Pierre-Aurélien Georges, Université Nice Sophia-Antipolis, CNRS[1]

Le THESOC : base de données et outils d'analyse consacrés à l'étude des dialectes occitans

Le THESOC – abréviation de THESAURUS OCCITAN – est une base de données informatique, compatible Mac et PC, destinée à l'étude des dialectes occitans. Elle est gérée au laboratoire CNRS Bases Corpus Langage (UMR 6039 Nice) sous la direction de Jean-Philippe Dalbera. Les faits linguistiques présents dans cette base de données correspondent aux deux critères suivants :

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

En outre, plus d'un millier d'enregistrements sonores et plus de 500 documents visuels (photos, vidéos, dessins) accompagnent et illustrent actuellement ces données linguistiques, 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.

Ce « tour d'horizon » présente les principales composantes du THESOC – à savoir la partie lexicale, le volet de microtoponymie, ainsi qu'un module plus spécifiquement conçu pour la syntaxe et la morphosyntaxe – sous différents aspects : données, traitements, et outils d'analyse seront présentés.

La base lexicale

La partie lexicale constitue le « cœur historique » du THESOC, développé depuis maintenant plus d'une quinzaine d'années. Actuellement, la base contient plus d'un million d'entrées lexicales, dont une partie est consultable en ligne sur Internet (Dalbera: 1992-).

Ces entrées sont réparties sur 831 localités recouvrant tout le domaine occitan. Comme la base contient notamment les réponses aux questionnaires des différents Atlas Linguistiques de la France par régions[2] qui couvrent la zone des dialectes occitans, une grande partie de ces 831 localités correspond aux points d'enquête de ces atlas. Chaque fiche localité de la base dispose donc de plusieurs champs qui permettent de noter, le cas échéant, le numéro associé à ce point d'enquête dans les Atlas régionaux et/ou dans l'ALF, ainsi que dans d'autres sources éventuelles, comme c'est le cas par exemple pour

[1] Laboratoire BCL, Université Nice Sophia-Antipolis, CNRS UMR 6039; MSH de Nice, 98 bd E. Herriot, 06200 NICE

[2] Pour une présentation de cette collection d'Atlas, *Cf.* (Séguy: 1973).

les Alpes Maritimes avec la série d'enquêtes *Parler des Alpes Maritimes (PAM)* réalisée sous la direction de Jean-Philippe Dalbera (cf. Dalbera: 1994). La consultation d'une fiche localité permet également d'avoir accès à la liste des enquêteurs et des informateurs associés aux différentes enquêtes qui se sont succédé dans cette localité.

Par souci d'organisation, les différentes entrées lexicales de la base sont regroupées de la manière suivante : chaque entrée lexicale, ou *réponse*, est associée à une *question*. Ainsi, si l'on veut visualiser par exemple toutes les réalisations phonétiques du terme « hirondelle » attestées dans les différentes localités de la base, on pourra facilement rechercher toutes les réponses associées à la question « hirondelle ». Le fichier des questions ou « responsaire »[3] contient à la fois les cartes et listes publiées par les différents atlas régionaux du domaine occitan, ainsi que des éléments relevés dans des monographies ou des résultats d'enquêtes non publiées. Il comporte à l'heure actuelle 8 338 questions. Ces questions sont à leur tour regroupées suivant les principaux thèmes traités dans les atlas, comme les cultures, l'élevage, la nature, l'espace, le temps, l'habitat et la vie quotidienne, ou l'homme.

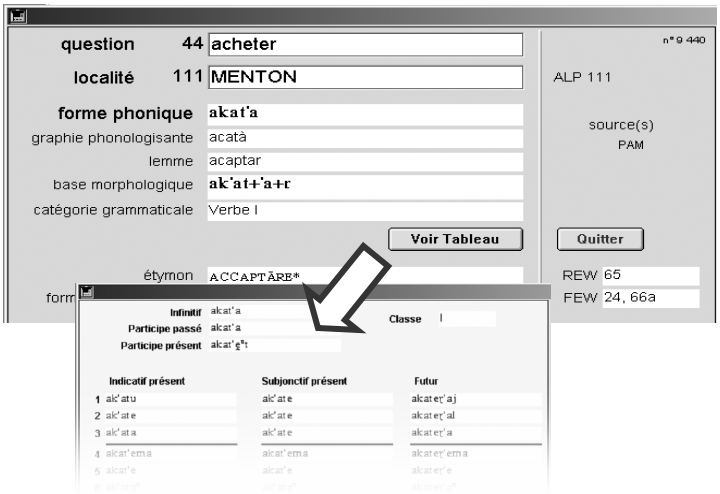
En ce qui concerne les fiches question des plantes, des champignons, et des animaux, la dénomination scientifique est également renseignée dans un champ prévu à cet effet. Du point de vue bibliographique, la consultation d'une fiche permet également d'avoir accès à la liste des cartes publiées et des carnets d'enquêtes, concernant cette question. Le cas échéant, d'autres sources éventuelles peuvent également y être consignées.

