Outils pour la géolinguistique automatisée

Gotzon Aurrekoetxea (UPV/EHU) & Charles Videgain (UPPA-Iker)

La géolinguistique a parcouru un vaste itinéraire durant ces dernières décennies, surtout dans l'utilisation de ressources informatiques comme aide au traitement et à l'analyse de données. Il existe actuellement un vaste éventail d'équipes de recherche qui utilisent quotidiennement de tels outils informatisés. L'équipe de recherche EUDIA (*EUskal DIAlektologia*) compte sur différents outils pour le traitement de l'information dialectale, outils qui fonctionnent *online*: l'outil "CorpusLem" est un élément qui transforme les textes en données. Il permet de travailler en *online* comme en local, en déchargeant un fichier *xls*, lequel peut ensuite être chargé sur le réseau. La base de données EDAK est un outil de gestion des données, qui supporte en lui-même diverses bases de données et qui, à partir des outils de gestion des données, est consultable sans avoir besoin de s'enregistrer. La base héberge actuellement près de 100.000 données sur toutes les variétés de l'euskara ou langue basque. Ces données peuvent servir à des analyses diatopiques ou diastratiques. La base de données héberge des données audiovisuelles, de telle forme que l'usager puisse en même temps lire et écouter la réponse intégrée dans la base de données. Enfin, l'outil cartographique et dialectométrique "DiaTech" possède divers modules: module de la base de données elle-même, module statistico-quantitatif et module cartographique. Il s'agit aussi d'un outil *online*. Le module cartographique a deux options: l'une de cartographie thématique avec possibilité d'un atlas parlant et une autre de cartographie thématique. L'outil permet l'importation de bases de données provenant d'autres projets géolinguistiques, la réalisation d'analyses statistiques et une cartographie quantitative.

1 Géolinguistique automatisée: état de la question

Dans le champ de la dialectologie on tient généralement pour assuré le fait que la quantification des données a commencé avec les travaux quantitatifs de Jean Séguy et son souci de trouver l'espace qui réunisse l'essence du gascon. Hans Goebl comme Henri Guiter ont poursuivi leurs travaux dialectométriques et c'est sans doute H. Goebl qui a su internationaliser cette problématique et pousser au développement informatisé de cette méthode quantitative. Peu de chercheurs doutent absolument de la valeur heuristique de la DM comme outil de l'analyse diatopique et le nombre d'applications informatiques ne fait qu'augmenter.

La dialectométrie a été l'une des avancées les plus significatives de la géolinguistique du XXème siècle. Sans doute l'une des avancées les plus importantes. La possibilité d'utiliser une grande quantité de données pour l'étude de la variation linguistique permet de dépasser la situation endémique de la dialectologie dès ses débuts, c'est-à-dire des études basées uniquement sur des critères et des variables linguistiques peu nombreuses. L'étude des frontières dialectales est beaucoup plus accessible et plus probante avec cette nouvelle méthodologie.

Dans un autre ordre de choses, les travaux de l'atlas parlant ou sonore ont dû attendre que la technologie de l'étude du signal audio soit d'accès facile pour les dialectologues. Les premiers travaux remontent aux années 1990, comme le sait fort bien Roland Bauer, un des pionniers dans le domaine.

Nous avons nous-même eu connaissance du projet d'atlas sonore en 1992 à l'occasion d'un congrès organisé au sein d'Euskaltzaindia / Académie de la langue basque et dont nous étions les responsables académiques. Lors de ce congrès, le Professeur H. Goebl donna deux conférences, l'une sur les atlas sonores et l'autre sur la DM.

Dans sa conférence sur les atlas sonores intitulée "L'atlas parlant dans le cadre de l'Atlas linguistique du ladin central et des dialectos limitrophes (ALD)" (cfr. Goebl 1992), il avait montré la possibilité de présenter les données directement depuis la réalisation

du locuteur interrogé jusqu'à l'oreille des enquêteurs-transcripteurs ou du grand public. Quiconque pouvait dorénavant, depuis son bureau ou son appareil, écouter les données audiovisuelles brutes de l'enquête. C'était là un pas de géant pour les études géolinguistiques et qui rendait possible la modernisation du savoir-faire dialectologique des atlas linguistiques.

Cette avancée souhaitée par les responsables techniques de l'Atlas *Euskal Herriko Hizkuntz Atlasa* (EHHA) basque n'a pas été possible en son temps au sein de l'Académie de la langue basque, mais nous avons pu développper d'autres projets qui prennent en compte cette dimension, comme le projet EDAK.

L'atlas EHHA poursuit sa publication de documents écrits (cartes, responsaires, index) et quatre volumes ont paru à ce jour dont le numéro 4 en septembre 2012, et nous travaillons à l'élaboration des cartes du tome 9. L'atlas est publié en version papier (en peu d'exemplaires) et en format numérique. La version *en ligne* se trouve à l'adresse web de l'Académie basque (www.euskaltzaindia.net/dialektologia). La matière des livres peut être téléchargée en pdf mais on peut y faire des recherches de cartes par réponses ou par lemme. Actuellement, une difficulté est à signaler aux non bascophones: l'atlas est à ce jour uniquement rédigé et consultable en basque.

La dialectométrie comme les atlas parlants et sonores se sont fondés sur les progrès de la technologie, en tirant profit de ces possibilités techniques : la dialectométrie avec une statistique automatisée et informatisée, les seconds par la facilité d'enregistrement et de digitalisation de la voix humaine.

