HAL
open science

Le Thesaurus occitan dans tous ses états

Michèle Oliviéri, Sylvain Casagrande, Guylaine Brun-Trigaud, Pierre-Aurélien Georges

To cite this version:

Michèle Oliviéri, Sylvain Casagrande, Guylaine Brun-Trigaud, Pierre-Aurélien Georges. Le Thesaurus occitan dans tous ses états. Revue Française de Linguistique Appliquée, 2017, XXII-1, pp.89-102. halshs-01633047

HAL Id: halshs-01633047

https://shs.hal.science/halshs-01633047v1

Submitted on 10 Jul 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Rev. franç. de linguistique appliquée, 2017, XXII-1 (89-102)

Le *Thesaurus Occitan* dans tous ses états

Michèle Oliviéri, Sylvain Casagrande, Guylaine Brun-Trigaud & Pierre-Aurélien Georges
Université Nice Sophia Antipolis / Université Côte d'Azur / CNRS

Résumé : *Le* Thesaurus Occitan (THESOC) *est une base de données multimédia qui vise à rassembler toutes les données dialectales recueillies sous forme orale en domaine occitan. Il est constitué de deux parties, l'une dédiée au lexique, l'autre composée de phrases et consacrée à la syntaxe. Divers outils et fonctionnalités sont associés aux données afin de permettre aux chercheurs de constituer des corpus de travail et d'émettre et de vérifier des hypothèses. Cet article a pour objectif de présenter les modalités de construction et de consultation du* THESOC*, dans ses développements les plus récents.*

Abstract: *The* Thesaurus occitan (THESOC) *is a multimedia database that aims at assembling all the dialectal data gathered in an oral form throughout the occitan-speaking region. It has two parts: one deals with the lexicon, and the other, composed of sentences, is devoted to syntax. Different tools and functionalities are associated with the data in order to allow researchers to constitute bodies of work and to formulate and verify hypotheses. This article presents the most recently upgraded form of the* THESOC*, its modalities of construction, and its methods of consultation.*

M**ots-clés :** Corpus oraux, Dialectes occitans, Lexique, Morphosyntaxe
Keywords: Oral corpora, Occitan dialects, Lexicon, Morphosyntax

1. La base de données THESOC

1.1. Présentation générale

Le *Thesaurus Occitan* (ou THESOC), développé dès 1992 sous la direction de Jean-Philippe Dalbera, a été une des premières entreprises de constitution de bases de données dialectales, pourvue d'outils d'analyse originaux, et ce pour une bonne raison : les progrès de la micro-informatique l'ont permis. D'une part, il s'agissait de rassembler l'ensemble des données dialectales recueillies en domaine occitan, qu'elles aient été publiées (atlas, monographies) ou non (enquêtes de terrain), et de les mettre à la disposition du plus grand nombre. Ce premier aspect du projet concerne l'aspect 'thesaurus' qui a notamment une fonction patrimoniale et culturelle. D'autre part, l'objectif était de fournir aux chercheurs un certain nombre d'outils automatisés pour le traitement de ces données, en conférant ainsi à l'objet informatique une fonction heuristique[1]. De la sorte, la base contient à la fois des données brutes et des données traitées.

Au fil des années, la base s'est ainsi progressivement enrichie et, en premier lieu, de données lexicales. En effet, ces données étaient les seules à avoir été cartographiées et publiées dans les atlas linguistiques et ethnologiques de la France par région, et il était naturel

[1] Pour une présentation plus détaillée du THESOC, cf. Dalbera *& al.* (2012).

de commencer par implémenter le lexique. Le THESOC compte à ce jour plus d'un million d'items lexicaux. Plus récemment, sous la direction de Michèle Oliviéri, s'est adjointe à la base lexicale une nouvelle base de données, comportant des textes et des phrases, destinée à l'étude morphosyntaxique et syntaxique. Elle est en cours d'implémentation. Ce 'Module Morpho-Syntaxique' (MMS) a bénéficié successivement de deux financements franco-allemands, dans un premier temps le projet PHC Procope 'Microvariation syntaxique : les pronoms clitiques dans les langues romanes' (2010-2012), et en 2012-2015 le projet ANR-DFG DADDIPRO[2], qui a permis notamment la mise en ligne d'une partie de nos enquêtes.

1.2. Principes

Les données qui figurent dans le THESOC répondent toutes impérativement à deux conditions : d'une part, ces données sont issues de l'oralité, et d'autre part chacune d'entre elles doit être localisable géographiquement et temporellement. L'on doit ainsi pouvoir, au minimum, fournir le nom de la commune dans laquelle ont été recueillies les données, ainsi que la date de recueil. Ces deux critères garantissent ainsi un aspect scientifique et non normatif au THESOC, aspect nécessaire à sa vocation patrimoniale, culturelle et scientifique. Le fichier des localités compte à ce jour 884 points d'enquête situés en terre de langue d'oc et dans les aires limitrophes, afin de permettre la comparaison, voire de tracer des isoglosses. Ainsi, les dialectes liguriens de la vallée de la Roya, à la frontière italienne, comme les parlers du Croissant ou certains parlers franco-provençaux, sont représentés dans le THESOC. Chaque fiche *Localité* donne toutes les informations pertinentes sur les sources des données (atlas, enquête, enquêteur, témoins, etc.).