Au final, chaque entrée lexicale, ou fiche *réponse*, est donc associée à un coupe localité / question. Il existe donc différentes possibilités d'interrogation de la base pour consulter les entrées lexicales : on peut notamment rechercher toutes les fiches réponses associées à une question précise, de manière onomasiologique, pour étudier la variation diatopique, ou bien rechercher toutes les fiches réponses associées à une localité donnée, pour établir une monographie, ou encore rechercher spécifiquement quel est le terme employé à Nice pour désigner « la chouette » par exemple.

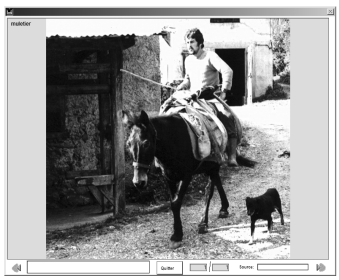
Le détail d'une fiche réponse, présenté sur la capture d'écran ci-après, fait apparaître un certain nombre de champs : transcription phonétique, transcription en graphie phonologisante, lemme, catégorie grammaticale, étymologie, etc. Les champs REW et FEW contiennent respectivement les numéros des entrées correspondantes dans les ouvrages de référence de (Meyer-Lübke: 1911) et (Wartburg: 1922-2002).

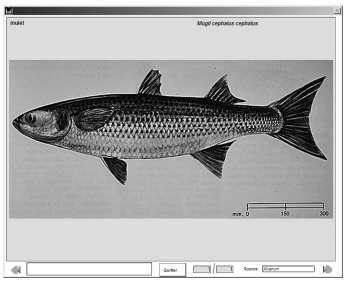
En cliquant sur le bouton « Voir Tableau », on peut également consulter le paradigme de conjugaison verbale ou nominale, lorsque celui-ci a été renseigné dans la base. Lorsque le paradigme morphologique d'un adjectif ou d'un article possède des variations phonétiques contextuelles, un tableau supplémentaire permet de visualiser les différentes formes attestées suivant le contexte phonétique qui suit et/ou qui précède ce terme, le tout étant agrémenté d'une série d'exemples concrets.

[3] Pour une présentation détaillée de ce responsaire, *Cf.* (Oliviéri: 12-13 juin 2002)

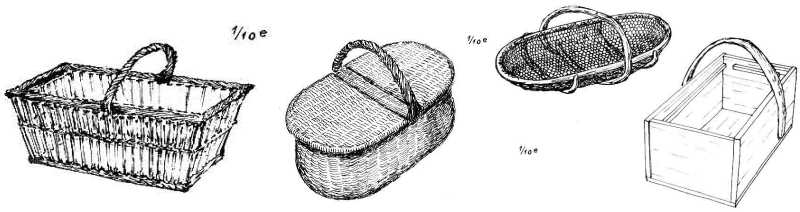


En ce qui concerne le caractère multimédia de la base, chaque fiche question peut-être associée à une ou plusieurs illustrations, ce qui peut s'avérer très utile dans certains cas :





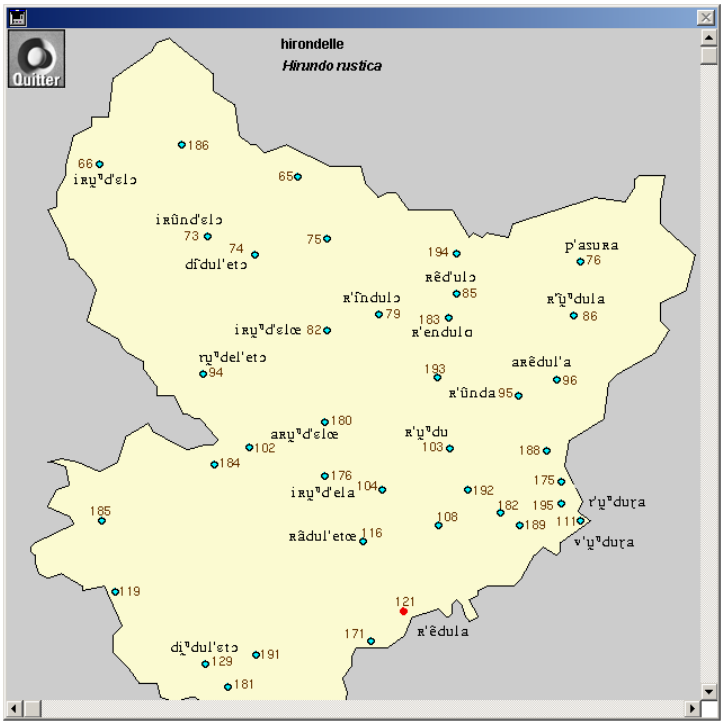
Dans l'exemple ci-dessus, les illustrations permettent de bien faire la distinction entre le mulet (photo de gauche) et le mulet (image de droite), respectivement un mammifère et un poisson. Au-delà de ce cas, certes un peu trivial, cela peu s'avérer très utile pour distinguer les différentes variétés, lorsque un terme quelconque peut correspondre à un grand nombre de variétés différentes. Les nombreuses variétés de paniers nous montrent ainsi tout l'intérêt des illustrations :