2 CorpusLem

Le groupe de recherche EUDIA, créé au sein de l'Université du Pays Basque, mais dont font partie des chercheurs d'autres universités (Université de Pau et des Pays de l'Adour (UPPA), Bamberg) a pour objectif l'étude de la variation linguistique (géo-et sociolinguistique) de la langue basque et la création d'outils nécessaires pour cette étude.

L'un de ces premiers outils créé a été "CorpusLem", consultable sur le site http://aholab.ehu.es/CorpusLem/login.html). Cet outil a été créé en collaboration avec le centre IKER (Centre de recherche sur la langue et les textes basques) UMR (Unité Mixte de Recherche) CNRS (Centre national de la recherche scientifique, Bayonne) et a voulu répondre à la nécessité de traiter des textes dialectaux en les transformant en un format "Donnée' lors du projet Bourciez (Aurrekoetxea/Videgain 2009, Aurrekoetxea 2011, Aurrekoetxea/Videgain/Iglesias 2004) puis lors du projet Sacaze (Aurrekoetxea 2011, Aurrekoetxea/Videgain 2012).

Dans les deux projets, nous traitons des textes dialectaux d'origine géographique homogène mais produits par des traducteurs différents et selon des orthographes non normées. Dans le projet Bourciez, il s'agit du texte bien connu de la Parabole de l'enfant prodigue, dans le projet Sacaze de deux légendes: La légende de Barbazan et La légende de Tantugou. Les textes ont été recueillis à la fin du XIXème siècle et concernent 150 localités bascophones, et nous n'avons pas pris à notre charge les textes rédigés en occitan ou catalan.

Notre intention de réaliser une exploitation géolinguistique de ces textes nous obligea à passer du texte original en format texte au format donnée pour introduire le texte dans une base de données et une visualisation cartographique des données. Ce passage du texte vers le format donnée est réalisé à partir du programme CorpusLem. Le programme sert aussi à la création de dictionnaires dialectaux.

L'outil fonctionne en 5 langues (basque, anglais, français, castillan et catalan) et il est d'une grande simplicité puisque ne comportant que deux écrans. Le premier écran sert à introduire l'utilisateur et son mot de passe, le second écran sert à convertir les textes de format *.tst* à une base de données au format MySQL), avec la possibilité de copie de données, élaboration de dictionnaires, etc.

2.1 Ecran de présentation

Il est indispensable d'être enregistré pour utiliser cet outil. La marche à suivre est rapide.

2.2 Ecran de travail

L'écran de travail est partagé en sections (fig. 1):

- Manuel d'instructions (à droite de l'écran)
- Champ de gestion des projets, des utilisateurs, des bases de données.
- Champ des projets (à droite sur l'écran):
- Copie de la base de données du projet
- Projet actuel
- Gestion du projet
- Gestion des règles internes au projet

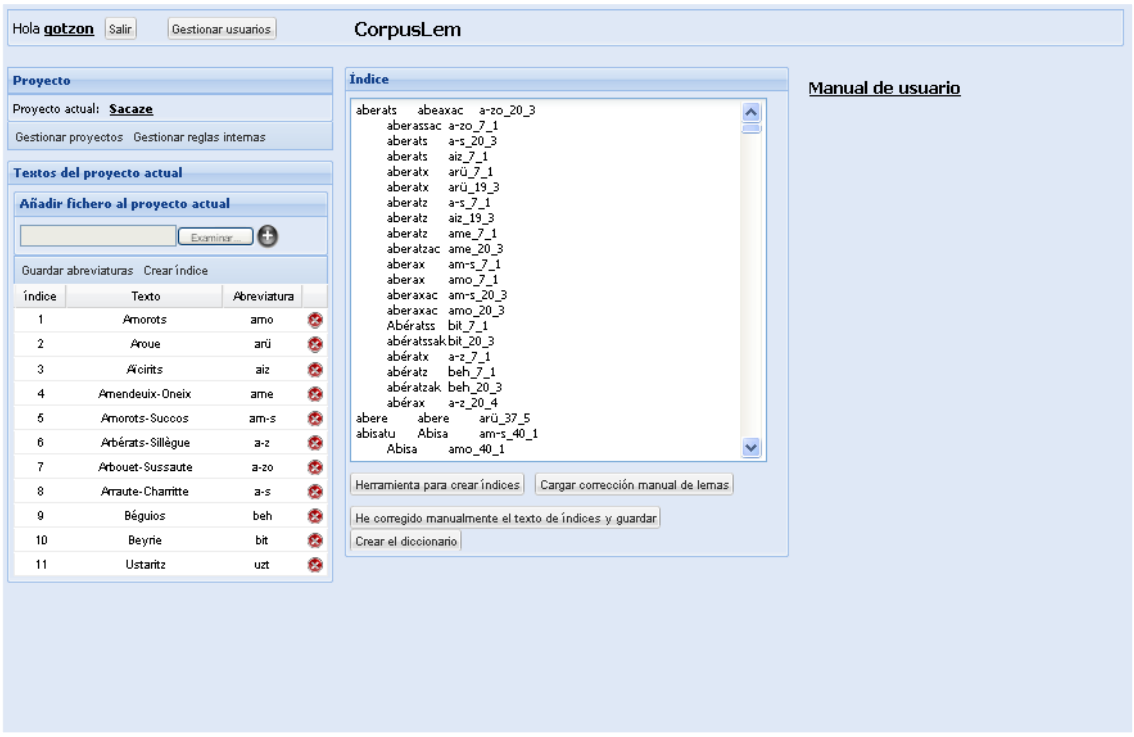

Fig. 1. L'outil *CorpusLem* : ecran de travail

Au moyen de ces règles internes, le chercheur peut adapter la graphie des textes en accord avec les buts du projet. Quatre options sont possibles (fig. 2):

a) 'Equivalent". Sous ce titre, si un vocable possède un grand nombre de variantes qui peuvent être représentées en une seule, le système choisit cette variante comme principale et y réunira les autres variantes. C'est ainsi que sous le factitif basque 'arazi', le système regroupera les variantes 'aazi' (avec perte de la vibrante intervocalique) ou 'aaci' ou 'aasi'.

