La Bretagne Linguistique

18 | 2014
Varia

Le Thesaurus Occitan, une base de données multimédiale dédiée aux dialectes occitans

Thesaurus Occitan: a multimedia database of Occitan dialects

Guylaine Brun-Trigaud

OpenEdition Journals

Édition électronique
URL : https://journals.openedition.org/lbl/1008
DOI : 10.4000/lbl.1008
ISSN : 2727-9383

Éditeur
Université de Bretagne Occidentale – UBO

Édition imprimée
Date de publication : 1 mars 2014
Pagination : 57-72
ISBN : 979-10-92331-07-3
ISSN : 1270-2412

Référence électronique
Guylaine Brun-Trigaud, « Le Thesaurus Occitan, une base de données multimédiale dédiée aux dialectes occitans », *La Bretagne Linguistique* [En ligne], 18 | 2014, mis en ligne le 01 mai 2021, consulté le 22 mai 2021. URL : http://journals.openedition.org/lbl/1008 ; DOI : https://doi.org/10.4000/lbl.1008

Guylaine BRUN-TRIGAUD*

Le Thesaurus Occitan, une base de données multimédiale dédiée aux dialectes occitans

Le THESAURUS OCCITAN (ou THESOC en abrégé) est une base de données informatique destinée à l'étude des dialectes occitans.

Elle est développée depuis 1992 et centralisée à Nice dans le cadre de l'UMR 7320 du CNRS « Bases, Corpus, Langage », sous la direction de Jean-Philippe Dalbera, il s'agit d'un programme inter-universitaire.

Le THESOC contient notamment :

– des données linguistiques et péri-linguistiques issues d'enquêtes de terrain, donc des données provenant des cartes et des carnets d'enquêtes des Atlas linguistiques régionaux et des monographies,

– des enregistrements sonores,

– des documents iconographiques,

– des données linguistiques procédant d'analyses déjà réalisées (lemmatisations, morphologie, étymologie, microtoponymie),

– des données bibliographiques,

– des outils d'analyse (représentations cartographiques, instru-

* Laboratoire Bases Corpus Langage, CNRS UMR 7320, Université Nice-Sophia-Antipolis, gbrun@unice.fr.

ments d'analyse diachronique, procédures de cartographie comparative, instruments d'analyse morphologique).

Il s'agit d'un objet à géométrie variable envisageant différents types d'exploitation grâce à des menus spécifiques qui intègrent toutes sortes de documents.

Le THESOC se présente comme un outil offrant à la fois, mais toujours séparément, des données linguistiques quasi brutes, des données ayant fait l'objet d'analyses et de traitements, ainsi que des outils d'investigation.

L'intérêt d'un tel outil réside également dans le fait qu'il est en permanence amené à évoluer selon les besoins des utilisateurs.

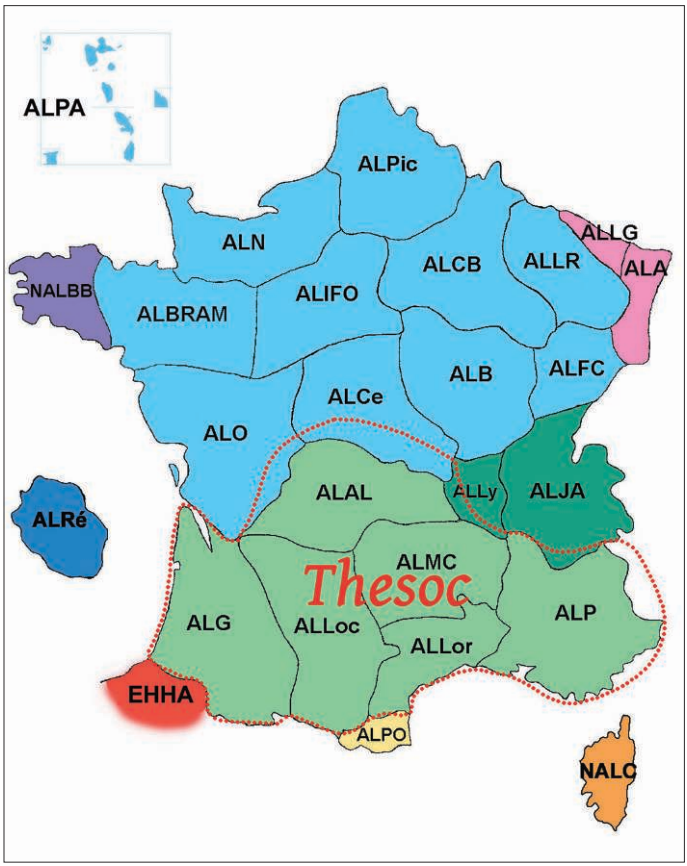

Fig. 1 : Domaine couvert par le Thesaurus Occitan.

