Traitement syntaxique pour l'occitan

Pierre-Aurélien Georges

Rencontres

19-20.02.15

Les technologies pour les langues régionales de France

Délégation générale à la langue française et aux langues de France

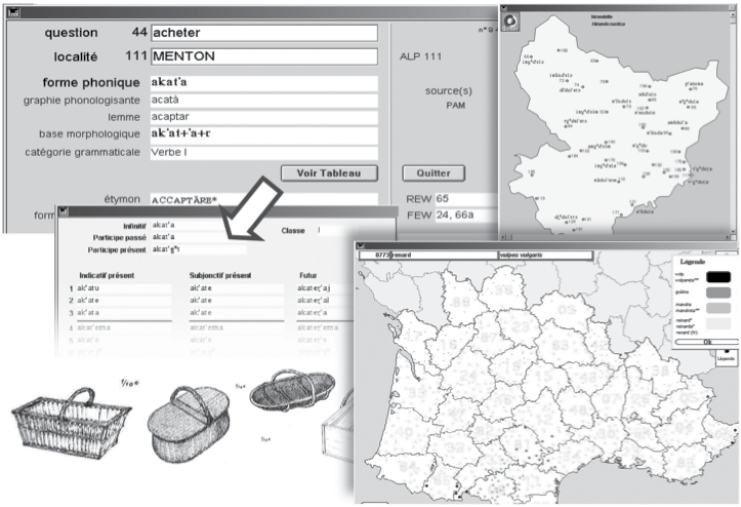
À l'occasion du colloque des 19 et 20 février 2015
Délégation Île-de-France Ouest et Nord du CNRS
Espace Isadora Duncan, Meudon

Traitement syntaxique pour l'occitan

Pierre-Aurélien Georges, université de Nice

Il ne s'agira pas ici d'évoquer le traitement syntaxique à proprement parler, mais plutôt la conception d'une base de données dédiée à la syntaxe et morpho-syntaxe des dialectes occitans. Ce sera l'occasion de présenter les pistes que nous avons suivies et les réflexions que nous avons eues concernant l'intégration d'un certain nombre d'outils de traitement linguistique sur cette base.

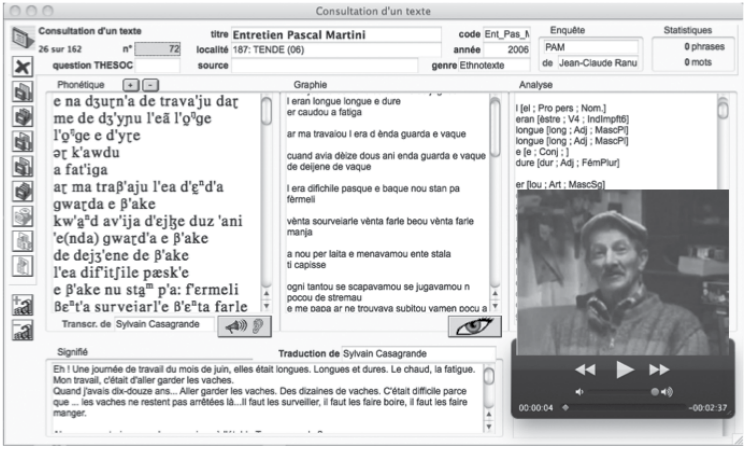
Le Thesaurus Occitan (ou Thesoc, en abrégé) est probablement plus connu pour sa base lexicale, développée au sein du laboratoire BCL[1] depuis 1992, et qui est d'ailleurs mentionnée dans l'inventaire des ressources linguistiques des langues de France (réalisé par l'ELDA, de mars 2013 à novembre 2014).



