LA BASE DE DONNÉES LINGUISTIQUE OCCITANE THESOC. TRÉSOR PATRIMONIAL ET INSTRUMENT DE RECHERCHE SCIENTIFIQUE

Jean-Philippe Dalbera
Michèle Oliviéri
Jean-Claude Ranucci
Guylaine Brun-Trigaud
Pierre-Aurélien Georges
Université Nice - Sophia Antipolis / CNRS
UMR 7320 (ex 6039) 'Bases, Corpus, Langage' – ISHSN

CARTE D'IDENTITÉ

Le *Thesaurus Occitan* (THESOC) se veut l'héritier et le continuateur des atlas linguistiques de la France par régions relevant du domaine occitan. Il réside à Nice, hébergé par l'UMR 6039 «Bases, Corpus et Langage» (BCL), plus précisément au sein de l'équipe «Dialectologie, diachronie, phonologie» de *BCL*, à l'adresse http://thesaurus.unice.fr. Sous cette étiquette de *continuateur* se trouve exprimée l'idée d'un double objectif assigné à cette base de données. Il s'agit d'une part de *poursuivre* la confection du trésor patrimonial initiée par les atlas, c'est-à-dire compléter et mettre à disposition du public (en en facilitant grandement l'accès) via Internet les matériaux dialectaux publiés par le CNRS dans la collection des atlas. D'autre part, nous souhaitons *prolonger* la démarche de sauvegarde de ce corpus de mémoire vivante en développant la recherche sur ces matériaux et en leur faisant dire, à l'aide de traitements appropriés, ce qu'ils renferment sur la langue, son évolution et sa partition dans l'espace géographique, notamment sur les réseaux sémantiques que le lexique donne à voir à travers la variation diatopique et qui structurent la vision des choses.

Le *Thesaurus Occitan* fait figure de travail de pionnier. Il a été l'une des premières bases de données dialectales réalisées, dotée d'outils conceptuels et d'analyse originaux. Nous nous efforcerons dans cette présentation du THESOC d'éclairer les choix qui ont été faits par la référence au contexte dans lequel ils sont intervenus.

Estudis Romànics [Institut d'Estudis Catalans], Vol. 34 (2012), p. 367-387
DOI: 10.2436/20.2500.01.108

HISTORIQUE

L'idée de constitution d'une telle base de données remonte au colloque de Wégimont organisé par l'Association Internationale d'Études Occitanes (AIEO) en avril 1989. Ce colloque est consacré aux outils de la recherche occitane. Au cours de la séance de synthèse, suite à un appel à projets lancé par Robert Lafont, Jean-Philippe Dalbera, fort de l'expérience de la Banque de Données Langue Corse (BDLC), esquisse l'idée de la construction d'une base de données linguistiques occitanes susceptible de fédérer les linguistes occitanistes et de fournir, à terme, matière à l'élaboration d'outils pédagogiques et de recherche.

Mais l'AIEO ne dispose pas des moyens de mettre en œuvre une telle entreprise. Tout au plus peut-elle jouer un rôle d'accompagnement et de caution. J.-Ph. Dalbera s'efforce alors de mettre en place une structure inter-universitaire souple à laquelle notamment les auteurs d'atlas du domaine d'oc ont adhéré; il définit à grands traits à cette occasion et soumet à la discussion les orientations de la base de données telle qu'il l'envisage et évoque les modalités de collaboration qui pourraient sans délais être mises en œuvre. Le contact est pris. Un axe Nice-Toulouse semble se dessiner; les autres adhésions sont essentiellement individuelles (telles, par exemple, celle de Kathryn Klingebiel, en charge d'un fichier autonome de bibliographie linguistique occitane). Néanmoins le travail peut commencer. J.-Ph. Dalbera obtient des soutiens financiers de la part de l'Europe (CEE), les Régions commencent à s'intéresser à ce programme et à y appor-

ter des contributions, notamment Midi-Pyrénées. La Délégation Générale à la Langue Française (DGLF) manifeste aussi son intérêt.

CONSTRUIRE ET IMPLEMENTER LA BASE DE DONNÉES

La base de données envisagée se voit assigner une double fonction: constituer un trésor patrimonial en prenant part à la sauvegarde de témoignages formant «mémoire vivante » d'une langue d'oc menacée dans son existence; et doter celle-ci d'un instrument de recherche permettant de faire progresser la connaissance de ce domaine linguistique. Plusieurs présentations, sous forme d'articles de revue ou de démonstrations publiques ont été effectuées au fil des ans, faisant la part belle à l'aspect patrimonial (cf. les références que nous donnons en fin d'article); nous insisterons ici un peu plus sur les perspectives heuristiques ouvertes.

De toute évidence la réalisation d'une telle base de données requiert une équipe stable, soudée, compétente et volontaire. Mais c'est avec «les moyens du bord» que sont entamées les premières démarches de la construction. Il ne saurait être question d'entrer dans les détails et de suivre pas à pas la progression, ronflante ou cahotique selon les périodes, du chantier: on en jugera d'après les résultats. En revanche, il semble légitime de s'attarder, l'espace de quelques lignes, sur les artisans qui ont façonné le THESOC. Le noyau dur ne comporte, au départ, que cinq personnes: trois linguistes dialectogues travaillant sur l'occitan, Jean-Philippe Dalbera, Michèle Oliviéri et Jean-Claude Ranucci, auxquels il faut associer deux autres acteurs importants: Marie-José Dalbera-Stefanaggi et Dominique Strazzabosco. Ces derniers se sont en effet engagés dans une entreprise analogue en Corse (la *Banque de Données Langue Corse*, BDLC) et le programme THESOC va bénéficier évidemment de l'expérience de la BDLC.

L'équipe de BCL organise à Nice en 1992 un colloque réunissant la plupart des acteurs de la linguistique occitane. Sont sollicités pour adhérer à ce projet tous les linguistes travaillant sur l'occitan; mais ceux-ci, au demeurant fort peu nombreux, hésitent à se détourner de leurs programmes propres et, en définitive, seuls quelques auteurs d'atlas appelés à être inclus dans le THESOC s'impliquent dans la démarche. Autour du noyau dur des fondateurs vont graviter à Nice plusieurs générations d'étudiants de DEA ou de Master qui, au titre de l'initiation à des tâches de recherche, vont mettre la main à la pâte et participer à des saisies de données de toutes sortes, à des séances de correction des fiches, à des tests sur les requêtes ou sur la lisibilité des formats... Au delà des étudiants, des renforts épisodiques sont accordés par les tutelles au programme en cours, sous la forme de rattachements temporaires au laboratoire BCL de personnels ITA ayant appartenu au GRECO des atlas et devenus «électrons libres» à mesure de la clôture par le CNRS des chantiers régionaux. Ces collaborations ponctuelles n'ont pas toujours été faciles à gérer. Néanmoins elles ont eu dans certains cas des retombées positives non négligeables; ainsi le rattachement de Jean-Claude Potte, auteur de *l'Atlas Linguistique et ethnographique de l'Auvergne et du Limousin*, au THESOC s'est traduit par le fait que

celui-ci a procédé lui-même à la saisie de ses données dans la base; ce qui, sans aucun doute, représente un volume de travail conséquent et confère *de facto* une garantie de fiabilité maximale des matériaux engrangés. Par ailleurs, la dynamique de recherche impulsée par les directeurs successifs du laboratoire BCL, Etienne Brunet, Sylvie Mellet et Tobias Scheer, profite à l'équipe du THESOC qui s'étoffe de manière significative. Dans un premier temps, le laboratoire obtient du CNRS un poste d'ingénieur d'étude sur lequel il recrute, en la personne de Guylaine Brun-Trigaud, la collaboratrice idéale dans la perspective de donner vie au THESOC. Les résultats ne se font pas attendre. L'enthousiasme, la compétence, l'esprit d'initiative de G. Brun-Trigaud ont raison des pesanteurs et des obstacles; la saisie des données dans la base connaît une accélération notable et l'on commence à voir le bout du tunnel. Dans un second temps, ce qui constitue désormais le maillon faible de l'opération, à savoir l'évolution de l'architecture proprement informatique de la base de données, va cesser d'être un problème, les choses ayant été remarquablement négociées par le laboratoire.