b) Un autre groupe de règles essaie de réguler les terminaisons, dont on sait qu'elles sont nombreuses de par le système de déclinaisons en langue basque. Sous la variante 'arazi' indiquée ci-dessous, on trouvera donc des formes déclinées comme 'arazia', 'aazia', 'arazten'.

c) "Joindre". Le système joint des termes qui sont séparés dans le texte traité. Ainsi 'aitasso amassoac' ('père et mère') seront réunis par un trait d'union.

d) "Séparer". Ce cas consiste inversement à séparer des termes qui sont écrits d'un seul tenant dans le texte traité. En particulier, les verbes périphrastiques voient souvent l'auxiliaire lié au verbe conjugué et notre règle opère la séparation.

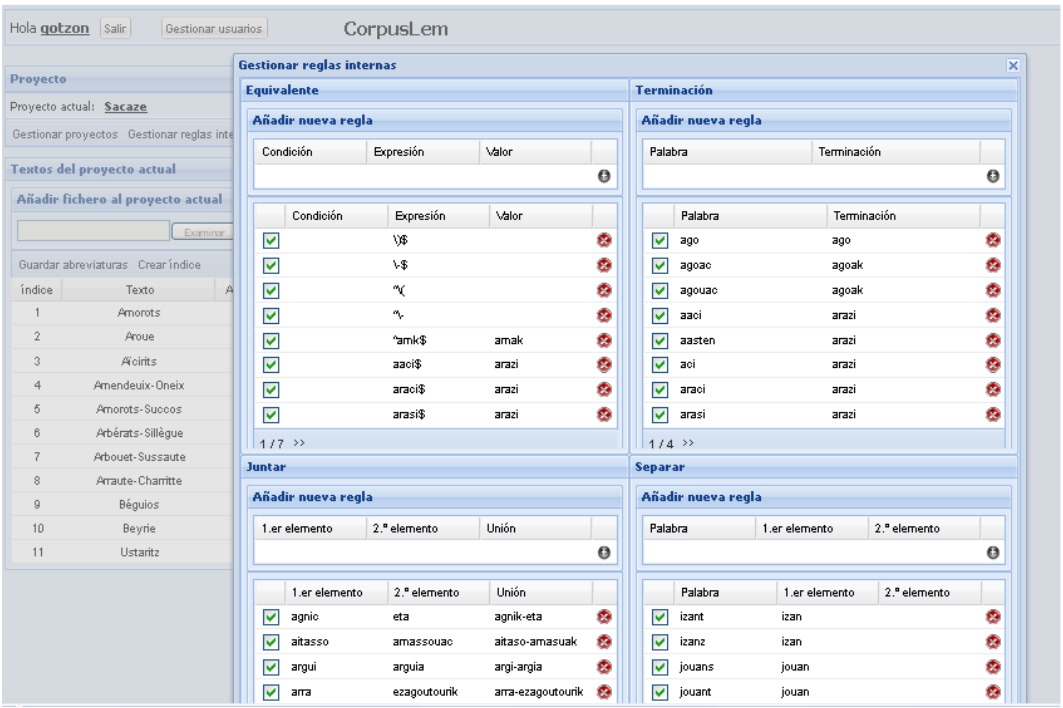

Fig. 2. L'outil *CorpusLem* : gestion des règles internes

- Textes du projet actuel, dans lesquels sont présents les textes actifs de chaque projet, avec la possibilité d'ajouter des documents, d'en effacer, etc.
- Champ de travail (centre de l'écran). C'est le centre de travail, à travers lequel le chercheur peut construire le dictionnaire des formes de tout le texte. Le chercheur peut le faire à partir de l'outil lui-même ou le charger en format *xls*, travailler en local et ensuite l'introduire de nouveau dans l'outil à partir du bouton "charger la correction manuelle des lemmes", et ensuite créer le dictionnaire des formes à partir de l'option "créer dictionnaire".
- Enfin, l'outil crée une base de données MySQL avec les données des textes.

3 Le corpus EDAK: l'atlas sonore basque

Le deuxième instrument créé par le groupe EUDIA est une base de données transcrites et sonores. Le corpus EDAK est un projet mené par le groupe de recherche EUDIA dont l'objectif est l'analyse de la variation. Cette analyse ne se limite pas à la variation géolinguistique, mais aussi la variation sociolinguistique et stylistique, chaque fois que la nature des données le permet.

L'un de ses projets est le "Corpus oral du basque dialectal-EDAK/Euskara Dialektalaren ahozko corpusa" hébergé sur le site http://aholab.ehu.es/edak/2/.