Pour pouvoir figurer dans la base, les données linguistiques brutes doivent satisfaire les deux critères suivants : d'une part, elles doivent être de nature orale (elles sont saisies dans la base avec leur transcription phonétique en Alphabet Phonétique International) et d'autre part, les données doivent être précisément localisées.

Actuellement la base complète comprend : 1 212 000 fiches réponse, 1 100 extraits sonores et plus de 500 documents visuels (photos et dessins), ce qui fait du THESOC un véritable recueil multimédia, qui peut s'inscrire tout autant comme outil de recherche pour les linguistes que comme outil pédagogique pour le grand public.

La base de données se trouve en consultation sur Internet à l'adresse : http://thesaurus.unice.fr

Présentation de la base lexicale

La partie lexicale constitue le « cœur historique » du THESOC, il est développé depuis maintenant plus d'une quinzaine d'années.

À l'intérieur de la base de données, chaque forme est identifiée par le couple LOCALITÉ-QUESTION, dont voici les deux fichiers principaux.

Le premier est le fichier des localités : il contient 845 entrées recouvrant tout le domaine occitan, y compris bientôt les données occitanes de l'Atlas Linguistique du Piémont oriental en Italie.

La consultation d'une fiche localité permet d'avoir accès à la liste des enquêteurs et des informateurs associés aux différentes enquêtes qui se sont succédé dans cette localité (fig. 2).

Le second est le fichier des questions que nous avons renommé « responsaire » : il est le résultat de la somme des cartes et listes publiées par les différents atlas linguistiques régionaux du domaine occitan, des éléments relevés dans les monographies et même des résultats d'enquêtes non publiés (par exemple les données inédites des atlas linguistiques régionaux).

Il comporte 8 200 *questions*, qui sont regroupées suivant les principaux thèmes traités dans les atlas linguistiques régionaux, comme l'élevage, la nature, l'espace, le temps, l'habitat et la vie quotidienne, etc.

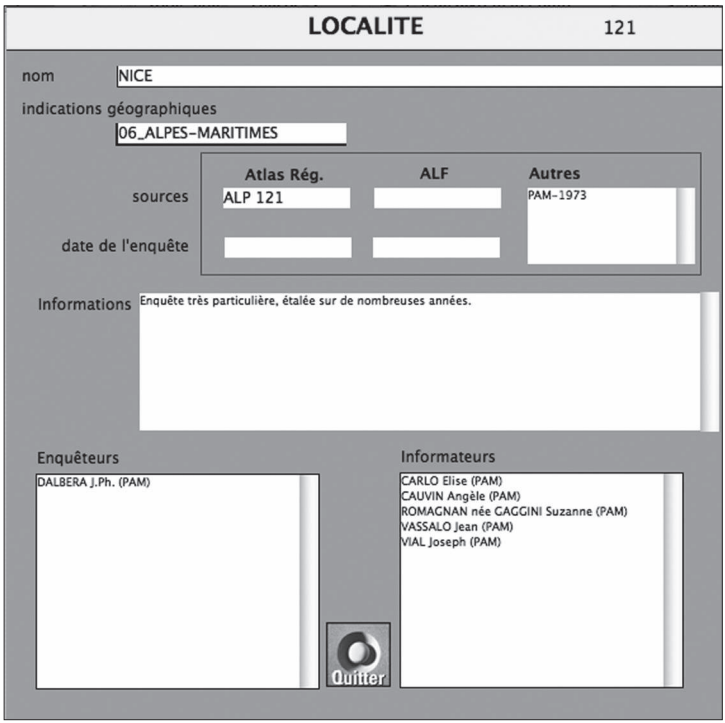

Fig. 2 : Fiche-localité n° 121 : Nice (Alpes-Maritimes).

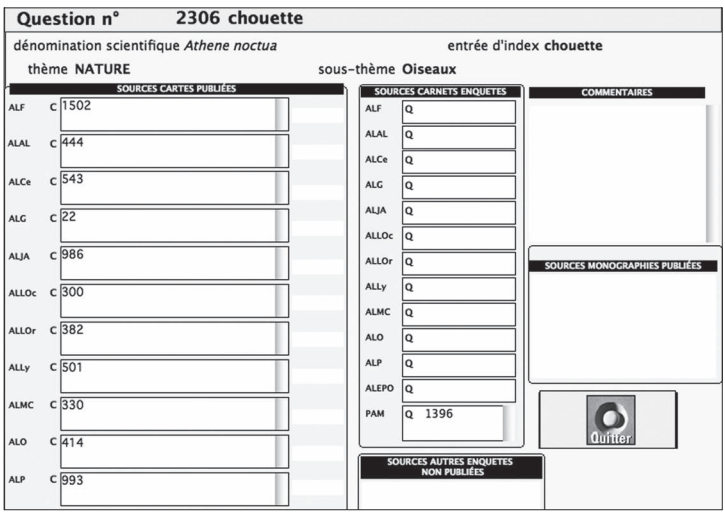

Fig. 3 : Fiche-question n° 2306 : chouette.