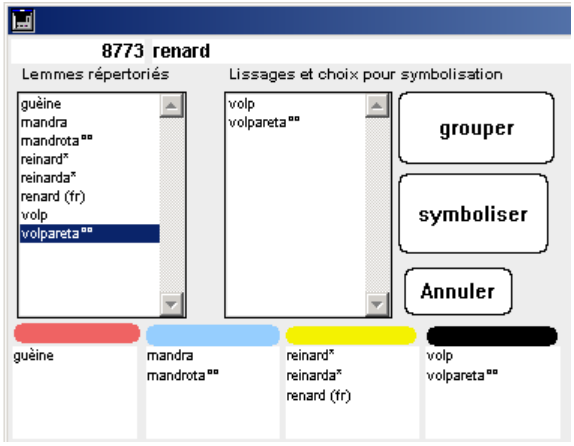
Cartographie interactive

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. 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 ne serait pas lisible d'afficher sur une carte de l'Occitanie toute entière l'ensemble des transcriptions phonétiques attestées dans les différentes localités, les cartes présentant les faits bruts sont disponibles à deux échelles, avec un système de zoom : au niveau de l'Occitanie toute entière, un simple point rouge signale les localités pour lesquelles la base contient une réponse à la question qui est cartographiée. Un clic sur un point rouge permet d'entendre l'enregistrement sonore associé à ce point d'enquête ou de visualiser sa transcription phonétique. Il est également possible de zoomer à l'échelle d'un département. La carte détaillée qui s'affiche alors se présente comme celle de l'exemple ci-dessous : la transcription phonétique associée à chaque réponse est affichée à côté du point de la localité concernée. Là aussi, on peut toujours cliquer sur un point pour écouter l'enregistrement sonore associé.



Voyons à présent un exemple de carte de synthèse, concernant la répartition géographique des différents types lexicaux pour le terme « renard ». Pour générer une telle carte, on sélectionne tout d'abord le terme que l'on souhaite cartographier (ici, la question « renard », qui porte le numéro 8773 dans la base). Le logiciel affiche alors à l'écran la liste des lemmes[4] répertoriés dans les différentes fiches réponses de la base concernant ce terme[5]. 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, à l'aide du formulaire présenté ci-dessous. Dans le cadre de notre exemple, nous avons simplement regroupé ensemble les lemmes qui sont similaires.



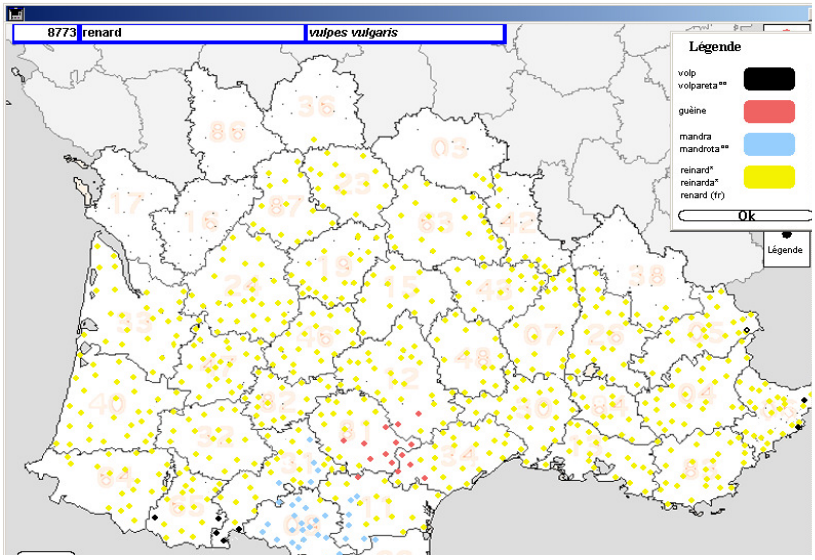
Un simple appui sur le bouton « symboliser » déclenche la construction et l'affichage quasi-instantanés de la carte ainsi demandée, telle qu'elle est représentée ci-après.

Sur cette carte aréale à symboles, on peut observer que les quelques réponses pour « renard » qui sont basées sur des formes du type *volp, volpareta*, représentées ici par des points noirs, se situent à l'extrême Est des Alpes-Maritimes, ainsi que dans une zone centrale des Pyrénées. Par ailleurs, la grande majorité du domaine occitan utilise des formes basées sur *reinard*, *renarda*, ou *renard*, la forme française. Il y a cependant une petite poche de formes du type *guèine*, située à cheval entre les départements de l'Hérault (34), de l'Aveyron (12), et du Tarn (81) ; ainsi qu'une zone avec des formes

[4] Le lemme regroupe toutes les variantes correspondant à une même entité. Cela permet de faire abstraction de la variation dans les réalisations rencontrées.