3.1 Caractéristiques du corpus

Ce corpus contient actuellement près de 100 000 registres. C'est le seul corpus oral en basque dialectal accessible sur le Web. Le corpus est assez réduit mais d'une grande richesse et ses caractéristiques essentielles sont les suivantes :

- 403 questions ainsi réparties:
- 120 questions relevant du lexique commun.
- 62 questions de morphologie nominale et verbale.
- 20 questions de syntaxe.
- 179 questions sur l'accent dans les mots monosyllabiques, bissyllabiques, trisssyllabiques, selon la déclinaison à l'indéfini, au singulier ou au pluriel.
- 22 questions sur l'intonation, 12 dans des phrases énonciatives, 6 interrogatives totales, 4 interrogatives partielles.
- Les points d'enquête sont au nombre de 100, 75 dans la partie péninsulaire, 25 dans la partie continentale.
- Les informateurs sont classés en adultes ou jeunes, donc de deux générations différentes, hommes pour le lexique, la morphologie, la syntaxe, et femmes pour la prosodie. On sait que l'étude de la variation sociolinguistique est d'autant plus importante que la langue basque est en situation de nivellement linguistique fort, depuis la création d'un basque unifié ou standard en 1968 et l'implantation récente d'un système d'enseignement de la langue basque, soit dans le réseau des ikastola (écoles privées immersives) soit dans le réseau public, sans oublier le rôle des médias en langue basque. Les diverses générations ont donc un profil fort différent; les adultes ont souvent été scolarisés selon le modèle espagnol ou français (scolarité exclusivement en espagnol ou français), tandis qu'une bonne part des jeunes sont passés par une scolarisation en langue basque à un degré variable. Le Pays Basque est donc un remarquable laboratoire pour l'étude du nivellement linguistique et les mécanismes qu'il génère mais aussi pour observer les formes et variétés plus conservatrices.
- Caractéristiques de la base de données :
    - Format: MySQL
    - L'information est retranscrite selon l'alphabet IPA à partir d'un écran incrusté dans la base de données, de telle sorte que l'utilisateur n'a pas

à saisir un signe phonétique mais à le choisir sur l'écran et à les copier dans la base.

L'information acoustique, sur l'annotation-étiquetage du signal sonore se fait à travers les programmes *SFSWin*. L'information acoustique peut être écoutée à partir d'un programme de son (Praat) et disposée pour être utilisée dans des analyses acoustiques.

Toute l'information indiquée ici est disponible sur le réseau à l'intention de tous les utilisateurs, sans nécessité de s'identifier dans l'accès à la base de données (creative common licencia)

La base de données EDAK est un outil de gestion des données, au format MySQL, qui supporte diverses bases de données et qui, à partir des outils de gestion des données, offre deux nouveaux modules: le module d'aide à la lemmatisation des données, et le module de conversion du format base de données au format corpus (format TEI).

Le module d'aide à la lemmatisation est nécessaire pour une exploitation linguistique ou géolinguistique des données: la lemmatisation permet une cartographie aréale, en utilisant divers outils de visualisation. Dans notre cas, nous avons recours de préférence à la couleur, mais d'autres options sont possibles.

Le module de conversion du format MySQL au format TEI est indispensable aujourd'hui. Il permet de passer d'un format privé, individuel à un format socialisé international. La dialectologie ne peut se dispenser de franchir ce palier en ayant recours à cette technologie.

Cette base de données héberge aussi bien des données écrites que des données audio, si bien que l'utilisateur peut lire ou écouter la réponse dans la base de données.

Les études facilitées par cette bases sont présentées dans les colloques ou congrès et publiées dans les diverses revues spécialisées.

4 L'outil dialectométrique DiaTech

L'outil cartographique et dialectométrique *DiaTech* utlise divers modules: module de données propres (Base de données EDAK), modèle statistico-quantitatif et module cartographique. Ce dernier propose l'option de préparer un atlas parlant et l'option de cartographie de faits synthétiques ou dialectométriques.

4.1 Menu principal

Tout au long de l'application Web, le menu principal reste visible en haut de l'écran et fournit les principaux champs de l'application. Outre ces différents champs, la possibilité est donnée de changer de langue de travail, de changer le mot de passe de l'utilisateur et celle de quitter le programme.

Les champs principaux sont les suivants:

- Début: dans le champ "début", est fourni l'objectif de l'application Diatech sur le Web.
- Projets: dans le champ "projets", apparaît le nom des projets à gérer et utiliser
- Aide: dans ce champ, est donné le guide d'utilisation de Diatech
- Compte: ce champ attribue un compte à chaque utilisateur
- Langues: le programme peut être utilisé en plusieurs langues, basque, espagnol, anglais, choix à enrichir ultérieurement.

4.2 Les projets

Le champ "Projet" est le champ principal du programme. Dans ce champ chaque utilisateur peut sélectionner le projet qu'il gère et en créer de nouveaux.

A chaque projet créé est attribué nécessairement un nom. Ce nom doit être propre au projet et avant d'être retenu, il faut que le système vérifie qu'il ne peut être confondu avec un autre projet existant.

Il faut aussi préciser si le projet est ouvert ou non en ce sens que si le projet est dit "ouvert" il donne la possibilité à tous les utilisateurs du programme Diatech d'utiliser ce projet mais non pas de le gérer. Ces utilisateurs pourront avoir accès à la recherche d'informations dans le projet, à l'élaboration de statistiques mais ne pourront pas modifier les données du projet.

Enfin est donnée la possibilité de fournir une description succincte du projet et de ses objectifs.

En bas de l'écran apparaissent les projets des différents utilisateurs, donc les projets créés par les utilisateurs, ceux ouverts et auxquels les utilisateurs peuvent avoir accès, et ceux soumis à autorisation pour accès et utilisation. Une table permet de détailler cette problématique pour rechercher, filtrer et sélectionner les divers programmes, connaître le nom du programme, le nom de son créateur ou propriétaire, les divers types d'autorisation accordés. De plus, pour chaque projet, des raccourcis sont mis en place. Un champ permet de chercher les réponses, un autre d'établir des statistiques, un autre de gérer la base de donnée, un autre de gérer le projet et enfin un autre d'effacer la sélection du programme.