La consultation d'une fiche question permet d'avoir accès à la liste des cartes publiées et des carnets d'enquêtes, concernant cette question. Le cas échéant, d'autres sources éventuelles peuvent également y être consignées (fig. 3).

Les différentes entrées lexicales de la base sont organisées de la manière suivante : chaque entrée lexicale, ou fiche *réponse*, est associée à une fiche *question* ET à une *localité* donnée.

Consultation de la base lexicale

Il existe différentes possibilités d'interrogation de la base pour consulter les données lexicales : on peut, d'une part, rechercher toutes les fiches *réponses* associées à une *question* précise, et ainsi, on peut visualiser sous forme de tableau toutes les réalisations lexicales d'un terme attesté dans les différentes localités de la base. Certaines fiches sont associées à des illustrations : par exemple "panier" (fig. 4)**.**

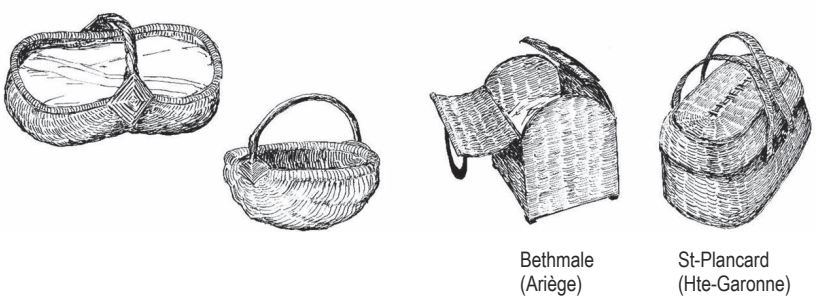

Fig. 4 : Illustrations : les paniers (1/10e).

Ces dessins proviennent de *l'Atlas de la Gascogne* et ont été réalisés par Jacques Allières.

On peut également rechercher toutes les fiches-réponses associées à une localité donnée, pour établir une monographie.

On peut enfin rechercher spécifiquement un certain couple *localité* / *question*. Ce qui permet de consulter en détail une fiche-réponse.

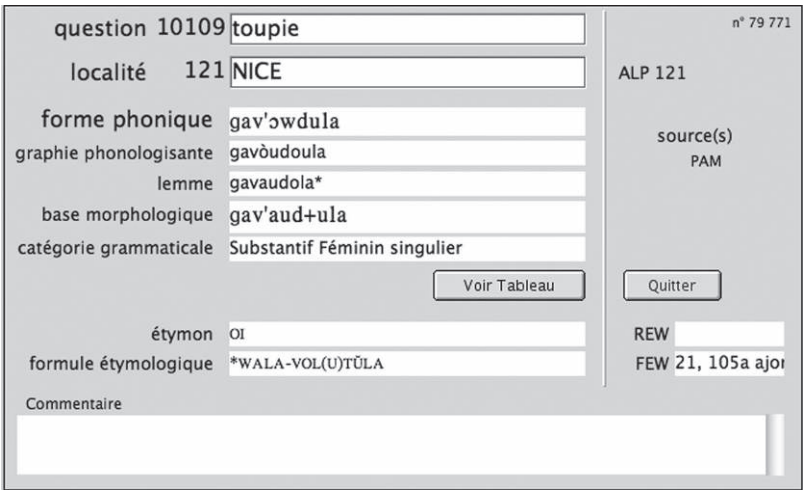

Fig. 5 : Fiche-réponse Nice / toupie.

Le détail d'une fiche réponse (fig. 5) fait apparaître un certain nombre de champs :

– question : comporte le numéro correspondant à son identification dans la base, complété de son intitulé,

– localité : comporte le numéro correspondant à son identification dans la base, complété de son nom officiel et de sa référence dans l'atlas linguistique,

– forme phonique : correspond à la transcription phonétique de la forme relevée en Alphabet Phonétique International (par souci d'harmonisation, nous avons transposé en API toutes les données des atlas régionaux qui étaient initialement en alphabet Rousselot),

– graphie phonologisante : il s'agit d'une transcription graphique dite « phonologisante », c'est-à-dire une sorte de forme intermédiaire entre la transcription phonétique et la graphie standardisante d'Alibert. Cette forme intermédiaire adopte pour l'essentiel les principes du *Tresor dóu Felibrige* de F. Mistral[1]. Elle peut être automatiquement générée par un algorithme, que nous verrons plus loin,