Les données étant de nature orale, elles sont d'abord naturellement transcrites en phonétique. Les différents auteurs des atlas ayant adopté diverses conventions de transcription, toutes cependant basées sur l'alphabet phonétique établi par Rousselot (1924) et traditionnellement utilisé par les romanistes, le THESOC a uniformisé le système de transcription phonétique en convertissant tout en Alphabet Phonétique International (API). Cela a permis d'en simplifier la lecture et de rendre les formes comparables, au risque parfois de perdre certaines distinctions mineures, voire mal établies[3]. Ce premier niveau de transcription (*Forme phonique*) est réservé aux spécialistes, les linguistes, qui ont ainsi accès aux données primaires pour leurs travaux. Un deuxième niveau de transcription est fourni, une graphie phonologique ou 'phonologisante', qui est surtout une graphie utilitaire, destinée à l'outil informatique et aux spécialistes. Elle est générée automatiquement et n'est pas sans rappeler la graphie établie par Mistral (1878). Cette graphie, relativement transparente pour les francophones, permet notamment à ceux qui ne sont pas familiers avec l'API et ne parlent pas le dialecte considéré d'en déduire la prononciation. Enfin, un troisième niveau de transcription, plus lissé, pan-occitan, est donné avec une lemmatisation basée sur les principes graphiques dits 'classiques', adoptés par Alibert (1966). Ce niveau trouve son utilité notamment dans les recherches de types lexicaux ou pour l'étiquetage morphosyntaxique des phrases.

Chaque forme est en outre bien documentée : toutes sortes d'informations figurent, notamment la source de la donnée, ce qui rend possible un contrôle permanent de sa fiabilité.

---

[2] *Dialectal, acquisitional, and diachronic data and investigations on subject pronouns in Gallo-Romance*, sous la responsabilité de M. Oliviéri (Nice) et G.A. Kaiser (Constance) <http://www.agence-nationale-recherche.fr/projet-anr/?tx_lwmsuivibilan_pi2%5BCODE%5D=ANR-11-FRAL-0007>.

[3] Cependant, chaque forme étant bien identifiée, il est toujours possible de se reporter à l'atlas pour vérifier la transcription figurant dans le THESOC.

Enfin, le THESOC offre l'avantage sur les atlas et ouvrages imprimés de pouvoir associer aux données linguistiques d'autres types de documents et a donc adopté une approche multimédia. Ainsi, les données brutes que sont les enregistrements audio ou vidéo des enquêtes dont nous disposons sont intégrées dans la base. C'est également le cas des images destinées à préciser la nature du référent, des reproductions d'illustrations accompagnant certains textes ou de tout autre document susceptible d'éclairer les faits.

2. La base lexicale

2.1. Les data

Les données ont été recueillies par enquête orale, et le plus souvent selon la tradition des atlas linguistiques à partir d'un questionnaire, i.e. une liste de mots dont la traduction est demandée aux locuteurs, appelés traditionnellement 'questions'. La base de données est organisée de la même manière, et le fichier central représente alors la réunion de ces différents questionnaires, le *responsaire* ou fichier *Questions*. Il faut cependant signaler que ce fichier n'est pas seulement une somme mais le résultat d'une élaboration (Oliviéri 2004, 2012). En effet, dans un certain nombre de cas, des questions d'atlas ont été regroupées, soit parce qu'elles étaient synonymes (ex. *peler* et *éplucher*), soit pour des raisons fonctionnelles. Ainsi, lorsqu'il existe un terme générique (ex. *cerise*) et différentes variétés (ex. *bigarreau, griotte*, etc.), deux entrées figurent dans le THESOC (*cerise* et *cerise (variétés de)*) Si l'on choisit l'entrée 'variétés de', on peut alors consulter les réponses obtenues pour une seule variété (ex. *bigarreau*) ou l'ensemble des termes pour toutes les variétés confondues. Ces 'questions' (au nombre de 8 200) sont indexées selon la thématique habituelle des atlas : agriculture, élevage, vie quotidienne, nature, etc[4].

Ainsi, chaque forme lexicale est identifiée par une double numérotation : numéro de la 'question' et numéro de la localité où elle a été recueillie (cf. figure 1).

Figure 1. *Fiche réponse par localité : la* BERGERONNETTE *à Pézenas*

On y accède par une fiche *Réponse* qui comporte, outre la source de la donnée et les trois transcriptions (*phonique, phonologisante, lemme*), des indications et des données d'ordre morphologique (catégorie, segmentation morphologique et formes fléchies lorsqu'elles

[4] Cette classification est organisée en deux niveaux : thèmes et sous-thèmes (ex. Elevage > ovins, bovins, caprins, etc.)

existent) et d'ordre étymologique (étymon et renvoi aux dictionnaires étymologiques de référence, REW et FEW).

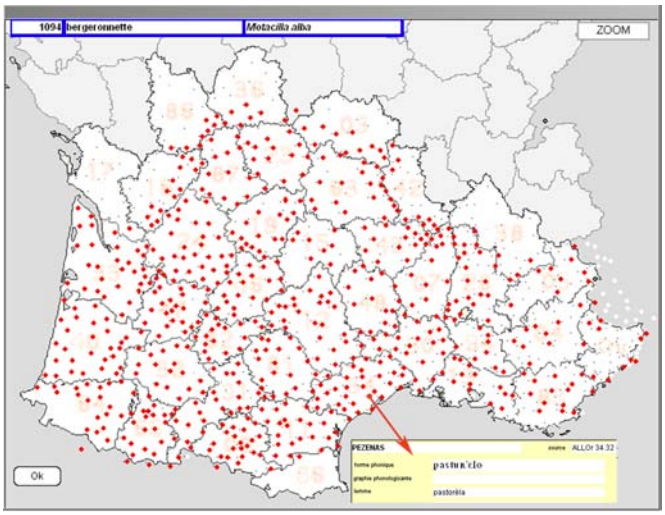
Mis à part les numéros de question et de localité, le seul champ obligatoire de la fiche, i.e. qui établisse l'existence de la donnée dans la base, est le champ *Forme phonique*, en API, ce qui garantit l'origine orale de l'item. Il faut noter que lorsque plusieurs réponses existent pour une question donnée, autant de fiches réponses sont implémentées.

