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

Ministère de la Culture et de la Communication

Délégation générale à la langue française et aux langues de France

Les technologies pour les langues régionales de France

En partenariat avec le **CNRS**, le laboratoire de recherche en informatique pluridisciplinaire (**LIMSI**), l'Institut des technologies multilingues et multimédia de l'information (**IMMI**), l'association européenne pour les ressources linguistiques (**ELRA**), l'agence pour l'évaluation et la distribution des ressources linguistiques (**ELDA**) et l'équipex **Ortolang** (Outils et Ressources pour un Traitement Optimisé de la Langue).

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

Cet ouvrage présente les travaux du colloque organisé les 19 et 20 février 2015 par le ministère de la Culture et de la Communication et le LIMSI-CNRS, en collaboration avec l'agence ELDA-ELRA et l'équipex Ortolang.

Pendant deux jours se sont réunis des experts, scientifiques, représentants des collectivités territoriales et du milieu associatif des différentes régions de France et d'Europe autour de la question de l'accompagnement technologique des langues régionales de France.

Nous nous sommes attachés à retranscrire aussi fidèlement que possible les présentations des intervenants et les débats avec la salle, à partir, notamment, des captations vidéos réalisées par les services du CNRS et intégralement disponibles sur le site internet : webcast.in2p3.fr/events-tlrf

Les transcriptions des présentations des intervenants reproduites dans cet ouvrage ont fait l'objet d'une relecture et d'une validation par leurs auteurs, et nous les en remercions.

Sommaire

Journée du 19 février

Approches scientifiques : traitement linguistique et automatique

Variation et norme : des transferts linguistiques aux transferts technologiques
Philippe Boula de Mareüil, laboratoire LIMSI-CNRS

Le projet RESTAURE
Delphine Bernhard, Laboratoire LILPA – université de Strasbourg
Marianne Vergez-Couret, Laboratoire CLLE-ERSS – université de Toulouse II

Le cas des créoles français : mutualisation des ressources pour des dialectes apparentés
Pascal Vaillant, université Paris XIII

Traitement syntaxique pour l'occitan
Pierre-Aurélien Georges, université de Nice

Développement de ressources syntaxiques : retour d'expérience et méthodologies
Eric de la Clergerie, INRIA, équipe Alpage

Traitement de la parole et traduction pour les langues sous-dotées
Fethi Bougares, université du Maine

Wikipedia comme ressource
Rémi Mathis, association Wikimédia France

Journée du 20 février

Traitement des langues régionales : que peuvent faire les acteurs publics ou privés en charge de l'accompagnement des langues régionales et les collectivités territoriales ?

Table ronde animée par ***Benaset Dazeàs***
(Lo Congrès permanent de la lenga occitana)

Olier Ar Mogn, Office public de la langue bretonne / Ofis publik ar Brezhoneg
Nourdine Combo, Conseil général de Mayotte
Gaëtan Crespel, Dastum
David Grosclaude, Conseiller régional d'Aquitaine
Sébastien Quenot, Collectivité territoriale de Corse / Capiserviziu di u Cunsigliu linguisticu

Ouverture du colloque

Président de séance : **Jean-François Baldi**

Introduction des débats

Jean-François Baldi, délégué général adjoint à la langue française et aux langues de France

Bonjour à tous.

Je crois qu'il est temps de commencer la première journée de ce colloque qui va se tenir sur deux jours. Je suis très heureux de parler devant la magnifique photo d'Isadora Duncan, puisqu'on m'a appris que c'était ici qu'elle répétait, preuve que la science et les arts se rejoignent dans cet endroit. Je suis très heureux ici de vous dire tout le plaisir que j'ai à ouvrir ce colloque au nom de la ministre de la Culture et de la Communication, Fleur Pellerin.

La délégation générale à la langue française et aux langues de France (DGLFLF), que je représente ici, vit un moment un peu particulier de son histoire administrative puisqu'il se trouve qu'elle n'a pas de délégué général, en raison de la nomination de Xavier North au poste d'inspecteur général des affaires culturelles et de l'absence de nomination de son successeur. Cette absence sera je pense très prochainement comblée. Il me revient donc à moi Jean-François Baldi, qui suis l'adjoint de Xavier North, d'ouvrir ce colloque, ce que je fais avec un infini plaisir.

En matière linguistique, notre organisation administrative en France a une double singularité. La première est de disposer au sein de l'État d'un service dédié à la promotion du français et de la diversité linguistique. Il faut y voir je crois le signe que l'État, qui est le garant de l'intérêt général, entend jouer un rôle moteur sur ces questions. Il y a un marché des langues, des concurrences entre les langues, les langues ont des « poids » différents, comme disent parfois certains linguistes. Il revient

à l'État de réguler ce marché, de faire en sorte qu'il n'y ait pas d'abus de position dominante et d'organiser une coexistence entre les langues aussi harmonieuse que possible.

Pour cela, l'État produit des normes, des textes, qui traduisent dans l'ordre juridique les droits accordés aux locuteurs des langues dans notre pays. Mais l'action de l'État ne se limite pas à ce domaine.

Il contribue également à créer des ressources, en particulier sur le plan numérique, raison pour laquelle nous nous réunissons aujourd'hui, pour que les langues et notamment les langues régionales de France puissent exprimer les réalités scientifiques, technologiques, culturelles, politiques de nos sociétés, en somme, dire le monde dans sa complexité.

Mais l'État n'est pas le seul acteur de cette politique, loin s'en faut : l'organisation de ce colloque en témoigne. Les collectivités territoriales, les organismes de recherche, qu'ils soient publics ou privés, les institutions chargées de la valorisation des langues régionales jouent un rôle éminent. Il est très heureux que cette manifestation soit l'occasion de réunir un grand nombre de ces acteurs pour faire un point sur ce que les technologies apportent aux langues régionales de France.

J'ai évoqué une double singularité. La première est celle d'un service, le nôtre, dédié à la promotion de la diversité linguistique au sein de l'État, mais la seconde singularité est d'avoir placé ce service au sein du ministère chargé de la culture. Il faut voir dans ce rattachement un sens politique et symbolique très fort. En effet, si les langues sont des outils de communication, elles sont aussi et d'abord des réalités culturelles. Si d'ailleurs ces langues n'étaient que des outils de communication, nous pourrions nous entendre pour communiquer dans une seule d'entre elles et faire l'économie de toutes les autres. Ce n'est pas la voie que nous avons choisie. C'est donc leur nature culturelle qui nous fait agir pour que les langues, et notamment les langues régionales, constituent un élément à part entière de notre patrimoine, soient des vecteurs pour la création, la transmission des savoirs et la circulation des idées.

Les ressources numériques précisément servent toutes ces dimensions propres aux langues. C'est la raison pour laquelle nous avons créé il y

a un peu plus de trois ans une mission des langues et du numérique au sein de la DGLFLF. Les chantiers qui ont été conduits, coordonnés ou soutenus par cette mission mais aussi par l'ensemble de la DGLFLF ces dernières années sont très nombreux. J'en citerai juste quelques-uns : la conservation et la valorisation du patrimoine linguistique à travers la base de données Corpus de la Parole[1], la mobilisation des ressources offertes par le web des données qui a débouché sur le projet JocondeLab[2], permettant la consultation en quatorze langues dont quatre langues régionales de France de la base de données des musées de France, l'identification des outils d'aide à la traduction les plus adaptés aux environnements de travail et aux besoins des traducteurs. Ce sont beaucoup de travaux qui ont été lancés ces dernières années à notre initiative en partenariat avec un grand nombre d'acteurs.

Sur le thème qui nous occupe plus précisément aujourd'hui, notre action s'inscrit dans un cadre stratégique qui a été plus précisément fixé par deux événements très importants pour nous : d'une part, les états généraux du multilinguisme dans les Outre-Mer à Cayenne en décembre 2011, où nous nous sommes penchés sur la question des technologies pour les langues d'Outre-Mer et où nous avons pu identifier dans ce domaine de très nombreuses lacunes, et d'autre part, le comité consultatif pour la promotion des langues régionales et de la pluralité linguistique interne. Je crois qu'il y a dans cette enceinte au moins un membre de ce comité consultatif, dont les conclusions ont été remises à la ministre de la Culture en juillet 2013. Au nombre de ces conclusions figure la nécessité d'équiper ces langues en technologies de nature à en conforter la fonctionnalité.

J'ajoute que notre politique s'inscrit également dans un cadre multilatéral : par exemple l'UNESCO suit avec attention la mise en œuvre par la France de la recommandation émise il y a quelques années déjà sur la promotion du multilinguisme dans le cyberespace.

Nous pourrions également citer la conférence internationale organisée à Iakoutsk l'année dernière, en juin 2014, sous l'égide notamment de l'UNESCO, conférence portant sur la diversité culturelle et

1 http://corpusdelaparole.huma-num.fr/
2 http://jocondelab.iri-research.org/jocondelab/

linguistique dans le cyberespace. Cette conférence a débouché sur une recommandation dont je vous cite un extrait qui fait très directement écho à ce que je disais au tout début de mon propos : « La relation entre langue et culture est celle de l'inter-dépendance, la culture est fonction de la langue, et la langue est vecteur de culture. L'une n'existe pas sans l'autre. »

J'en citerai également deux autres extraits, caractérisant notre action et ayant tout à fait leur place dans ce colloque :

« La mise au point des politiques linguistiques doit englober les langues officielles, les langues nationales et régionales, et aussi les langues des migrants. »

« Le développement des technologies de traitement du langage humain constitue un pas décisif vers la garantie de chances numériques égales pour toutes les langues. »

Les nombreux travaux, études, expériences qui vont être présentés durant ces deux journées s'inscrivent donc parfaitement dans ces préconisations. Ils ont vocation pour nous à éclairer la politique publique en faveur de la pluralité linguistique, car nous concevons la recherche comme telle.

Je terminerai mon propos en remerciant chaleureusement les organisateurs de cette manifestation : le CNRS bien évidemment, et en particulier le laboratoire de recherche en informatique pluridisciplinaire (LIMSI), l'agence pour l'évaluation et la distribution des ressources linguistiques (ELDA), l'équipe de la mission des langues et du numérique de la DGLFLF, Thibault Grouas, la mission des langues de France, dirigée par Michel Alessio et l'observatoire des pratiques linguistiques.

Les langues de France aujourd'hui

Michel Alessio, délégation générale à la langue française et aux langues de France

L'intitulé du colloque associe aux langues régionales de France le terme de « technologies ». C'est un rapprochement qui pourrait paraître surprenant, tant on est encore habitué à lier ces langues aux modes les plus traditionnels de la communication, mais la rencontre d'aujourd'hui vient heureusement invalider les représentations dépassées, de langues trop souvent perçues comme inaptes à un traitement de type moderne.

Non que les langues soient toutes également avancées dans leur développement numérique, il s'en faut de beaucoup, mais ce que nous confirment les travaux techniques et scientifiques des uns, l'expérience des autres sur le terrain, c'est que ces disparités ne tiennent évidemment pas à je ne sais quelles qualités intrinsèques à telle ou telle langue, mais à des circonstances d'ordre politique, économique ou démographique.

Or, comme le disait Félix Castan, « Toutes les langues sont égales entre elles, comme les citoyens d'une même république ». C'est une déclaration de principe, et si l'on veut faire entrer dans les faits l'égalité formelle et théorique des langues, s'attaquer au retard numérique des moins pourvues d'entre elles est une voie d'action privilégiée.

La DGLFLF a pour mission de faire entrer la pluralité linguistique interne, les langues de France, dans le combat pour le plurilinguisme en général, à l'échelle de l'Europe et du monde. Tout ce qui peut être fait pour la défense et l'illustration du basque, du breton, du catalan, du créole martiniquais, de l'occitan ou du mahorais est un pas vers le maintien de l'indispensable diversité des ressources langagières de la planète.

Dans cet esprit, il nous appartient de veiller à ce que les savoir-faire acquis et les outils élaborés dans le traitement du français puissent être réemployés dans le maniement des langues régionales et des autres langues en usage dans notre pays.

En 2013, un comité consultatif avait été mis en place par la ministre de la Culture pour éclairer les pouvoirs publics dans la construction d'une politique publique en faveur de la pluralité linguistique. Il recommandait entre autres de travailler à mieux connaître la situation concrète des langues en France.

Il va de soi que tout ce que vous nous ferez découvrir du traitement automatique des langues régionales, de ses lacunes et de ses progrès, ira dans ce sens et contribuera à la consolidation d'une véritable politique publique des langues en France.

Parallèlement à ce chantier, de notables avancées ont été réalisées : l'Assemblée nationale a voté des textes qui renforcent la place de nos langues dans le système éducatif, ou qui envisagent l'adhésion de la France à la Charte européenne des langues régionales ou minoritaires ; le Ministère a publié un recueil de textes législatifs et réglementaires relatifs aux langues régionales, et réaffirmé leur rôle dans la politique linguistique qu'il est chargé d'animer et de mettre en œuvre.

Le colloque d'aujourd'hui et demain vient donc s'inscrire dans cette visée d'ensemble de promotion de la pluralité des langues, qui forme le fond de la grande réflexion et de l'action dans lesquelles les services de l'État sont actuellement engagés aux côtés des institutions et associations que vous représentez.

On peut considérer qu'il marque une étape importante dans ce processus. Que les organisateurs en soient donc remerciés.

Enjeux des technologies du langage pour les langues régionales

Thibault Grouas, délégation générale à la langue française et aux langues de France

Merci à tous d'avoir répondu présent pour ce colloque, inédit je crois, car il n'y a jamais eu en France de colloque sur les technologies pour les langues régionales. On peut s'en réjouir, et également espérer que ce ne sera pas le dernier.

Qu'entend-on par « technologies du langage » ?