– lemme : conçu comme forme de référence ou de convention, comme dans les dictionnaires, il sous-tend tout le faisceau de variantes phonétiques consignées dans la base. Le choix du lemme

1. Frédéric Mistral, *Tresor dou Felibrige*, Paris, Raphèle-lès-Arles, 1878.

est effectué en s'appuyant sur le *Dictionnaire occitan-français* de Louis Alibert et sa notation respecte donc les principes de la graphie alibertine[2],

– base morphologique : reconstitue la base morphologique à l'aide de la racine et des suffixes et/ou préfixes,

– catégorie grammaticale : un clic sur le bouton « Voir Tableau » permet d'accéder aux différentes variations sur le genre et le nombre quand elles ont été relevées,

– étymologie : renseigne sur l'étymon du terme (ici OI signifie Origine Inconnue dans le FEW[3]), avec renvoi au REW[4] et au FEW,

– formule étymologique : contient ici une proposition de Jean-Philippe Dalbera pour le terme *gavaudola*,

– commentaire : reprend éventuellement les commentaires trouvés en marge des atlas afférents à cette réponse.

Dans le cas des fiches contenant un verbe, en cliquant sur le bouton « Voir Tableau », on peut également consulter le paradigme de conjugaison verbale, lorsque celui-ci a été renseigné dans la base (fig. 6).

Cartographie

La consultation de la base ne permet pas seulement d'obtenir des glossaires, comme nous venons de le voir, elle permet également de cartographier des faits lexicaux de différentes manières.

Deux types de cartes sont disponibles dans le Thesoc : d'une part, des cartes présentant les faits bruts, et d'autre part, des cartes de synthèse.

Aucune carte n'est cependant stockée dans la base de données : elles sont toutes générées dynamiquement à partir des données linguistiques présentes dans la base, en fonction des requêtes demandées par l'utilisateur.

---

2. Louis Alibert, *Dictionnaire occitan-français*, Toulouse, Institut d'Études Occitanes, 1966.
3. FEW = Walther von Wartburg, *Französisches Etymologisches Wörterbuch*, Bonn, Schroeder, 1922-
4. REW = Wilhelm Meyer-Lübke, *Romanisches etymologisches Wörterbuch*, Heidelberg, Winter, 1935.

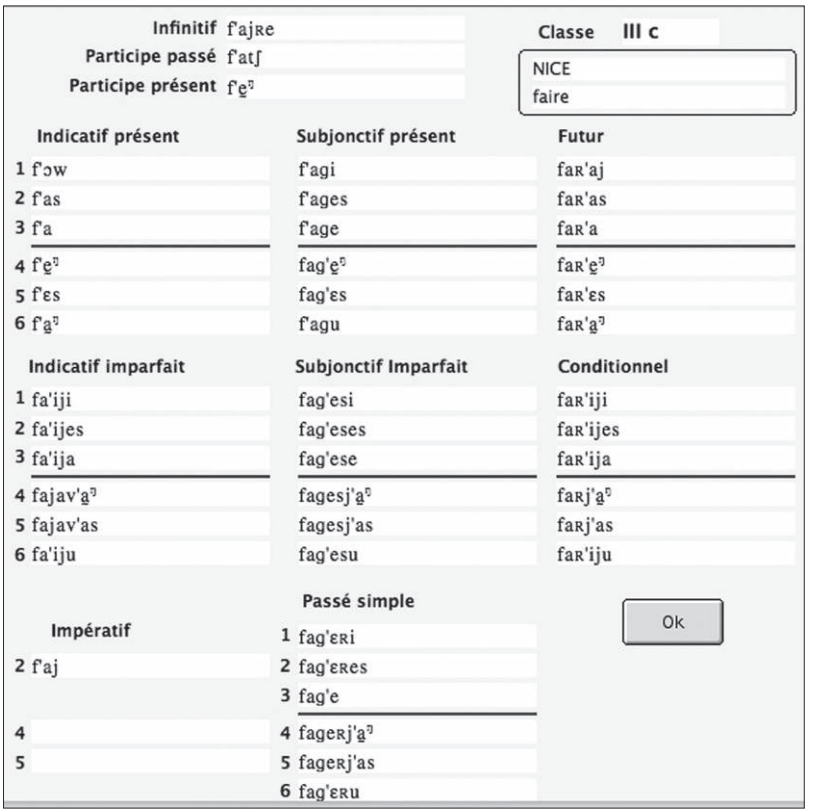

Fig. 6 : Fiche-réponse-morpho : Nice / faire.

Chaque fois que l'utilisateur modifie sa requête, une nouvelle carte est générée en temps réel. C'est en ce sens que l'on peut dire qu'il s'agit d'une cartographie interactive.