La construction de la base de données a longtemps été le fait de Dominique Strazzabosco, un mathématicien féru d'informatique et sans *a priori* linguistique, ayant tout à découvrir en matière de langue et *a fortiori* de dialectes. Tout le travail s'est opéré dans le cadre d'une réelle interdisciplinarité, sous forme d'échanges permanents entre des linguistes, porteurs d'un contenu et détenteurs de modèles en matière de langage, et un informaticien, chargé de traduire en termes informatiques les structures et les requêtes des linguistes. Ces dialogues intervenant au fil de la construction de la base sans que la matière soit toujours dominée, on conçoit aisément que la base puisse aujourd'hui apparaître parfois comme un empilement de fichiers liés par des liens *ad hoc*, que des redondances indues alourdissent sans nécessité certaines programmations, etc. Cela a été aggravé par le fait que la construction de l'édifice est intervenue dans une période d'évolution extrêmement rapide de l'informatique elle-même de telle sorte que ce qui faisait figure un jour d'obstacle infranchissable, de véritable montagne (qu'il s'agisse de problèmes relatifs à la capacité de mémoire, aux temps de réponse, à l'utilisation de polices de caractères spécifiques, commodes à usage interne en circuit fermé mais véritable poison dès lors qu'une libre diffusion était envisagée) cessait de faire difficulté le lendemain ; et que la solution adoptée à un moment *t* comme la meilleure, eu égard aux instruments informatiques disponibles, était susceptible d'apparaître comme désuète au moment *t+1* ou *t+2*.

Lorsqu'on regarde en arrière, on en arrive même à s'étonner, compte tenu des conditions difficiles de sa mise en œuvre (tâches sur vacations, aléas des collaborations précaires...) de la qualité des résultats obtenus. On aura compris néanmoins que, sans renier le moins du monde le travail effectué sur ce plan informatique, la nécessité s'est fait sentir à un moment donné de réécrire la base. Et il se trouve que, dans le cursus de sciences du langage de Nice, avait été remarqué un étudiant en informatique brillant, issu de l'ESSI, en passe d'achever (brillamment) un cursus complet de linguistique. Le recrutement en 2007 de Pierre-Aurélien Georges en tant qu'ingénieur de recherche (CNRS) est l'avant-dernière touche apportée à la composition de l'équipe THESOC; elle est d'une grande importance dans la mesure où elle garantit une assise stable et une sécurité précieuse à l'entreprise dans son ensemble. Les linguistes-dialectologues peuvent désor-

mais envisager de modifier ou d'étendre certaines de leurs hypothèses de travail sans inquiétude; l'implémentation des ajustements informatiques suivra.

L'ultime retouche, à ce jour, concernant l'équipe en charge du THESOC revêt également une grande importance; c'est la nomination à l'Université Nice–Sophia Antipolis d'Elisabetta Carpitelli sur un poste de Professeur (sciences du langage-dialectologie); cette mesure traduit la reconnaissance, sur le plan de l'Institution, du pôle de dialectologie-diachronie, constitue un renfort de poids pour cette formation et laisse augurer, du fait de l'implication d'E. Carpitelli dans les programmes supranationaux de l'*Atlas Linguistique Roman* et de l'*Atlas Linguarum Europae*, un nouvel essor des travaux comparatistes et un nouvel élan du renouveau disciplinaire amorcé en syntaxe et en étymologie.

LES OBJECTIFS SCIENTIFIQUES

Le temps de l'exploitation de la base de données THESOC à des fins de description de la langue occitane dans toutes ses variétés, d'établissement d'une typologie des parlers, d'une analyse qui conduise à un essai de reconstruction de la langue d'oc est sans doute venu. Et, juste retour des choses, ces résultats devraient alimenter des modélisations nouvelles du changement linguistique et peser dans les discussions, à l'intérieur des sciences du langage, sur l'évolution de certains secteurs disciplinaires.

L'UMR 6039 s'est fait une spécialité d'aborder un certain nombre de domaines à travers le prisme de la dialectologie. L'idée directrice est simple. Les parlers qui se développent sur une aire vraisemblablement homogène à partir d'un système vraisemblablement commun illustrent un cas de figure d'école: un développement presque *toutes choses égales par ailleurs*; or l'évolution y produit toutes sortes de divergences, engendre des variations multiples, donne à voir des phases de convergence spécifique (sous forme d'innovations partagées) autant que de divergences. Tout cela donne à penser que l'on peut essayer de saisir la dynamique spécifique du langage en analysant les phénomènes qui s'y produisent et en tentant de reconstruire le système qui en constitue la source et les changements qui, à travers des phases successives, l'ont altéré; bref, le processus modelant les dialectes.

La reconstruction du phonétisme, c'est-à-dire la phonologie diachronique, et celle des structures morphologiques, la morphologie diachronique, ne sont plus à découvrir. En revanche, la syntaxe et plus encore le lexique semblent se dérober à une telle approche.

Forts d'un instrument puissant pour accéder aux données dialectales, les trier, les représenter dans l'espace, les superposer et disposant avec celui-ci d'un moyen de tester des hypothèses, nous avons tenté l'aventure d'un essai de reconstruction lexicale. Cela conduit notamment, une fois établi que, dans le domaine roman, la part de reconstruction dans la démarche étymologique ne concerne que la forme phonique (aboutissant aux formes à astérisques dans les dictionnaires étymologiques) et que le sens ne fait pas systématiquement l'objet d'une reconstruction sur base comparative, à montrer qu'il est

possible de reconstruire le signifié. Et la méthode pour parvenir à ce résultat consiste à pratiquer une analyse motivationnelle comparative sur les variantes que constituent les divers types lexicaux par lesquels les dialectes expriment une notion donnée.

Ces recherches excèdent largement le cadre de cette présentation du THESOC. Nous entendons seulement, en les évoquant, souligner que, dès lors, la base de données n'est plus seulement une masse inerte de matériaux ni même une vitrine technologique mais remplit une fonction heuristique.

LES OBJECTIFS FONCTIONNELS

Le cahier des charges de notre base de données comporte quatre exigences majeures. Il s'agit de faire en sorte que ces données soient (1) *exhaustives*, (2) *commensurables*, (3) *consultables* et (4) *exploitables*. Encore faut-il préciser le sens que nous donnons à ces quatre termes.