[5] On remarquera au passage que, selon (Dalbera: 2006), c'est le motif de la « mère » qui semble être à l'origine de tous ces termes : la grande majorité des réponses, qui sont du type *mandra*, *mandrota*, correspondent à *madre*, « la mère » ; celles du type *reinard*, *reinarda*, *renard*, correspondent à la « reine-mère » ; quand à *guèine*, il s'agit de « la gaine », le vagin ; et enfin les formes basées sur *volp*, *volpareta*, « le loup » évoquent la louve « mère nourricière » de Romulus et Remus, à l'origine du mythe fondateur de Rome.

du type *mandra*, *mandrota*, un peu plus au Sud, essentiellement sur le département de l'Ariège (09).



L'utilisateur peut ainsi modifier les critères et les regroupements à volonté pour générer autant de cartes qu'il le souhaite. Signalons enfin que ces cartes peuvent être exportées au format vectoriel pour être intégrées dans un document PDF ou pour servir de document de travail.

Dictionnaire inversé

A partir des données lexicales présentes dans la base, le THESOC propose également un dictionnaire inversé occitan / français. Celui-ci permet de rechercher, à partir d'un lemme donné, les différents termes français susceptibles de traduire un terme occitan. Par exemple, les formes occitanes basées sur le lemme *cuca* peuvent désigner, suivant les cas, un asticot, une chenille, une chouette, un hanneton, un puceron, une vipère, etc. On peut ensuite sélectionner l'un de ces termes français pour faire apparaître la liste des termes occitans renvoyant à la même notion. Il s'en suit donc un véritable va-et-vient entre approche sémasiologique, et approche onomasiologique. Signalons qu'il est également possible au passage de consulter la liste des localités dans lesquelles un de ces termes occitans est attesté.

Bibliographie

Il ne faut pas oublier de mentionner également toute la partie bibliographique de la base, qui est divisée en deux composantes : d'une part, il est possible de consulter toutes les sources qui ont permis de constituer la base de données, et d'autre part, une bibliographie générale suggère un certain nombre d'ouvrages supplémentaires auquel on pourra également se référer.

Traitements des données

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 l'objet d'une procédure de lemmatisation : le lemme est conçu comme une forme de référence conventionnelle ; il sous-tend tout le faisceau de variantes consignées dans la base. En ce qui concerne le THESOC, le choix du lemme est effectué en se basant sur le *Dictionnaire Occitan-Français* de (Alibert: 1966) ; sa notation respecte donc les principes de la graphie alibertine.

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.

Lorsque cela est pertinent, certaines questions sont regroupées, ou disons « compactées » sous un intitulé commun, afin de faciliter la navigation dans la base. Par exemple, dans le cas des champignons, cela permet de rechercher soit une variété particulière de champignon (cèpes, girolles, amanites, etc.) soit de pouvoir consulter les réponses pour toutes les variétés de champignons. Un outil informatique permet de faciliter cette procédure de compactage.

Par ailleurs, le THESOC dispose de fonctionnalités d'importation et d'exportation des fiches de la base, afin d'assurer l'interopérabilité et de pouvoir échanger des données avec d'autres sources de données disponibles.

Au final, la base contient donc des données linguistiques quasi brutes, ainsi que des données ayant déjà fait l'objet d'analyses et de traitements ; et des outils d'investigation que nous allons maintenant présenter.

Outils d'analyse et de recherche

Nous avons déjà eu l'occasion d'aborder plus haut les différentes fonctionnalités proposées sur le plan cartographique, le dictionnaire inversé occitan / français, ainsi que quelques possibilités de recherches concernant les fiches réponses.

Parmi les autres fonctionnalités de recherche de la base que nous n'avons pas encore abordées, il faut noter qu'il est possible d'effectuer une recherche par étymons : si l'on saisit par exemple l'étymon latin *APĬCŬLA*, apparaît alors à l'écran la liste des questions en rapport avec cet étymon[6], la liste des réponses correspondantes, ainsi que la liste des lemmes associés, tels que présentés sur la capture d'écran ci-après.

[6] 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.

Le volet de microtoponymie

Le volet de toponymie du THESOC consigne les différents micro-toponymes recueillis lors des enquêtes. Une fiche toponyme contient généralement les informations suivantes : localité dans laquelle le micro-toponyme a été recueilli, formes graphique et phonétique du toponyme, signifié associé à ce terme dialectal, formes graphique et phonétique de sa traduction en français, type de toponyme (oronyme, hydronyme, etc.), type de référent (quartier de la commune, chemin, cours d'eau, forêt, etc.), étymon et références étymologiques du type REW, commentaires, et éventuelles références bibliographiques complémentaires. Lorsque ce toponyme fait par ailleurs l'objet d'attestations écrites, il est également possible de consigner les formes provenant des différentes sources écrites, avec une identification précise de ces sources et leur date.

A l'instar du reste du THESOC, diverses fonctionnalités de recherche et d'analyse proposées par la base permettent ensuite de consulter ces données : on retrouve là encore la recherche par localité, pour établir une monographie par exemple, la recherche par lemme, et la recherche par étymon ; ainsi que d'autres types de recherche comme la recherche par type de toponyme, par type de référent, par langue d'origine de l'étymon, par type de construction morphologique (avec déterminant ou non, mot composé, syntagme) ou encore par formule étymologique. En outre, le type de toponyme et le type de référent permettent de proposer deux grilles d'analyse : une grille d'analyse du signifié, et une grille d'analyse du référent.