Comme il n'était pas possible d'afficher sur une carte de l'Occitanie tout entière l'ensemble des transcriptions phonétiques attestées dans les 845 localités, les cartes présentant les faits bruts sont disponibles à deux échelles, avec un système de zoom :

1) au niveau de l'Occitanie tout entière, un simple point rouge signale les localités pour lesquelles la base contient une réponse à la question qui est cartographiée. Un clic sur l'un des points permet de visualiser la ou les réponses transcrites en phonétique dans une petite fenêtre (fig. 7).

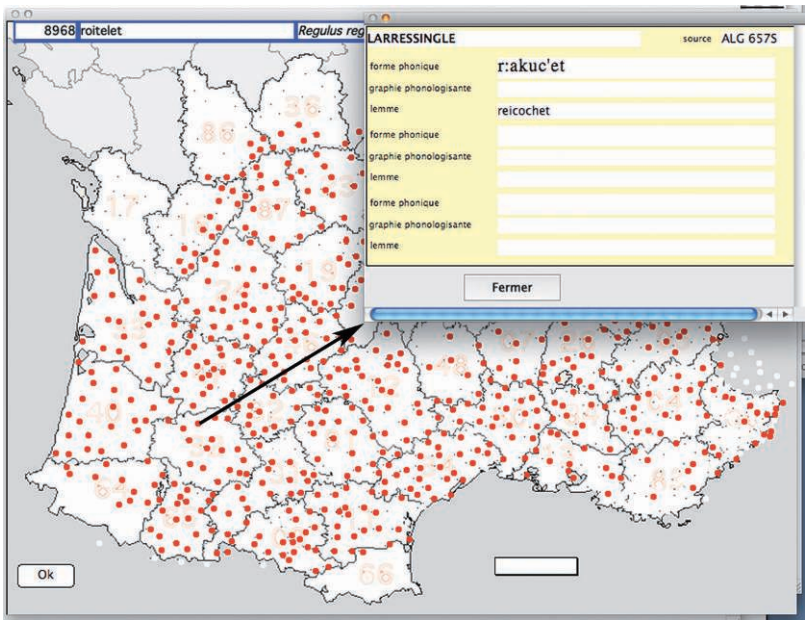

Fig. 7 : Carte à symboles (roitelet) et mini-fiche (Larressingle, Gers).

2) Il est donc également possible de zoomer à l'échelle d'un département. La carte détaillée affiche la transcription phonétique associée à chaque réponse, à côté du point de la localité concernée, comme sur les cartes des atlas linguistiques. Un point rouge indique que l'on peut écouter l'enregistrement sonore associé (fig. 8).

Voyons à présent un exemple de carte de synthèse, concernant la répartition géographique des différents types lexicaux d'une notion sur le domaine occitan :

Le logiciel affiche à l'écran la liste des lemmes répertoriés dans les différentes fiches réponses de la base concernant ce terme. On peut alors effectuer des groupes contenant un ou plusieurs de ces lemmes, selon nos souhaits, et affecter une couleur à chacun de ces groupes (fig. 9 et 10).

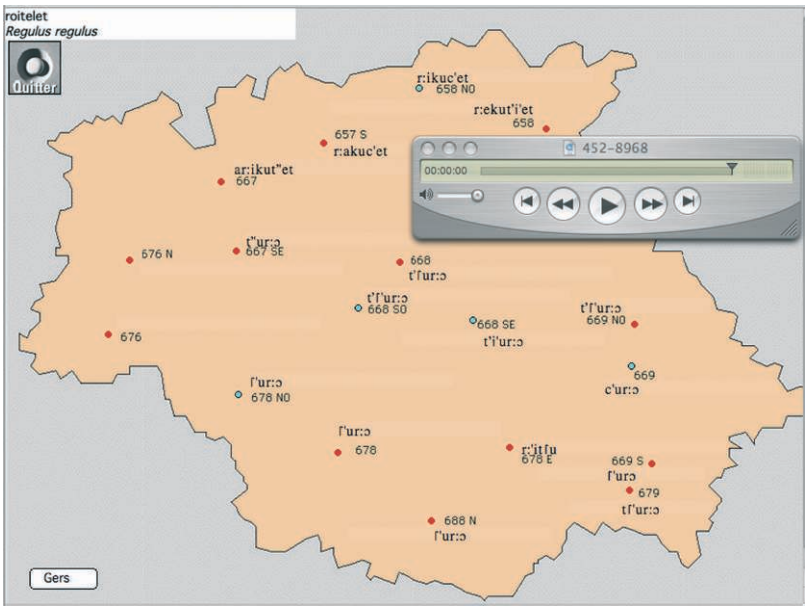

Fig. 8 : Carte des réponses : roitelet (Gers) et enregistrement sonore associé.