(1) exhaustivité

Les matériaux que nous avons décidé de prendre en compte dans la base de données sont définis par deux propriétés: être issus de l'oralité, et donc avoir été recueillis par voie d'enquête de terrain auprès de locuteurs natifs, et être localisables dans l'espace géographique. Ces deux contraintes –consubstantielles à nos données puisqu'il s'agit là du format propre aux atlas linguistiques– sont complètement assumées par nous et restent non négociables, du moins pour la partie lexicale de la base. Cela signifie que l'exhaustivité que nous avons en point de mire, se réduit, si on ose dire, à l'ensemble des données accumulées au cours du siècle dernier par les dialectologues qui ont réalisé les atlas et complétées par les monographies satisfaisant aux mêmes critères.

Les atlas n'ont pas tous été publiés complètement et ne sont pas toujours disponibles. La base de données inclut, outre les matériaux figurant dans les cartes publiées, ceux des volumes restés en préparation ainsi que des données issues des carnets d'enquêtes. Sont également intégrées dans le THESOC les résultats d'autres enquêtes, publiées ou non, souvent postérieures à celles des atlas, notamment celles menées par l'équipe de BCL dans les Alpes-Maritimes (indiquées par le sigle *PAM*).

(2) commensurabilité

Le système de transcription utilisé dans les atlas n'est pas toujours homogène et les modalités de transcription elles-mêmes peuvent varier d'un auteur à l'autre. On ne peut empêcher que tel enquêteur accorde plus d'importance que tel autre à la précision de la transcription ou à la variation ; il ne faut pas perdre de vue que consigner sur une carte la forme phonique entendue dans une localité en réponse à telle question n'est pas un geste neutre, ne serait-ce que par exemple parce que l'on dispose souvent de plusieurs témoins et que leurs productions phoniques ne sont pas toujours rigoureusement superposables. Certes, les enquêteurs sont libres de noter ces variations mais il n'en reste pas moins que

le seuil de variation à partir duquel on prend la peine de consigner plusieurs réalisations ou, à l'inverse, de n'en retenir qu'une seule en introduisant une dose, fût-elle minime, de normalisation, n'est jamais défini à l'avance. D'où des distorsions inévitables dans la notation. Nous avons, dans la mesure du possible et en restant fidèle au transcripteur, gommé certains écarts dont l'artificialité nous semblait manifeste. De même, nous nous sommes efforcés de ramener les formes figurant les aires dégagées dans les cartes au format ordinaire (une forme phonique dans chaque localité), en recourant aux carnets d'enquête.

(3) consultabilité

C'est bien là la moindre des choses; mais cette exigence apparemment tautologique est là pour rappeler que le lecteur doit pouvoir accéder commodément aux matériaux bruts et non pas seulement aux résultats des analyses proposées par les auteurs de la base. Les fichiers de transcriptions sont (autant que possible) doublés par des fichiers de sons numérisés, ayant le double objectif de sauvegarder les data brutes et d'autoriser une réécoute de contrôle. De même, la classification des verbes n'est pas un donné mais une représentation, qui reste hypothétique, de l'organisation du système verbal; celle-ci est destinée à éclairer les faits verbaux et à mettre de l'ordre dans le fouillis, mais elle ne doit pas remplacer les faits. Le lecteur doit pouvoir la réenvisager en fonction d'un principe éventuellement différent.

(4) exploitabilité

La base de données a été conçue, non seulement comme banque de dépôt à titre conservatoire, mais aussi et surtout comme outil de recherche. Les présentations cartographiques caractéristiques des atlas sont évidemment conservées mais elles deviennent actives. Les cartes figées sur le papier cèdent la place à des représentations cartographiques «à la demande». Globalement, la philosophie du THESOC est de fournir à l'utilisateur à la fois séparément et conjointement des données brutes et des données traitées.

ARCHITECTURE DE LA BASE

L'architecture de la base autorise cette double fonction. Elle peut livrer des séries de faits en réponse à une demande particulière mais elle peut aussi proposer sous différentes formes des résultats d'analyses stockés, notamment des lemmatisations, des découpages morphologiques, des cartes traitées (à aires, à symboles, etc.) ou même des tableaux figurant des corrélations ou des cartes de synthèse. Elle autorise également des modules périphériques.

Nos données de départ sont pré-formatées. Nous n'en sommes pas maîtres et nous ne pouvons en aucun cas les modifier. Notre seule liberté est d'en tirer le meilleur parti. Ces données sont de divers ordres: lexicographiques au premier chef, mais aussi

ethnographiques (chacun sait que les enquêteurs ont recueilli toute une iconographie en relation avec leur enquête), textuelles (ethnotextes) ou encore métalinguistiques (commentaires des témoins sur les mots eux-mêmes, sur les realia, sur le sentiment linguistique, etc.).

VISITE DU THESOC

Pour s'orienter au cours d'une visite de la base on dispose d'un tableau de bord :

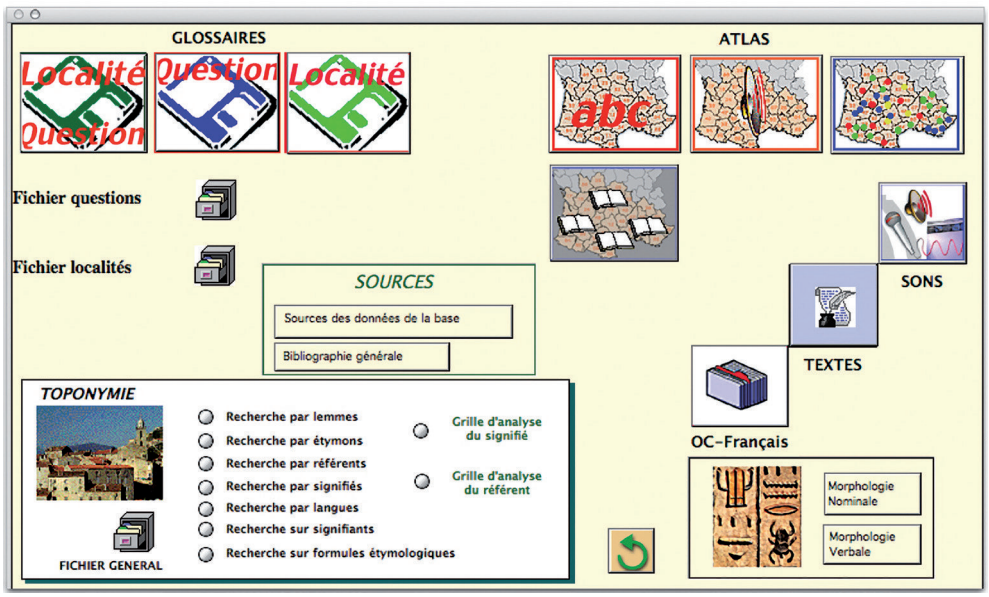
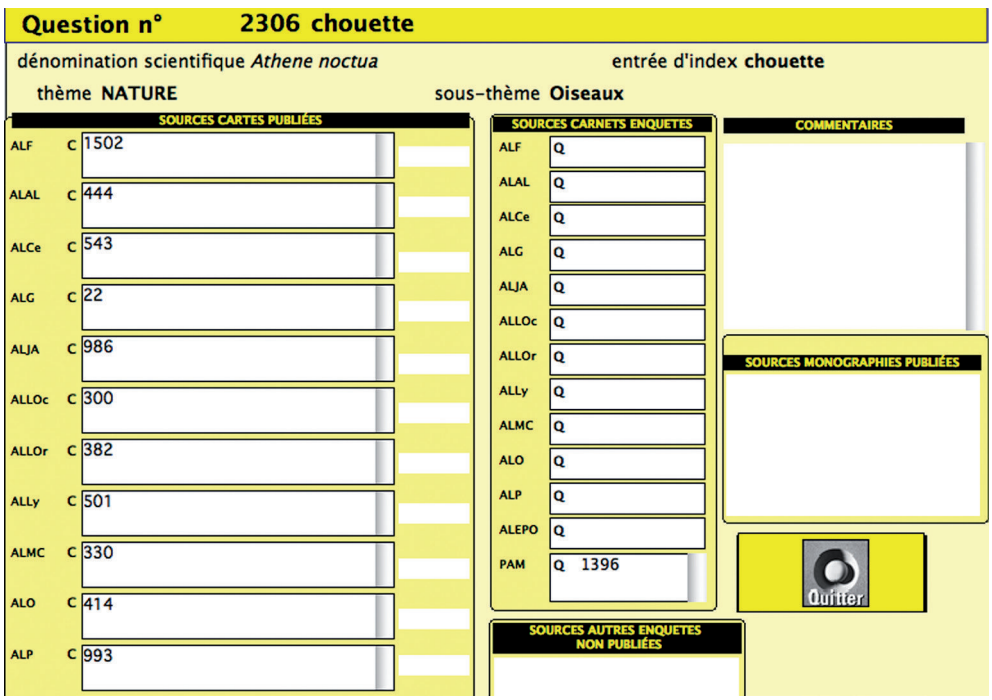