Nous prévoyons actuellement d'ajouter dans la base la possibilité de joindre des illustrations à une ou plusieurs fiches toponyme (pour pouvoir illustrer par exemple un oronyme par une photo du sommet de la montagne en question).

Le Module MorphoSyntaxique (MMS)[7]

Ce dernier module, qui a été mis en place depuis quelques années, est plus particulièrement dédié à l'étude de la syntaxe et de la morphosyntaxe. Il s'agit d'un recueil de textes et de phrases en dialectes occitans, qui sont généralement accompagnés de leur enregistrement sonore, ce qui permet au linguiste utilisateur de la base d'avoir un véritable contrôle des données puisqu'il peut à tout moment avoir un accès direct aux sources, ce qui permet de vérifier par exemple une transcription phonétique douteuse ou de revenir sur un détail particulier quelconque. Il est en outre là-aussi possible d'attacher des documents visuels tels que des photos ou des vidéos à une phrase ou un texte de la base, ce qui lui procure tout son caractère multimédia.

Par rapport au reste du THESOC, les conditions d'oralité et de localisation précise que nous avions évoquées en introduction ont été quelque peu assouplies : si les textes et les phrases sont bien associés à une localité précise, ce sont cependant des aires géographiques qui sont utilisés dans le cadre des traitements réalisés par les différents outils linguistiques de la base, tel que le lemmatiseur ou l'analyseur syntaxique.

Textes

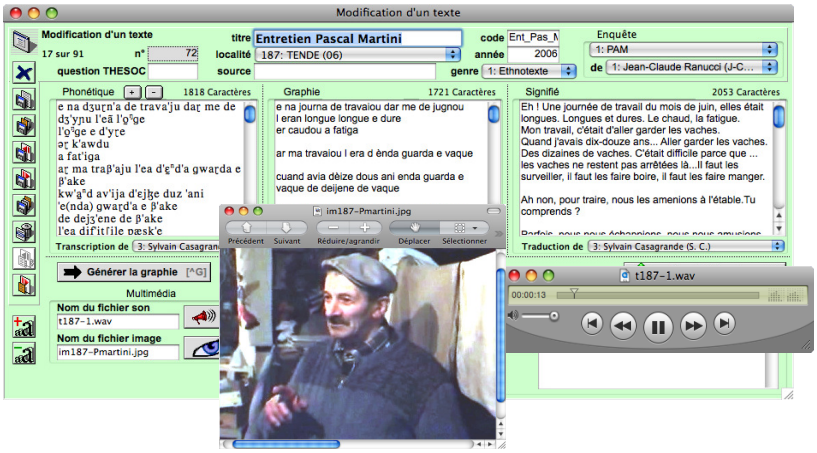
En ce qui concerne les textes présents dans la base, la condition d'oralité a été étendue afin d'intégrer non seulement des ethnotextes et des émissions de radio, qui respectent scrupuleusement la condition d'oralité, mais aussi un certain nombre de textes écrits (ou lus) dans une langue « orale », familière, tels que des articles de presse populaire, ou certaines pièces de théâtre, qui constituent ainsi une sorte d'« oral-écrit ».

Un champ de la base permet cependant de caractériser le « genre » de chacun de ces textes : on peut donc très facilement séparer par exemple d'un côté les contes, chansons, ethnotextes, et émisisons de radio ; et de l'autre les poésies, pièces de théâtre, articles de presse. Si bien que l'utilisateur est libre au final de se cantonner au strict respect de la condition d'oralité, en filtrant un certain nombre de textes sur la base de ce critère de « genre », ou bien d'utiliser la totalité des textes présents dans la base afin d'englober également l' « oral-écrit ». Il est donc tout à fait possible de ne sélectionner que certains genres de textes, ou d'en éliminer certains dans les résultats d'une recherche afin d'obtenir un corpus de travail homogène et cohérent. Par ailleurs ce champ « genre » est consultable à tout moment, ce qui permet de savoir de quel type de texte les données proviennent.

---

[7] Au départ, ce module était intitulé la « base TEXTES » du THESOC, car il contenait essentiellement des textes et ethnotextes. Suite aux évolutions et améliorations successives, 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. Cela permet également de tenir compte de la présence de phrases seules dans la base, qui viennent compléter les textes.

En plus de ce champ « genre », chaque fiche texte contient généralement la transcription phonétique, l'enregistrement sonore associé, une transcription graphique, la traduction en français, la localité associée à ce texte[8], l'année durant laquelle ce texte a été produit, ainsi que quelques commentaires et fichiers multimédia éventuels, comme on peut le voir sur l'exemple ci-dessous :



Phrases

La base contient également un certain nombre de phrases. Il peut s'agir de phrases isolées, ou bien de phrases issues de questionnaires d'enquête morphosyntaxique ou provenant de carnets d'enquête ou de certains atlas linguistiques tel que l'ALMC[9]. Selon un schéma similaire à la structure *réponses – questions* qui est utilisée pour la base lexicale du THESOC et que nous avons déjà évoquée plus haut, chaque phrase est ainsi associée à une question. Chaque question peut faire partie d'un ou plusieurs questionnaires. Ce qui permet tout à la fois de retrouver les questionnaires d'origine et de pouvoir consulter leurs listes de questions, mais aussi de pouvoir factoriser les questions communes à plusieurs questionnaires, selon un mécanisme de « compactage » similaire à celui utilisé pour les questions de la base lexicale.