Cette expression peut recouvrir la traduction automatique, la synthèse vocale, la reconnaissance vocale, mais aussi des technologies plus classiques que l'on oublie un peu mais qui entourent notre quotidien, permettant par exemple d'aider à la rédaction, la correction orthographique ou grammaticale, la suggestion automatique de mots pour accélérer la saisie lorsque l'on écrit avec un téléphone…

Toutes ces technologies constituent un enjeu culturel considérable. En effet, une langue est beaucoup plus facilement accessible, traduisible, manipulable quand l'on dispose de ces technologies que quand l'on n'en dispose pas. Il importe donc de s'assurer qu'elles soient disponibles pour une langue donnée. Si l'on considère le français, en termes technologiques, c'est à peu près satisfaisant, même si nous restons beaucoup moins dotés que l'anglais. Mais quand nous regardons le niveau des ressources disponibles pour les langues régionales, l'on s'aperçoit qu'il est extrêmement faible, très inégal, voir inexistant pour certaines langues. Il y a donc un intérêt, nous allons voir lequel, à encourager le développement de technologies pour ces langues.

Quels sont les objectifs que l'on poursuit lorsque l'on s'intéresse aux technologies pour les langues régionales ?

Le premier enjeu est à la fois culturel et pédagogique. Le développement des technologies pour les langues régionales permet de valoriser un patrimoine régional colossal, sur le plan culturel et linguistique, qui existe aujourd'hui et qui pourrait être moins accessible demain, voir même disparaître. C'est donc une priorité culturelle très forte.

C'est aussi une priorité pédagogique, puisque, lorsque l'on dispose d'outils pour une langue, par exemple des outils relatifs à l'orthographe ou à la grammaire, l'on peut l'enseigner plus facilement, et cette langue peut par ailleurs se structurer de manière un peu plus formelle. C'est donc une aide pédagogique pour la transmission de cette langue.

Un autre intérêt que l'on peut constater concerne la numérisation de documents. Si l'on dispose des technologies adaptées de reconnaissance de texte dans une langue donnée, on peut plus facilement passer d'un corpus scanné à un corpus écrit en mode texte, facilement accessible sur l'internet via les moteurs de recherche par exemple. Cela permet ainsi d'améliorer la visibilité d'un patrimoine.

Un autre objectif, de nature différente, est l'objectif social. Comme l'on a pu le constater lors des États Généraux du Multilinguisme dans les Outre-mer[1], qui se sont déroulés à Cayenne en décembre 2011, une langue fait face à des réalités sur le terrain, réalités qui doivent absolument être prises en compte. Par exemple, la relation entre le service public et les usagers se révèle très critique, et, dans le cas de services tels que les services d'urgence, si la personne qui appelle ne maîtrise pas bien le français et qu'elle s'exprime dans une langue régionale telle qu'une langue créole, cela pose un problème. Il faut en effet que les pompiers puissent intervenir même s'ils ne comprennent pas bien le sens de l'appel, le lieu d'intervention ou la situation urgente. Il s'agit donc là d'un réel besoin social de réponse à des attentes de la population.

Il y a d'autres objectifs, qui sont peut-être un peu plus en retrait, mais qui

1 http://www.culturecommunication.gouv.fr/Politiques-ministerielles/Langue-francaise-et-langues-de-France/Politiques-de-la-langue/Langues-de-France/Langues-des-Outre-mer/Rencontres-2011-etats-generaux-du-multilinguisme-dans-les-outre-mer

sont néanmoins intéressants à souligner, comme des objectifs d'emploi et de croissance économique. En effet, ces technologies de traitement automatisé du langage peuvent permettre d'améliorer la possibilité de créer du travail et de commercer pour les entreprises. Leur utilisation peut permettre de faire tomber des barrières linguistiques et de constituer des avantages commerciaux pour les entreprises qui les utilisent, même pour les langues régionales. Les technologies qui peuvent aujourd'hui analyser les opinions, les sentiments, et qui ont le vent en poupe, pourraient, appliquées aux langues régionales, contribuer à développer de nouveaux marchés et susciter de nouvelles clientèles.

Il y a peut être là des pistes en termes économiques à exploiter, et il est intéressant de souligner cet enjeu puisque la France est l'un des pays à la pointe des technologies du traitement de la langue en Europe. Nous disposons en effet d'un tissu de petites et moyennes entreprises (PME) françaises compétentes dans le domaine. Il y a par ailleurs des cas de réussite, comme Systran par exemple, qui a développé des outils mondialement utilisés. On trouve donc en France un vivier économique d'entreprises sur ces sujets-là qu'il serait intéressant de conduire sur le chemin des langues régionales.

Voilà donc les différents enjeux qu'il nous a semblé utile de rappeler à l'ouverture de ce colloque. J'aimerais également citer un programme qui a beaucoup aidé la recherche dans le champ des technologies pour la langue, il s'agit du programme Technolangue[1]. Celui-ci a eu lieu entre 2003 et 2005, il a été coordonné notamment par la DGLFLF et le CNRS et a donné lieu à un certain nombre de pépites d'innovation dans ce domaine.

Plus récemment, un rapport remis par Jacques Attali au Président de la République au mois d'août 2013, intitulé « La francophonie et la francophilie, moteurs de croissance durable »[2], recommandait de renouveler ce programme Technolangue, sous la coordination de la DGLFLF pour continuer à investir sur les technologies du langage. Cela paraît en effet très pertinent. L'objectif de Technolangue étant le français, pourquoi ne pas y ajouter un nouveau volet sur les langues régionales ?

1 http://www.technolangue.net/
2 http://www.elysee.fr/assets/Uploads/Rapport-Jacques-Attali-la-francophonie-conomique.pdf

Pourquoi ce colloque ?

Le premier objectif que nous avons est de dessiner les contours d'un éventuel futur programme d'accompagnement visant à obtenir des moyens pour travailler sur les technologies du langage pour les langues régionales. Il faut avoir en tête ce que pourrait être un tel programme d'aide nationale.

Par ailleurs, pour l'élaborer, il faut savoir de quoi l'on a besoin. J'ai évoqué tout à l'heure le besoin social. Il faut essayer de cartographier les besoins pour ces langues régionales : faut-il travailler davantage sur l'écrit, sur l'oral, sur l'apprentissage, sur les outils de traduction, de reconnaissance vocale ? Faut-il encourager des projets de type collaboratif ? Mettre l'accent sur les ressources linguistiques plutôt que sur les technologies ? Est-ce qu'il faut contribuer à Wikipédia, encourager le Wiktionnaire, et si oui, dans quelle mesure ? Quelles sont les priorités, selon les langues ? Est-ce que l'on souhaite mettre en avant certaines langues régionales plutôt que d'autres, certains types de langues ? Enfin, faut-il plutôt privilégier des technologies de type interlingue, ou des technologies centrées sur une seule langue ? Ces questions méritent d'être posées, et feront débat, je pense, lors de ce colloque.

Enfin, le dernier point important pour nous pendant ces deux jours sera d'identifier des moyens que l'on pourra mettre en place pour répondre aux besoins qui se feront jour : moyens financiers, auprès des collectivités territoriales, de l'État, auprès de partenaires publics ou privés ou même du côté de l'Europe. Il y a beaucoup de possibilités que nous aborderons ensemble.

Je terminerai mon propos en citant plusieurs questions qui se sont posées lors de la réunion préparatoire que nous avons tenue sur le périmètre du traitement des langues et qu'il me paraît utile de rappeler. Il semble tout d'abord évident qu'il faudra se concentrer sur certaines langues, que l'on ne pourra pas tout traiter, notamment dans le cas des langues de France qui ne sont pas territoriales, comme la langue des signes ou le yiddish. Dans la mesure où nous espérons obtenir un financement des collectivités territoriales, il paraît en effet difficile de financer une aide pour des langues qui ne sont pas rattachées à un territoire. C'est la raison pour laquelle nous avons volontairement délimité le périmètre

de ce colloque aux langues régionales de France et que nous en avons exclu les langues non-territoriales.

D'autres questions ont été posées : quel outil ou technologie allons nous choisir de développer en priorité ? Comment évalue-t-on les besoins ? Est-il nécessaire de mettre en place un comité pour évaluer ces besoins ? Comment financer des enrichissements de corpus, dans la mesure où l'aide à la recherche ne concerne pas directement l'enrichissement de corpus existants ? Comment traiter la question de la variabilité des langues, notamment pour prendre en compte les différentes graphies ? Comment assurer une diffusion la plus large possible des ressources et des technologies que l'on met à disposition ? Faut-il choisir de construire une nouvelle plate-forme, ou faut-il plutôt s'appuyer sur une plate-forme existante, et laquelle ?

À ce titre, le choix des licences de diffusion paraît très important, et une des recommandations que nous pourrions avoir est de choisir des licences ouvertes, reconnues, largement utilisées aujourd'hui pour développer ce genre de plate-formes.

Voila quelques pistes de réflexion pour entamer le programme de ce colloque, je pense que bien d'autres apparaîtront pendant les débats, que j'espère fructueux.

Les langues de France dans le programme Corpus de la Parole

Olivier Baude, délégation générale à la langue française et aux langues de France (DGLFLF) et Laboratoire Ligérien de Linguistique (LLL) – université d'Orléans

J'ai le plaisir de travailler à la fois pour la DGLFLF et le CNRS, et cela va me permettre de vous parler d'une expérience menée depuis une dizaine d'années maintenant dans le cadre d'un partenariat entre la DGLFLF, le CNRS et les universités. Cette expérience est celle du programme *Corpus de la parole* consacré au français et aux langues de France.

Nous sommes partis d'un triple constat auquel nous étions confrontés il y a une dizaine d'années :

• Les étagères sont vides : il y avait très peu de ressources disponibles sous forme de corpus en langues régionales de France pour une réexploitation bien que depuis de nombreuses années les linguistes enregistrent les langues.

• L'armoire s'abîme. Il y a depuis les années soixante énormément de travaux en dialectologie, en linguistique descriptive, en socio-linguistique, etc. fondés sur des enregistrements sonores. Or ces enregistrements sont conservés sur des supports analogiques périssables et il était important de les sauvegarder

• De nouvelles technologies sont mûres. Nous sommes à l'époque d'un tournant technologique fort : le numérique, le développement d'outils d'exploitation doit favoriser la conservation et l'accès aux ressources. Il favorise aussi l'essor de nouvelles pratiques dans le cadre des humanités numériques

Face à ce triple constat, la délégation a pris la décision de développer un programme autour d'une approche patrimoniale et scientifique des langues de France. Si l'on considère les corpus oraux comme des objets à la fois patrimoniaux et scientifiques, la chaîne de traitement est repérée :

il faut que la collecte soit possible, que ces corpus et enregistrements collectés soient exposés, disponibles, accessibles, qu'on puisse les archiver, et enfin les réutiliser et les diffuser.

C'est dans cette perspective que le programme Corpus de la parole s'est développé autour de quatre grands axes :

• la réalisation d'un guide de « bonnes pratiques » (techniques et juridiques) ;
• la contribution au plan de numérisation du Ministère de la Culture et de la Communication ;
• l'élaboration d'un entrepôt de corpus en langues de France, dans un objectif d'archivage et de diffusion ;
• la diffusion par l'intermédiaire d'un site de valorisation.

Cette expérience sur une dizaine d'années est intéressante, car elle est en accord avec l'évolution à la fois du cadre institutionnel, des outils et du cadre scientifique autour des technologies de la langue.

La première étape a consisté, en partant des pratiques des chercheurs qui collectent, analysent et diffusent ces données, à produire un guide des bonnes pratiques. Outre l'importance d'un travail sur l'interopérabilité des données, les bonnes pratiques revêtent des aspects juridiques que l'on peut distinguer en deux grandes catégories : la protection des données personnelles et la gestion de la propriété intellectuelle y compris les actions pour des données libres et disponibles dans lesquels se sont engagés les ministères de la Recherche et de l'Enseignement supérieur et de la Culture et de la Communication.

En 2008, le CNRS a développé un très grand équipement dédié à l'archivage scientifique, devenu la Très Grande Infrastructure de Recherche (TGIR) des humanités numériques (Huma-Num) en 2012. C'est dans ce cadre qu'a été instauré un projet pilote sur l'archivage des données orales dont a bénéficié le programme Corpus de la parole. Il s'agissait notamment de constituer le premier entrepôt de données orales en langues de France. La maîtrise d'ouvrage a été confiée au Centre de ressources et description de l'oral de Paris (Centre de Ressources sur la Description de l'Oral (CRDO) devenu la plateforme Collection de Corpus Oraux Numériques (CoCoON)),

sous la responsabilité de Michel Jacobson. L'objectif était de constituer un entrepôt permettant d'une part une gestion des opérations d'archivage et d'autre part l'accès aux données pour tous les utilisateurs.

La réalisation de cet entrepôt et du site de diffusion a entraîné un travail sur les aspects techniques : utilisation des normes internationales pour l'audio et la vidéo, élaboration de normes pour les traductions et transcriptions, renseignement des métadonnées selon un objectif d'interopérabilité et juridiques : usages de licences Creative Commons. Toute cette phase a été réalisée en collaboration avec le service des archives sonores de la Bibliothèque Nationale de France (BnF) dirigé par Pascal Cordereix. L'objectif était de permettre aux chercheurs, aux responsables de projets et de laboratoires de déposer leurs corpus dans un entrepôt d'archives ouvertes, disposant d'une base de données permettant une exposition, un affichage et un signalement des données à des fins de conservation et de diffusion.

Cet entrepôt comporte aujourd'hui plus de 2 000 enregistrements sonores, une centaine d'enregistrements vidéos (notamment en langue des signes), des centaines de transcriptions pour un total de presque 1 000 heures de données en 45 langues représentatives de la diversité des langues de France. Ainsi, ces documents oraux obtiennent le même statut que les ouvrages et documents écrits sur les langues de France.