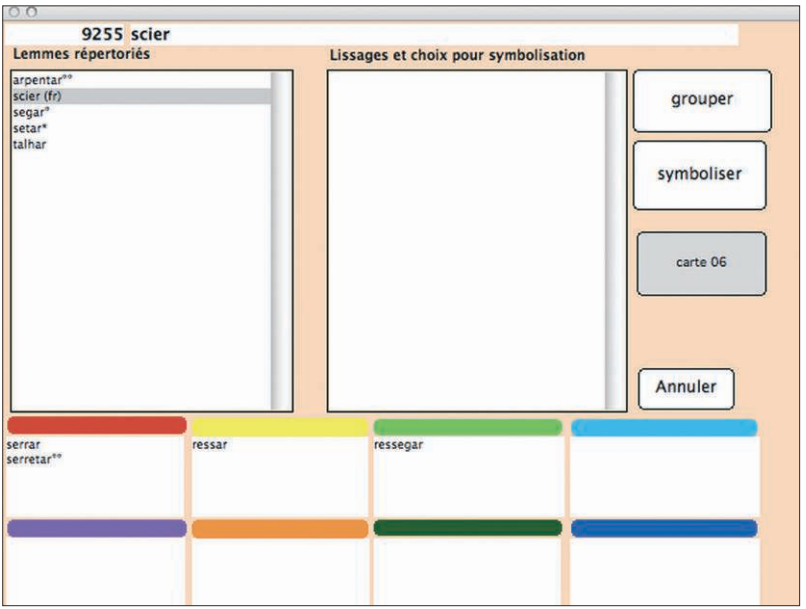

Fig. 9 : Tableau de regroupement des lemmes (scier).

L'utilisateur peut modifier les critères et les regroupements à volonté pour générer autant de cartes qu'il le souhaite.

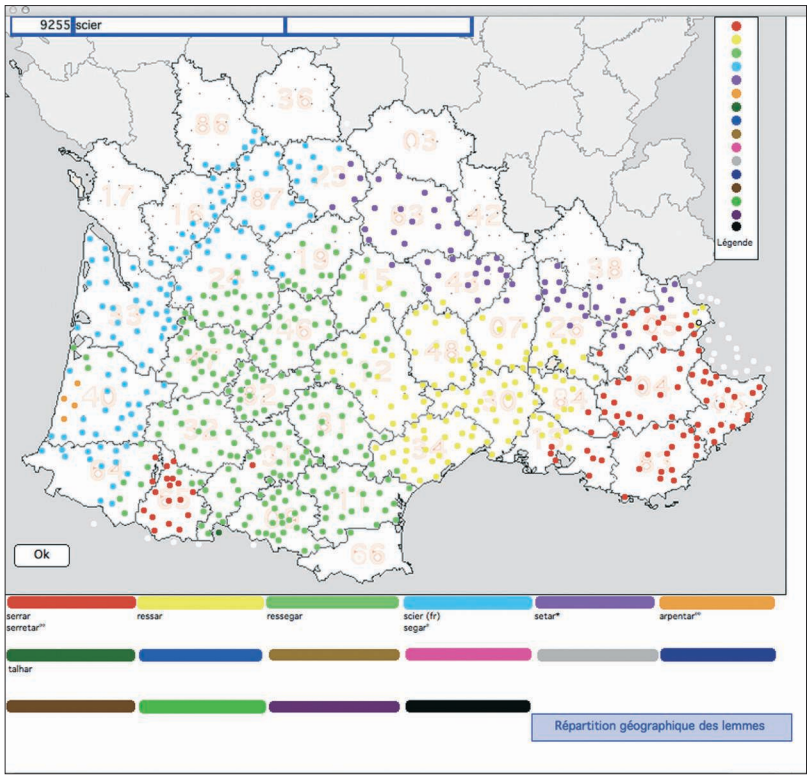

Fig. 10 : Carte de répartition des lemmes (scier).

Autres ressources du Thesoc

Le THESOC ne se limite pas aux données purement lexicales mais il permet également d'autres recherches, à divers niveaux : notre "couteau suisse" contient aussi quelques outils bien utiles...

1) Un dictionnaire inversé oc-français :

À partir des données lexicales présentes dans la base, le THESOC propose également un dictionnaire inversé occitan / français, sur la base des lemmes présents dans la base de données.

Il permet ainsi de trouver les différentes notions correspondant à un terme occitan donné. Ce type de requête est particulièrement utile pour des recherches en sémantique lexicale et en reconstruction étymologique.