De la même manière que pour les réponses de la base lexicale, chaque fiche phrase de la base MMS est donc associée à un couple localité / question. Elle contient également la transcription phonétique, une transcription graphique, d'éventuels commentaires, et il est possible d'y adjoindre l'enregistrement sonore lorsque celui-ci est disponible.

---

[8] Pour les ethnotextes, il s'agit de la localité dans laquelle il a été recueilli. Pour les textes de type « oral-écrit », il s'agit généralement de la localité d'origine où le texte a été rédigé.

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

Traitements des données

Les transcriptions graphiques présentes dans les textes et phrases de la base peuvent avoir deux origines possibles, suivant que :

- il s'agit d'un enregistrement sonore recueilli sur le terrain, pour lequel on ne dispose pas *a priori* d'une transcription graphique préalable. Dans ce cas, le même transcripteur que pour la base lexicale du THESOC est utilisé pour générer automatiquement une graphie phonologisante à partir de la transcription phonétique. Cette graphie, proche de la graphie mistralienne utilisée par (Mistral: 1979), permet un premier niveau de « lissage » qui gomme en quelque sorte la variation phonétique, et permet d'accéder au contenu du texte avec une plus grande lisibilité que la transcription phonétique. [10]
- il s'agit d'un texte « oral-écrit » pour lequel on dispose *a priori* d'une transcription graphique déjà existante. Dans ce cas, nous la conservons, et nous utilisons directement cette transcription graphique dans la base, quelque soit la convention qui a été utilisée par son auteur : graphie mistralienne, alibertine, ou italianisante. Un algorithme informatique permet par ailleurs de détecter automatiquement la graphie qui est utilisée dans la transcription graphique d'un texte.

Dans les deux cas, cette transcription graphique servira de point de départ à tous les autres traitements linguistiques effectués par la suite.

Le Module MorphoSyntaxique (MMS) est en effet lui aussi doté d'un certain nombre d'outils linguistiques spécifiquement conçus pour effectuer ces traitements linguistiques.

Le premier d'entre eux est un lemmatiseur, qui identifie chaque élément lexical d'une phrase ou d'un texte en se basant sur un dictionnaire intégré à la base, ce qui permet d'annoter chaque élément lexical avec sa catégorie grammaticale et sa flexion.

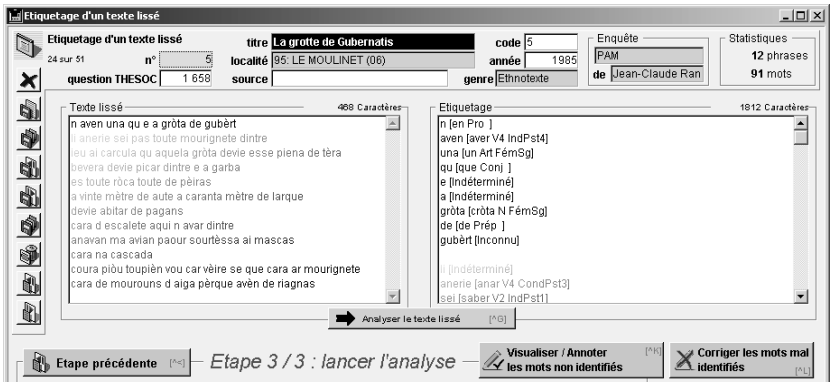
Afin de pouvoir gérer convenablement la variation (graphique, dialectale, flexionnelle), le dictionnaire sur lequel se base le lemmatiseur est structuré sur deux niveaux hiérarchiques : les variantes – qui enregistrent individuellement chaque forme avec la localité dans laquelle elle est attestée – sont 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.

---

[10] 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.

Les données du dictionnaire proviennent à la fois de la base lexicale du THESOC et de différents dictionnaires papier tel que (Eynaudi: 1932) pour lesquels un travail conséquent de numérisation, de balisage, et d'intégration dans la base a été réalisé.

La capture d'écran ci-après représente le formulaire qui permet de faire appel au lemmatiseur : on peut y visualiser le résultat de la lemmatisation générée automatiquement[11].



