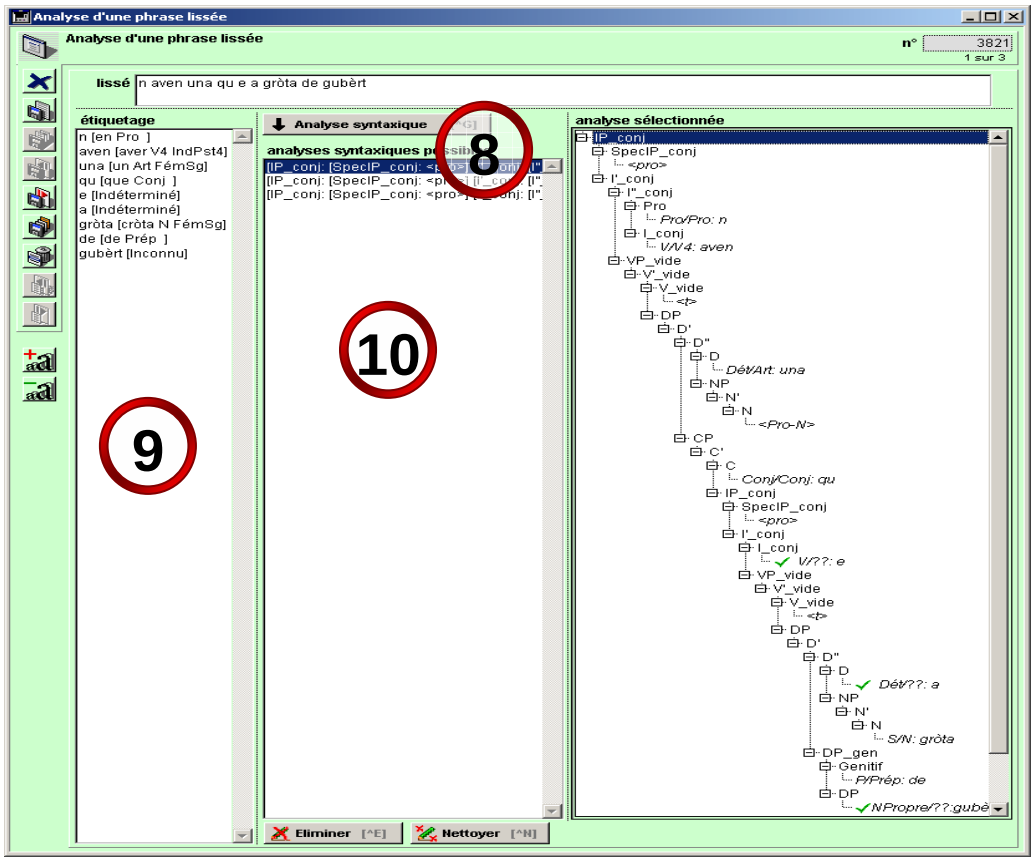
Cette structure à deux niveaux nous permet ainsi de traiter la variation lexicale dans toutes ses dimensions, et d'intégrer des dictionnaires occitans imprimés ou informatisés, tels que ceux de Alibert (1966), de Castellana (1983), de Eynaudi (1932), de Mistral (1979) et de Palay (1991) ; afin de compléter les données provenant de la base lexicale du THESOC qui ont initialement alimenté le dictionnaire de la base TEXTES.

Analyse syntaxique

Enfin, la dernière étape consiste à exécuter phrase par phrase l'analyseur syntaxique (8) sur la base des résultats fournis par l'étiqueteur morphosyntaxique (9) afin d'obtenir une ou plusieurs analyses syntaxiques (10) pour chaque phrase :

7. Cette forme est extraite de l'entrée *fòl* de Alibert, 1966.

Capture d'écran du formulaire d'analyse de phrases : analyse syntaxique[8]