Tableau de bord

La base lexicale est dotée d'un fichier central, celui des «réponses» aux questionnaires élaborés pour l'enquête, et de fichiers périphériques contenant notamment les images et les sons. Une «réponse» dans la base est repérée par le couple «localité-question». Les questions sont organisées en un «responsaire» et les localités sont stockées dans un fichier spécifique.

LE « RESPONSAIRE » (OU FICHIER QUESTIONS)

Le fichier des questions ou «responsaire» (car il s'agit de l'ensemble des notions, des réponses prévisibles) est le résultat de la somme des notions ayant fait l'objet de questions dans les différents atlas régionaux du domaine occitan. Cependant, cette «somme» est beaucoup plus qu'une réunion de questionnaires divers: un travail considérable a été réalisé pour éclairer les relations entre les mots et les choses et gérer les discordances observées.

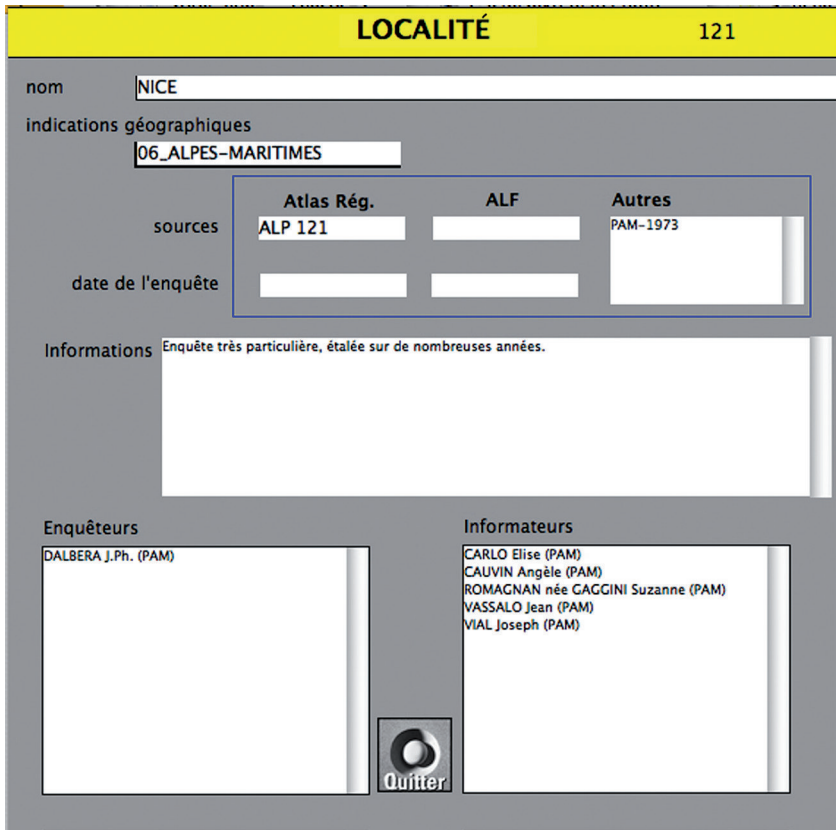


Fiche question : 2036 CHOUETTE

Il comporte plus de 8000 questions regroupées autour des principaux thèmes traités dans les atlas: les cultures, l'élevage, la nature, l'espace, le temps, l'habitat et la vie quotidienne, l'homme... Chaque fiche indique les références de cette question dans chaque atlas.

LE FICHIER LOCALITÉS

Le réseau comprend 831 localités et chaque fiche précise toutes les indications sur l'obtention des données (date, enquêteur, témoins,...).

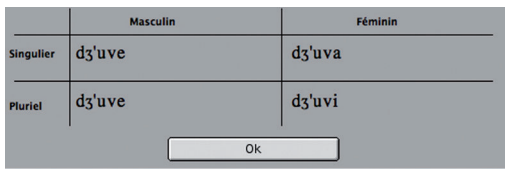


Fiche localité : 121 NICE

LE FICHIER RÉPONSES

La fiche-réponse, si elle constitue une unité de compte commode, ne se borne pas à enregistrer la réponse. Celle-ci est affichée sous trois formes qui correspondent à trois niveaux de description de la langue et qui ont chacun une validité propre: (i) transcrip-

tion phonétique API, (ii) transcription lissée (d'ordre phonématique) qui renvoie à la notation de certains dictionnaires classiques (TDF notamment), (iii) transcription lemmatisée qui gomme au moins partiellement les variations phonétiques que l'espace géographique donne à voir et qui renvoie à la graphie préconisée par Alibert.



Fiches réponse par localité : la TOUPIE à Nice *JEUNE à Nice*

Ces trois niveaux de transcription ne sont nullement décoratifs, chacun possède sa pertinence en fonction des requêtes à envisager. Toute recherche d'ordre lexical s'opère préférentiellement sur les lemmes mais toute considération sur les systèmes phonologiques ne peut prendre en compte que la transcription phonétique. Différentes informations figurent également sur cette fiche, morphologiques (avec accès aux tableaux) et étymologique (avec références aux deux grands dictionnaires étymologiques que sont le *Romanisches Etymologisches Wörterbuch* (REW) et le *Fransozisches Etymologisches Wörterbuch* (FEW). A ce jour, le nombre des fiches-réponses saisies dépasse largement le million.

Divers types de consultations des données lexicales sont alors proposés, par question ou par localité, afin de constituer des corpus onomasiologiques ou monographiques.