Je vais maintenant vous donner quelques exemples d'accès à ces données. Le travail conjoint entre le ministère de la Culture et celui de l'Enseignement supérieur et de la Recherche, dans un cadre de réutilisation et de diffusion des données, nous permet actuellement d'avoir ces données entièrement accessibles grâce à l'outil ISIDORE développé par le CNRS (disponible depuis peu de temps en version multilingue), qui permet aussi l'accès aux travaux, outils et ressources de sciences humaines et sociales. Ces corpus sont également disponibles sur le site de la BNF, sur le site de la plateforme CoCoON, sur le site dédié de la DGLFLF « Corpus de la parole »[1], ainsi que sur le site de l'équipement d'excellence (équipex) Outils et ressources pour un Traitement optimisé de la Langue (Ortolang), qui a pour mission de mettre à disposition des ressources et des outils pour la linguistique.

1 http://corpusdelaparole.huma-num.fr/

Par l'intermédiaire du site Corpus de la Parole, nous pouvons entendre le son, voir la transcription, la traduction et les métadonnées, et récupérer l'ensemble de ces ressources pour tout autre usage relevant de la licence Creative Commons utilisée. L'interface de la plateforme CoCoON est plus orientée vers les chercheurs mais propose également des usages plus culturels comme par exemple l'accès aux atlas linguistiques à travers des cartes.

J'aimerais vous donner trois exemples de nouveaux usages qui vont être faits de Corpus de la Parole dans un contexte culturel en 2015.

D'une part, l'interface du site Corpus de la Parole va être revue sur des aspects graphiques, techniques, mais aussi éditoriaux avec un nouveau comité de pilotage, un nouveau comité technique et un nouveau comité scientifique.

À côté de ça, nous développons dans la continuité de Jocondelab un projet avec l'Institut de recherche et d'Innovation (IRI) du Centre Pompidou sur la sémantisation des données de Corpus de la Parole. C'est un projet innovant dans une perspective du web sémantique appliquée à des usages scientifiques et culturels.

Le ministère de la Culture et de la Communication soutient également la réalisation d'une exposition sur les langues de France, intitulée *Le cabinet de curiosités*, qui se tiendra à l'automne 2015.

Enfin nous avons également confié à des artistes numériques la création d'une œuvre, avec une consigne simple : en partant de Corpus de la Parole, jouez avec les transcriptions, les documents sonores, les métadonnées disponibles et construisez une œuvre. Cette œuvre, *Les paysages du larynx*, visible sur le site corpus de la parole sera aussi exposée dans d'autres configurations scéniques dans différents lieux en France.

Pour conclure, je préciserai que ce qui est important dans ce projet, c'est de construire les relations entre le ministère de l'Enseignement supérieur et de la Recherche et le ministère de la Culture et de la Communication, dans le cadre de projets sur les langues. Le projet Corpus de la Parole se trouve d'ailleurs inscrit dans l'accord-cadre qui gère les relations ces deux

ministères. La différence de développement de ce partenariat sur une dizaine d'années est saisissante. Le paysage de la recherche a changé, il y a maintenant en soutien des projets un équipement d'excellence, des plates-formes technologiques, une Très grande infrastructure de recherche en sciences sociales qui peut apporter des services notamment en termes d'exposition des données et d'archivage, et un réseau européen sous la forme d'infrastructures européennes. Celles qui concernent le ministère de la Culture et de la Communication et celui de l'Enseignement supérieur et de la Recherche sont les deux *European Research Infrastructure Consortium* (ERIC) très récents : *Common Language Resources and Technology Infrastructure* (CLARIN) et *Digital Research Infrastructure for the Arts and Humanities* (DARIAH).

Voilà pour une présentation très rapide illustrant certains développements présentés lors des introductions. Je vous remercie pour votre attention.

Technologies de la langue : état des lieux

Joseph Mariani, Laboratoire de recherche en informatique pluridisciplinaire (LIMSI) - Centre national de la recherche scientifique (CNRS) et Institut des technologies multilingues et multimédia de l'information (IMMI)

Mon propos sera de montrer que les technologies de la langue sont nécessaires voire indispensables pour permettre le multilinguisme, et de faire une présentation de la situation de la langue française à l'heure du numérique.

Les enjeux du multilinguisme comportent deux volets

Le premier est de veiller à la préservation des cultures à travers les langues, en permettant aux citoyens de continuer à s'exprimer dans leur langue maternelle. Une étude de la commission européenne montre que 90 % des citoyens européens préfèrent avoir accès à un site web rédigé dans leur langue maternelle. On peut également signaler que moins de 30 % du web est en anglais aujourd'hui, alors que ce pourcentage s'élevait à 50 % en 2000, pour diminuer à 35 % en 2004. Je pense que Daniel Prado tout à l'heure nous présentera des chiffres plus précis.

Il faut également signaler que 50 % des citoyens européens ne parlent qu'une seule langue, et lorsqu'ils en parlent une deuxième ce n'est pas nécessairement l'anglais. À peine 3 % des Japonais parlent une langue étrangère et, contrairement aux idées reçues, seulement 5 % des citoyens indiens parlent anglais.

Le deuxième volet est de permettre la communication entre humains parlant différentes langues.

Dans le cas de l'Union européenne, pour 28 états-membres, nous comptons 24 langues officielles, ce qui représente 552 paires de langues à traduire, et 35 langues nationales. On estime que 60 à 220 langues sont parlées en Europe. 2 500 traducteurs sont employés par la commission européenne, traduisant un peu moins de deux millions de pages par an.

Pour une commission européenne réellement multilingue, on estime qu'il faudrait pouvoir disposer de 8 500 traducteurs, qui traduiraient 7 millions de documents par an. Au niveau du parlement européen, 30 % du budget est consacré au multilinguisme, ce qui représente 300 millions d'euros par an et l'emploi de 500 traducteurs. Le coût estimé du multilinguisme pour l'Union européenne est d'un peu plus d'un milliard d'euros par an, ce qui peut paraître beaucoup, mais ramené au nombre de citoyens européens, cela ne représente jamais que 2,20 € par citoyen.

Seuls 30 % des citoyens européens sont prêts à acheter des produits sur un site internet qui n'est pas dans leur langue. Cela montre bien que les langues sont un obstacle au développement d'un marché commun européen. A contrario, 80 % d'entre eux pensent que les sites web rédigés dans leur langue maternelle devraient être traduits dans des langues étrangères, particulièrement européennes, pour en faire profiter les autres citoyens de l'Union Européenne.

Voici maintenant quelques chiffres relatifs à la mondialisation de l'information à l'échelle internationale : il y a actuellement 100 nouvelles heures de vidéos placées sur YouTube chaque minute (équivalent à 4 jours de vidéo, dans toutes les langues), et si l'on considère les plus de 6 000 langues parlées dans le monde, cela revient à 36 000 000 de paires de langues à traduire si l'on voulait toutes les couvrir.

Les enjeux sont donc importants. Considérons maintenant les besoins, qui se retrouvent dans des domaines très différents.

En Europe, la bibliothèque européenne Europeana est constituée de 23 millions de documents en 26 langues (2013). Cela suppose donc de pouvoir disposer d'outils d'accès multilingues et interlingues à ces documents. L'office européen des brevets, qui détient un portefeuille de 10 millions de brevets rédigés en 32 langues, a réduit le nombre de langues de travail à trois, pour des questions de coûts (anglais, allemand et français, même si c'est l'anglais qui est principalement utilisé). Si l'on voulait tout traduire, il faudrait réaliser 300 millions de traductions, ce qui représente une estimation de 1500 ans de travail pour 1000 traducteurs.

Ces besoins se retrouvent à la Commission européenne, à la cour

européenne de justice et au parlement européen, en ce qui concerne la traduction de documents ou de rapports, et la traduction simultanée au cours des réunions. En 1997, 45 % des documents source à traduire par la commission européenne étaient en anglais, et 40 % en français, ce qui constituait un équilibre relatif. Dix ans après, en 2007, 72 % des documents source à traduire étaient en anglais, et 12 % seulement en français.

À l'international, on trouve également des besoins, comme la bibliothèque mondiale de l'UNESCO (10 000 documents de 80 pays), ce qui engendre également la nécessité d'outils d'accès interlingues. Plus généralement, les besoins se retrouvent aussi pour les notices techniques dans des domaines multiples comme l'aéronautique, l'électroménager, l'automobile, ainsi que dans le commerce international sur internet, le doublage et le sous-titrage des œuvres audio-visuelles, la traduction des textes et vidéos sur internet, l'interprétation des conférences, des cours en ligne ouverts à tous (MOOC), l'interprétation pour les opérations militaires et sanitaires (comme lors du tremblement de terre en Haïti, par exemple), et aussi la rédaction des articles scientifiques dans la langue maternelle : à l'heure actuelle, une analyse faite par le Science Citation Index montre une évolution vers une pratique de plus en plus hégémonique de la langue anglaise pour les publications scientifiques : le taux est passé de 85 % de publications répertoriées en anglais en 1980 à 96 % en l'espace de vingt ans, les autres langues (français, allemand, espagnol, japonais, russe) se partageant les 4 % restant. Pour être reconnu, un auteur doit être cité ; pour être cité, il doit être lu et pour être lu par le plus grand nombre, il doit publier en anglais.

À partir de ces analyses des besoins et des enjeux, nous pouvons faire un certain nombre de constats.

• Le premier est que j'espère vous avoir convaincu qu'il est impossible de répondre rapidement (voire de répondre tout court) aux nombreux besoins actuels et futurs liés au multilinguisme avec les ressources humaines actuelles voire futures, si l'on voulait former des traducteurs pour répondre à tous ces besoins.

• Deuxième constat : le multilinguisme n'est pas la priorité des entreprises. Mais la somme des petites priorités que l'on peut trouver

dans de nombreux secteurs économiques, elle, est importante. Il est donc nécessaire de mener une réflexion et d'entreprendre une action politique, pour faire cette somme et répondre à l'aune de son importance.

• Troisième constat : le multilinguisme est nécessaire, mais son coût est très important. Il y a donc intérêt à disposer de technologies de la langue pour le faciliter, afin de diminuer les coûts, de manière à généraliser son usage, mais si, et seulement si, les performances atteintes répondent aux besoins des utilisateurs. Les technologies doivent être de qualité suffisante.

• Quatrième constat : les langues ne disposant pas de technologies seront de moins en moins utilisées : si on ne peut pas utiliser son smartphone, son GPS, naviguer sur internet appeler un service d'urgence ou interroger une base de connaissances dans sa langue, on sera obligé d'utiliser une autre langue que sa langue maternelle, et son usage diminuera. A contrario, les langues qui bénéficient des technologies de traduction seront de plus en plus utilisées, puisque le passage vers d'autres langues devient transparent.

• Cinquième constat : les technologies de la langue n'ont pas encore atteint leur maturité. La traduction automatique, par exemple, est loin d'être d'une qualité suffisante pour traduire des œuvres littéraires ou des textes nécessitant une traduction de qualité (sites webs officiels, par exemple). En revanche, la traduction automatique peut aider le traducteur humain dans son activité et sa qualité est suffisante pour apporter une traduction approximative, rapide et gratuite, au grand public.

Que recouvrent les technologies de la langue ?

On trouve à la fois des composantes, comme des analyseurs syntaxiques ou morpho-syntaxiques, des extracteurs d'entités nommées, de mots-clés, de terminologie. On trouvera également pour la langue écrite des systèmes complets pour la correction orthographique et grammaticale, de compréhension et de génération de textes, de résumé automatique, de classification automatique, d'analyse d'opinion, de recherche d'information (moteurs de recherche), et des nouveaux systèmes comme les systèmes

d'interrogation de bases de connaissances, par exemple le système Watson d'IBM aux États-Unis, qui a récemment gagné le jeu télévisé *Jeopardy*. On trouve également des technologies interlingues, permettant de passer d'une langue à une autre, comme la traduction automatique ou assistée par ordinateur, ainsi que des systèmes de recherche d'information interlingues, pour avoir accès à l'information quelle que soit la langue dans laquelle elle a été produite.

On trouve également des technologies pour le traitement de la langue parlée : reconnaissance et compréhension de la parole, synthèse vocale, dialogue oral (*Siri* d'Apple, *Cortana* de Microsoft, *Google Now*), système de reconnaissance du locuteur, et également des systèmes interlingues d'identification de la langue et de traduction vocale simultanée.

Des technologies sont aussi développées pour le traitement de la langue signée, aussi bien en analyse, en synthèse ou en traduction entre langues signées. J'aimerais insister sur le fait que ces technologies sont essentielles pour permettre l'accessibilité : des technologies que l'on peut qualifier d'intermédias, synthèse vocale à partir du texte pour l'aide au handicap visuel, systèmes de transcription de la parole, ou de traitement de langue des signes pour l'aide au handicap auditif, systèmes de commande vocale pour l'aide au handicap moteur, ainsi que des technologies de type interlingue, qui vont aider à supprimer la barrière des langues, que l'on peut également considérer comme un handicap.

Dans ce cadre de développement des technologies, il est important de pouvoir bénéficier d'une infrastructure où l'on puisse trouver à la fois des ressources linguistiques, comme cela a été mentionné par Olivier Baude tout à l'heure, et des moyens d'évaluation permettant de mesurer la qualité des technologies. On trouvera dans les ressources linguistiques un ensemble de données, de corpus, de lexiques, de dictionnaires, de bases terminologiques ou encyclopédiques comme Wikipédia. Elles seront nécessaires bien sûr pour mener des recherches en linguistique, mais elles seront aussi essentielles pour effectuer l'apprentissage automatique des systèmes, qui souvent sont basés sur des approches statistiques. Plus massives seront les données, meilleurs seront les systèmes. Bien sûr, pour diffuser ces ressources linguistiques il est nécessaire de bénéficier de standards permettant de les échanger.

L'évaluation des technologies de la langue consiste à comparer les performances de systèmes provenant de différents laboratoires, pouvant être fondés sur différentes approches. Il faut donc les évaluer sur des données communes, avec un protocole commun, dans le cadre de campagnes d'évaluation. C'est un indicateur de la qualité des recherches et du progrès des technologies, à mi-chemin entre compétition et coopération internationale : on utilise maintenant le terme de « coopétition » pour qualifier ces campagnes d'évaluation. Cela permet également de comparer les performances de ces technologies avec les besoins des applications, en apportant une indication sur le niveau de maturité technologique.