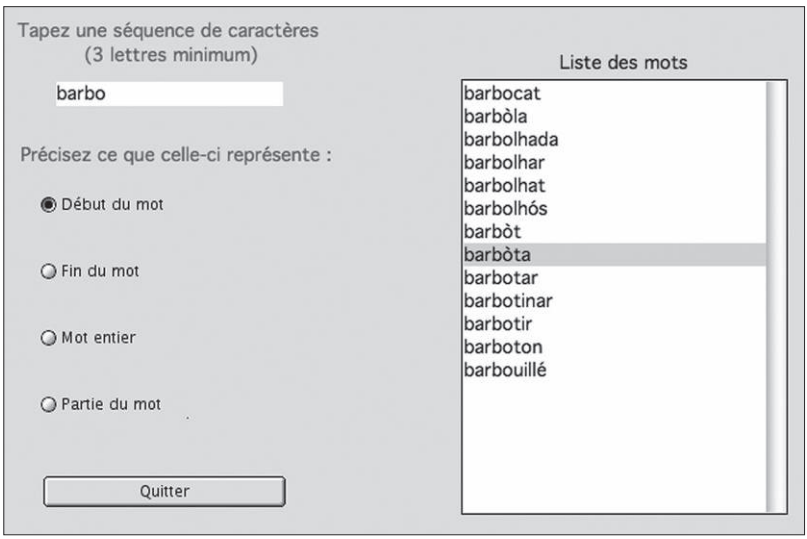

Fig. 11 : Dictionnaire inversé oc-français : fenêtre de recherches.

Si on choisit, *barbo-* dans le cadre en haut à gauche et que l'on utilise "Début de mot" comme critère de sélection (fig. 11), alors, dans le cadre de droite, s'affichent tous les termes commençant par cette séquence. On peut dès lors sélectionner un mot (*barbòta*) et afficher dans la fenêtre suivante, les différents sens relevés parmi les données saisies dans les atlas linguistiques régionaux (fig. 12).

Dans la colonne de gauche, lorsque le lemme en question n'a pas de signe, c'est que le sens existe dans le dictionnaire d'Alibert qui est notre dictionnaire de référence, ici, "araignée", "blatte", "bousier" et "hanneton" avaient été relevés. Mais lorsqu'il est accompagné d'un signe "°", c'est que le sens n'avait pas été consigné et on remarque, alors, que les sens de "couleuvre" et "serpent" ne figuraient pas.

En cliquant sur l'une des lignes de la colonne de gauche, toutes les localités qui utilisent ce terme pour désigner le référent correspondant apparaissent dans la fenêtre inférieure. Par exemple, *barbòta* au sens de "couleuvre" a été essentiellement recueilli en Limousin.

Tandis qu'en sélectionnant un signifié (le mot français) dans la colonne centrale, on accède à la liste de tous les autres termes dialectaux (les signifiants) renvoyant à la même notion (*bòba*, *cinglant*, *cingla*, etc.).

Enfin, chacune des formes est localisée. Un clic sur l'un des termes dialectaux fait apparaître la liste des localités dans lesquelles ce terme est attesté (Blasimon et Varaignes).

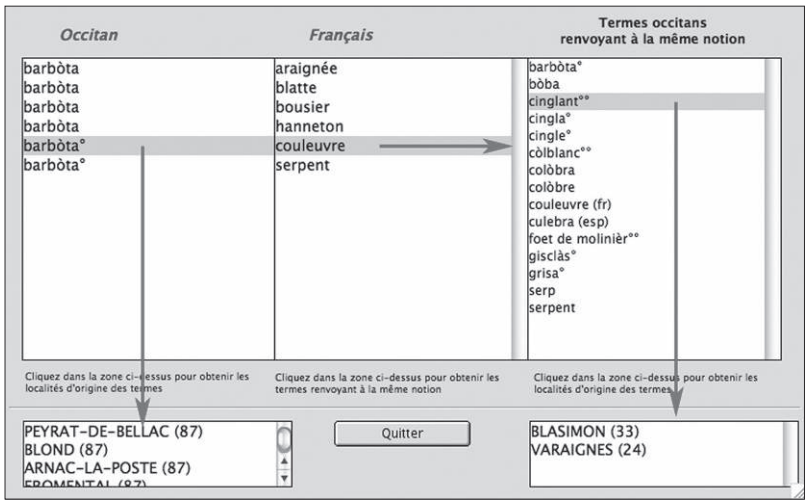

Fig. 12 : Dictionnaire inversé oc-français : fenêtre des résultats trouvés.

2) Le transcripteur automatique

Afin de pouvoir générer des cartes de synthèse, comme celle que nous venons de présenter ci-dessus, et de pouvoir effectuer différents types d'analyses et de recherches dans la base, les réponses lexicales font notamment l'objet d'une procédure de lemmatisation.

Un certain nombre d'outils informatiques permettent de faciliter le traitement des données.

Ainsi, un transcripteur permet de générer automatiquement une graphie phonologisante à partir de la transcription phonétique.