Captures d'écran de la base lexicale du Thesoc (version hors-ligne pour Windows)

1 UMR 7320 : Laboratoire Bases, Corpus, Langage ; CNRS / université Nice Sophia-Antipolis

Une grande partie des données de cette base sont disponibles sur le site internet (thesaurus.unice.fr), mais le Thesoc dispose également d'autres modules, tels qu'un volet de micro toponymie ainsi qu'un module de cartographie interactive, qui ne sont pour l'heure pas encore disponibles en ligne. C'est le cas également du module morpho-syntaxique (MMS), que nous allons maintenant évoquer. Il s'agit d'une base dédiée au monde de la recherche, pour les linguistes qui souhaitent travailler sur la morphosyntaxe et la syntaxe des dialectes occitans. Elle contient un corpus de phrases et d'ethnotextes, issus d'enquêtes linguistiques sur le terrain, qui sont constitués (entre autres) d'un enregistrement audio ou vidéo, d'une transcription phonétique associée (en alphabet phonétique international), d'une transcription graphique, ainsi que d'une traduction en français.



Capture d'écran du module morpho-syntaxique (MMS) du Thesoc

Notons au passage que ces données pourraient éventuellement intéresser quiconque souhaiterait mettre en place des systèmes de traduction automatique pour l'occitan, puisque le texte en occitan et la traduction en français sont ici disponibles et qu'il s'agit la plupart du temps d'une traduction phrase par phrase (l'alignement des deux ne devrait donc pas constituer de problème majeur et pourrait être envisagé).

Dans la base MMS, un certain nombre de traitements sont réalisés pour ajouter des annotations sur ces données brutes. Il y a par exemple un lemmatiseur, qui identifie chaque terme du texte et lui associe un lemme, une catégorie et une flexion. Par ailleurs, les données dans la base sont systématiquement géo-référencées, ce qui permet ensuite de générer des cartes à la demande, pour, par exemple, visualiser des zones de transition linguistique entre différents dialectes. Cet aspect diatopique intéresse tout particulièrement les chercheurs, car cela leur permet d'étudier la variation entre les dialectes, sur le plan syntaxique et morpho-syntaxique. Il est donc important que les outils linguistiques utilisés pour annoter les données sachent gérer cette dimension spatiale.

Au final, trois types de variations ont du être prises en compte dans la conception de cette base de données :

1. Puisque l'objectif du module MMS est de travailler sur la syntaxe et la morphosyntaxe de l'occitan, les fines variations phonétiques enregistrées dans la prononciation locale de tel ou tel dialecte ne sont pas au centre de nos préoccupations. C'est pourquoi nous avons pris le parti, dans cette base, d'opérer tous les traitements linguistiques à partir de la transcription graphique (présente à côté de la transcription phonétique) plutôt qu'à partir de la transcription phonétique (ce qui permet dans un premier temps de « gommer » cette variation phonétique) mais en laissant la possibilité de retrouver par la suite les différentes prononciations phonétiques d'un même mot, que ce soit dans les résultats de recherche ou dans la consultation des données.

2. Au niveau des traitements linguistiques effectués, les outils utilisés doivent être capables de gérer la variation dialectale. Par exemple, pour « bien », en provençal on observera plutôt le terme « *ben* » alors que du côté languedocien ce sera plutôt « *plan* » ; le lemmatiseur doit donc être capable de gérer ces variantes dialectales.

3. Une des particularités de l'occitan est qu'il y coexiste plusieurs systèmes graphiques différents (contrairement à des langues comme le français ou l'anglais, où l'orthographe y est fixée depuis longtemps). Ainsi, deux graphies se partagent la part du lion : il s'agit de la graphie dite « classique » ou « alibertine » et de la graphie mistralienne (celle de

l'école du *Felibrige*) ; mais il existe également d'autres graphies utilisées localement ou historiquement. Citons par exemple la graphie italianisante, du côté de Nice, qui n'est plus tellement utilisée aujourd'hui mais que l'on retrouve dans la majorité des textes en *nissart* écrits jusque dans les années 50. Si l'on souhaite pouvoir étudier tous ces textes, Il est donc intéressant d'avoir des outils de traitements linguistiques qui sachent gérer ces différentes graphies. La graphie utilisée dans un texte peut d'ailleurs être détectée automatiquement grâce à un outil que nous avons développé et intégré dans MMS.