Prenons l'exemple de l'évaluation du *National Institute of Standards and Technology* (NIST) pour les technologies de l'oral. Depuis les années 80 et jusqu'à un passé récent, on voit que fort heureusement les taux d'erreurs diminuent dans le temps, pour des tâches de plus en plus difficiles (reconnaissance d'un vocabulaire d'un millier de mots, dictée vocale, transcription automatique d'émissions de radio ou de télévision, transcription de conversations téléphoniques, et enfin transcription de réunions où plusieurs personnes peuvent parler en même temps). C'est rassurant pour les agences de financement qui se rendent compte que les moyens consentis ont permis d'améliorer la qualité des technologies, jusqu'à atteindre le niveau d'un auditeur humain pour certaines tâches (transcription audio). Cependant, pour des tâches plus difficiles comme la transcription de conversations ou de réunions, les taux d'erreur stagnent à des niveaux encore assez importants (de l'ordre de 50 %), ce qui montre qu'il faut continuer à soutenir la recherche.

C'est la même chose pour la traduction automatique. Je vous présente ici une matrice produite dans le cadre du projet européen Euromatrix, comportant les mesures de qualité des systèmes de traduction automatique entre les paires de langues européennes. La mesure de qualité est présentée sous la forme d'un score *Bilingual Evaluation Understudy* (BLEU), qui correspond au taux de ressemblance mot à mot entre la traduction automatique et un ensemble de traductions de référence qui ont été produites par des traducteurs humains. On voit que pour certaines langues les résultats sont plutôt bons, notamment l'anglais, avec des taux supérieurs à 50 %. En revanche, pour certaines langues comme le hongrois ou le maltais, les scores sont inférieurs à 30 %.

Cela va de pair avec la disponibilité des ressources, ce qui me permet d'insister une nouvelle fois sur ce point. Sur cette autre matrice produite dans le même projet présentant l'existence de corpus parallèles (corpus traduits dans deux langues différentes permettant de développer le système de traduction), nous voyons que si pour certaines langues de tels corpus existent en grand nombre (l'anglais, le français, l'allemand, l'italien), pour d'autres langues comme le maltais ou l'irlandais, il n'y a quasiment pas de ressources, ce qui entraîne des systèmes de qualité médiocre.

Après ce panorama général, j'aimerais maintenant faire un état des lieux, et en particulier signaler une analyse menée par le réseau d'excellence européen META-NET, l'alliance technologique pour une Europe multilingue (une soixantaine de membres dans 34 pays européens), financé par la commission européenne et regroupant quatre sous-projets qui couvrent l'ensemble des pays de l'Union Européenne.

META-NET a produit un ensemble de livres blancs sur les langues européennes en septembre 2012, présentant d'une part la situation de ces langues, en termes de locuteurs, langue maternelle ou seconde, les états pour lesquels ces langues sont des langues officielles, les médias qui existent dans ces langues, les organismes, particulièrement internationaux, qui les pratiquent, leur présence sur internet ou Wikipedia, et d'autre part, l'état des technologies qui permettent de les équiper. 31 livres blancs ont été produits pour 30 langues (du fait de deux variantes pour le norvégien), aussi bien sur des langues nationales que régionales (basque, catalan, galicien, gallois) dont les 24 langues de l'Union Européenne. Ces livres blancs sont bilingues (langue concernée-anglais).

Ces guides présentent l'état des technologies et des ressources existantes (orales, écrites, corpus parallèles, en termes de quantité, disponibilité, qualité, couverture, etc., pour chacune de ces technologies et chacune de ces ressources). On a ensuite essayé d'interclasser la situation pour les différentes langues, sur cinq niveaux, allant d'excellente à inexistante, dans quatre grands domaines : traitement de la parole, traitement de l'écrit, traduction automatique et ressources linguistiques. Cela a donné lieu à de grandes discussions à Berlin en octobre 2011 avec les représentants de ces 30 langues pour finaliser ce classement.

Pour le traitement de la parole, aucune langue n'est dans la catégorie « base excellente » : aucune langue ne dispose actuellement de technologies de traitement de la parole qui soient de qualité suffisante pour couvrir l'ensemble des applications. On trouve dans la catégorie « bonne base » l'anglais, assez bien doté, ensuite un ensemble de langues dans la catégorie « base modérée », dont le français, puis un nombre de langues encore plus important dans la catégorie « base fragmentée », dont le basque, le catalan et le galicien (le gallois n'était pas encore disponible à ce moment-là) et enfin dans la catégorie des bases faibles voir inexistantes, un certain nombre de langues européennes.

Pour le traitement de l'écrit, le classement est à peu près similaire.

Pour la traduction automatique, même chose : rien dans la première catégorie, l'anglais dans « bonne base », le français et l'espagnol dans les bases modérées, et nous verrons tout à l'heure que ce n'est pas anodin. Quelques langues sont dans « base fragmentée », et de très nombreuses langues dans « base faible » ou « intermédiaire ».

Enfin, de manière transversale pour les ressources linguistiques, le classement est similaire.

On peut noter dans l'ensemble de ces analyses une bonne place pour la langue française, liée à l'effort continu des laboratoires de recherche et des associations comme l'Association pour le Traitement Automatique des Langues (ATALA) ou l'Association Francophone de la Communication Parlée (AFCP), au fil des ans, ainsi que quelques grands programmes nationaux qui ont été menés et financés (GRECO du CNRS, réseau francophone des industries de la langue (FRANCIL) au sein de l'Agence Universitaire Francophone, programme Techno-Langue, QUAERO, les projets de l'Agence nationale de la recherche (ANR), les équipex, les laboratoires d'excellence (labex), les initiatives d'excellence (idex)…). Cela permet de disposer de ressources linguistiques pour la langue française ainsi que d'avoir une bonne connaissance de notre niveau technologique. On peut dire que le niveau de la recherche en France pour la langue française est bon. Les laboratoires français sont bien placés dans les campagnes d'évaluation internationales. En revanche, il est regrettable que le tissu industriel soit essentiellement constitué de PME,

la plupart des grands groupes ayant quitté ce domaine de recherche et de développement il y a quelques années, contrairement aux industriels américains qui se retrouvent sur le devant de la scène : Apple, Google, Facebook, Amazon, Microsoft, IBM…

Le programme des technologies de la langue (Technolangue) fait suite à un rapport remis au premier ministre en 2000 et organisé par le comité pour le traitement informatique des langues présidé par André Danzin pour le conseil supérieur de la langue française, alors présidé par Bernard Cerquiglini en tant que délégué général à la langue française et aux langues de France. Suite à cela, en 2001 une réunion interministérielle à Matignon a décidé de trois actions :

• une action de veille technologique et d'évaluation des outils de traitement de la langue française (ministère de la Culture et de la Communication, ministère de l'Industrie, ministère de l'Enseignement supérieur et de la Recherche) : il s'agit de Techno-Langue ;

• une action sur le développement des usages du traitement informatique de la langue française (ministère de la Culture et de la Communication, ministère de la Fonction publique et de la Réforme de l'État), avec une action sur la simplification du langage administratif ;

• une troisième pour la formation de professionnels en ingénierie documentaire (ministère de l'Éducation Nationale).

Technolangue était organisé autour d'un comité de pilotage composé d'une quinzaine de personnes représentant l'industrie, la recherche et les administrations.

L'appel à propositions concernait des projets de trois ans couvrant la période 2002-2005, financés par les ministères de l'enseignement supérieur et de la recherche, de l'industrie, et de la culture et de la communication. Sur une cinquantaine de propositions, 28 ont été retenues et 21 financées, regroupant une centaine de participants de la recherche publique, des industriels, des associations ainsi que des laboratoires étrangers qui pouvaient participer à condition d'apporter leurs propres financements. Le budget total se montait à 20 millions d'euros, dont 7,5 millions d'aides publiques sur trois ans. Parmi ces 21 projets financés, sept concernaient la production de ressources linguistiques, trois projets de boîtes à outils, deux projets sur les standards (écrits et oraux), un

projet de veille technologique, avec la mise en place d'un portail, et huit projets sur l'évaluation de technologies (cinq pour le traitement de l'écrit, trois pour le traitement de l'oral).

Au plan européen, en plus de la rédaction des livres blancs, META-NET s'est également chargé de la mise en place d'un agenda stratégique à la demande de la commission européenne pour aider à la préparation du futur programme-cadre. Cela a mené à la proposition d'un programme européen coordonné sur les technologies de la langue, concernant à la fois ce nouveau programme-cadre de recherche et développement intitulé Horizon2020, et un autre projet en parallèle se rapprochant plus d'une commande publique, intitulé *Connecting Europe Facility* (CEF). Cet agenda proposait un effort partagé entre la commission européenne, qui apporterait un soutien générique, et les états-membres qui pourraient financer les besoins spécifiques à leur(s) langue(s), ceci selon 4 volets : traduction & localisation, médias & services d'information, systèmes interactifs, ressources linguistiques. Présenté à Berlin en janvier 2013, cet agenda stratégique est disponible en ligne.

Comment ces propositions ont-elles été déclinées par la commission européenne ? Dans le programme-cadre, on trouve certes une ligne, intitulée « *Cracking the language barrier* », dotée d'un financement relativement modeste, et qui ne traite que de la traduction automatique (et pour l'écrit uniquement) pour 21 langues officielles européennes. Trois langues ont donc été exclues de façon aberrante : le français, l'anglais et l'espagnol, car elles sont supposées être suffisamment bien dotées, aussi bien comme langue source que comme langue cible de traduction, ainsi que toutes les langues régionales. Le premier appel à propositions a eu lieu en décembre 2013. Un seul projet a été retenu pour l'aspect recherche, intitulé QT21, qui traite de la production de systèmes de traduction automatique de qualité entre sept paires de langues (anglais-allemand dans les deux sens, anglais-tchèque dans les deux sens, et anglais vers le roumain, l'estonien et le letton). Il est donc étonnant de constater que malgré son exclusion initiale, l'anglais apparaît comme une langue quasiment obligatoire dans les projets soutenus. Le projet s'attache à faire intervenir des informations de type sémantique pour améliorer la qualité de la traduction, et à traiter les langues peu dotées et à morphologie riche comme l'allemand. Une action particulière est

également menée pour faire en sorte d'obtenir de meilleures mesures de la qualité de la traduction que BLEU. Sur les trois projets retenus en innovation, l'un concerne la production de traduction automatique générique, à base de MOSES (logiciel libre), et deux la traduction de textes médicaux et de cours massivement en ligne (MOOC). Enfin, on compte aussi deux projets d'infrastructures : Cracker et Language Technologie (LT)-Observatory.

Il est à signaler dans le cadre d'Horizon2020 un prochain sommet à Riga sur le marché commun numérique multilingue, en avril 2015, incluant une table ronde avec les représentants nationaux.

Au-delà du périmètre européen, l'UNESCO a diffusé en octobre 2014 une déclaration sur le multilinguisme dans le cyberespace, où il est dit que l'organisation s'attachera à promouvoir la diversité linguistique et le multilinguisme, en particulier en appliquant des solutions technologiques innovantes et en proposant l'organisation d'un sommet mondial sur le multilinguisme.

En conclusion, je dirais qu'il existe des technologies et des ressources pour le français, mais qu'elles se placent très en dessous de ce qui existe pour l'anglais, et que certaines langues européennes nationales ou régionales sont en danger d'extinction numérique voire en danger d'extinction tout court. On peut noter un manque de continuité dans le soutien des pouvoirs publics à l'effort scientifique et industriel (en France comme dans l'Union Européenne), alors qu'il est indispensable de disposer de technologies de la langue pour traiter le multilinguisme, qui est une donnée fondamentale de l'Union Européenne et des états-membres. On ne peut que continuer à souhaiter qu'il y ait la mise en place d'un large programme coordonné au niveau européen, associant la commission européenne et les états-membres, qui puisse être étendu au plan régional, avec la participation et le soutien des collectivités régionales, voire au plan international avec d'autres pays partenaires et l'appui de l'UNESCO.

Débat avec la salle

Un intervenant
M. Mariani a mentionné entre 60 et 220 langues parlées. Pourquoi un tel écart ?

Joseph Mariani
C'est toute la discussion qu'il y a entre la définition d'une langue et la définition d'un dialecte : si l'on considère les langues, on en compte 60, si l'on considère les « dialectes », on en compte 220.

Un intervenant
Concernant les atlas linguistiques, s'agit-il d'une remise au goût du jour des atlas papier, ou bien d'une mise à disposition du grand public d'une visualisation des usages des langues et dialectes, où l'on pourrait envisager les nouvelles technologies par forcément au service des langues en tant que système de synthèse, reconnaissance, etc., mais pour illustrer la dynamique et la variation au sein du territoire français par exemple ?

Olivier Baude
Je vais faire une moitié de réponse. Il y a dans Corpus de la parole des enregistrements sonores provenant des atlas donnés par les auteurs de ces enregistrements qui sont donc archivés dans le cadre du programme Corpus de la Parole. Je suis allé un peu plus loin dans la présentation en vous montrant une interface qui est réalisée dans le cadre de la plate-forme CoCoON[1] par Michel Jacobson, en lien directement avec certains producteurs d'atlas. Donc sur la deuxième partie de la réponse, je vais plutôt le laisser répondre directement sur ce qu'il fait avec ces atlas.

Michel Jacobson, Laboratoire ligérien de linguistique – CNRS (depuis la salle)
Comme l'a expliqué Olivier, c'est juste une représentation des ressources disponibles dans Corpus de la Parole sur une présentation géographique qui donne une bonne représentation du maillage du territoire, fait pour faire ces enquêtes dialectologiques réalisées dans le cadre de ces atlas. Mais il s'agit juste de visualiser ces ressources disponibles qui ont été collectées dans ce cadre-là.