Ces données lexicales, dont le nombre est de 1 212 256 à ce jour[5], peuvent être consultées en utilisant diverses requêtes, selon l'objectif recherché : par question (toutes les réponses figurant dans la base pour une question donnée, ex. *Quels sont les termes qui désignent la chouette ?*) ; par localité (toutes les réponses obtenues dans une localité donnée, ex. *Quels sont les termes recueillis à Nice ?*) ; par couple localité-question (ex. *Comment dit-on* chouette *à Nice ?*) ; par thème/sous-thème (ex. *l'élevage des ovins*) ; ou par atlas (ex. l'*Atlas Linguistique de la Gascogne*, ALG). A partir de ce module de type *Glossaire*, l'utilisateur peut alors extraire du THESOC le corpus qu'il souhaite.

2.2. Cartographies

Dans la tradition dialectologique et celle des atlas linguistiques, l'aspect cartographique occupe une place importante dans l'architecture de la base. Trois types de cartes peuvent être générés automatiquement : cartes onomasiologiques, cartes sonores et cartes aréales.

Les cartes onomasiologiques correspondent à une requête par question et ne diffèrent des cartes d'atlas que par l'étendue du domaine considéré. A chaque localité où figure une réponse apparaît un point rouge cliquable qui donne accès à la réponse (ou aux réponses en cas de variation). On peut ainsi visualiser directement les items lexicaux à une plus grande échelle que les atlas régionaux imprimés (carte 1).

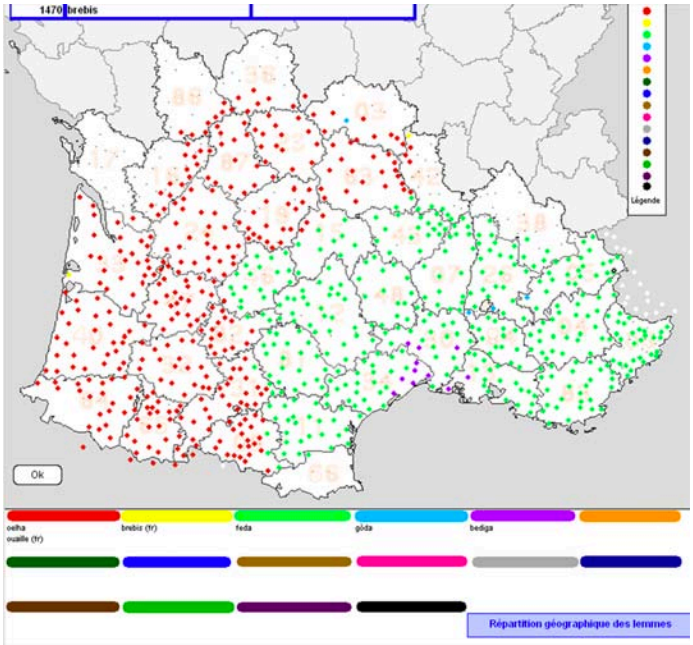


Carte 1. *Réponses par question : 1094* BERGERONNETTE.

[5] D'autres données (environ 150 000 items) sont en cours de traitement et seront implémentées sous peu.

Un système de zoom permet cependant de se focaliser sur un département donné. Dans ce cas, les formes phoniques apparaissent sur la carte et un point rouge s'affiche si l'enregistrement sonore est disponible. En cliquant sur ce point rouge, on peut alors l'entendre, et ainsi comparer les attestations et contrôler les transcriptions. Il est d'autre part possible d'accéder directement aux enregistrements audio sans passer par la transcription, avec les cartes sonores.

Si ces deux premières cartographies ne concernent que les données brutes, la troisième consiste en un traitement des données qui conduit à représenter automatiquement la distribution géographique des lemmes, i.e. des types lexicaux, pour une notion donnée (une 'question'). Pour ce faire, l'utilisateur effectue une classification en regroupant les lemmes et en associant chacune de ces classes à une couleur différente, et le THESOC produit alors une carte où apparaissent en couleurs les aires de diffusion des différents types lexicaux (carte 2).



Carte 2. *Répartition des lemmes correspondant à la question* BREBIS.

2.3. Les outils

Le THESOC contient naturellement de nombreuses fonctionnalités autorisant et facilitant l'implémentation et l'enrichissement de la base. C'est le cas notamment du transcripteur automatique (phonie > graphie), mais aussi d'autres procédures permettant l'organisation et l'indexation du fichier *Questions*, la lemmatisation, l'affectation des étymons, l'intégration des documents image, audio et vidéo et toutes sortes de corrections et de vérifications.

Mais, afin de fournir au chercheur les outils dont il a besoin, d'autres fonctionnalités sont également disponibles dans le THESOC, offrant la possibilité de naviguer dans le lexique, en synchronie comme en diachronie.

2.3.1. Dictionnaire inverse

Ainsi, avec le module *Oc-Français*, on peut interroger la base de données en y entrant, non par le mot français (la question), mais par le lemme occitan. Une telle requête ouvre à une navigation dans le lexique qui permet d'appréhender la variation référentielle à travers un jeu tour à tour sémasiologique et onomasiologique. Par exemple, si *cuca* désigne l'*asticot* ou la *chenille* dans certaines localités données, le même terme peut désigner la *chouette* ou la *vipère* dans d'autres localités. A partir de là, on peut voir également comment cette notion ou ce référent (par ex. la *vipère*) est désigné dans les autres parlers occitans. Ce module, qui dépasse le cadre strictement dialectologique en posant des questions de linguistique générale, notamment sur la relation signifiant-signifié et le rapport entre le mot et la chose, est donc conçu pour l'élaboration d'hypothèses et de pistes de recherche, et illustre au premier chef la fonction heuristique du THESOC.

2.3.2. Etymologie

Les données d'ordre étymologique associées aux différents items lexicaux ne se limitent pas à indiquer l'étymon d'une forme donnée mais sont aussi exploitées par le module étymologique dans une perspective de reconstruction. Ce module permet d'accéder aux données de la base à partir d'un étymon donné, sélectionné dans une liste. Si les outils nécessaires à la recherche en matière de phonétique historique restent encore en cours de développement, les travaux en sémantique lexicale ont déjà bénéficié de cet outil, en montrant les continuateurs de tel ou tel étymon dans les différents parlers modernes.