Contrairement à d'autres projets de corpus occitans (tel que la base BaTelÒC, qui est plus dédiée à la littérature écrite), nous cherchons plutôt à travailler sur des données orales et dialectales, même si en théorie rien n'empêcherait d'utiliser notre module MMS pour traiter des données textuelles. Pour cela, il suffirait de ne pas remplir le champ phonétique, puisque de toutes façons tous les traitements sont réalisés à partir de la transcription graphique.

Le lemmatiseur que nous avons développé et qui est utilisé pour annoter les textes de la base fonctionne à partir de ressources lexicales. À terme, l'objectif, pour pouvoir gérer les différentes variantes dialectales, serait donc de disposer de ressources lexicales pour chacune des grandes variantes dialectales de l'occitan ainsi que pour chacune des principales graphies utilisées. Cela fait beaucoup de combinaisons, et *a priori* il peut paraître peu raisonnable de vouloir procéder de la sorte : cela risquerait en effet d'entraîner des problèmes au niveau de la performance du lemmatiseur, en causant un trop grand nombre d'ambiguïtés et donc une baisse des performances. Mais la géolocalisation des données permet de répondre à cette problématique, en évitant par exemple d'avoir recours à un dictionnaire des patois gascons lorsqu'il est question de lemmatiser un texte en niçois. Ainsi, il s'agit d'essayer d'utiliser, parmi les ressources lexicales déjà disponibles dans la base, plutôt celles qui sont si possible dans la même graphie et le plus géographiquement approprié. Les dictionnaires les plus éloignés du texte à lemmatiser se retrouvent donc en quelque sorte « temporairement désactivés » pour l'occasion.

Concernant l'origine des ressources lexicales utilisées par notre lemmatiseur, nous nous sommes essentiellement tournés vers la numérisation de

dictionnaires, que nous avons réalisée en interne, avec tout ce que cela implique de passage à l'OCR, correction et balisage pour obtenir un fichier XML qui a ensuite pu être intégré dans MMS. Une autre source est constituée par la base de données lexicales du Thesoc, qui contient plus d'un million d'entrées lexicales, qui sont là aussi géolocalisés. Sur ce dernier point, on peut parler en quelque sorte d'enrichissement mutuel des deux bases : en effet, nous utilisons les informations de la base lexicale pour le bon fonctionnement du lemmatiseur, et *in fine* les données ainsi traitées dans MMS ont vocation à venir alimenter en retour la base lexicale. À chaque fois que l'on trouve dans un texte lemmatisé une occurrence attestée d'un terme dans un lieu donné qui ne se trouve pas déjà dans le Thesoc, l'idée est d'ajouter cette attestation à la base lexicale. On retrouve également une telle boucle de rétroaction au niveau des dictionnaires : lorsqu'un terme est lemmatisé, la prononciation phonétique attestée dans telle ou telle localité lui est associée, ce qui permet petit à petit de venir rajouter les prononciations phonétiques dans un dictionnaire qui, à l'origine, n'en contenait pas. Au final, lors d'une recherche d'un terme dans un des dictionnaires intégrés dans la base, l'ensemble des prononciations phonétiques attestées pourra ainsi être listé.

Malheureusement, un des problèmes d'utilisation des dictionnaires est que les jeux d'étiquettes et de catégories grammaticales ne sont pas toujours les mêmes d'un dictionnaire à un autre. Nous sommes confrontés à des situations, où par exemple certains dictionnaires utilisent « verbe transitif » alors que d'autres utilisent « verbe du premier groupe ». La question est donc de savoir comment recouper ces jeux d'étiquettes pour obtenir un jeu homogène et cohérent pour le fonctionnement du lemmatiseur.