1 Collection de Corpus Oraux Numériques, http://cocoon.huma-num.fr/exist/crdo/

Un intervenant

C'est plus une remarque qu'une question. Sur ce que Joseph Mariani vient de présenter, on voit que le français qui est plutôt pas trop mal doté en ressources. Il se trouve que je fais des ressources langagières pour le français pour le Traitement automatique du langage (TAL) depuis quelques années maintenant, et ce que je peux dire c'est qu'il faut nuancer cet « optimisme » par le fait que de nombreuses ressources pour le français ne sont pas libres, ce qui les rend inutilisables ou très peu utilisables par les chercheurs en général puisqu'il faudrait avoir les moyens pour les acheter, on ne peut pas forcément les redistribuer comme on voudrait, etc. Donc plutôt que des problèmes de production, on a des problèmes liés à la diffusion sur le français. Aujourd'hui c'est en train de changer évidemment, c'est un modèle en évolution très large, mais cela explique qu'aujourd'hui on a encore beaucoup de retard sur l'anglais. L'effort n'est pas terminé. Il y a encore beaucoup de travail à faire pour libérer les ressources, les diffuser de manière plus large et en créer de nouvelles.

Joseph Mariani

Effectivement, c'est un des éléments qui ont été pris en compte, puisque j'ai montré les différents paramètres qui avaient été considérés, et le paramètre de disponibilité est l'un des paramètres traités pour la langue française et pour les autres langues. On est loin de disposer de ce dont on aurait besoin pour le français. La situation est plutôt meilleure que pour d'autres, où elle est catastrophique.

Des premières études menées dans le domaine

Président de séance : **Jean-François Baldi**

Étude sur la place des langues de France sur l'internet

Daniel Prado, réseau Maaya

Je viens vous présenter l'étude que Maaya, le réseau mondial pour la diversité linguistique, vient de mener grâce au soutien de la Délégation générale à la langue française et aux langues de France, concernant la place des langues de France dans le cyberespace. Les résultats de cette étude ont été publiés en novembre 2014 et sont accessibles sur le site internet de la délégation[1].

Mais avant tout, il conviendrait d'évoquer les difficultés pour mesurer la présence des langues dans l'internet et d'expliquer la différence des approches à appliquer selon qu'on le fait pour les langues les plus répandues ou pour les moins répandues.

C'est par la suite que je décrirai le matériel objet de l'étude pour les langues régionales de France et les résultats de celle-ci.

Nous avons récupéré, pour une partie de ce travail, des données utilisées dans le rapport sur l'usage de la langue française produit par l'Organisation Internationale de la francophonie (OIF) en 2014[2].

1 http://www.culturecommunication.gouv.fr/Politiques-ministerielles/Langue-francaise-et-langues-de-France/Politiques-de-la-langue/Langues-et-numerique/La-diversite-linguistique-et-la-creation-artistique-dans-le-domaine-numerique/Etude-sur-la-place-des-langues-de-France-sur-l-internet

2 http://www.francophonie.org/Rapports-Publications.html

Les difficultés pour mesurer les langues dans le cyberespace

Jusqu'il y a quelques années il y avait deux méthodes qui fonctionnaient assez bien en matière de mesure de la présence des langues dans le cyberespace :

1. La méthode proposée par le *Language Observatory Project* (Japon), basé sur le ratissage par noms de domaine nationaux et concernant les langues minoritaires de l'Asie, de l'Afrique, de l'Amérique Latine et des Caraïbes. Cela donnait des résultats nets sur le nombre de pages dans telle ou telle langue.

2. D'autre part, l'Union latine et Funredes avaient utilisé, depuis 1996, les moteurs de recherche usuels (Google, Altavista, etc.) pour faire un travail statistique grâce à certains algorithmes linguistiques propres[1]. Nous comptions pour cela sur une large couverture de la toile par ces moteurs de recherche.

Aujourd'hui, alors que ces deux méthodes ne sont plus utilisables, nous pouvons néanmoins utiliser le travail de l'entreprise W3tech[2] qui fait des statistiques intéressantes sur la présence des langues sur la toile, mais se limitant aux 10 millions de sites les plus visités. C'est un biais puisque cela ne représente que 2 % des sites existants sur l'ensemble du réseau, et, de plus, ce sont les sites ayant le plus grand nombre de visites. Il n'y a donc sans doute pas de place pour les langues minoritaires ou moins répandues.

La première difficulté aujourd'hui est due à la taille de l'internet qui est devenue trop importante pour les méthodes séquentielles de ratissage. L'Union Latine n'existe plus, et Funredes ne peut plus faire un travail sur l'ensemble des sites, étant donné que les moteurs de recherche n'indexeraient aujourd'hui que 5 % du total des pages internet (dans les années 90, c'était 80 %). De plus, le système de comptage des moteurs de recherche n'est plus fiable, l'évolution vers le web dynamique, la vidéo et les réseaux sociaux compliquant encore plus la tâche, autrefois centrée sur des pages textuelles.

1 Les algorithmes linguistiques des moteurs de recherche n'étaient pas fiables à l'époque, donc un travail important a été mené par les auteurs de l'étude.
2 http://w3techs.com/technologies/overview/content_language/all

Pour donner une idée de la taille du réseau Internet, selon des statistiques de la fin 2014, il y aurait environ 180 millions de sites – sur 900 millions de noms de domaine inscrits – dont la plupart seraient de simples doublons ou des sites fantômes sans contenu. De cet ensemble, seuls 5 à 10 % des sites seraient indexés par les moteurs de recherche, et seuls 10 millions de sites peuvent fournir des statistiques linguistiques relativement fiables (sur lesquels W3tech s'appuie). Le travail de ratissage de l'ensemble du cyberespace est donc colossal.

L'Union latine, l'OIF, l'UNESCO, l'Union Internationale des Télécommunications et d'autres institutions ont, à deux reprises, présenté auprès de la Commission européenne des projets de veille permanente sur la présence des langues dans l'Internet, mais sans succès.

Le Réseau Maaya a donc proposé des méthodes alternatives, en attendant une issue favorable à ces projets ambitieux, et qui ne donnent pas nécessairement de résultats nets, mais des résultats relatifs permettant d'obtenir des approximations et des tendances.

La première de ces méthodologies, applicables aux langues les plus répandues, consiste à prendre un ensemble large d'espaces et d'applications à quantifier. Des critères de sélection sont appliqués à ces espaces ou applications[1], notamment en matière de grande utilisation. On procède ensuite à une collecte des données issues de différentes statistiques sur le cyberespace que l'on croise avec des statistiques démo-linguistiques, permettant de pondérer chaque langue étudiée à l'intérieur de chaque espace ou application. Il y a évidemment de grandes différences entre les espaces et applications et donc la place de la langue sera différente en fonction de l'usage de chaque espace ou application dans chaque pays, région, culture, etc.
Nous essayons de croiser l'ensemble de ces informations, pour voir dans ce très large espace quelle est la place approximative de chaque langue. La langue française serait la 4e langue utilisée sur l'internet, si l'on considère l'ensemble des francophones, 7e si l'on ne considère

1 Une centaine d'espaces et d'applications ont été étudiées, concernant les infrastructures, les bibliothèques numériques, les ordiphones, les systèmes d'exploitation, les navigateurs, les chats et conversations par internet, les applications de bureau, le web 2.0, les moteurs de recherche, les courriels, le pair-à-pair. Exemple : Wikipedia, Twitter, Wordpress, Skype, Facebook, etc.

que les locuteurs natifs. Elle est en compétition rapprochée, selon les espaces et les applications prises en compte, avec l'espagnol, l'allemand, le japonais, le portugais, le russe, et l'arabe. Cela va changer, pour le japonais et l'allemand, à cause de certaines langues comptant un grand nombre de locuteurs, qui progressent très vite sur l'internet.

Langues moins répandues

Les langues minoritaires ou moins répandues étant moins présentes sur l'Internet, elles n'ont pas de corpus permettant d'élaborer des statistiques croisées et de ce fait, leur étude demande une méthodologie différente.

Nous partons du constat qu'il y a aujourd'hui une dizaine de langues très bien représentées sur l'internet, environ 75 à 80 langues qui le sont partiellement, entre 270 à 300 qui sont insuffisamment présentes et environ 900 qui auraient une « présence ponctuelle ». Les autres (environ 6 000 à 6 500) sont des langues simplement « référencées », par exemple, dans le *Projet langues en danger* de Google ou dans l'*Atlas UNESCO des langues en danger dans le monde* ou encore dans Sorosoro ou Portalingua. Rappelons-nous que la plupart des langues de France dont nous traitons aujourd'hui ne sont pas suffisamment représentées sur Internet.

Le matériel d'étude

Voici comment nous avons procédé pour les langues de France :

En premier lieu, nous avons cherché les articles ou sources qui pouvaient traiter spécifiquement de la place que chacune de ces langues avait dans le cyberespace. On n'en a pas trouvé, sauf pour le corse, le basque et le catalan, et pour ces deux dernières langues, l'origine est surtout espagnole. On constate dès le départ qu'il y a une faible occupation des espaces et des applications pour certaines langues.

Par la suite, il a été procédé à une recherche des références de sites qui seraient en « relation étroite » avec chaque langue de l'étude. On entend par « relation étroite » tout site, livre ou article traitant de près ou de loin

de la situation de cette langue sur internet ou offrant des données sur le sujet. Les autres références étudiées ont été des bases de données, des méta-sites, des ressources linguistiques, des références au sujet de cette langue tant culturelles (littérature, poésie, etc.) que concernant des cours de langues, ainsi que les blogs et réseaux sociaux dans cette langue ou au sujet de cette langue.

Tout d'abord, nous avons fait une recherche simple afin de repérer les sites principaux et pris note systématiquement de tous les liens externes, faisant une analyse systématique de ces liens et les incorporant si cela était pertinent dans notre corpus d'étude. Le ratissage des liens externes se fait tant que l'on ne trouve plus de références pertinentes ou dès que l'on trouve systématiquement des liens déjà étudiés. Ce corpus a été ensuite évalué, référencé et noté.

Il a été pris note de toutes les références pouvant apporter des données démographiques sur toutes les langues de France.

Nous nous sommes concentrés sur les langues territoriales au sens strict, c'est-à-dire les langues qui sont nées ou partagées sur le territoire de la France actuelle, et qui ne sont pas venues d'Europe ou d'ailleurs. Nous avons pris en compte uniquement celles comptant plus de 50 000 locuteurs ou bien celles qui, tout en ayant moins de locuteurs sur le territoire national, étaient enseignées, c'est-à-dire l'alsacien, le basque, le breton, le catalan, le corse, les créoles, le francique, le franco-provençal, le futunien, les langues kanakes, les langues de Mayotte, les langues d'oïl, l'occitan, le tahitien et le wallisien.

Après une analyse d'une grande quantité de documents, ont été sélectionnées quelque 1 500 références pertinentes et non récursives. Les résultats n'étant pas exhaustifs, leur abondance permettait de donner des chiffres précis sur la présence sectorielle de chaque langue. Il est à noter que nous n'avons pas pris en compte la masse importante de sites espagnols concernant le basque et le catalan, à l'exception de quelques références majeures en matière linguistique ou démo-linguistique.

Des aspects qualitatifs ont également été notés : les données collectées concernent l'année de mise à jour (ou, en son absence, celle de mise

à jour), l'actualisation, l'origine du producteur (acteur gouvernemental, ONG, etc.), le type (blog, article, portail, réseau social, etc), la langue de l'interface, les données chiffrées (démographiques, sociolinguistiques, etc., souvent à profusion), les commentaires et les particularités.

Dans les conclusions de l'étude, nous constatons que certaines langues sont très dynamiques sur la toile, avec un déploiement homogène multi-sectoriel, où plusieurs acteurs participent au déploiement sur la toile, y compris les collectivités locales.

Ceci nous a permis de noter que, pour le breton par exemple, c'est la société civile qui est plus active dans la production, et que, pour le franco-provençal, ce sont les ONG qui sont dynamiques, alors que pour le futunien, les langues kanakes, les langues créoles et le wallisien c'est le monde académique qui est plus actif que le secteur public ou les individus. Pour le tahitien, par contre, c'est le monde commercial qui est le plus actif, évidemment lié au tourisme.

Il est intéressant de noter que certaines de ces langues très parlées par la population (pour lesquelles le français est souvent langue seconde) ont une présence en ligne peu marquée, et c'est le secteur académique qui est beaucoup plus productif que la société civile et le secteur public : c'est le cas déjà mentionné des langues kanakes, des créoles, du futunien, des langues de Mayotte, du tahitien et du wallisien.

Le breton, le franco-provençal et l'occitan arrivent en tête des langues pour lesquelles les citoyens ou la société civile sont plus actifs que les collectivités locales en matière de production. Par contre, le francique et les langues d'oïl connaissent des difficultés particulières quant à leur présence sur la toile.

Enfin, il y a des langues avec une bonne présence sur la toile, soutenue de manière équilibrée par tous les acteurs du secteur, notamment l'alsacien et le catalan, mais dans ce dernier cas, l'appui public est bien supérieur du côté espagnol.

Nous avons aussi regardé ce qui était le plus fréquent comme matériel suivant les différentes langues. Par exemple, pour le breton et le corse, ainsi

que pour le tahitien, on trouve beaucoup plus de portails que pour les autres langues. On trouve énormément de sites qui concernent les ressources linguistiques pour les créoles, le franco-provençal, les langues kanakes, l'occitan que d'autres types de ressources liées à la langue ou la culture.

Il est aussi intéressant de se pencher sur la langue principale de l'interface ou celle qui accompagne la langue étudiée. Dans 48 % des cas étudiés, le français est la langue principale de l'interface. Pour les sites internet relatifs au corse la langue locale, c'est à dire le corse, est la langue d'interface la plus fréquente. Le breton et le corse arrivent en tête des langues pour lesquelles la langue d'interface est autant le français que la langue locale.