4.3 Gestion du projet

Ce champ est consacré à la gestion du projet. Apparaissent le nom du projet, l'information sur le fait que le projet soit ouvert ou non, et la possibilité de modifier la description du projet. Il est possible d'importer le projet, de gérer les invitations au projet, et d'entrer dans un champ réservé aux commentaires produits par différents utilisateurs du projet.

4.4 Importation de la base de données.

Quand un utilisateur crée un projet, la base de données est vide. Il peut y introduire les données l'une après l'autre dans le programme ou bien les importer à partir de la base de données dans son ensemble si cette dernière est déjà créée.

Chaque projet crée une base de données adaptée à ses données. En conséquence, les diverses bases de données de l'application peuvent être différentes. Aussi pour favoriser leur importation, les fichiers texte au format CSV sont acceptés. Quatre fichiers peuvent être acceptés dans un seul fichier comprimé par ZIP (fig. 3).

Chaque fichier texte est ainsi organisé.

4.4.1 Localisation.csv:

On donne ici la liste des communes relatives aux données. Outre les données propres à chaque lieu d'enquête, on pourra donner leurs coordonnées géographiques. Le fichier comprendra les lignes suivantes:

- **Id**: identificateur de la commune, signalé par un code numérique
- **Localisation**: nom de la commune

- **Latitude**: coordonnée de la latitude de la commune, en chiffre jusque la décimale
- **Longitude**: coordonnée de la longitude de la commune, en chiffre jusque la décimale

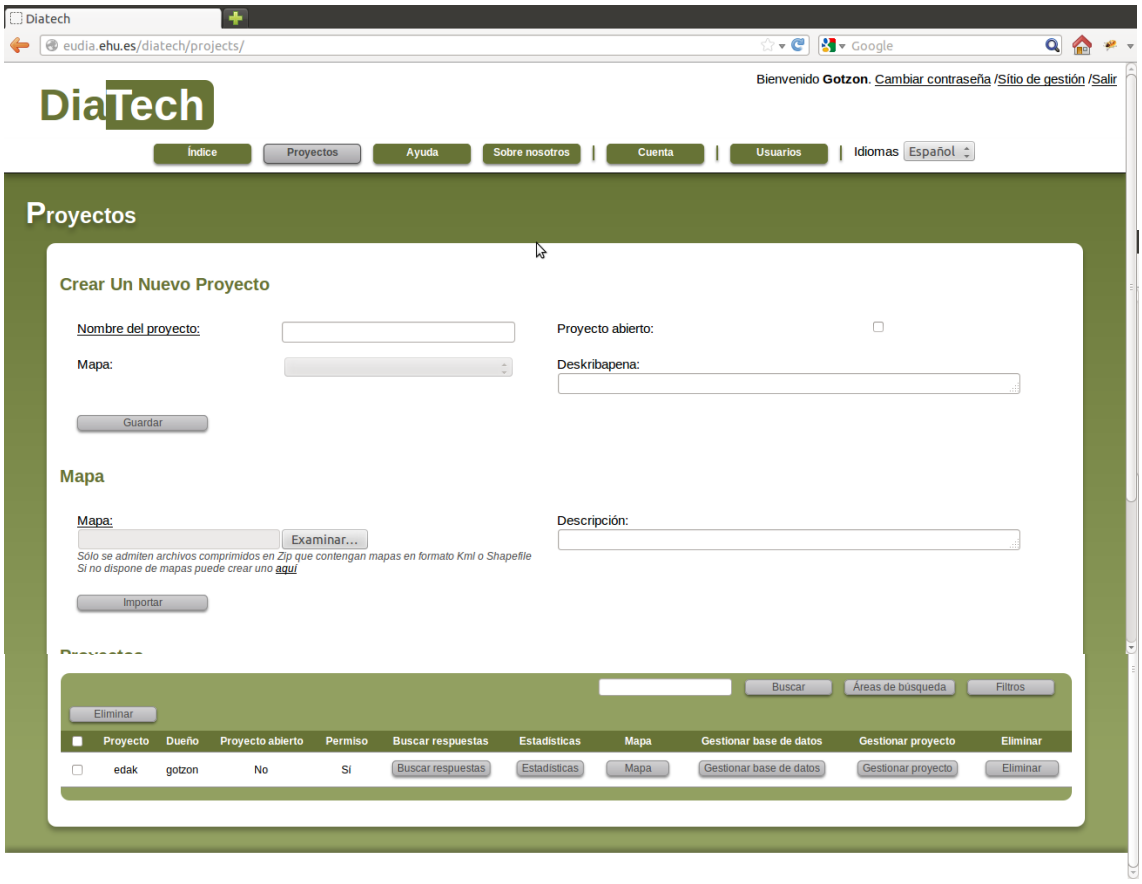

Fig. 3. L'outil *DiaTech:* Gestion de nouvels projects.

4.4.2 Informant.csv (locuteur):

Ici on donne la liste des personnes ayant servi de témoins d'enquêtes, avec l'identification de leur commune. Les lignes suivantes sont à renseigner:

- **Id**: identificateur du locuteur, signalé par un code numérique
- **Nom**: patronyme du locuteur
- **Prénom**: prénom (s) du locuteur
- **Localisation**: identificateur de la commune d'origine du locuteur, en code numérique (qui doit être le même que dans location.csv).
- **Sexe**: sexe du locuteur, signalé en chiffres. 0: ignoré. 1: masculin. 2: féminin.
- **Classe d'âge:** catégorie d'âge, en chiffres (0: non choisi. 1: jeune. 2: adulte. 3: âgé.