Celui-ci est basé sur un ensemble de règles de réécriture qui peuvent être configurées par l'utilisateur et qui peuvent varier d'une localité à l'autre pour prendre en compte les systèmes phonologiques des différents dialectes occitans.

Ainsi [tʃaʀlˈẽⁿkɔ] est bien transcrit *charlenco* ou [kɥˈɛtʃ], *cuèch*.

3) Recherches étymologiques

Parmi les autres fonctionnalités de recherche de la base, il faut noter qu'il est possible d'effectuer une recherche par étymons (fig. 13) :

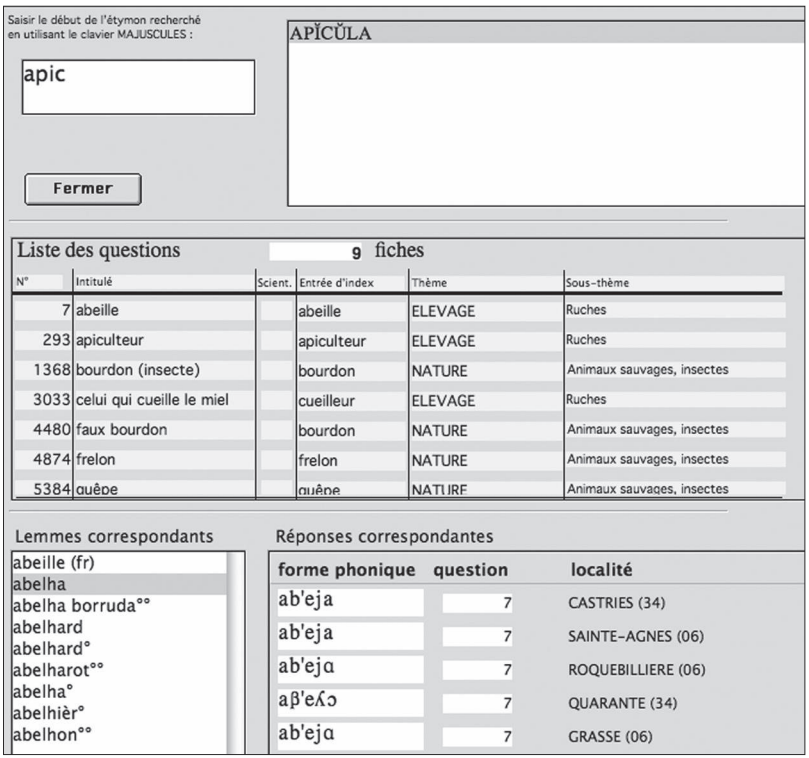

Fig. 13 : Recherches étymologiques (APICULA).

Si l'on saisit par exemple l'étymon latin *APĬCŬLA*, apparaît alors à l'écran la liste des questions en rapport avec cet étymon[5] (ex. "abeille", "apiculteur", "bourdon (insecte)", etc.), la liste des lemmes correspondants associés (ex. *abeille (fr)*, *abelha*, etc.), ainsi que la liste des réponses correspondantes en phonétique pour chaque lemme avec leur localisation (ex. [ab'eja] à Castries (Hérault)).

D'autres outils sont également à disposition :

4) On peut consulter à part les enregistrements sonores, soit à partir d'une carte comme nous l'avons vu précédemment ou soit à partir d'une liste de localités.

5) De même pour les paradigmes morphologiques, ces derniers peuvent être consultés soit à partir d'un fiche-réponse soit à partir d'un module spécifique "Morphologie nominale et verbale".

Nous avons aussi, en cours de développement un Module Morpho-Syntaxique, tout particulièrement dédié à l'analyse syntaxique et morpho-syntaxique des dialectes occitans. Il contient à la fois un ensemble de textes oraux (notamment des ethnotextes) et de phrases isolées telles que les réponses à des questionnaires d'enquêtes. La localisation de ces textes et phrases permet d'envisager à terme une comparaison des dialectes sur le plan syntaxique.

Il comporte un étiqueteur et un analyseur syntaxique capables de gérer la variation linguistique dans toutes ses dimensions.

6) La carte des atlas permet de visualiser les domaines des différents atlas linguistiques dont les données figurent dans la base.

7) Des bibliographies donnent les références des ouvrages concernés par les travaux du THESOC.

8) Le THESOC comporte aussi un volet de toponymie qui consigne les différents micro-toponymes recueillis lors des enquêtes.

---

5. Il s'agit en fait de la liste de toutes questions pour lesquelles la base contient au moins une fiche réponse qui possède l'étymon demandé.

Conclusion

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 (lemmatisation, recherches étymologiques, etc.), ainsi que l'élaboration du Module Morphosyntaxique 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.