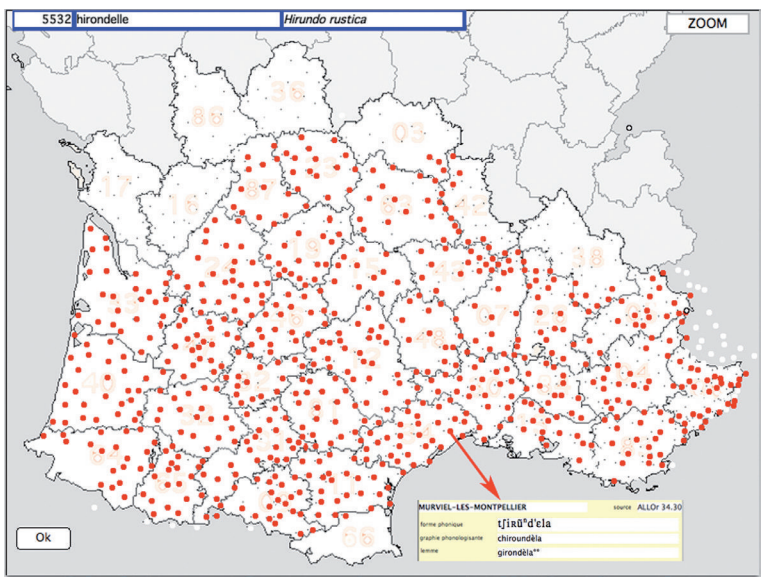
LA TROUSSE À OUTILS

Parallèlement à la consultation des données brutes, la base lexicale du THESOC est dotée de plusieurs fonctionnalités conçues en prévision de divers traitements, notamment pour l'étude de la typologie lexicale, pour l'analyse motivationnelle et pour la reconstruction étymologique.

CARTOGRAPHIE

Outre la consultation de glossaires, le THESOC permet également de cartographier des faits lexicaux, de diverses manières.

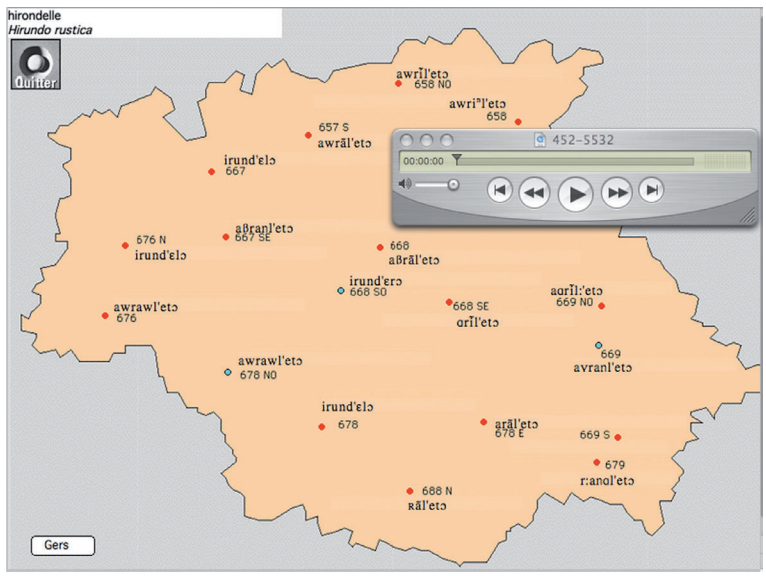
La carte des réponses par question correspond à une carte d'atlas classique, les données pour une question donnée étant réparties dans l'espace. Compte tenu du peu de place disponible pour consigner l'ensemble des termes recueillis sur l'aire considérée, les informations sont contenues dans des fiches auxquelles on accède en cliquant sur le point rouge de la localité que l'on a choisie. Le point, quand il apparaît en rouge, indique qu'une réponse à cette question dans cette localité a été effectivement saisie. Pour chaque forme, les trois niveaux de transcription sont disponibles.



Carte des réponses par question : 5537 HIRONDELLE

Estudis Romànics [Institut d'Estudis Catalans], Vol. 34 (2012), p. 367-387
DOI: 10.2436/20.2500.01.108

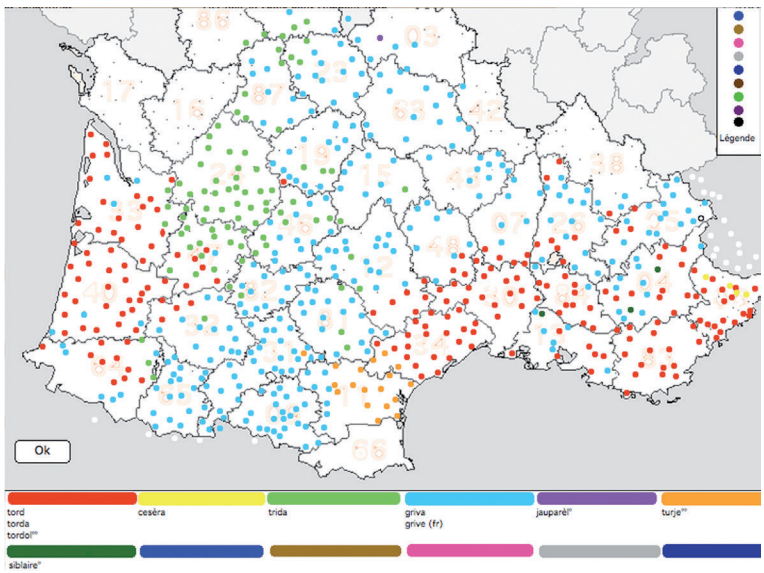
On peut cependant visualiser les termes comme sur une carte d'atlas, mais par département avec un système de «zoom». Lorsqu'un point rouge accompagne la forme phonique, le fichier-son est disponible[1]. On peut alors vérifier la transcription ou comparer directement les différents termes.



Carte des réponses par question et par département :
HIRONDELLE dans le Gers

Un des apports remarquables de l'informatique est la possibilité de produire des cartes traitées, notamment celles qui montrent la répartition des lemmes. A partir d'une entrée donnée du responsaire, on accède à la liste des lemmes répertoriés pour cette question. C'est le linguiste qui effectue alors les choix permettant ensuite au programme d'établir une carte de synthèse montrant les aires de diffusion des différents types lexicaux:

1. On peut également accéder directement aux données sonores de la base, par listes ou par cartes.

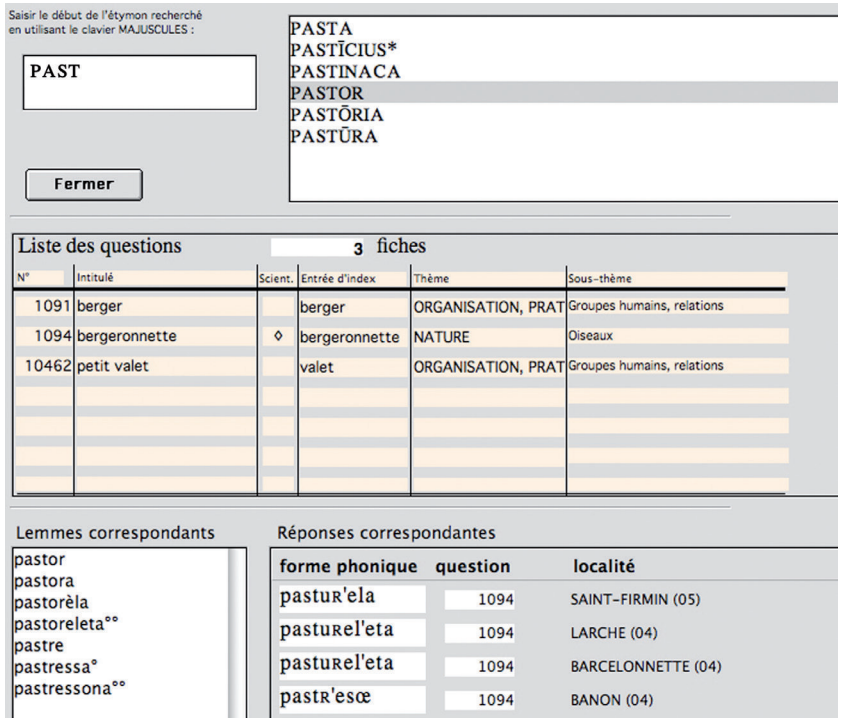


Carte de synthèse : la GRIVE

ÉTYMOLOGIE

Le *module étymologique* procède du désir d'instituer un pont, sous forme d'un simple système de renvois entre le THESOC et les grands dictionnaires étymologiques FEW et REW.