4.4.3 Question.csv

Ici on donne les informations relatives aux questions posées au locuteur. Outre les identificateurs et le domaine linguistique, les traductions en chaque langue doivent apparaître. Les lignes suivantes sont donc à construire en fonction de chaque langue et sont:

- **Id**: identificateur de la question, signalé par un code numérique
- **Champ_linguistique**: domaine linguistique choisi, signalé par code numérique. 0: non choisi. 1. phonologie. 2. morphologie nominale. 3. morphologie verbale. 4. syntaxe. 5. lexique.
- **Domaine linguistique**: basque, anglais, castillan... Une ligne est consacrée à la traduction dans la langue concernée. Le titre de la ligne sera le nom de la langue traduit en anglais.

4.4.4 Réponse.csv

Ici on donne les informations relatives aux réponses fournies par les locuteurs. Outre les informations relatives à la réponse, apparaissent aussi les identificateurs du locuteur et de la question. Les lignes suivantes sont à renseigner.

- **Id**: identificateur de la réponse, signalé par code numérique.
- **Question**: identificateur de la question, signalé par code numérique (doit être le même que celui donné dans Réponse.csv).
- **Locuteur**: identificateur du locuteur, signalé par code numérique (doit être le même que celui donné dans Locuteur.csv).
- **Contexte**: contexte de la question
- **Réponse orthographique**: réponse en code orthographique
- **Réponse-phonétique**: réponse en code phonétique
- **Lemme**: lemme de la réponse.

4.4.5 Exportation de la base

Le gestionnaire du projet a la possibilité de faire une copie de sécurité de la base de données et de la déposer sur son ordinateur. L'information sera tenue dans quatre fichiers comprimés par Zip. Les signaux audio et leurs liens avec les réponses ne seront pas sauvegardés sur cette copie.

4.4.6 Localisation.csv:

Ici on donne la liste des communes relatives aux données. Outre les données propres à chaque lieu d'enquête, on pourra donner leurs coordonnées géographiques. Le fichier comprendra les lignes suivantes:

- **Id**: identificateur de la commune, signalé par un code numérique
- **Localisation**: nom de la commune

- **Latitude**: coordonnée de la latitude de la commune, en chiffre jusque la décimale
- **Longitude**: coordonnée de la longitude de la commune, en chiffre jusque la décimale

4.4.7 Informant.csv (locuteur):

Ici on donne la liste des personnes ayant servi de témoins d'enquêtes, avec l'identification de leur commune. Les lignes suivantes sont à renseigner:

- **Id**: identificateur du locuteur, signalé par un code numérique
- **Nom**: patronyme du locuteur
- **Prénom**: prénom (s) du locuteur
- **Localisation**: identificateur de la commune d'origine du locuteur, en code numérique (qui doit être le même que dans location.csv).
- **Sexe**: sexe du locuteur, signalé en chiffres. 0: ignoré; 1: masculin. 2: féminin
- **Classe d'âge:** catégorie d'âge, en chiffres (0: non choisi. 1. jeune. 2. adulte 3. âgé.

4.4.8 Question.csv

Ici on donne les informations relatives aux questions posées au locuteur. Outre les identificateurs et le domaine linguistique, les traductions en chaque langue doivent apparaître. Les lignes suivantes sont donc à construire en fonction de chaque langue et sont:

- **Id**: identificateur de la question, signalé par un code numérique
- **Champ_linguistique**: domaine linguistique choisi, signalé par code numérique. 0: non choisi. 1. phonologie. 2. morphologie nominale. 3. morphologie verbale. 4. syntaxe. 5. lexique.
- **Domaine linguistique**: basque, anglais, castillan…. Une ligne est consacrée à la traduction dans la langue concernée. Le titre de la ligne sera le nom de la langue traduit en anglais.

4.4.9 Réponse.csv

Ici on donne les informations relatives aux réponses fournies par les locuteurs. Outre les informations relatives à la réponse, apparaissent aussi les identificateurs du locuteur et de la question. Les lignes suivantes sont à renseigner:

- **Id**: identificateur de la réponse, signalé par code numérique.
- **Question**: identificateur de la question, signalé par code numérique (doit être le même que celui donné dans Réponse.csv).
- **Locuteur**: identificateur du locuteur, signalé par code numérique (doit être le même que celui donné dans Locuteur.csv).
- **Contexte**: contexte de la question
- **Réponse orthographique**: réponse en code orthographique

- **Réponse-phonétique**: réponse en code phonétique
- **Lemme**: lemme de la réponse

4.5 Invitations

Le gestionnaire du projet peut procéder ici à des invitations. La liste des utilisateurs apparaît ainsi que celle des autorisations. L'utilisateur sélectionné peut être invité.

4.6 Commentaires

Un espace est réservé ici aux commentaires que le ou les gestionnaires du projet souhaitent formuler. Quand un commentaire est rédigé, il est ensuite envoyé par mail aux utilisateurs du projet.

4.7 Gestion de la base

On peut ici sélectionner les possibilités de sauvegarder les données, de les actualiser ou de les supprimer. Cet espace est divisé en quatre parties: communes, locuteurs, questions, réponses.

4.8 Communes

On peut ici définir la latitude et longitude de la commune. Une carte permet d'obtenir ces coordonnées.