La même problématique se pose lorsqu'on s'intéresse aux questions d'interopérabilité, que ce soit pour travailler avec d'autres équipes de recherche, pour échanger des corpus, ou encore pour mettre en place un portail de recherche sur internet qui fédérerait plusieurs bases de données. Bien sûr il y a eu quelques tentatives de standardisation par le passé, comme dans le cadre du programme *Expert Advisory Group on Linguistic Engineering Standards* (EAGLES), où un jeu d'étiquettes relativement canonique a été établi, mais les chercheurs adaptent souvent ce jeu en fonction de leurs besoins scientifiques. Ils y apportent quelques modifications, et ces spécificités locales rendent ensuite difficile toute correspondance entre

les jeux d'étiquettes des différents projets de recherche. C'est en tout cas ce à quoi nous avons été confrontés à plusieurs reprises.

La question se pose alors de savoir pourquoi on observe une telle disparité dans les jeux d'étiquettes utilisés, d'un projet de recherche à un autre. Les enjeux du choix des étiquettes, pour savoir par exemple s'il vaut mieux constituer un jeu d'étiquettes réduit ou un jeu plus complexe, reposent sur différents critères, à savoir notamment :

- **l'interopérabilité** (avec d'autres logiciels, d'autres bases de données, d'autres équipes de recherche) ;
- **l'efficacité** des outils de traitement automatique qui reposent sur ces catégories grammaticales ;
- **la flexibilité** (laisser le choix à l'utilisateur ou au contraire lui imposer un cadre théorique) ;
- **l'ergonomie**, notamment dans le cadre d'une utilisation grand public (avoir par exemple des fonctionnalités de recherche qui soient facilement accessibles sans nécessiter la lecture d'un fastidieux manuel de prise en main) ;
- **la pertinence**, pour les chercheurs, des catégories ainsi retenues.

Les différents objectifs listés ici entre bien souvent en compétition les uns les autres (à titre d'exemple, d'un côté le souci d'efficacité des outils de traitement automatique utilisés et les besoins exprimés par les chercheurs voudraient parfois que l'on augmente le nombre de catégories, mais d'un autre côté, pour des questions d'interopérabilité et d'ergonomie pour les autres utilisateurs, il vaudrait mieux se limiter à un nombre réduit de catégories). On essaye donc généralement de trouver un compromis entre ces deux extrémités, ce qui aboutit alors à un jeu d'étiquettes consensuel, mais qui, dans les détails, ne satisfait réellement personne.

Après réflexion, nous avons donc opté pour une organisation hiérarchique du jeu d'étiquettes utilisé pour les catégories grammaticales, avec autant de niveaux de hiérarchie que l'on souhaite : au premier niveau, les différentes parties du discours puis, en descendant l'arborescence, les sous-catégories, et enfin, tout en bas de l'arbre, les spécificités locales voulues par les chercheurs.

Cela permet à la fois :

1. de détailler et de mieux préciser certaines choses pour le fonctionnement optimal de l'analyseur syntaxique (niveaux inférieurs de l'arbre) ;
2. d'assurer l'interopérabilité avec d'autres bases de données (niveaux supérieurs de l'arbre) ;
3. tout en permettant à l'utilisateur de la base de choisir le niveau de détail qu'il souhaite avoir pour effectuer ses recherches (niveaux intermédiaires), de manière similaire au principe du dictionnaire de la base MMS, qui est structuré en deux niveaux : lemmes et variantes.

L'on peut même envisager des héritages multiples : à titre d'exemple, l'étiquette « participe passé » hérite à la fois de « Adj » et de « V ». Ainsi, si l'on recherche (hors contexte) des verbes dans le dictionnaire, toute flexion confondue (temps/mode/personne), les résultats de recherche contiendront également les participes passés de ces verbes. Et réciproquement, si l'on recherche dans le dictionnaire tous les adjectifs, les participes passés employés en tant qu'adjectifs apparaîtront eux aussi dans les résultats de recherche. Dans cet exemple il s'agit bien évidemment d'une recherche « hors-contexte » (c'est-à-dire que l'on recherche un **terme** dans le dictionnaire, et non une **occurrence** dans une phrase ou un texte) : en revanche, une fois en contexte dans une phrase, chaque occurrence d'un participe passé fait partie ou bien d'un groupe adjectival ou bien d'un groupe verbal, mais pas les deux en même temps. Il faut donc bien distinguer la position syntaxique occupée par une occurrence dans une phrase de la catégorie morphosyntaxique référencée dans le dictionnaire.