2.3.3. Toponymie

Ce module, qui a été développé sous la direction de Jean-Claude Ranucci, est consacré à l'étude des micro-toponymes, i.e. les noms de toutes les entités du paysage recueillis au cours des enquêtes. Ces toponymes sont classés, indexés et organisés selon une série de filtres et de critères permettant de les étudier. Comme les autres items lexicaux du THESOC, la forme phonique est à la base de la fiche et les deux autres niveaux de transcription y figurent également. Suivent ensuite des indications sur le référent, sur le signifié et la manière dont le perçoit le locuteur, sur le signifiant et enfin, une discussion sur l'étymologie du terme considéré, si celle-ci n'est pas transparente et clairement établie. Le référent et le signifié sont indexés selon une liste de catégories prédéfinies, qui permettent ainsi la recherche et la comparaison de termes apparentés.

Outre l'intérêt patrimonial de cette entreprise, ces toponymes ne figurant parfois sur aucun cadastre et n'existant qu'à l'oral, l'objectif est de constituer des corpus, de comparer ces formes et de les traiter de manière à en tirer des hypothèses sur les motivations et remotivations qui ont présidé à ces dénominations ainsi que sur la manière dont l'être humain perçoit son environnement.

2.4. La mise en ligne

Le site web du THESOC a vu le jour dès 2005-2006, d'abord avec les données lexicales et s'est progressivement enrichi. Actuellement, environ 80 % des entrées lexicales présentes dans la base sont consultables en ligne, sur le site <thesaurus.unice.fr> qui propose des fonctionnalités de recherche par question, par localité, ou encore par département.

Au départ, le choix s'est porté sur la technologie Flash qui permet la consultation des transcriptions phonétiques en API par tous, sans avoir besoin d'installer une police particulière

ou un quelconque logiciel sur l'ordinateur de l'internaute, si ce n'est le *plugin* Flash qui, à l'époque, était disponible sur la plupart des ordinateurs. Aujourd'hui, cependant, cette technologie est devenue quelque peu obsolète et il s'agit d'effectuer une profonde refonte du site web, avec pour objectif d'utiliser les nouvelles technologies du web (HTML 5) afin de s'affranchir définitivement de toute contrainte et de permettre la consultation du site par tous, y compris depuis un *smartphone* récent. C'est ce qui a déjà été réalisé pour les données issues du projet DADDIPRO (cf. *infra*), et cela sera étendu à la base lexicale dans un deuxième temps, une conversion préalable des données au standard Unicode étant nécessaire.

3. Le Module Morpho-Syntaxique (MMS)

Dans la perspective des travaux sur la microvariation syntaxique, qui connaissent depuis quelques années un essor considérable, le module MMS a été conçu pour traiter de la description et de la (micro-)variation syntaxiques et morphosyntaxiques dans les dialectes occitans, dans une perspective à la fois synchronique (géolinguistique), diachronique (reconstruction) et théorique (grammaire générative). Cet objectif répond à un réel besoin dans ce domaine, notamment pour les romanistes car, si une partie du lexique occitan était déjà disponible dans les atlas, il n'existait pas de corpus de phrases et dans le paysage linguistique européen, les données dialectales occitanes manquaient cruellement. Elles restent d'ailleurs encore mal connues et nettement sous-exploitées.

3.1. Les data

L'objectif de MMS est de disposer de phrases pour l'étude de la syntaxe occitane. Ces phrases proviennent soit de textes dont elles ont été extraites, soit de réponses à des questionnaires d'enquête. Les atlas, pour leur part, n'offrent que peu de phrases : certaines (très brèves) sont cartographiées et ont pu être implémentées, d'autres (rares) figurent dans les marges des cartes d'atlas et sont en cours de recensement.

3.1.1. Les textes

Il convient de distinguer ici deux types de textes. Lors des enquêtes, nous recueillons souvent des récits ou des commentaires (par exemple lors des enquêtes de toponymie) qui constituent des témoignages spontanés de la langue orale. Ces textes sont désignés dans la base comme *ethnotextes*[6] et représentent les données les plus fiables de MMS, mais ce sont aussi les moins nombreux et bien souvent les moins productifs. Il est en effet difficile d'obtenir dans un tel contexte des structures syntaxiques particulières ou complexes. C'est pourquoi, afin d'augmenter le volume de données, nous avons choisi d'intégrer dans MMS une autre catégorie de textes. La condition d'oralité stricte est alors assouplie pour accepter d'implémenter des textes relevant de ce que l'on pourrait appeler de l'*oral-écrit*. Il s'agit de textes écrits mais ayant une réalité orale : pièces de théâtre populaire, presse populaire, émissions de radio, récits, etc. Cependant, les sources des phrases sont toujours identifiées et il est toujours possible de limiter ses recherches à un type de textes.

La plupart du temps, les ethnotextes sont, comme les items lexicaux, d'abord implémentés en API, puis convertis automatiquement en graphie phonologique, l'analyse syntaxique ne nécessitant pas une transcription phonétique précise. C'est pourquoi, selon les cas et les contingences, il arrive aussi qu'un ethnotexte soit transcrit directement dans une graphie phonologique inspirée des principes mistraliens.

---

[6] Le terme d'*ethnotext* est utilisé ici dans une acception qui s'éloigne quelque peu de celle de Bouvier (1992). Il s'agit en réalité de discours libres en dialecte recueillis au cours des enquêtes de terrain.

Les autres textes, 'oraux-écrits', sont implémentés dans leur forme originale, le plus souvent dans une graphie mistralienne, mais certains dans d'autres systèmes graphiques, tels que la graphie italianisante dans le pays niçois ou la graphie alibertine en occitan central. MMS accepte donc toutes les graphies et les outils fournis dans la base gèrent toute cette variation, tout en conservant les particularités et l'originalité de chaque texte.