A titre d'exemple, si l'on veut explorer le champ des continuateurs de lat. PASTOR / PASTORE, il suffit de saisir les premières lettres de ce mot: apparaît alors la liste des mots latins commençant par PAST- répertoriés dans le fichier Etymons. Une fois PASTOR sélectionné, on accède aux questions qui ont donné lieu à l'occurrence d'un continuateur de PASTOR, tandis que dans les cadres inférieurs s'affichent les lemmes puis les formes phoniques rapportées respectivement à chacun des sens trouvés et à leur localisation dans l'espace géographique.

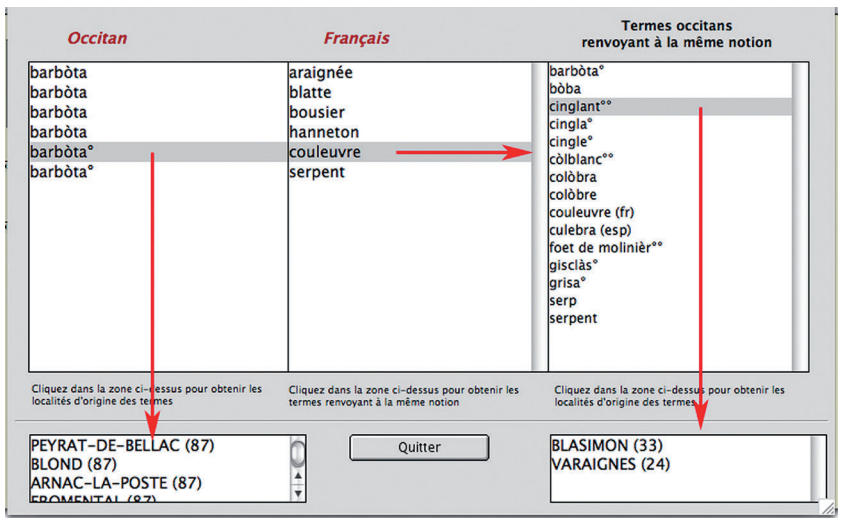


Consultation des continuateurs de l'étymon : PASTOR

Nous n'avions pas, au moment où nous avons décidé de corréler les entrées lexicales du THESOC au FEW et au REW, d'idée précise sur le mode d'utilisation qui pourrait en être fait. Nous étions éblouis à l'idée que la phonétique historique ou, du moins, l'établissement des correspondances métachroniques entre formes latines et formes dialectales modernes deviendrait presque ludique à partir d'une base de données munie de quelques outils. Et nous avions imaginé, pour faciliter les requêtes, d'élaborer et d'inclure dans la base un état «latin tardif aréalisé» (LTA), c'est-à-dire un système abstrait, une sorte de galloroman commun et de le tester. Mais nous n'avions pas explicitement prévu que c'est surtout la sémantique lexicale et tout particulièrement la sémantique lexicale diachronique qui bénéficieraient le plus de ces essais de reconstruction. C'est pourtant le cas.

LEXIQUE INVERSE

Le *module oc-français* se présente comme une ébauche de lexique inverse (occitan-français); la recherche part d'un mot occitan et vise à en faire apparaître, à travers le mot français donné comme équivalent, ce qu'on appelle habituellement le sens. Mais là, autant l'aspect de traduction est familier au lecteur (les dictionnaires bilingues sont monnaie courante), autant ce qu'on peut lire dans le tableau ci-dessous est susceptible d'intriguer. On n'a pas conscience habituellement que la variation dans la relation qui lie le mot au référent puisse prendre une telle ampleur. Que le mot occitan *barbòta* puisse désigner une araignée aussi bien que la blatte ou le bousier semble déjà difficile à admettre. S'agit-il d'imprécision, de confusion? Mais que *barbòta* puisse aller jusqu'à référer à la couleuvre ou au hanneton..., cela dépasse l'entendement. Au point que l'on est prêt à mettre en cause la fiabilité de ces matériaux issus de l'oralité. Pour dissiper ce doute, il reste la possibilité de consulter des dictionnaires de manière à établir si cette pluralité s'y trouve consignée. En l'occurrence, le premier coup d'œil jeté sur le dictionnaire est quelque peu rassurant: il n'y a certes pas unicité de référent de *barbòta* mais la liste est très limitée (*blatte*, *cloporte*); le répit est toutefois de courte durée car figure à la suite un fâcheux *divers insectes;* et pour peu que le regard se promène sur l'article du dictionnaire, il ne tarde pas à croiser un *barbòt* poisson («la loche»). Le témoignage de l'oralité n'est donc nullement démenti.



Consultation du dictionnaire inverse : barbòta

Et donc le problème demeure de cette variation référentielle. Ce n'est certes pas ici le lieu de le traiter mais cela met en évidence le fait qu'une manipulation innocente sur les matériaux de la base de données amène à poser ou re-poser des questions fondamen-

tales en matière de sciences du langage. La situation lexicale créée par un croisement de requêtes dans THESOC oblige le chercheur à constituer un corpus illustrant la question de savoir si les mots seraient autres choses que des étiquettes posées sur des objets. Et, à la lumière des modélisations suggérées par l'interaction des variations diatopiques et diachroniques, à réviser les concepts de sens, *signifié*, *motif*, *emploi*, *arbitraire, référent, étymon*... ainsi que les champs respectifs de la sémantique lexicale, de l'étymologie, des structures lexicales.