Avec un tel système d'étiquettes hiérarchiques, le fonctionnement est alors le suivant :

• à l'importation d'une nouvelle ressource lexicale dans notre base de données, l'on fait correspondre le jeu d'étiquette de cette ressource lexicale externe avec certaines de nos étiquettes, préférentiellement situées le plus possible dans les niveaux inférieurs de notre hiérarchie d'étiquette, mais il y a bien souvent « sous-spécification » : lorsque le jeu d'étiquettes de cette ressource externe n'est pas suffisamment détaillé pour pouvoir être mis en correspondance avec le niveau le plus détaillé de notre hiérarchie d'étiquettes, on doit alors se contenter de catégories

grammaticales un peu plus vagues, et les raccrocher aux branches qui sont situées plus haut dans notre arborescence, faute d'avoir pu « rentrer dans les détails » pour préciser la sous-catégorie exacte ;

• en interne, on utilise, chaque fois que possible, les étiquettes situées plutôt dans les niveaux inférieurs de la hiérarchie d'étiquette (c'est le jeu d'étiquettes le plus détaillé que possible), mais lorsque l'information n'est pas disponible (par exemple car il s'agit de données importées depuis une autre base de données, qui utilise un jeu d'étiquette plus réduit), le lemmatiseur ou l'analyseur syntaxique savent aussi utiliser en dernier recours les niveaux supérieurs de la hiérarchie d'étiquettes, quitte à ce que les résultats fournis par ces outils de traitement automatique soient alors de moins bonne qualité (mécanisme de *fall-back*) ;

• à l'exportation de données (provenant de notre base MMS et à destination d'autres bases) ou lorsqu'il y a nécessité d'interopérabilité (par exemple pour s'interfacer avec un moteur de recherche qui fédère plusieurs bases de données), on peut choisir le jeu d'étiquettes que l'on souhaite utiliser dans l'export (ou dans l'interfaçage), de manière à le faire correspondre à celui de la base de données dans laquelle on souhaite intégrer *in fine* ces données : on pioche, dans les différents niveaux hiérarchiques de notre jeu d'étiquettes interne, les catégories à exporter (celles qui possèdent une correspondance exacte dans l'autre base de données).

Dalbera (1980) a par exemple montré que pour pouvoir discriminer les groupes adverbiaux correctement formés de ceux qui sont agrammaticaux (« très bien » versus « *bien très »), il est nécessaire de distinguer en français six classes d'adverbes. Malheureusement, la plupart du temps, dans les ressources lexicales à notre disposition, tous les adverbes sont classés dans une seule et même catégorie fourre-tout « adverbe ». L'objectif de cette organisation arborescente des catégories grammaticales est donc de pouvoir répondre à cette problématique, à savoir : améliorer l'efficacité des outils de traitement automatique (lorsque les données à notre disposition le permettent) tout en conservant une bonne interopérabilité, à la fois au niveau des imports et des exports depuis/vers d'autres bases de données (notamment pour se laisser la possibilité d'utiliser toutes les ressources lexicales disponibles).

Sur ce dernier point, nous nous situons à la fois en tant que consommateurs et producteurs, pour les trois aspects suivants :

• **Ressources lexicales** ; en effet nous numérisons des dictionnaires papier au format XML pour les besoins de fonctionnement de notre lemmatiseur, et nous serions intéressés à échanger ces dictionnaires avec d'autres pour compléter les ressources lexicales disponibles dans notre base.

• **Corpus annotés** ; puisque l'objectif in fine de cette base MMS est la constitution d'un corpus oral pour l'ensemble des dialectes occitans, mis à disposition de la communauté scientifique et du grand public, et que si d'aventure des corpus oraux seraient localement disponibles pour tel ou tel dialecte, nous serions ravis de pouvoir les intégrer dans la base MMS.