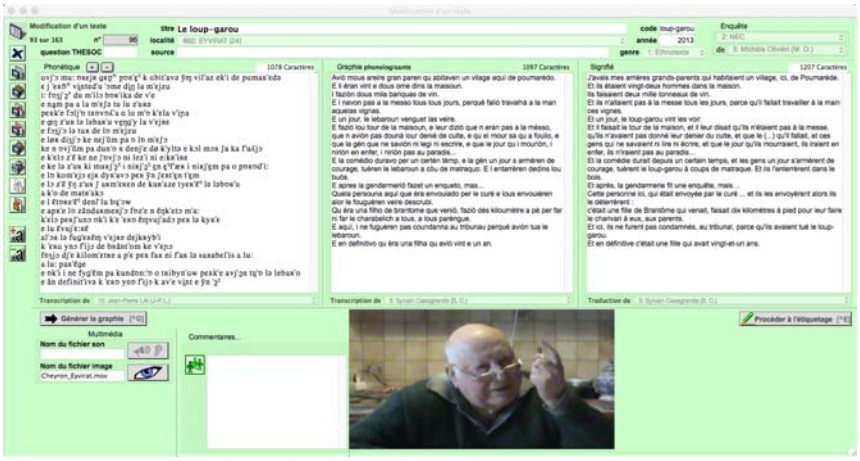


Figure 2. *Un ethnotexte dans MMS.*

3.1.2. Les questionnaires

Afin de permettre la comparaison entre les systèmes linguistiques, il est nécessaire de disposer d'énoncés comparables et donc de procéder à des enquêtes de terrain en utilisant un questionnaire, i.e. une liste de phrases à faire traduire par le témoin.

Selon les époques et les besoins des chercheurs, plusieurs questionnaires ont été successivement élaborés et coexistent dans la base, certains se recoupant partiellement, une même question pouvant être présente dans plusieurs questionnaires. Chaque 'question' est alors identifiée par un numéro unique et donne lieu à plusieurs réponses, selon les localités et les locuteurs. Trois de ces questionnaires sont particulièrement productifs. Le premier est le questionnaire PAM (*Parlers des Alpes-Maritimes*), conçu par Jean-Philippe Dalbera, qui a connu deux versions successives (PAM1 et PAM2), et qui couvre l'essentiel de la morphologie et de la syntaxe. Récemment, dans le cadre du projet DADDIPRO, un second questionnaire a vu le jour, NEC (*Nouvelles Enquêtes Complémentaires*), d'abord dans le but d'étudier l'apparition des clitiques sujets dans les dialectes du nord de l'aire occitanophone (NEC1), puis enrichi d'autres structures syntaxiques, plus complexes (NEC2). Encore plus récemment, dans le cadre du projet ANR SyMiLa[7], un troisième questionnaire a été élaboré par Anne Dagnac et Patrick Sauzet, qui interroge de manière plus approfondie et plus systématique les mécanismes syntaxiques. Les questions sont en outre annotées, selon le(s) objectif(s) recherché(s) par l'auteur du questionnaire : ex. *proposition relative, interrogation,*

[7] Projet ANR SyMiLa 2013-2015 *Syntactic Microvariation in the Romance Languages of France*. UMR 5462 CNRS UTM / CLLE ERSS (Toulouse) & UMR 8129 CNRS EHESS ENS / Institut Jean Nicod (Paris), sous la responsabilité de P. Sauzet (Toulouse) & D. Sportiche (Paris) <http://www.agence-nationale-recherche.fr/projet-anr/?tx_lwmsuivibilan_pi2%5BCODE%5D=ANR-12-CORP-0014>.

clitique datif, etc. Cet étiquetage des questions, en cours d'implémentation dans MMS, permettra une recherche plus efficace des structures syntaxiques.

Ces enquêtes sont aussi l'occasion de tester la grammaticalité de certaines structures, de proposer des phrases alternatives aux locuteurs et de recueillir leurs jugements linguistiques. Elles donnent aussi lieu à des enregistrements audio et/ou vidéo, qui sont naturellement intégrés à la base de données et permettent ainsi d'accéder à la prosodie et de contrôler les transcriptions.

3.2. Les outils

3.2.1. Traitement des données

Si les données des questionnaires et les *ethnotextes* en API sont facilement implémentables par importation automatique et ne nécessitent qu'une conversion en graphie phonologique[8], les autres textes reçoivent au préalable un traitement particulier. Ils doivent également être 'nettoyés', de manière à exclure les passages en français, les didascalies des pièces de théâtre, etc. Tous les textes sont ensuite 'lissés', de manière à préparer les opérations suivantes, en traitant les problèmes liés à la ponctuation, aux apostrophes, aux élisions, à la forme des chiffres, etc.

Une fois ces données primaires implémentées (transcriptions phonétique et graphique, audio et/ou vidéo), les phrases, qu'elles proviennent de textes ou de questionnaires, sont soumises à plusieurs opérations.

En premier lieu intervient l'étiquetage morphosyntaxique. Cette opération repose sur un dictionnaire intégré qui est structuré en deux niveaux. Le premier niveau est celui des *variantes*, i.e. toutes les formes enregistrées avec leurs flexions et les localités où elles sont attestées, et dans les différentes graphies où elles se présentent. Le second niveau est celui des *lemmes*. Comme dans la base lexicale, les lemmes (en graphie alibertine) constituent la forme pan-occitane qui permet de regrouper les différentes variantes (dialectales, flexionnelles ou graphiques) sous une forme unique. L'étiquetage consiste alors à attribuer à chaque mot de chaque phrase une étiquette indiquant le lemme[9], la catégorie grammaticale[10] et la flexion. Cette tâche est partiellement automatisée car le dictionnaire ne suffit pas à désambiguïser les homonymes : l'utilisateur peut ensuite intervenir pour corriger une annotation erronée ou pour identifier un terme ambigu ou inconnu du dictionnaire. Cependant, une partie de ces désambiguïsations peuvent être effectuées par l'analyseur syntaxique et il est possible d'éviter dans une certaine mesure cette procédure manuelle.