Lorsque plusieurs entrées du dictionnaire sont susceptibles de correspondre à une occurrence lexicale, celle-ci ne peut pas être directement identifiée et sa nature est donc temporairement considérée comme « indéterminée. ». De même, les termes qui sont absents du dictionnaire, comme les noms propres par exemple, sont considérés comme « inconnus ». On peut alors les identifier manuellement, bien que cela ne soit pas indispensable, comme nous allons le voir. Enfin, il peut arriver qu'une occurrence lexicale soit mal identifiée, lorsque l'on est en présence d'un cas d'homonymie et que seul un des deux homonymes est présent dans le dictionnaire de la base. Il faut alors corriger l'identification de cette occurrence lexicale et créer une nouvelle entrée dans le dictionnaire pour y ajouter cet homonyme. A cet effet, les boutons situés en bas du formulaire permettent d'intervenir sur une lemmatisation lorsque de tels ajustements sont nécessaires.

A ce stade, les éléments lexicaux d'un texte ou d'une phrase n'ont pas pu être tous identifiés car il subsiste certaines ambiguïtés et/ou certains termes inconnus du dictionnaire.

C'est alors qu'intervient ensuite l'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, qu'il s'agisse d'une phrase appartenant à un texte ou d'une phrase seule. Les arborescences ainsi générées se situent dans un cadre génératiViste,

[11] Pour plus de détails concernant le processus de lemmatisation, *Cf.* (Georges: 2006)

mais les règles syntaxiques sur lesquelles l'analyseur se base pour effectuer son analyse sont modifiables par l'utilisateur.

Même si l'utilisateur n'a pas forcément besoin d'avoir accès à une analyse syntaxique des phrases et textes de la base, le recours à un analyseur syntaxique permet de compléter l'identification de chaque occurrence lexicale en fonction du contexte syntaxique ; ce qui n'avait pas pu être réalisé dans sa totalité par le lemmatiseur : les éléments lexicaux « indéterminés » sont désambiguïsés et l'analyseur « devine » la catégorie syntaxique de ceux qui étaient « inconnus » du dictionnaire, ce qui évite de devoir les lemmatiser ou les désambiguïser manuellement, opération qui peut se révéler assez fastidieuse quand le volume de données à traiter devient conséquent.

Ainsi, ces outils permettent d'automatiser en grande partie le travail de lemmatisation et d'annotation des données de la base.

Les phrases sont soumises aux mêmes traitements informatiques que les textes, ce qui permet d'utiliser les mêmes outils d'analyse et de recherche pour exploiter ces deux corpus complémentaires : les fonctionnalités de recherche de la base font apparaître dans leurs résultats à la fois les phrases et les textes correspondants aux critères demandés, et ce sur un même plan. Il reste cependant tout à fait possible de ne retenir que les textes ou que les phrases dans les résultats de la recherche, de la même manière que l'on peut éliminer certains textes en fonction de leur « genre », comme nous l'avons vu précédemment.

Outils d'analyse et de recherche

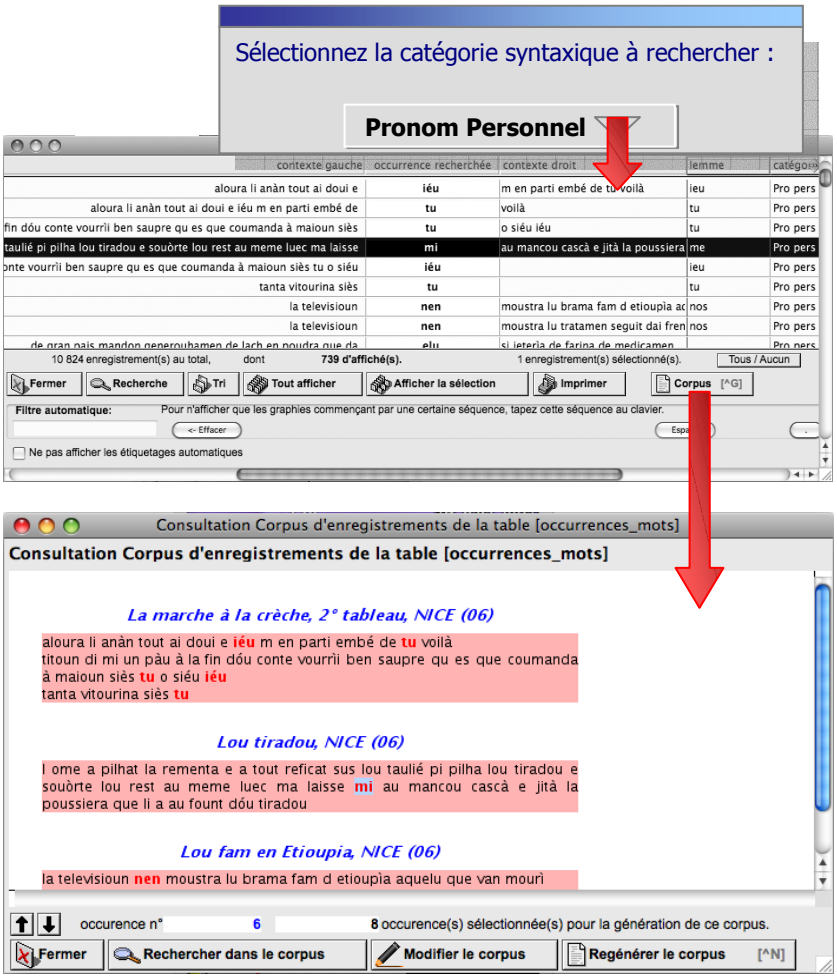
Une fois que les phrases et les textes ont ainsi été annotés, différentes fonctionnalités de recherche de la base permettent de sélectionner des données selon divers critères. On peut ainsi rechercher par exemple :