Finalités d'exploitation

Fonctionnalités de recherche

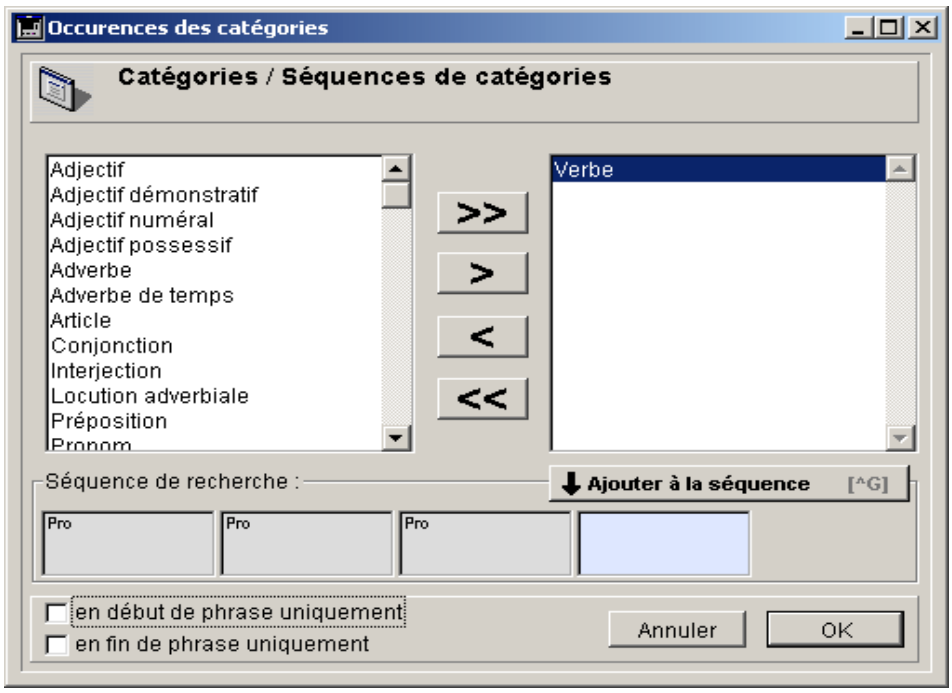
Les différentes annotations générées lors des étiquetages morphosyntaxiques de textes ou de phrases sont ensuite exploitées par les fonctionnalités de recherche. On peut notamment rechercher toutes les occurrences de :

- une variante donnée, ou de toutes les variantes rattachées à un lemme.
- toutes les variantes d'une ou plusieurs catégories syntaxiques données.
- toutes les occurrences d'une séquence de catégories syntaxiques donnée.

8. Pour des raisons techniques, l'arborescence syntaxique est ici présentée de manière verticale au lieu de l'habituelle représentation horizontale. A terme, nous envisageons d'effectuer un développement informatique pour pouvoir afficher les arbres syntaxiques de manière traditionnelle.

On peut également consulter la liste des variantes attestées dans une localité donnée.[9]

Exemple de recherche de toutes les occurrences de trois Pronoms consécutifs



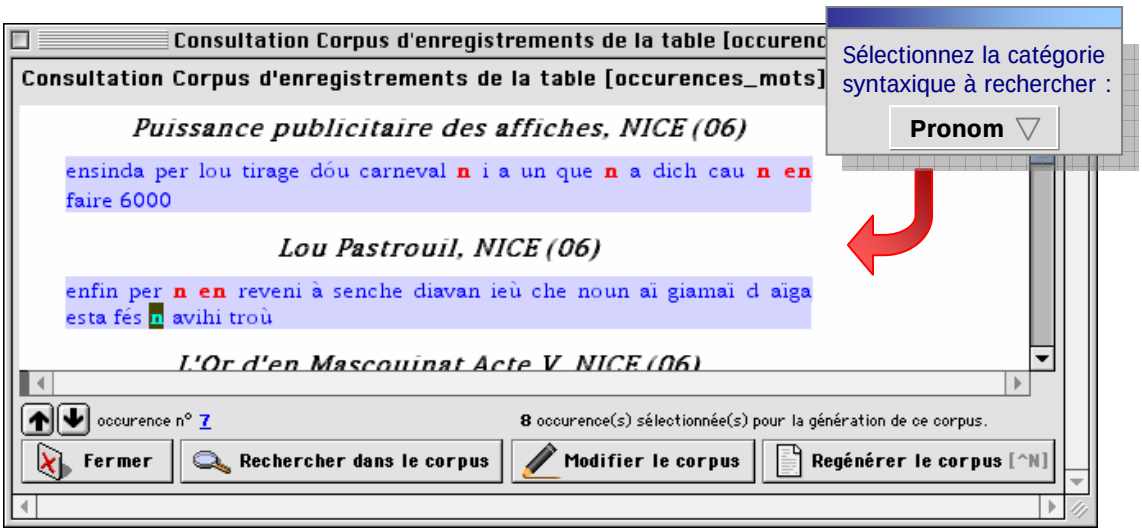
Il est également possible d'effectuer des recherches sur les résultats des analyses syntaxiques précédentes. On peut ainsi rechercher toutes les phrases contenant une certaine structure syntaxique (par exemple toutes les phrases contenant un DP Génitif à l'intérieur d'une Small Clause).