Le second traitement consiste en une analyse syntaxique de chaque phrase, sur la base de l'étiquetage précédent, sous forme d'un arbre syntaxique[11]. Le programme propose alors plusieurs représentations (de la plus probable à la moins probable) parmi lesquelles l'utilisateur peut choisir. Il peut éventuellement en conserver plusieurs, lorsque c'est pertinent. C'est là que les homonymes sont désambiguïsés, l'analyseur syntaxique étant à même de distinguer par exemple dans la forme [e] le verbe *être* de la conjonction de coordination.

---

[8] Toutes les conversions de l'API à la graphie phonologique sont effectuées au moyen du transcripteur automatique du THESOC.

[9] Ces lemmes sont partagés par la base lexicale et MMS, chacune des deux bases permettant d'enrichir l'autre.

[10] Les catégories grammaticales sont organisées selon une hiérarchie qui permet notamment une recherche plus ou moins large (par ex. Déterminant > Article > Article défini).

[11] Le modèle choisi ici pour cette opération est celui de la *Théorie du Gouvernement et du Liage (GB)*, qui était le modèle dominant en grammaire générative lorsque MMS a débuté, mais tout autre modèle peut être implémenté.

Enfin, les phrases peuvent encore être identifiées par un autre système d'étiquetage personnalisé, où l'utilisateur leur attribue un *tag* en fonction d'une configuration syntaxique particulière. Par exemple, on peut identifier par des *tags* les phrases qui manifestent l'ordre 'Accusatif-Datif' des clitiques objets *vs* celles où l'ordre est inverse, ou encore celles qui ont un sujet réalisé *vs* celles qui n'en ont pas.

3.2.2. Interrogation de la base

Plusieurs types de recherches peuvent être effectués dans MMS, qui permettent de constituer des corpus de travail. On peut bien sûr rechercher un lemme, voire une variante, si l'on veut étudier une occurrence précise. Mais on peut également faire une recherche par catégorie, ou par séquence de catégories, avec quelques options supplémentaires. Par exemple, on peut rechercher les adverbes, une séquence de plusieurs pronoms, ou encore un verbe en début de phrase.

Il est aussi possible d'interroger la base par structure syntaxique, i.e. en formulant une requête sous la forme d'une structure parenthétisée. Dans un proche avenir, il est envisagé d'améliorer cette fonctionnalité en proposant une liste de requêtes possibles, telles que 'phrase à pro drop' ou 'phrase à V2'. Mais d'ores et déjà, si l'on a auparavant affecté des *tags* aux phrases qui présentent tel ou tel phénomène syntaxique, on peut les sélectionner à partir d'une recherche par *tag*. A la différence de la base lexicale, MMS n'a pas encore d'outil de cartographie automatique, mais ce système de *tags* constitue une première étape vers cette fonctionnalité. Une fois que les faits syntaxiques présents dans les phrases sont identifiés par des *tags*, il devient en effet possible de les cartographier (ex. cartographie des occurrences 'pro drop' *vs* 'pro réalisé').

Une fois les données recherchées sélectionnées, il suffit d'un clic pour générer un corpus de phrases dans lequel est surlignée chaque occurrence qui répond au critère de recherche. On peut finalement exporter ce corpus vers un traitement de texte ou un tableur, pour l'exploitation scientifique des données.

3.3. La mise en ligne

Le second volet du site web du THESOC destiné aux données morphosyntaxiques (MMS) est en cours de développement <thesaurus.unice.fr/daddipro/>. Sa conception a été permise par le projet DADDIPRO et il propose, accompagnées de leurs enregistrements vidéo, une sélection de phrases issues d'enquêtes de terrain réalisées dans le cadre de ce projet et intégrées à MMS.

Les données étant géolocalisées, la page d'accueil affiche une carte interactive du domaine occitan permettant de visualiser l'ensemble des localités du THESOC, avec, en couleur, les différents points d'enquêtes DADDIPRO pour lesquels figurent d'ores et déjà des données, ce qui évolue au fur et à mesure de la mise en ligne des enquêtes.

Pour chaque point d'enquête, figurent les différentes questions (issues du questionnaire NEC) avec la transcription phonétique des réponses fournies par le témoin, ainsi que la vidéo associée[12], le nom du témoin et le lieu de l'enquête (avec son numéro THESOC). Un minutage précis a été réalisé phrase par phrase, afin d'obtenir une synchronisation automatique entre la transcription et la vidéo : un clic sur l'une des transcriptions phonétiques place la lecture de la vidéo à l'endroit précis où la phrase sélectionnée démarre, et la lecture s'arrête automatiquement à la fin de cette phrase, ce qui permet de comparer rapidement les réponses données par les informateurs. Il est prévu également d'implémenter à moyen terme l'opération

[12] Les vidéos sont diffusées sous licence Creative Commons CC-BY-NC-ND, donc librement diffusables et réutilisables sur le web. Cf. <http://creativecommons.fr/licences/>.

inverse, i.e. de faire défiler les transcriptions phonétiques au fur et à mesure de la lecture de la vidéo, avec un surlignage de la transcription API de la phrase prononcée par le témoin dans la vidéo.

Contrairement à la partie lexicale, plus ancienne, ici les données sont déjà en Unicode, et la consultation du site internet ne requiert aucun prérequis sur la machine de l'internaute : il lui suffit d'avoir un navigateur web suffisamment récent. En effet, aucun *plugin* n'est requis : au moment de la première consultation du site, la police phonétique Unicode est automatiquement téléchargée et mise en place côté client par le navigateur web.