Sur cette carte apparaissent les communes qui ont déjà été saisies dans la base de données mais on peut y ajouter un symbole pour intégrer de nouvelles communes. En activant ce symbole on peut saisir les coordonnées de la commune correspondante.

Si les coordonnées sont déjà connues, la commune sélectionnée peut être représentée sur la carte en actionnant le bouton. Au-dessous de la carte apparaît la liste des communes à gérer. Des communes peuvent être aussi supprimées.

4.9 Locuteurs

Pour identifier un locuteur nouveau, il faut informer les champs relatifs à la commune, le nom, le prénom, le sexe et la classe d'âge du locuteur puis les sauvegarder.

Sur la table en bas d'écran apparaît la liste des locuteurs de la base. Des locuteurs peuvent aussi être supprimés.

4.10 Questions

Pour identifier une question, il faut informer le domaine linguistique. On le saisit ensuite dans la langue de travail choisie.

Sur la table en bas d'écran apparaît la liste des questions. Des questions peuvent aussi être supprimées.

4.11 Réponses

Avant de saisir une réponse, il est possible d'enregistrer l'enregistrement sonore correspondant. Cet enregistrement peut être un extrait dans lequel est intégrée la réponse. Dans ce cas il faut saisir tout l'enregistrement. A la réponse sont assignés le nom du locuteur, la question, la réponse orthographique, la réponse phonétique, le contexte et le lemme. Si on veut ajouter un extrait de l'enregistrement sonore, il faut choisir parmi la

liste des enregistrements qui ont été saisis celui dont on veut obtenir l'importation. Par le jeu des boutons, on fixe le début et la fin de chaque enregistrement correspondant à la réponse. Ensuite on saisit la réponse. L'enregistrement sélectionné peut être supprimé.

Sur la table en bas d'écran apparaît la liste des réponses de la base. Des réponses peuvent aussi être supprimées.

4.12 Recherche

Ici on peut opérer une recherche parmi les informations de la base. Pour opérer cette recherche on peut jouer sur les éléments suivants: question, réponse orthographique, réponse phonétique, commune, sexe, classe d'âge. Quand on lance la recherche, les réponses apparaissent en bas d'écran.

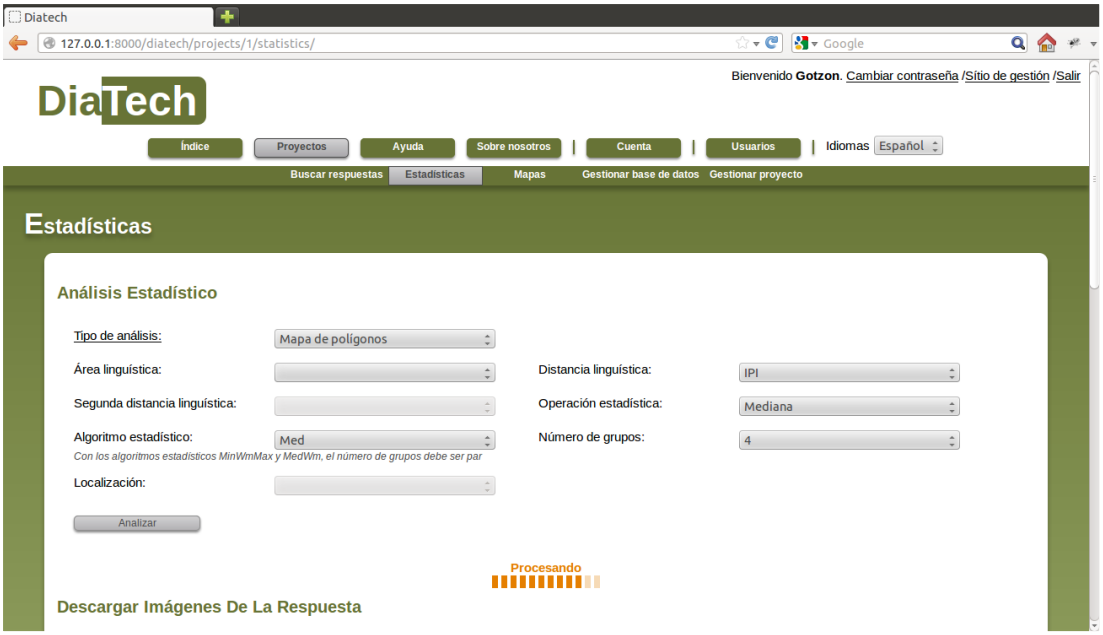

Fig. 4. L'outil DiaTech: statistiques

4.13 Statistique

Ici on peut opérer un travail statistique reposant sur quatre éléments: cartes à polygones, carte à barycentres, carte des limites et arbre. Il faut préciser les choix à opérer pour réaliser ces opérations. Ces choix dépendent pour une part d'entre eux du type d'analyse recherchée, d'autres sont indépendants de ce choix (fig. 4) :

- Champ linguistique: dans toutes les opérations il est possible de choisir un domaine linguistique ou d'analyser toutes les réponses fournies.
- Indice de distance: on peut chercher les distances IPI, IRI, et Levensthein.
- Opérations statistiques: on peut chercher dérive standard, asymétrie, moyenne, maximale, minimale, et corrélation.
- Deuxième champ linguistique: si la corrélation est sollicitée comme opération statistique, un deuxième domaine linguistique peut être activé.
- Algorithmes statistiques: on peut étudier les algorithmes Med, MinWmMax, MedWm.

- Analyse multidimensionnelle: on peut construire un arbre à partir des choix de Ward, Average et Complete.
- Nombre des groupes; on peut choisir de un à dix groupes.