• **Outils de traitement automatique** ; qu'il s'agisse d'outils de détection automatique de la graphie utilisée dans un texte, de transcripteur automatisé « phonétique vers graphie », de lemmatiseur, d'analyseur syntaxique, nous avons développé en interne un certain nombre d'outils et nous serions également intéressés de pouvoir intégrer dans cette base MMS d'autres outils de traitement automatique pour l'occitan qui seraient développés par la communauté, afin d'enrichir ou d'améliorer les fonctionnalités déjà disponibles.

Indications bibliographiques communiquées par l'auteur

BRAS, Myriam, et Marianne VERGEZ-COURET, 2014, BaTelÒc : a Text Base for the Occitan Language, *First International Conference on Endangered Languages in Europe, Oct 2013, Minde, Portugal*.
https://hal.archives-ouvertes.fr/hal-00987241

DALBERA, Jean-Philippe, 1980, Esquisse d'une Classification Syntaxique des Adverbes Français, in *Travaux du Cercle Linguistique de Nice*, Université de Nice, Nice, p. 39-60.

DALBERA, Jean-Philippe, Guylaine BRUN-TRIGAUD, Michèle OLIVIERI et Jean-Claude RANUCCI, 2012. La base de données linguistique occitane Thesoc. Trésor patrimonial et instrument de recherche scientifique. Estudis Romànics 34, 367-387.

GEORGES, Pierre-Aurélien, 2010. The Thesaurus Occitan: a multimedia database dedicated to occitan dialects. Presentation of its morphosyntax module. Actes du colloque *Tools for Linguistic Variation (EUDIA-2)*, édité par Jose Luis ORMAETXEA et Gotzon AURREKOETXEA, p. 107-118. Bilbao : ASJU-ren gehigarriak, LIII, UPV-EHU.

LEECH, G., R. BARNETT, et P. KAHREL, EAGLES Recommendations for the Syntactic Annotation of Corpora, 1996. EAGLES DOCUMENT EAG–TCWG–SASG/1.8

LEIXA Jérémy, Valérie MAPELLI et Khalid CHOUKRI, 2014, Inventaire des ressources linguistiques des langues de France, ELDA (Agence pour la Distribution des ressources linguistiques et l'Évaluation).

Actes du colloque « Les technologies pour les langues régionales de France », organisé les 19 et 20 février 2015 à l'espace Isadora Duncan, Meudon, France par la délégation générale à la langue française et aux langues de France, le laboratoire de recherche en informatique pluridisciplinaire (LIMSI) - Centre national de la recherche scientifique (CNRS) et l'Institut des technologies multilingues et multimédias de l'information (IMMI).

Ministère de la Culture et de la Communication
Délégation générale à la langue française et aux langues de France
6, rue des Pyramides 75001 Paris
téléphone : 01 40 15 73 00 / télécopie : 01 40 15 36 76
courriel : dglflf@culture.gouv.fr
www.culturecommunication.gouv.fr/Politiques-ministerielles/Langue-francaise-et-langues-de-France

Délégué général
Loïc Depecker

Délégué général adjoint
Jean-François Baldi

Organisation du colloque
Gilles Adda, IMMI-CNRS
Lucie Gianola, DGLFLF
Thibault Grouas, DGLFLF
Joseph Mariani, LIMSI-CNRS
Quentin Samier, ELDA-ELRA

Captation des débats
Cellule Webcast – Centre de calcul IN2P3 / CNRS

Transcription des débats et constitution des actes
Lucie Gianola, DGLFLF

Coordination générale du projet
Thibault Grouas, DGLFLF

Coordination éditoriale
Pauline Chevallier, DGLFLF

Graphisme
Claire Méry, Micaela Neustadt, DGLFLF

Délégation générale à la langue française et aux langues de France
6, rue des Pyramides
75001 Paris
téléphone : 01 40 15 73 00
télécopie : 01 40 15 36 76
courriel : dglflf@culture.gouv.fr
www.culturecommunication.gouv.fr/Politiques-ministerielles/Langue-francaise-et-langues-de-France