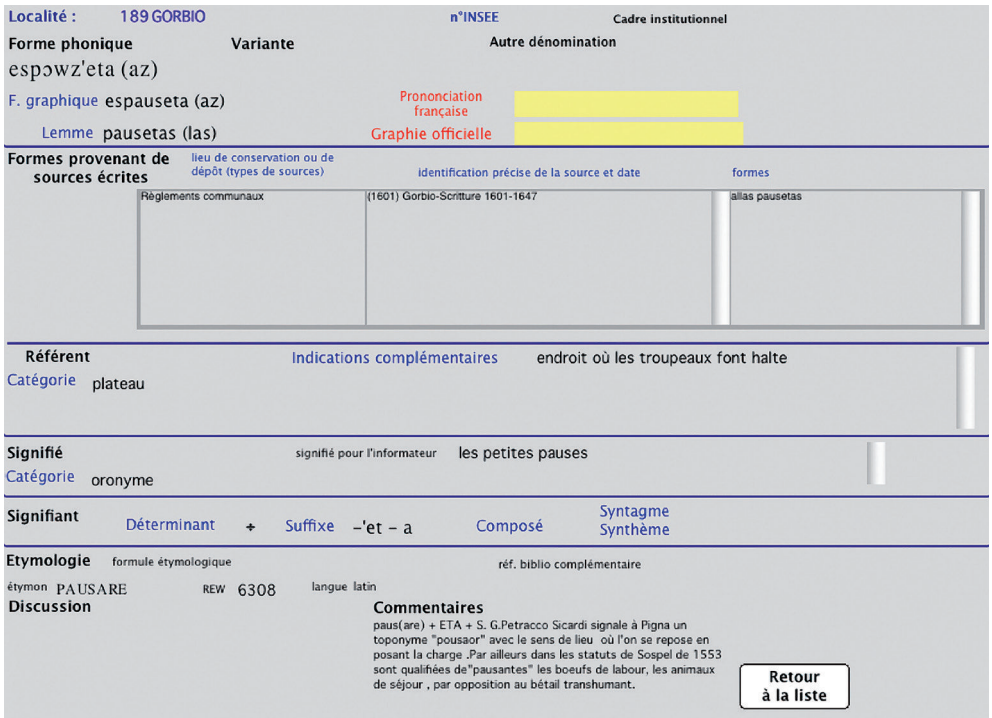
Il est donc patent que la base de données, à côté de ses fonctions de recueil, de stockage, de lissage et d'étiquetage des données est de nature à susciter des hypothèses et des modélisations nouvelles.

LE MODULE DE TOPONYMIE

Ce volet permet de consigner des corpus micro-toponymiques importants. Tous les noms des entités du paysage y ont leur place, qu'ils figurent ou non au cadastre, qu'ils soient anciens ou récents, quelle que soit leur forme (dialectale, française, ou autre), pour autant qu'ils aient fait l'objet d'une collecte orale auprès d'informateurs.

La fiche de saisie se compose de deux parties. La première zone est consacrée à la présentation des faits bruts tels qu'ils ont été recueillis sur le terrain sans qu'aucune correction ni modification ne soient apportée. La forme orale recueillie est transcrite en API, et là encore, sont fournis les deux autres niveaux de transcription: graphie pré-phonologique et lemme. Lorsqu'une forme écrite officielle ou administrative est attestée, elle est également présente, ce qui permet de mesurer l'écart entre les formes orales et les formes écrites actuelles.

La seconde zone du masque de saisie est plus particulièrement affectée à l'analyse. Deux approches sont alors proposées: l'une essaie de cerner le référent actuel auquel s'attache le toponyme en fonction des renseignements donnés par les informateurs, de documents photographiques ou descriptifs ou bien par une visite sur le terrain lorsque c'est possible. L'autre perspective de classement vise à faire ressortir la motivation des toponymes lorsqu'elle est encore visible. En fonction de leur signifié, les toponymes sont classés selon une typologie bien établie et le champ «signifié pour l'informateur» permet de recueillir tout le cotexte qui s'attache au nom de lieu et qui traduit la vision que les informateurs, usagers du paysage, ont de leur espace de vie. C'est le lieu des étymologies populaires et des remotivations qui assurent la pérennité du système toponymique. Enfin, le signifiant est lui aussi analysé selon sa morphologie.



Saisie d'un micro-toponyme

La base de données micro-toponymique du Thesaurus Occitan est donc à la fois un outil de présentation du matériau toponymique qui s'adresse à un public large et varié et un outil d'analyse qui permet d'appliquer une série de filtres à la réalité des faits bruts tels qu'ils ressortent des enquêtes orales de terrain.

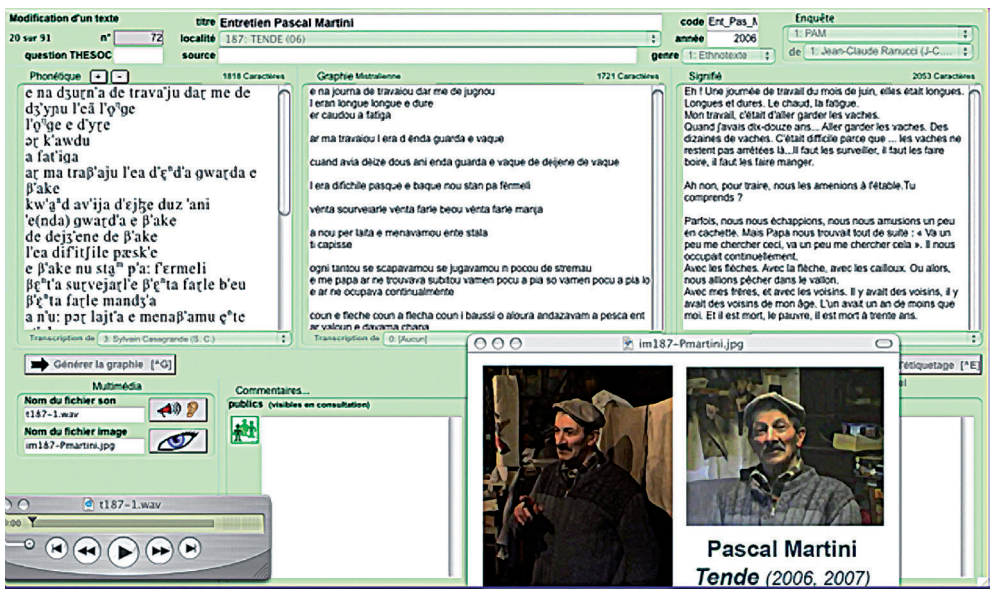
LE MODULE MORPHO-SYNTAXIQUE

Ce module du THESOC (MMS), en cours de développement, a pour objet l'analyse morphologique et syntaxique des dialectes occitans, dans une perspective d'étude de la (micro-)variation, tant synchronique que diachronique. Il est conçu pour traiter des phrases, provenant soit d'ethnotextes recueillis notamment au cours des enquêtes lexicales, soit de réponses à des questionnaires spécifiques, soit de textes qui bien qu'écrits, ont aussi une existence orale (émissions de radio, pièces de théâtre populaire).

Comme dans la base lexicale, les enregistrements sonores (ou vidéo, pour les enquêtes les plus récentes) sont associés aux transcriptions. Les trois niveaux de transcrip-

tion (API, graphie lissée, graphie lemmatisée) sont également disponibles ici, mais dans une perspective un peu différente de celle en vigueur pour les items lexicaux. En effet, l'objectif essentiel étant la syntaxe, l'aspect phonologique revêt moins d'importance.

Ainsi, les textes peuvent être saisis soit en API (c'est le cas des ethnotextes), soit dans une graphie. La base est conçue pour pouvoir gérer toutes les graphies possibles (mistralienne, alibertine, italianisante, etc.); elle comporte à cette fin un dictionnaire associé qui englobe toutes les variations de chaque lemme, qu'elle soit graphique, dialectale ou flexionnelle. Ce dictionnaire est structuré en deux niveaux: (i) le niveau des «variantes» qui enregistre individuellement chaque forme avec sa flexion ainsi que la localité dans laquelle elle est attestée; (ii) le niveau des «lemmes» qui regroupe toutes les variantes correspondant à un même lemme.