Génération d'un corpus de travail

Les résultats de ces recherches peuvent ensuite faire l'objet d'une génération automatique de corpus de travail. Il est possible, par exemple, de rechercher tous les mots portant l'étiquette « pronom » et de générer automatiquement un corpus de travail dans lequel les occurrences sont présentées en surlignage dans leur contexte d'origine, à savoir le titre du texte, sa localisation, et la phrase ou le texte complet dans lequel apparaît chaque occurrence :

9. C'est-à-dire toutes les variantes pour lesquelles on dispose dans la base d'au moins une occurrence dans un texte ou une phrase attesté(e) dans cette localité.

Capture d'écran d'un corpus de pronoms généré avec la base de textes



De la même manière, on peut générer un corpus à partir de recherches basées sur les résultats des analyses syntaxiques précédemment effectuées.

Conclusion

La base Textes se révèle donc être un excellent outil informatique pour les chercheurs qui s'intéressent à la variation dialectale et dans ses dimensions morphosyntaxique et syntaxique, et nous pensons qu'elle devrait permettre une meilleure étude de ces aspects-là des dialectes occitans, à la fois sur le plan diatopique, mais également sur le plan diachronique, si l'on se réfère à l'hypothèse formulée par Dalbera (1992), qui 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.*[10]

[10] Dalbera (1992), p. 138.

Par ailleurs, il est envisagé, dans le cadre des développements futurs de cette base, d'y intégrer un vol et cartographique, comme c'est déjà le cas dans la base lexicale du THESOC, ce qui permettra d'automatiser et donc de faciliter l'élaboration d'une cartographie des mécanismes identifiés.

Bibliographie

Alibert, Louis. 1966. *Dictionnaire Occitan-Français d'après les parlers languedociens*, Toulouse, Institut d'Etudes Occitanes.

Castellana, Georges. 1983. *Dictionnaire Nicois-Français*, Nice, Serre.

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

Dalbera, Jean-Philippe. 1997. *Langue Occitane - La base de données THESOC : Brève présentation - Etat des travaux*, Nice.

Delais-Roussarie, Elisabeth. 2002. "Constituer des Corpus Oraux: Méthodes et Outils", *Carnets de Grammaire* n°10, Equipe de Recherche en Syntaxe et Sé mantique, UMR 5610, CNRS & Université de Toulouse-Le Mirail.

Eynaudi, Jules. 1932. *Dictionnaire de la Langue Niçoise*, Nice, Imprimerie de "l'Eclaireur de Nice".

Mistral, Frédéric. 1979. *Lou Tresor Dóu Felibrige ou Dictionnaire Provençal-Français*, Raphèle-lès-Arles, Culture Provençale et Méridionale.

Oliviéri, Michèle. 2003. *Constitution d'une base de textes occitans*, 36ème colloque de la Societas Lin guistica Europaea: Linguistique et Corpus, Lyon, 4-7 septembre 2003.

Palay, Simin. 1991. *Dictionnaire du Bearnais et du Gascon Modernes*, Paris, Éditions du Centre National de la Recherche Scientifique.