- toutes les occurrences d'une variante donnée
- toutes les occurrences d'un lemme donné
- toutes les occurrences d'une catégorie syntaxique donnée
- toutes les variantes attestées dans une localité donnée (monographie)
- une séquence de catégories syntaxiques particulière
- toutes les phrases contenant une certaine structure particulière dans leur arborescence syntaxique (par exemple, toutes les phrases contenant une subordonnée)
- par similitude[12] graphique du lemme et/ou de la variante

Quelque soit les critères de recherche sélectionnés, les résultats de la recherche peuvent être affichés sous la forme d'un concordancier, mais on peut également visualiser directement les différentes occurrences dans leur

[12] Les règles qui précisent quel est le degré de similitude entre deux formes graphiques sont par ailleurs configurables et modifiables par l'utilisateur de la base.

contexte d'origine[13], comme le montre la capture d'écran suivante : seules les phrases répondant aux critères de recherche sont affichées ici[14], et les occurrences recherchées sont colorées en rouge afin de pouvoir les repérer plus facilement.



[13] Il existe une seule exception : dans le cas d'une recherche de toutes les variantes attestées dans une localité donnée, l'affichage sous la forme d'un concordancier ne présente gère d'intérêt puisque tous les éléments lexicaux d'une phrase issue de cette localité-là seront systématiquement présents dans les résultats de recherche. Cette présentation se trouve donc remplacé par un affichage en tableau. En cliquant sur une des lignes du tableau, on accède à l'affichage en contexte.

[14] Il est également possible d'afficher le contexte « complet » des occurrences, c'est-à-dire d'afficher le texte complet. Les phrases d'un texte qui répondent aux critères de recherche sont alors surlignées en rose, et à l'intérieur de ces phrases, on retrouve en rouge les occurrences recherchées.

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.

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. Bien qu'il y ait un grand nombre d'utilisations possibles de cette fonctionnalité, une application intéressante de ces « étiquettes » permet de rechercher la présence ou absence de corrélations entre deux ou plusieurs paramètres linguistiques, notamment dans un cadre génératiViste.[15]

Conclusion

Au-delà de la simple consultation des faits, le THESOC est un outil « à géométrie variable », qui permet d'envisager toutes sortes d'exploitations et de recherches, grâce à des outils spécifiques et des modalités d'accès très diverses.

Le module cartographique de la base lexicale permet d'exploiter toute la richesse des données lexicales et s'avère d'une aide précieuse pour qui s'intéresse à la variation diatopique. Aussi, nous envisageons d'améliorer, d'étendre et de généraliser ces fonctionnalités cartographiques au reste du THESOC, et notamment à la base MMS, pour pouvoir à terme générer également des cartes linguistiques concernant des phénomènes morphosyntaxiques ou syntaxiques par exemple.

Nous envisageons également de publier progressivement sur le site Internet une proportion de plus en plus importante des données de la base, et à terme de donner accès via ce site Internet à un certain nombre des outils présents dans la base qui ont été évoqués ici.

Bibliographie

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

Dalbera, Jean-Philippe, *Les parlers des Alpes-Maritimes, étude comparative, essai de reconstruction*, Association Internationale d'Etudes Occitanes, 1994.

Dalbera, Jean-Philippe, *Des dialectes au langage: Une archéologie du sens*, Paris: Champion, 2006.

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

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

[15] Pour plus de détail, Cf. (Georges: 2009) pour une application concrète de cette fonctionnalité.

Georges, Pierre-Aurélien, "Présentation de la base Textes associée au THESOC", in: *actes du colloque 1906-2006: La dialectologie hier et aujourd'hui* (Lyon, 2006), à paraître.

Georges, Pierre-Aurélien, "Les chaînes de clitiques: l'outil informatique au service de l'analyse comparative", in: *actes du colloque Mémoires du terrain: enquêtes, matériaux, traitement des données* (Lyon, 2009), à paraître.

Meyer-Lübke, Wilhelm, *Romanisches etymologisches Wörterbuch*, Heidelberg: Carl Winter's Universitätbuchhandlung, 1911.

Mistral, Frédéric, *Lou Tresor Dóu Felibrige ou Dictionnaire Provençal-Français*, Raphèle-lès-Arles: Culture Provençale et Méridionale, 1979[3].

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

Oliviéri, Michèle, "Le responsaire du THESOC", in: *Actes du 8e colloque de dialectologie et littérature du domaine d'oïl occidental* (Université d'Avignon, 12-13 juin 2002), 2004.

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

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

Wartburg, Walther von., *Französisches Etymologisches Wörterbuch. Eine Darstellung des galloromanischen Sprach-schatzes*, Leipzig - Bonn - Basel: Schrœder - Klopp - Teubner - Helbing & Lichtenhahn - Zbinden, 1922-2002.