Le travail nécessaire pour les divers traitements (annotation, transcription, segmentation, et minutage des vidéos) étant considérable, il s'agit pour le moment d'un échantillon établi à partir d'un choix de questions illustrant quelques faits syntaxiques bien connus (tels que la négation, l'interrogation, la subordination, etc.) et d'une partie des localités enquêtées. A terme, toutes les données de MMS sont destinées à figurer sur le site, à l'exception de celles pour lesquelles les informateurs en ont refusé ou retardé la diffusion.

4. Perspectives

4.1. Enrichissement du THESOC

4.1.1. Les données

Parallèlement à la tâche de saisie des données encore inédites provenant de l'exploitation systématique des carnets d'enquêtes des atlas linguistiques régionaux, devenus disponibles et facilement manipulables depuis leur numérisation, le travail de lemmatisation et de recherches étymologiques se poursuit sur le quart restant des données lexicales. Cela permettra à terme de mettre en ligne la quasi-totalité des données lexicales provenant de l'ensemble des enquêtes effectuées en domaine occitan, ce qui représente environ 1 500 000 items. De nouvelles enquêtes sont cependant effectuées régulièrement et, ces dialectes étant malheureusement en voie de disparition, priorité est donnée à cette collecte. Ces nouvelles enquêtes viendront naturellement et progressivement enrichir la base de données.

En outre, de nouvelles données complémentaires concernant la morphologie verbale sont en cours d'implémentation.

4.1.2. Les outils

La morphologie verbale fait ainsi l'objet d'une attention particulière. En effet, il n'est pas possible de disposer de toutes les formes verbales de tous les verbes dans toutes les localités. Aussi, afin de pouvoir étiqueter efficacement toutes les formes verbales dans MMS, y compris celles qui n'ont pas encore été recensées dans la base, un conjugueur automatique est en cours de conception. Cet outil devrait permettre de conjuguer n'importe quel verbe occitan dans tous les points d'enquêtes pour lesquels cela est possible. Les atlas linguistiques ne se sont malheureusement pas toujours penchés sur la question de la morphologie verbale, ce qui engendre de grandes lacunes dans ce domaine. L'originalité de ce conjugueur sera, à terme, de pouvoir reconstituer la conjugaison d'un verbe dans différents systèmes graphiques (y compris API) à partir de n'importe quelle forme conjuguée de ce verbe (et non à partir de l'infinitif seul). Par ailleurs, la numérisation de tous les dictionnaires occitans se poursuit, notamment en collaboration avec le milieu associatif provençal, ce qui permettra une annotation plus performante des occurrences.

Au fil du temps, l'utilisation de la base a fait apparaître plusieurs améliorations possibles à apporter aux fonctionnalités du THESOC. Ainsi, le transcripteur automatique doit être révisé

afin de permettre une meilleure transcription phonologique des parlers nord-occitans, le traitement des textes dans MMS pourra être simplifié en ajoutant des procédures automatisées (supprimant alors l'étape du 'lissage'), l'analyseur syntaxique sera amélioré (*via* l'implémentation d'un algorithme à base de programmation dynamique), en y ajoutant la gestion des achoppements (*disfluences*), qui constituent une spécificité de l'oralité que les analyseurs traditionnels, conçus pour traiter des textes écrits, ne peuvent traiter efficacement.

D'autres fonctionnalités sont également en perspective, telles que des outils permettant l'analyse phonologique diachronique[13] ou d'appliquer à MMS la présentation des phrases et des vidéos telle qu'elle figure sur le site web, de sorte qu'un clic sur une phrase dans MMS déclenche la lecture du passage de la vidéo contenant cette phrase.

Le THESOC devrait aussi bénéficier des résultats du projet ANR ECLATS[14] dans lequel une partie de l'équipe est engagée. L'objectif de ce projet est d'apporter un outillage logiciel et méthodologique facilitant l'extraction, l'analyse, la visualisation et la diffusion des données contenues dans l'*Atlas Linguistique de la France* (Gilliéron & Edmont 1902-1910) dont, à terme, la partie occitane sera intégrée au THESOC ainsi que les outils cartographiques élaborés. Tout cela permettra d'offrir une cartographie plus moderne et plus performante et d'améliorer les fonctionnalités cartographiques du THESOC, tant pour le lexique que pour les données morphosyntaxiques.

En outre, il s'agit maintenant de renforcer les interactions entre la base lexicale et la base MMS, qui à terme doivent fusionner et ne plus faire qu'une seule base. Cela concerne d'une part les référentiels communs aux deux bases, qui sont principalement les lemmes et les localités, dont la mise à jour s'en trouvera simplifiée. D'autre part, les textes de MMS peuvent compléter et enrichir la base lexicale, car ils contiennent souvent des termes que les auteurs des atlas n'ont pas cartographiés, voire qui ne figuraient même pas dans leurs questionnaires[15]. Enfin, le module de morphologie, qui est alimenté à la fois par les données du lexique et par les étiquetages de MMS, et qui est utilisé dans les deux bases, devra alors être complètement révisé dans une perspective de meilleure efficacité.

4.2. Le site web

Le site du THESOC, pour sa part, va bénéficier d'une refonte totale dans une perspective de modernisation et d'amélioration, tant pour la présentation des données que pour leur diffusion. Ainsi, dans un souci constant d'interopérabilité et d'ouverture des données, la mise en place en cours d'une interface dite 'API REST' permettra à terme aux développeurs du monde entier d'interroger à distance le THESOC afin de télécharger des résultats de recherche, afin d'en assurer une diffusion plus large. Plusieurs partenariats sont actuellement en cours dans cette perspective, avec des associations travaillant sur l'occitan dont les sites internet renverront à celui du THESOC, ou encore avec la plateforme EDISYN (*European Dialect Syntax* <http://www.dialectsyntax.org>) développée par le Meertens Institute (Université d'Amsterdam) qui a pour objectif de mettre à la disposition de la communauté scientifique toutes les données dialectales européennes en vue de l'étude syntaxique. A plus long terme, il peut être aussi envisagé de proposer une participation aux internautes en ajoutant à la consultation un champ

[13] Cf. Dalbera & *al.* (2012, 381).