4.14 Compte

Ici chaque utilisateur peut modifier son compte, et donc modifier son nom, son prénom et son e-mail. Le statut de l'utilisateur en relation avec ses autorisations d'accès est indiqué, un utilisateur ne pouvant s'attribuer davantage d'accès que ce qui lui est assigné par les autorisations qu'il a obtenues.

Cet outil fournit une solution au problème des réponses multiples, à ce jour casse-tête de bien d'applications informatiques en dialectologie. Nous pouvons cartographier les cas où plus d'une réponse a été fournie sur un point d'enquête et y apporter un traitement statistique. D'autre part, cet outil permet d'utiliser les bases de données d'autres projets géolinguistiques de telle sorte que n'importe quel chercheur puisse charger ses données dans le système, et puisse cartographier ces données ou en réaliser un traitement dialectométrique qu'il peut ensuite conserver sur son propre ordinateur.

5 L'avenir de la dialectologie

Il est peu concevable que, dans les années qui viennent, la dialectologie puisse se passer du support informatique et n'ait pas recours aux divers outils qui sont mis en place dans le domaine.

Le dialectologue doit nécessairement créer les outils indispensables à ce champ de la connaissance. Ce travail devrait prendre la voie de grandes équipes de telle sorte que nous puissions fournir des outils de travail satisfaisants pour les situations les plus diverses, à partir des outils actuels et en réunissant les forces créatrices capables de les améliorer.

Nous pensons que la phase de groupes réduits peut laisser place à des consortiums plus puissants pour développer des outils de plus grande ampleur et relativement meilleur marché, ce qui n'est pas négligeable en situation économique difficile.

Aussi avons-nous décidé de présenter un projet européen à partir d'un consortium auquel nous invitons tous les chercheurs intéressés à participer autour des points suivants:

a) Création d'un programme informatique sur la base de ceux existant sur le marché.

b) Engagement à importer des données de projets propres dans la base de l'outil à créer.

Nous remercions d'avance tous ceux et celles qui voudront participer à cet évènement.

Bibliographiques

Aurrekoetxea, G. (2011) : "CorpusLem" una herramienta para la conversión de corpus textuales en datos". In M.L. Carrió Pastor & M. A. Candel Mora (eds.), *Las TIC: Presente y futuro en el análisis de Corpus*, Valencia: Universitat Politècnica de València, 611-618. http://www.upv.es/upl/U0547372.pdf (http://alfpro.cc.upv.es:8080/alf

resco/d/d/workspace/SpacesStore/189a8fff-c6da-4c79-bfc1-ad645b17ac38/index.html#/611/zoomed.

Aurrekoetxea, G./Videgain. X. (2004) : *Seme Prodigoaren Parabola Ipar Euskal Herriko 150 Bertsiotan [La Parábola del hijo pródigo en 150 versiones vascas recogidas en el País Vasco-francés]*, ASJUren gehigarriak, EHU, Bilbo (2004) [también en http://klasikoak.armiarma.com/testuak/testuakBourciez001.htm].

Aurrekoetxea, G./Videgain, X. (2009) : "Le projet Bourciez: Traitement géolinguistique d'un corpus dialectal de 1895", in : *Dialectologia* 2, 81-111. (http://www.publicacions.ub.es/revistes/dialectologia2/).

Aurrekotxea, G./Videgain, X./Iglesias, A. (2004): *Bourciez Bildumako Euskal Atlasa (BBEA-1): 1. Lexikoa. [El atlas lingüístico Bourciez: 1. Léxiko]*, ASJU 38:2 [ed. 2007].

Aurrekotxea,G./Videgain, X/Iglesias, A. (2005) : *Bourciez Bildumako Euskal Atlasa (BBEA-2): 2. Gramatika. [El atlas lingüístico Bourciez: 2. Gramática]*, ASJU 39-1 [ed. 2008].

Aurrekoetxea, G. et al. (2012): "DiaTech": A New Tool for Dialectology" (aceptado para su publicación en *Literary and Linguistic Computing*).

Aurrekoetxea, G., Sánchez, J./Odriozola, I. (2009): "EDAK: A Corpus to Analyze Linguistic Variation", in : Cantos Gómez, P./Sánchez Pérez, A. (arg.), 2009, *A Survey on Corpus-based Research Panorama de investigaciones basadas en corpus,* Asociación Española de Lingüística del Corpus, 489-503. (http://www.um.es/lacell/aelinco/contenido/pdf/34.pdf).

DiaTech: http://eudia.ehu.es/diatech/login/?next=/diatech/index/.

EDAK: http://aholab.ehu.es/edak/2/.

GABMAP: http://www.gabmap.nl/.

Goebl, Hans (1992): "L'atlas parlant dans le cadre de l'Atlas linguistique du ladin central et des dialectes limitrophes (ALD) ", in : Aurrekoetxea, G./Videgain, X. (eds.) : *Nazioarteko Dialektologia Biltzarra*, Iker, 7, Bilbao.

Goebl, Hans (2010): "Introdución a los problemas y métodos según los principios de la Escuela Dialectométrica de Salzburgo (con ejemplos sacados del 'Atlante Italo-Svizzero', AIS)", in : Aurrekoetxea, G./Ormaetxea, J. L. (eds.), *Tools for Linguistic Variation*, Bilbao: UPV/EHU, 3-39.

MySQL: http://www.mysql.com/.

Praat: http://www.fon.hum.uva.nl/praat/.

SFSWin: http://www.phon.ucl.ac.uk/resource/sfs/.

TEI: http://www.tei-c.org/release/doc/tei-p5-doc/en/html/TS.html.