Il est à noter que certaines applications ou espaces ont des statistiques propres qui peuvent ne pas être pertinentes. C'est le cas notamment de Facebook, qui propose une fonctionnalité permettant à l'utilisateur d'indiquer la langue qu'il parle, mais apparemment un nombre négligeable de personnes l'utilisent, raison pour laquelle ces statistiques seraient inexploitables pour l'instant. Il ne s'agit pas de la langue de l'interface choisie par l'utilisateur, mais d'une catégorie du type « *je parle cette langue* ». Nous constatons que les personnes ne parlent pas toujours la langue qu'ils indiquent, et c'est parfois juste un intérêt pour la langue, ou une sorte d'acte militant. L'intérêt est donc symbolique. Cela montre une intention de dire « Moi, j'aime, je parle cette langue, je veux m'exprimer dans cette langue », mais rien de plus.

Il s'agit d'une première étude, qui reste donc partielle puisque nous ne nous sommes pas occupés de l'ensemble des langues, que nous n'avons pas exploité toutes les ressources de manière exhaustive. Mais ces premiers résultats sont intéressants, car ils portent sur un sujet qui n'est pas encore exploré de manière systématique. Nous avons donc besoin d'élargir la recherche à d'autres langues. Nous prévoyons déjà la création d'un méta-site, avec le soutien de la DGLFLF. Nous avons besoin du soutien de la communauté pour participer au recensement des données.

Conclusions

Les conclusions générales auxquelles nous arrivons après ces recherches, aussi bien pour les langues bien répandues que moins répandues, c'est que le web statique est le reflet de la place des langues dans la connaissance pour les langues dites « équipées ». Pour les langues disposant déjà de ressources multimédias très importantes (audio, vidéo, papier, etc.), le web 1.0 représente un équilibre : les langues présentes dans notre vie quotidienne sont très présentes sur le web. La numérisation d'œuvres amplifie cet effet et favorise une fracture entre les langues bien équipées et les autres.

Le web 2.0 semble changer la donne, et les langues moins bien équipées semblent investir cet espace. L'audiovisuel bien sûr pourrait favoriser les langues moins pré sentes, notamment les langues peu ou pas utilisées à l'écrit. Mais il est important de rappeler que l'accès à internet ne suffit pas. Il faut mener des actions d'alphabétisation numérique : rendre les populations actives dans la création de pages, plutôt que passives, dans une démarche qui se limite à la consultation. Les nouveaux outils audiovisuels mobiles sont fondamentaux dans cette démarche. Si les langues reprennent de la valeur vis-à-vis de leurs locuteurs, le cyberespace peut devenir un espace plus équitable linguistiquement.

Inventaire des ressources linguistiques des langues de France

Valérie Mapelli, Agence pour l'évaluation et la distribution des ressources linguistiques (ELDA)

Je suis Valérie Mapelli et je travaille pour la société ELDA, qui est l'agence pour l'évaluation et la distribution des ressources linguistiques. Nous avons l'habitude de réaliser des inventaires de ressources linguistiques, mais c'était la première fois que nous le faisions pour les langues régionales de France.

Commissionnés par la DGLFLF, nous avons cherché à réaliser un inventaire des ressources disponibles dans le milieu des technologies de la langue, ainsi que des ressources potentielles, c'est-à-dire qui pourraient être utiles pour produire des ressources linguistiques. La tâche principale était d'identifier les principaux canaux de production et de diffusion, dont ELDA et l'Association européenne pour les ressources linguistiques (ELRA) font bien sûr partie, et d'obtenir ainsi une liste non-exhaustive des ressources linguistiques pour les langues régionales de France. Le but de l'étude n'était pas forcément de faire seulement un inventaire, c'était aussi d'évaluer l'adéquation des ressources identifiées pour le développement des technologies de la langue, et par conséquent d'en promouvoir le développement pour les langues régionales de France.

L'étude en elle-même était constituée de six grandes étapes : la définition du périmètre, l'inventaire des différentes sources d'information, la collecte des informations et différentes ressources, la focalisation sur quelques langues (en consultation avec la DGLFLF), la réalisation d'une analyse en fonction des besoins des acteurs clés, et enfin la production d'un rapport final.

La définition du périmètre de l'étude visait en particulier à identifier les principaux critères permettant de classer les différentes langues régionales de France. Nous nous sommes basés notamment sur les études et informations disponibles sur le site Ethnologue[1], auprès de la

1 http://www.ethnologue.com/

DGLFLF, etc. Nous avons retenu comme critères les différentes familles de langues, le nombre de locuteurs, les différentes modalités (tradition orale par exemple). Nous avons essayé d'axer le thème autour de six technologies : la traduction automatique, la reconnaissance et la synthèse de la parole, la correction orthographique, l'analyse sémantique, l'analyse grammaticale, et la génération automatique de textes.

Nous nous sommes donc basés sur le site Ethnologue et sur les données de la DGLFLF pour identifier les langues régionales. Nous en avons distingué 84 réparties en 21 familles. Nous avons donc également traité la Langue des signes française (LSF) et dans une moindre mesure les langues non-territoriales. L'inventaire en lui-même est axé non seulement sur la recherche des ressources linguistiques disponibles, mais aussi sur les sources pouvant permettre la production de ressources : journaux, radios, télévisions, sites institutionnels ou culturels pouvant être traités informatiquement pour créer des données multimodales. Les ressources linguistiques quant à elles ont été identifiées à travers les grands canaux de diffusion : le catalogue ELRA[1], *European Language Resources Association* (LRE) Map[2], Meta-Share[3], *Linguistic Data Consortium*[4], l'initiative OLAC[5].

Nous avons aussi établi des critères pour définir l'information en vue de sa compilation : la langue, la description des ressources, leur emplacement sur internet, les fournisseurs, et les droits éventuels associés à leur usage. En ce qui concerne les sources, une ontologie a été définie pour récupérer l'information (nom, description, URL, etc.), la langue, les applications potentielles et le contact nécessaire pour produire ces ressources.

Après nous être demandé comment présenter l'inventaire, nous avons pris la décision de le faire dans une base *MySQL*, permettant non seulement de recenser les informations mais également d'obtenir des statistiques et éventuellement de les enrichir, à la suite du projet. On trouve dans la base le nom de la source, une description très brève, des informations sur la source internet, le type de ressource linguistique (corpus oral

1 http://catalog.elra.info/
2 http://www.resourcebook.eu/
3 http://www.meta-share.eu/
4 https://www.ldc.upenn.edu/
5 *Open Language Archives Community*, http://www.language-archives.org/

ou écrit, grammaire, ontologie, thésaurus, etc.), la langue associée, les applications potentielles, et enfin les informations sur la disponibilité éventuelle associées au fournisseur et aux informations juridiques qui s'y rapportent.

Parmi les dix langues les plus représentées dans le rapport, le breton arrive en premier avec 420 ressources identifiées, le gascon en deuxième position avec 286 ressources identifiées, le languedocien en troisième avec 202 ressources, jusqu'au catalan en dixième position avec 47 ressources.

Concernant les différentes appellations de l'occitan et de ses variantes, des correctifs ont été apportés par la suite.

Le classement par familles de langues est pratiquement le même. Après les langues celtiques et les langues d'oc viennent les langues non-territoriales en troisième position, celles de Polynésie française en quatrième, et les langues de Grande Terre en cinquième.

En concertation avec la DGLFLF, nous avons décidé de nous focaliser sur trois groupes de langues : les langues d'Outre-mer, le breton et l'occitan. Nous avons aussi souhaité ne pas nous concentrer uniquement sur les ressources, mais considérer également les usages en termes de technologies. Nous nous sommes donc concentrés sur la traduction automatique, la reconnaissance et la synthèse vocale, et la correction orthographique, le but final étant d'évaluer la faisabilité de ces technologies pour ces groupes de langues.

Pour les trois langues du focus, le type de ressource majoritaire est du type corpus de parole.

En plus du focus, nous avons réalisé une analyse de notre inventaire sur la base d'autres études similaires, comme *Basic LAnguage Resource Kit* (BLARK)[1] (volonté d'établir un kit de ressources linguistiques de base), *Network for Euro-Mediterranean LAnguage Resources* (NEMLAR) , *Fostering Language Resources Network* (FlareNet)[2] et particulièrement le livre

1 http://www.blark.org/
2 http://www.flarenet.eu/

blanc Meta-Net[1], afin d'estimer les manques en termes de ressources linguistiques pour les langues régionales.

Le livre blanc Meta-Net permet d'évaluer sur une échelle de 1 à 5 la présence de ressources linguistiques selon les langues. Nous souhaitions faire la même chose pour les langues régionales, en comparant avec le français. Cela nous a permis de voir que l'on trouve beaucoup de ressources pour le basque et le catalan, beaucoup moins pour les autres langues régionales.

En ce qui concerne le focus de l'étude, l'occitan et le breton sont placés une nouvelle fois au milieu de l'échelle, alors que les autres langues régionales sont beaucoup moins représentées.

Enfin, nous sommes allés plus loin dans le focus en cherchant à évaluer la faisabilité des technologies selon les langues. Les trois technologies choisies ont été évaluées. Le breton est de nouveau la première langue outillée, l'occitan est bien placé pour la traduction automatique et la correction orthographique mais pas pour la synthèse et la reconnaissance vocale, et enfin les autres langues régionales apparaissent à la traîne.

Le rapport final reprend la méthodologie, l'état de l'art, l'inventaire des ressources, le focus, l'analyse, le bilan et les recommandations.

Il y a une grande disparité entre les langues en termes de représentativité des ressources linguistiques. Nous avons également pu constater que les ressources identifiées sont très variées en ce qui concerne le volume du contenu : une ressource peut aussi bien correspondre à un enregistrement de deux minutes comme à un corpus parallèle de deux millions de mots. Il est donc nécessaire de faire un nouveau travail d'analyse pour affiner l'étude et regrouper les informations pour avoir une meilleure perception de l'exploitabilité des ressources. De plus, nous avons trouvé des problèmes propres à chacune des langues, notamment des problèmes de normalisation. Nous avons pu remarquer certaines initiatives pilotes pour développer des ressources et des technologies associées, par exemple l'initiative de base de données Watreng par l'Académie des Langues Kanak. Il y a une nécessité de cibler les applications aux besoins locaux.

1 http://www.meta-net.eu/whitepapers/volumes/french

Le rapport final est disponible sur le site d'ELDA[1] et sur le site de la DGLFLF[2], ainsi que l'inventaire complet au format tabulaire.

Débat avec la salle

Jean-François Baldi
Nous avons là deux études qui identifient des lacunes aussi bien au niveau des acteurs que des ressources linguistiques. J'ai l'impression qu'il va y avoir un certain nombre de questions. Le temps du débat est désormais ouvert.

Benaset Dazéas (depuis la salle)
Je remercie Valérie Mapelli pour son intervention. Vous avez sans doute senti qu'il y avait des réactions un peu spontanées.

Effectivement, sur nos territoires nous sommes face à des réalités qui sont d'ordre politique, presque idéologique. La question de l'occitan en Aquitaine est un peu sensible, notamment sur ce que l'on considère comme une langue, comment une langue est composée de plusieurs dialectes, tout en souhaitant ne pas laisser d'ambiguïté sur son unité. Malheureusement sur notre territoire c'est un combat, car quelques minorités de personnes font le choix de s'opposer à cette notion d'unité. Pour nous, il est très clair qu'il y a une langue occitane, avec certes une variété. C'est intéressant de voir comment on peut s'emparer de vos données, et encore une fois merci, car c'est riche et constructif, et comment on arrive, dans un second temps peut-être – je ne me permettrai pas du tout de dire de corriger – mais de se mettre d'accord sur des termes partagés et communs pour une utilisation et une vulgarisation plus large face à un grand public. Mais c'est très utile pour nous, donc merci encore.

Valérie Mapelli
C'est une étude qui était assez riche, comme vous l'avez exprimé, très limitée en temps et en financement aussi. Pour rappel, il nous a quand

1 http://www.elra.info/en/projects/archived-projects/review-existing-lrs-france/
2 http://www.culturecommunication.gouv.fr/Politiques-ministerielles/Langue-francaise-et-langues-de-France/Politiques-de-la-langue/Langues-et-numerique/Les-technologies-de-la-langue-et-la-normalisation/Inventaire-des-ressources-linguistiques-des-langues-de-France

même fallu inventorier toutes les ressources correspondantes aux 84 langues. C'était déjà un pari important. Effectivement, il est toujours possible de s'améliorer. Je pense aussi que l'outil créé nous permettra de coopérer sur l'amélioration de ces informations.

Thibault Grouas (depuis la salle)
Je souhaiterais apporter une précision pour faire suite à ce que dit Valérie en ce qui concerne le rapport sur lequel nous avons en effet travaillé avec vous. La question de l'occitan s'est posée, Xavier North avait beaucoup parlé de ce sujet, avec l'apparition de différentes familles, différentes langues occitanes dans le rapport. Effectivement, les documents scientifiques étaient catalogués sous les différentes appellations régionales de la langue occitane. Nous avions donc essayé de regrouper. Dans le document que vous avez, vous trouvez un regroupement par famille où vous voyez bien l'occitan. C'est très important, sans quoi en effet c'est un peu confus, notamment le doublon entre « occitan auvergnat » et « auvergnat » qui n'a aucun sens, surtout si « occitan » apparaît à côté. Il ne faut donc pas le prendre en compte mais plutôt regarder le tableau où l'on voit le rassemblement par famille, ce qui était notre souhait. C'est clair, en tout cas pour nous.

Khalid Choukri (depuis la salle)
Je voudrais ajouter un mot à cela : c'est un inventaire, c'est-à-dire que l'on répertorie ce que l'on a trouvé sous différentes nomenclatures. Il y a des cas où « auvergnat », si je reprends le terme, est indiqué. Cela permet de savoir qu'il s'agit d'occitan auvergnat. Nous aurions dû trouver une façon de l'indiquer. Parfois c'est indiqué « occitan », au même titre que parfois on trouve « anglais », alors que l'on a l'anglais d'Angleterre, l'anglais d'Australie, etc. Nous sommes donc obligés de faire une catégorie « anglais ». Nous aurions pu l'appeler occitan « générique », c'est-à-dire là où les gens n'ont pas précisé la famille, mais si l'on veut aller plus loin, et je suis ravi de voir les différentes réactions, car cela promet qu'on ira assez loin, on peut regarder site par site et essayer de les catégoriser de nouveau avec le savoir-faire des uns et des autres.