[14] ANR ECLATS 2015-2019 *Extraction automatisée des Contenus géolinguistiques d'Atlas et analyse Spatiale : application à la dialectologie*, Institut Polytechnique de Grenoble (UMR 5217) & Université Stendhal, Grenoble (UMR 5216), sous la responsabilité de P.A. Davoine <http://www.agence-nationale-recherche.fr/projet-anr/?tx_lwmsuivibilan_pi2%5BCODE%5D=ANR-15-CE38-0002>.

[15] La philosophie des auteurs des atlas pour la publication a été de ne retenir que les cartes qui présentaient de l'intérêt du point de vue de la variation lexicale. De ce fait, une partie du lexique pourtant ordinaire et fréquent (par ex. le mot *famille*) ne figure pas dans les données provenant des atlas.

Commentaire, dans lequel ils pourraient soumettre des propositions ou des remarques concernant une entrée lexicale, une question, une occurrence, etc. L'intérêt de cette participation, contrôlée par le biais de modérateurs, devrait être de diffuser plus largement l'outil que représente le THESOC (en relation avec son caractère éminemment culturel et patrimonial), d'augmenter et de préciser les données, et d'offrir une aide ponctuelle à l'équipe développant la base de données.

L'interactivité du volet morphosyntaxique du site (pour l'instant dédiée aux données DADDIPRO) sera également progressivement améliorée, d'une part entre la vidéo et les transcriptions affichées sur la page (en faisant en sorte que soient successivement surlignées les transcriptions phonétiques des phrases produites dans la vidéo), et d'autre part entre la cartographie et les données elles-mêmes (en facilitant la navigation entre la carte des points d'enquête et les données d'une localité déterminée).

Les nouvelles technologies du web d'aujourd'hui permettront en outre de mettre en ligne progressivement de nouveaux outils et de nouvelles fonctionnalités présentes dans le logiciel monoposte mais non encore disponibles sur Internet, comme un module de cartographie interactive, ce qui est rendu possible par la technique du *Webmapping*.

Enfin, il s'agira également de renforcer l'interopérabilité entre la partie *lexique* et la partie *phrases*, en ajoutant chaque fois que possible des liens entre ces deux volets du site, et en donnant la possibilité de superposer, sur une même carte interactive, les données issues de la base lexicale et celles issues de la base MMS.

5. Conclusion

Imaginé il y a désormais 25 ans, le Thesaurus Occitan s'est tout d'abord imposé comme un pionnier dans l'aventure du corpus linguistique informatisé. Si l'implémentation informatique de la base de données est désormais entrée dans une phase de restructuration rendue nécessaire par l'évolution technologique, le principe qui a présidé à la conception du THESOC n'a pas changé : regrouper en un seul outil l'ensemble des données occitanes issues de l'oralité. Du reste, la fidélité au principe fondamental n'empêche pas l'adaptation aux nouveaux centres d'intérêts de la linguistique, comme le montre notamment la création de MMS : objet de deux financements successifs, le petit frère de la base lexicale a déjà permis de nombreuses avancées dans le domaine de la syntaxe dialectale. Enfin, la refonte du site Internet, actuellement en cours, reste le chantier le plus ambitieux : une fois celle-ci achevée, le grand public aura accès au million et demi de données que comporte la base lexicale, ainsi qu'aux nombreuses enquêtes dialectologiques contenues dans la base MMS.

Le THESOC poursuit ainsi son objectif d'être un outil à la fois ancré dans l'actualité et irremplaçable pour quiconque s'intéresse à l'oralité de la langue d'oc.

Michèle Oliviéri, Sylvain Casagrande, Guylaine Brun-Trigaud, Pierre-Aurélien Georges
Université Nice Sophia Antipolis-CNRS / UMR 7320 'Bases, Corpus, Langage'
Campus Saint Jean d'Angély
24 Avenue des Diables Bleus
06357 Nice Cedex 4, France
Tél : 04 89 88 14 49
<Michele.OLIVIERI@unice.fr>
<Sylvain.CASAGRANDE@unice.fr>
<Guylaine.BRUN-TRIGAUD@unice.fr>
<Pierre-Aurelien.GEORGES@unice.fr>

Références

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

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

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

Bouvier, J.C. (1992). La notion d'ethnotexte. In Pelen, J.N. & Martel, C. (éds), Les voies de la parole, ethnotexte et littérature orale, approches critiques. *Les cahiers de Salagon*, 1, 12-21.

Dalbera, J.P., Oliviéri M., Ranucci J.C. & *al.* (2012). La base de données linguistique occitane THESOC. Trésor patrimonial et instrument de recherche scientifique. *Estudis Romànics*, 34, 367-387.

Gilliéron, J. & E. Edmont (1902-1910). *Atlas Linguistique de la France*. Paris, Champion.

Mistral, F. (1878). Lou Tresor dou Felibrige ou Dictionnaire Provençal-Français. Aix-en-Provence, Veuve Remondet-Aubin.

Oliviéri, M. (2004). Le responsaire du THESOC. In Brasseur, P. (éd.), *Proceedings of the 8e Colloque de dialectologie et littérature du domaine d'oïl occidental*, Avignon, 12-13 juin 2002, Avignon, Université d'Avignon, 23-33.

Oliviéri, M. (2012). Le Mot et la chose : Réflexion sur le responsaire du THESOC. In Oliviéri, M., Brun-Trigaud, G. & Del Giudice, P. (éds), *La Leçon des dialectes. Hommages à Jean-Philippe Dalbera*, Alessandria, Edizioni dell'Orso, 13-32.

Rousselot, P. J. (1924). *Principes de phonétique expérimentale*. Paris, Didier.