Un ethnotexte dans MMS

Chaque phrase est d'abord traitée par le module morphologique qui attribue à chaque occurrence une «étiquette» comprenant le lemme auquel elle est rattachée, sa catégorie morpho-syntaxique et sa flexion. Cette première étape permet notamment de générer des corpus de travail «à la demande», selon les objectifs du chercheur, sur la base des catégories (ex. les pronoms), des lemmes (ex. *non*) ou des variantes (ex. le nissart *nen*) et de les exporter commodément dans un format texte. Le deuxième traitement concerne plus particulièrement la syntaxe, avec un analyseur syntaxique qui propose des représentations arborescentes de la phrase. Plusieurs analyses concurrentes peuvent être conservées dans la base, de manière à laisser toute latitude au chercheur et à ne pas trancher *a priori*. Les recherches peuvent alors également être effectuées sur des configurations

syntaxiques précises (ex. sujet vide), et permettent ainsi de constituer un corpus de travail sur un problème de syntaxe.

UNE BASE TOUJOURS EN DEVENIR

L'énorme avantage de ce type d'outil informatique est qu'il peut évoluer en permanence. Le THESOC offre ainsi de nombreuses fonctionnalités, qui restent cependant à approfondir, perfectionner et multiplier. La base dans son ensemble doit être refondée dans un format plus simple, plus cohérent et plus moderne et le site web est en passe d'être renouvelé. Les projets ne manquent pas et les tâches sont multiples. Ainsi, outre la poursuite des travaux entrepris, notamment pour l'élaboration de MMS et des différents outils du THESOC, un important chantier se profile, qui concerne le développement de la cartographie et son amélioration, tant sur le plan esthétique que sur le plan fonctionnel.

RÉFÉRENCES

DOF: Alibert, L. (1966), *Dictionnaire occitan-français d'après les parlers languedociens*, Toulouse, IEO.

FEW: Wartburg, W. von (1922), *Französisches Etymologisches Wörterbuch*, Basel, Zbinden Druck und Verlag AG.

REW: Meyer-Lübke, W. (1953), *Romanisches Etymologisches Wörterbuch*, Heidelberg, Winter.

TDF: Mistral, F. (1878), *Lou Tresor dóu Felibrige ou Dictionnaire Provençal-Français*, Aix-en-Provence.

Brun-Trigaud, G. (2010), «Le Thesaurus Occitan: une base de donnée multimedia consacrée aux dialectes occitans», *Tools for Linguistic Variation (EUDIA-2)*, J.L. Ormaetxea et G. Aurrekoetxea (éds), Bilbao, ASJU-ren gehigarriak, LIII, UPV-EHU, 91-108.

Brun-Trigaud, G. (2010), «Thesaurus occitan: traitement et valorisation des données orales», *Projet Biens culturels africains. L'IFAN face à la virtualisation de son patrimoine. Actes du Colloque international. Université de Toulouse II-Le Mirail. 26-28 nov. 2007*, Y. Sylla et D.H. Zidouemba (éds), Dakar, IFAN/Université Ch. A. Diop, 189-197.

Brun-Trigaud, G. (à paraître), «Valorisations et exploitations des enquêtes dialectales: présentation du Thesaurus Occitan, une base de données multimédiale», *Langues, communautés et territoires en France aujourd'hui. Recherches et enquêtes en ethnologie et en linguistique*, J.-R. Trochet (éd), Paris, CTHS.

Brun-Trigaud, G. et Molinu L. (2004), «Présentation du Logiciel Multimédial Thesaurus Occitan (Thesoc)», *Terres et hommes du Sud. 126ème Congrès des sociétés historiques et scientifiques (CTHS)*, G. Hasenohr (éd), Paris, CTHS, 199-207.

Estudis Romànics [Institut d'Estudis Catalans], Vol. 34 (2012), p. 367-387
DOI: 10.2436/20.2500.01.108

Brun-Trigaud, G. et Sauzet, P. (à paraître), «Des atlas au dictionnaire: le Thesaurus Occitan». *Cahiers Corpus*.

Dalbera, J.-Ph. (1990), «Bases de données et recherches dialectologiques», *Bulletin de l'AIEO* 8, London, 35-40.

Dalbera, J.-Ph. (1996), «La base de données dialectales du Thesaurus Occitan», *Bollettino dell'Atlante Linguistico Italiano*, III[ème] série, Torino, Istituto dell'ALI, 187-202.

Dalbera, J.-Ph. (1996), «Strates et représentations dans une base de données dialectales», Actes du Colloque international *«Bases de données linguistiques: conceptions, réalisations, exploitations»,* G. Moracchini (éd.), Corte, Université de Corse, 103-116.

Dalbera, J.-Ph. (1998), «La base de données THESOC. Etat des travaux», *Toulouse à la croisée des cultures*, J. Gourc et F. Pic (éds), Pau, AIEO, 403-417.

Dalbera, J.-Ph. (2006), *Des dialectes au langage: Une archéologie du sens*, Paris, Champion.

Dalbera, J.-Ph. (2009), «Quel avenir pour la dialectologie?», Actes du colloque *La dialectologie hier et aujourd'hui (1906-2006)*, B. Horiot (éd.), Lyon, Centre d'études linguistiques Jacques Goudet, 455-468.

Dalbera, J.-Ph. et Dalbera-Stefanaggi, M.-J. (1993), «Bases de données dialectales et diasystèmes», *Verhandlungen des Internationalen dialekto-logenkongresses*, Stuttgart, Franz Steiner Verlag, 286-301.

Georges, P.-A. (2009), «Présentation de la base *Textes* associée au THESOC», Actes du colloque *La dialectologie hier et aujourd'hui (1906-2006)*, B. Horiot (éd.), Lyon, Centre d'études linguistiques Jacques Goudet, 81-93.

Georges, P.-A. (2010), «The Thesaurus Occitan: a multimedia database dedicated to occitan dialects. Presentation of its morphosyntax module», *Tools for Linguistic Variation (EUDIA-2)*, J.L. Ormaetxea et G. Aurrekoetxea (éds), Bilbao, ASJU-ren gehigarriak, LIII, UPV-EHU, 107-118.

Oliviéri, M. (2001), «Le mot et la chose: réflexion sur le *responsaire* du THESOC», *Cahiers Corpus*, à paraître.

Oliviéri, M. (2004), «Le responsaire du THESOC», Actes du colloque *8ème Colloque de dialectologie et littérature du domaine d'oïl occidental*, Avignon, 12-13 juin 2002, P. Brasseur (éd.), 23-34.

Oliviéri, M. et Brun-Trigaud, G. (2009), «Présentation du logiciel Thesaurus Occitan», Actes du colloque *La dialectologie hier et aujourd'hui (1906-2006)*, B. Horiot (éd.), Lyon, Centre d'études linguistiques Jacques Goudet, 61-80.

Oliviéri, M. (à paraître), «Le Thesaurus Occitan: une base de données dialectales multimédias», Actes du *Curso de verano de la Fundació Germà Colón*, L. Gimeno (éd.), Castelló de la Plana, Universitat Jaume I.

Ranucci, J.-C. (2004), «Microtoponymie et bases de données; méthodes et problèmes: l'exemple de la base de données microtoponymique du Thesaurus Occitan», *Etudes corses*, vol. 59, 65-76.

Ranucci, J.-C. (2009), «Microtoponymie et dialectologie: le terrain en partage», Actes du colloque *La dialectologie hier et aujourd'hui (1906-2006)*, B. Horiot (éd.), Lyon, Centre d'études linguistiques Jacques Goudet, 325-336.