Michel Alessio (depuis la salle)
Les catégoriser de nouveau en ayant le souci de la rigueur et de la cohérence interne de votre travail, parce que c'est ce qui fait parfois

défaut. Au-delà du cas occitan qui est le plus net, je pense aux langues kanakes, qui apparaissent sous différentes orthographes.

Valérie Mapelli
Il me semble que cette information a été corrigée dans le rapport.

Michel Alessio
Oui, mais on trouve parfois une catégorie « Grande-Terre », je pense que cela fait référence aux langues kanakes de Nouvelle-Calédonie, mais on ne sait pas pourquoi Grande-Terre apparaît à ce moment-là. Il faudrait peut-être veiller à une plus grande cohérence des données ou de leur rendu.

Valérie Mapelli
Pour Grande-Terre, nous avions parlé de famille de langues, donc pas de langues au pluriel.

Michel Alessio
Oui, mais il est difficile de concevoir, par exemple, une famille de langues « Grande-Terre ». Non, il y a les langues kanakes, que ce soit en Grande-Terre ou dans les quelques îles des archipels de Nouvelle-Calédonie, il n'y a pas de différence. Comment se justifie cette répartition en plusieurs appellations ? Il faudra peut-être les retravailler, tout simplement dans le sens d'une plus grande clarté.

Abraham Bengio (depuis la salle)
Que penserait-on d'une enquête statistique où l'on trouverait les catégories australien, néo-zélandais, canadien, anglais de diverses régions d'Angleterre ? On découvrirait de ce fait que l'anglais n'est plus du tout ni de près ni de loin la première langue ! Les forces seraient divisées en autant de catégories. Ce serait absurde, évidemment personne ne songerait à le faire pour l'anglais, mais pour l'occitan cela paraît naturel de ne pas regrouper, de ne pas gommer ce qui n'est que des différences d'appellation des différents dialectes d'une seule langue. Ou alors, si l'on n'est pas d'accord sur le fait que c'est une seule langue, il y a un vrai sujet de débat de politique linguistique, mais pas seulement. Mais il me semble que ce n'était même pas ça : aujourd'hui on prend les mots tels qu'ils figurent sur la toile et on ne cherche pas à les consolider.

Jean-François Baldi
Oui, c'est ce que M. Choukri indiquait, c'est un inventaire effectivement, il a recueilli ce qu'il a collecté.

Nourdine Combo (depuis la salle)
En ce qui concerne les langues de Mayotte, nous avons répertorié quatre ressources. Maintenant avec l'arrivée du haut débit, de plus en plus de textes sont écrits et mis en ligne par des locuteurs natifs. Nous avons aussi des ressources plus fiables. Mayotte 1re, la première chaîne radio, propose chaque semaine une émission appelée Za Lada, où des poètes locaux, des jeunes qui s'amusent à écrire et à s'exprimer en mahorais sont invités, et ces textes sont traduits en français. En plus de cela, il y a des émissions de télévision : par exemple les informations sont diffusées chaque soir dans une version en mahorais, depuis quelques années, pas encore en shibushi. Tout cela forme des ressources qui peuvent être exploitées.

Traitement de langues régionales en Europe : quelques exemples marquants

Président de séance : **Benaset Dazeàs**

Ressources et technologies de la langue catalane

Asuncion Moreno, Universitat Politècnica de Catalunya

Cette présentation concernera les ressources linguistiques et technologiques du catalan.

Après l'introduction, je présenterai les ressources linguistiques, puis les applications développées, et enfin une discussion des manques.

Comme vous le savez tous, le catalan est parlé en Catalogne, dans la communauté valencienne, aux Baléares, dans le sud de la France (ce que nous appelons la Catalogne du nord), en Andorre et à Alghero (Sardaigne). Nous formons plus ou moins une communauté de 10 millions de personnes.

Les principaux acteurs de ce travail sont l'Institut d'Estudis Catalan, qui dépend de la *Generalitat de Catalogne*, les universités catalanes et du pays valencien, des entreprises multinationales qui ne se trouvent pas en Catalogne mais qui font des choses pour le catalan (Microsoft, Nuance, Google), et des entreprises locales plus petites (Verbio, Clic, Incyta), etc.

Les financements proviennent de trois origines : les projets de recherche

du gouvernement catalan, les projets de recherche nationaux espagnols, et des financements européens.

Nous avons des ressources orales et écrites, des terminologies, et des ressources multimédias.

Les ressources orales sont des bases de données annotées pour la reconnaissance automatique de la parole, la synthèse vocale, l'identification du locuteur, etc. La qualité est plutôt bonne, car elle correspond aux normes européennes reconnues et que le contrôle de qualité est réalisé en externe. Toutes ces bases de données sont libres.

Pour la reconnaissance de la parole, nous avons des bases pour les communications téléphoniques, les applications automobiles, les applications grand public. Tout est annoté. Cela représente environ 7 000 personnes enregistrées, et près de 100 heures d'enregistrement pour chaque base.

En plus de ces bases, nous disposons de dialogues spontanés, comme le corpus parallèle *TALP tourism dialogues* (enregistrements de 160 dialogues pour une durée de 22 heures dans le domaine du tourisme, avec transcription orthographique des propos et annotation du bruit environnant et des disfluences de la parole spontanée), des enregistrements de la télévision (34 programmes de la chaîne TV3 enregistrés pour 68 fichiers audio d'une durée totale de 43 heures) et de la radio catalanes (80 heures de journaux télévisés dont 19 retranscrites, les données restantes sont segmentées et annotées en fonction de la taille, de l'état de l'enregistrement et du locuteur).

Nous avons aussi une base pour la synthèse vocale, *FESTCAT Catalan TTS baseline speech database*, avec deux locuteurs professionnels, un homme et une femme, de dix heures chacun. Les corpus proposent une couverture phonétique et prosodique. Les annotations orthographiques, phonémiques et prosodiques ont été réalisées manuellement. Les marques de hauteur ont été annotées manuellement sur deux heures puis automatiquement. La segmentation phonétique a également été réalisée manuellement sur deux heures. Le lexique comprend une transcription phonétique de chaque énoncé.

Enfin nous avons aussi *Technical meetings*, une base de données de conférence, et Glissando, un corpus bilingue établi par le gouvernement espagnol de quarante heures d'enregistrements de 28 locuteurs (professionnels et non-professionnels), transcrit orthographiquement et phonétiquement et comportant l'annotation prosodique des contours de la F0[1].

Pour les ressources textuelles, nous disposons de corpus de référence, de corpus sémantiques, de corpus parallèles, de lexiques, d'ontologies, de ressources terminologiques et de dictionnaires.

Les corpus de référence sont le *Corpus Textual Informatizat de la Llengua Catalana* (CTILC), 52 millions de mots lemmatisés et annotés morpho-syntaxiquement), le Corpus technique de l'IULA[2] (plusieurs spécialités rassemblées, 23 millions de mots lemmatisés et annotés morpho-syntaxiquement, sections de textes parallèles en catalan, espagnol, anglais), et *Annotated Corpora-Catalan* (AnCora-CA) et *Annotated Corpora in Dependency-based representation-Catalan* (AnCora_DEP-CA), moins importants mais avec plus d'annotations (500 000 mots issus de textes journalistiques avec divers types d'annotations : lemme, catégorie morphologique, constituants et fonctions syntaxiques, structure de l'argument, rôles thématiques, classe sémantique verbale, etc.). Le premier est financé par le gouvernement catalan, le deuxième par le gouvernement catalan et l'Europe, et le troisième par le gouvernement espagnol.

Nous avons des corpus de références très volumineux tirés du web. Il s'agit du Wikicorpus (750 millions de mots en catalan, espagnol et anglais, tirés d'une partie de Wikipedia et annotés automatiquement avec des informations morphologiques) et du CUCWeb (Corpus d'utilisation du catalan sur le web, 166 millions de mots automatiquement compilés à partir du web et annotés avec le lemme, la catégorie morphologique et la fonction syntaxique).

Les corpus parallèles sont assez rares, puisque les langues régionales comme le catalan ne sont pas reconnues par le Parlement Européen. Nous avons le *Corpus Lingüístico da Universidade de Vigo* (CLUVI), qui regroupe

1 Fréquence de vibration des cordes vocales
2 *Institut Universitari de Lingüística Aplicada*

5,5 millions de mots dans quatre langues officielles d'Espagne : espagnol, galicien, catalan et basque. Le deuxième est un corpus parallèle espagnol-catalan de plus de 100 millions de mots issus de dix ans d'articles du *Periodico de Catalunya* automatiquement post-édités et traduits. Les données sont alignées au niveau de la phrase et stockées dans des fichiers de texte brut. Le dernier, *TALP Tourism Dialogue*, est un corpus parallèle en espagnol, catalan et anglais de transcriptions de 100 heures de dialogues dans le domaine du tourisme.

Ensuite, en ce qui concerne les lexiques, qui sont très importants, nous avons le lexique phonétique catalan LC-STAR, utilisé pour la reconnaissance et la synthèse vocale, composé de 50 000 noms communs fréquents et 50 000 noms propres avec leurs transcriptions phonétiques et leurs catégories syntaxiques. Nous avons également deux lexiques morphologiques, El corrector (640 000 entrées, 71 000 lemmes) et Apertium (11 800 lemmes avec des informations morphologiques), et enfin le lexique PAROLE, étiqueté morpho-syntaxiquement et sémantiquement, que vous connaissez certainement puisqu'il est disponible dans douze langues européennes.

En ce qui concerne les corpus sémantiques, nous avons *Sentence Semantics* (SenSem), qui contient 100 phrases pour chacun des 250 verbes espagnols les plus fréquents traduits en catalan. On y trouve la structure de l'argument, les modèles de sous-catégorisation (avec la fréquence), les rôles sémantiques et l'information sémantique de la phrase pour chaque verbe.

Pour les ontologies, l'université de Barcelone a développé un WordNet dans le cadre d'un projet européen (41 991 synsets), et une version élargie de ce premier avec la participation de l'IULA, qui fait partie du dépôt central multilingue et qui intègre des versions de WordNet pour l'espagnol, le basque, le galicien et l'anglais (46 442 synsets). Enfin, le référentiel *Multilingual Center Repository* (MCR) relie des WordNets dans différentes langues ou versions avec d'autres ontologies (*Suggested Upper Merged Ontology*, FrameNet, VerbNet, etc.).

Dans le domaine des ressources terminologiques, qui sont également très importantes, nous disposons de TERMCAT, un centre de terminologie

catalane, qui propose une consultation gratuite en ligne avec Cercaterm. En plus de ça, nous avons d'autres banques de terminologie comme *Banc de dades terminològiques de l'Universitat Pompeu Fabra Barcelona (*UPF_ Term), un vocabulaire de l'énergie, un vocabulaire de base du génome humain, un lexique pour la prévention des risques professionnels.

En ce qui concerne les dictionnaires, nous avons d'abord ceux de la famille d'Apertium (anglais, français, italien, portugais, espagnol et aranais, allant de 9 000 à 33 000 entrées), un dictionnaire d'anciens textes catalans (PDC) qui fournit des informations sur le thème, la catégorie grammaticale, les formes, etc., et le Dacco, un dictionnaire libre catalan-anglais (21 000 entrées) et anglais-catalan (28 000 entrées) contenant des fichiers audio, des listes de prêts importants, des moteurs de recherche, des réseaux sémantiques, la conjugaison des verbes et des informations sur la fréquence relative de chaque entrée.

Maintenant, pour les applications et les outils, nous disposons de FreeLing[1], développé par l'université polytechnique de Catalogne, qui réalise des analyses grammaticales et sémantiques dans treize langues différentes. Nous avons un correcteur de textes appelé El Corrector, libre et multi-platesformes, développé par l'université Pompeu Fabra (UPF) de Barcelone avec l'aide de Barcelone Media. En ce qui concerne le traitement de l'oral, nous avons des systèmes de reconnaissance de la parole, de synthèse de la voix, d'identification du locuteur, des systèmes de dialogue. Ces outils sont utilisés pour les centres d'appels, le sous-titrage automatique. C'est très utile puisqu'en Catalogne il est obligatoire de sous-titrer les films et les informations en catalan, ce qui est très coûteux à faire réaliser par des êtres humains.

Nous avons aussi des systèmes de traduction automatique, notamment celui d'Apertium, entre différentes langues (espagnol-aranais, catalan-aranais, breton-français).

En conclusion, dans le cas de la langue catalane, nous sommes modérément optimistes quant à l'état actuel de soutien aux technologies langagières. La communauté de recherche en Catalogne, soutenue par les programmes de recherche espagnols et catalans, est viable. Toutefois, la

1 http://nlp.lsi.upc.edu/freeling/

portée des ressources et la gamme d'outils sont encore très limitées par rapport à celles existant pour la langue espagnole (et évidemment pour la langue anglaise), et elles ne sont tout simplement pas suffisantes en termes de qualité et de quantité pour développer le type de technologies nécessaires pour soutenir une société véritablement multilingue.

En ce qui concerne l'industrie des technologies de la langue catalane dédiée à la transformation des résultats de la recherche en produits, elle est actuellement très faible. La plupart des grandes entreprises ont soit arrêté soit fortement réduit l'élaboration de technologies de la langue, reléguant les langues parlées seulement par un petit nombre de personnes à un objectif secondaire.

Il n'y a aucun programme de recherche, ni au niveau national ni au niveau régional pour les technologies de la langue. Il y a un manque de continuité dans les financements de la recherche et du développement. Des programmes coordonnés à court terme ont tendance à alterner avec des périodes de financements rares ou inexistants. Il y a un manque général de coordination avec les programmes d'autres pays de l'UE ainsi qu'au niveau de la Commission Européenne.

Nos résultats montrent que la seule alternative est de produire un effort considérable pour créer des ressources de technologies linguistiques pour le catalan, et de les utiliser pour faire avancer la recherche, l'innovation et le développement. Vu la nécessité d'avoir de grandes quantités de données et l'extrême complexité des systèmes de technologies de la langue, il est essentiel de développer une nouvelle infrastructure et une organisation de recherche plus cohérente afin de stimuler un plus grand partage et une meilleure coopération. Nous pouvons donc conclure qu'il y a un très grand besoin d'une large initiative coordonnée destinée à surmonter les différences de maturité dans les technologies de la langue pour les langues européennes dans leur ensemble.

Offre active et « choix » linguistique : le cas du Pays de Galles

Jeremy Evas, Prifysgol Caerdydd

La diglossie, en socio-linguistique, désigne une situation où deux langues coexistent dans le même territoire, utilisées par les mêmes personnes mais à des fins différentes, l'une des langues ayant un statut plus élevé que l'autre (utilisation dans le cadre public et utilisation dans le cadre privé par exemple).

Cette situation a mené Steve Eaves, poète, compositeur et chanteur d'expression galloise, à préciser que « Il n'y a aucune tradition d'attente de service en langue galloise » et par conséquent, le taux d'utilisation des services en langue galloise peut souvent être très bas.

Dans cette situation, quelles sont les implications de la technologie pour les langues ?

Les questions posées dans cette présentation sont les suivantes :

• la sensibilisation peut-elle être un moyen d'accroître l'utilisation d'interfaces ou de services en langue galloise, bretonne, catalane, etc. ? Ou bien s'agit-il d'un changement d'attitude à adopter ?

• Peut-on changer le comportement qui s'est construit pendant plusieurs siècles ? Si oui, comment, et qui doit le faire ?

Les économistes comportementaux Fogg, Chanasyk, Quihuis, Moraveji, Hreja et Nelson posent un postulat selon lequel « croire que l'information mène à l'action est l'une des dix erreurs les plus courantes des techniques de changement du comportement »[1].

1 Fogg, B.J., Chanasyk, K., Quihuis, M., Moraveji, N., Hreja; J., & Nelson, M. (2010). *Top 10 Mistakes in Behaviour Change*. Persuasive Tech Lab, Stanford University : http://www.slideshare.net/captology/stanford-6401325

Les points à prendre en compte pour la conduite d'un changement de comportement linguistique peuvent comporter les caractéristiques suivantes (tirées du modèle britannique *Messenger Incentives Norms Defaults Salience Priming Affect Commitment Ego* (*MINDSPACE))* :

• le messager : la source d'une information a une influence importante sur la façon dont elle est considérée ou prise en compte par les usagers ;
• l'incitation : les réponses aux incitations sont définies par des processus mentaux prévisibles, comme chercher à tout prix à éviter la perte ;
• les normes : ce que les autres font a une influence très forte sur notre propre comportement ;
• les manques : nous nous laissons entraîner par les options pré-définies;
• la prépondérance : notre attention est attirée par ce qui est nouveau et ce qui nous semble pertinent ;
• l'amorçage : nos actes sont souvent influencés par des preuves inconscientes ;
• l'affect : les associations émotionnelles peuvent influencer lourdement les actions ;
• les engagements : nous cherchons à tenir nos engagements ;
• l'ego : nous agissons de façon à nous sentir mieux avec nous-mêmes.

À travers cette présentation nous analysons une méthodologie particulière pour mener des changements de comportements linguistiques qui consiste à modifier les réglages linguistiques par défaut de certains services en ligne, et d'introduire la notion d' « offre active », une solution largement éprouvée au Canada, qui est en situation de bilinguisme.

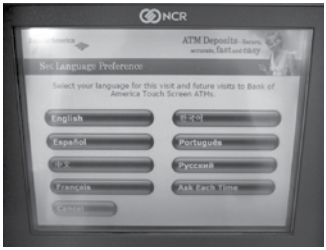
Cette présentation analyse les résultats d'une étude portant sur différentes façons et modalités de proposer un choix linguistique de manière proactive à l'usager. Nous constatons que plus le choix proposé est clair, compréhensible et sans équivoque, plus les usagers se tourneront vers l'alternative linguistique qui leur est proposée.

Voici quelques exemples d' « offre active » :

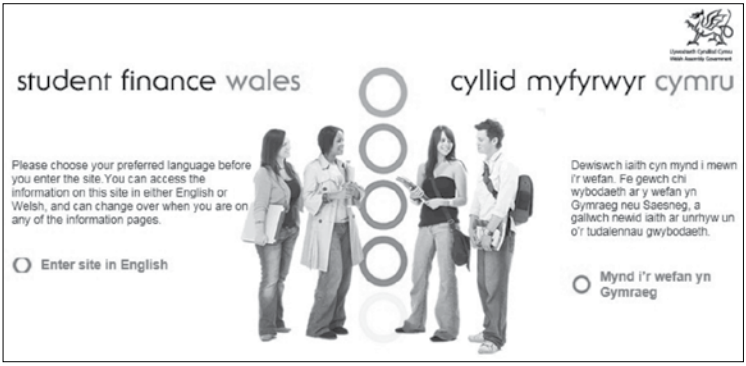
Site internet du Cardiff Council

Site internet Canada Games

Site internet du Cyngor County council



Distributeur automatique



Site internet de l'Assembly government gallois

La technologie de la langue comme outil efficace pour la promotion des langues avec peu de ressources : le cas de la langue basque

Kepa Sarasola, Euskal Herriko Unibertsitatea

Les technologies nous aident dans la construction d'une société multilingue. Une personne lambda, avec l'aide des technologies du langage, est capable de comprendre et de communiquer dans plus de langues qu'il y a 20 ans. Alors, on peut dire que l'usage de ces technologies est une bonne chose pour les langues minoritaires dotées de peu de ressources. Cependant, l'aide apportée n'est pas suffisante. Parmi les points à améliorer figurent le cadre juridique, l'enseignement (à l'université notamment), les médias, les sphères publique, sportive et commerciale, la justice, et, bien sûr, l'impulsion citoyenne. Sans tout cela, c'est impossible. Mais qu'apportent les technologies de la langue ? Je vais vous présenter l'expérience que nous avons menée avec le basque.

Le premier point de la présentation concerne les langues dans le contexte des technologies de l'information et de la communication et des technologies de la langue. Ensuite, je vais vous présenter le travail de recherche sur le basque réalisé au sein du groupe Ixa de l'Université du Pays Basque.

Concernant les langues et dans le contexte des technologies de l'information et de la communication, le livre blanc de Meta-Net conclut que le basque dispose de ressources, d'outils et d'applications, mais que des recherches plus poussées sont nécessaires pour que les solutions technologiques soient prêtes pour une utilisation quotidienne. Le développement de technologies de haute qualité pour le basque est urgent et d'une importance capitale pour la préservation de la langue.

Si l'on compare avec l'anglais et le français, le basque vient en dernier, mais nous sommes quand même satisfaits des résultats qui ne sont pas si mauvais.

Il n'est pas facile d'obtenir des chiffres sur la quantité de ressources disponibles pour les langues sur Internet, et donc d'identifier les langues minoritaires.

Selon des statistiques de 2010[1], 536 millions d'usagers d'Internet, soit 27%, utilisent l'anglais. Le top 10 des langues représente 1 616 millions d'usagers, soit 82% (anglais, chinois, espagnol, japonais, portugais, allemand, arabe, français, russe et coréen). Le reste des langues représente 351 millions d'usagers, soit 18%.

En ce qui concerne le nombre de documents sur le web, on trouve peu de statistiques fiables pour les différentes langues. Celles dont je dispose aujourd'hui datent de 2007, elles sont donc assez anciennes. Il s'agit d'une étude sur la présence des langues romanes (Union Latine) qui indique que 45% des pages web sont écrites en anglais, 7,80% en espagnol, 5,9% en allemand, 4,41% en français, 2,66% en italien, 1,39% en portugais, 0,28% en roumain, 0,14% en catalan. L'anglais se démarque donc fortement.

Le nombre d'entrées dans Wikipedia constitue aussi une statistique intéressante. En juin 2014, on trouve des articles écrits dans 286 langues. L'anglais se classe en premier avec 4,54 millions d'articles, suivi du hollandais (1,78 millions) et de l'allemand (1,73 millions). Les langues régionales sont aussi représentées, mais pas en tête : 17e position pour le catalan, avec 429 000 articles ; 35e position pour le basque, avec 181 000 articles ; 54e pour l'occitan, avec 87 000 articles ; et 71e place pour le breton, avec 50 000 articles.

Parmi les référentiels publics, ELRA propose plus de 1 000 ressources pour 60 langues (certaines d'entre elles sont gratuites pour la recherche), dont 6 produits pour le basque, mais la recherche par langue n'est pas possible. De la même manière, le *Linguistic Data Consortium* (LDC) concentre plus de 500 ressources pour 82 langues ; la recherche par langue n'est pas possible non plus, et il n'y a pas de produits disponibles pour le basque. L'*Association for Computational Linguistics* (ACLWiki) rassemble des ressources pour 73 langues dont 15 pour le basque ; cette fois, la recherche par langue est possible. Le *Natural Language Software Registry* du *Deutsches Forschungszentrum für Künstliche Intelligenz* (DFKI)

1 Internet World Stats

rassemble des ressources pour 30 langues dont trois produits pour le basque ; la recherche par langue est possible aussi dans ce cas. De plus, 59 produits sont utilisables pour n'importe quelle langue.

Yourdictionary.com rassemble des ressources lexicales pour plus de 300 langues ; la recherche par langue est possible, et on trouve 9 liens vers des dictionnaires du basque.

La présence ou l'absence des langues dans les technologies les plus populaires (traitements de texte, moteurs de recherche, traduction automatique, etc.) est également significative. Pour les logiciels de traitement de texte, le basque est présent dans les deux programmes les plus utilisés : Microsoft Word (91 langues en tout) et LibreOffice (104 langues). Pour les moteurs de recherche, Google identifie 50 langues, et son service Translate traduit environ 80 langues (dont le basque). Babelfish en traduit quatorze.

Après ce bilan chiffré, on peut se demander quelles sont les langues minoritaires, c'est-à-dire, celles avec le moins de ressources. La réponse est relative, et l'on peut distinguer six niveaux différents. Au premier niveau on trouve l'anglais, qui est de loin la langue la mieux dotée en ressources et pour lequel pratiquement tous les types d'applications du langage existent. Le deuxième niveau constitue le top 10 des langues les plus représentées sur le web, qui regroupe 82 % des usagers de l'Internet, anglais y compris, suivi de l'allemand, du français, de l'espagnol, de l'italien, du portugais… Pour ces langues, le développement actif de ressources langagières continue, même si la plupart des applications de technologies de la langue sont représentées. De même, la plupart des ressources décrites dans des référentiels comme LDC ou ELRA sont disponibles pour ces langues. Le troisième niveau est celui des langues qui possèdent une ou plusieurs applications de technologie de la langue : celles de ELRA, LDC, ACLWiki, etc. Le quatrième niveau est constitué par des langues qui possèdent des ressources lexicales voire des dictionnaires en ligne : 307 langues sont représentées sur yourdictionary.com, ce qui correspond pratiquement au même ensemble de langues présentes dans Wikipedia. Le cinquième niveau est celui des langues qui possèdent un système d'écriture (plus de 2 000 langues), et, finalement, au sixième niveau on trouve des langues non-écrites (plus de 4 500 langues).

Les appellations de « langues avec le moins de ressources » ou « minoritaires » sont donc relatives. Si l'on compare avec l'anglais, toutes les langues sont minoritaires. En fait, lorsque l'on dit « langues minoritaires », ce sont les langues de troisième ou quatrième niveau (celles qui possèdent une ou plusieurs applications langagières). Les langues relevant du cinquième et du sixième niveau sont vraiment en danger (du point de vue de leur utilisation dans les technologies de l'information et de la communication).

Cette classification n'est pas stricte, mais elle peut être utile pour reconnaître des domaines d'application et pour dessiner des stratégies de développement des ressources langagières.

La stratégie de développement des technologies linguistiques pour le basque menée par le Groupe Ixa est conduite selon l'idée principale qu'il ne faut pas mettre la charrue avant les bœufs : il faut d'abord placer les fondations avant de développer des applications. Au début, nous souhaitions travailler sur la traduction automatique, mais nous nous sommes rendu compte que c'était impossible sans avoir au préalable étudié, par exemple, la morphologie du basque, qui est très compliquée. Le chemin emprunté est différent de celui qui a été choisi pour le développement des technologies du langage pour l'anglais, pour lequel les ressources langagières n'ont pas évolué à la suite d'un plan unique et coordonné, et qui ont été produites par beaucoup d'efforts indépendants, afin de répondre aux besoins spécifiques de projets concrets.

Au sein du Groupe Ixa, le développement de ressources pour le traitement de la langue ont été planifiées et organisées, car nous sommes peu à travailler sur la question et, donc, il est essentiel de travailler en équipe, en coordonnant les efforts et en partageant les tâches. Dans le groupe de recherche, fondé en 1988, nous sommes maintenant une cinquantaine de personnes, informaticiens et linguistes, la plupart enseignants à l'université (donc le temps de travail n'est pas consacré uniquement à la recherche).

La stratégie de recherche que nous menons est de concevoir et de développer les bases des technologies langagières d'une façon progressive et planifiée, afin d'en tirer le meilleur bénéfice. Nous cherchons à normaliser les ressources afin de les utiliser dans des recherches

variées, pour développer des outils divers, des applications et de produits différents, avec l'adoption, par exemple, des recommandations de *Text Encoding Initiative* (TEI) et de standards comme *Extensible Markup Language* (XML) comme base pour l'étiquetage aux différents niveaux du traitement de la langue (méthodologie générale pour l'annotation des corpus).

En prenant comme référence notre expérience dans la conception et le développement de ressources et d'outils, nous proposons une stratégie générale à quatre phases pour le développement d'une infrastructure du traitement automatique d'une langue. La priorité stratégique est d'aller de la recherche fondamentale vers le développement d'applications.

Dans la phase 1 on pose les fondations. Pour les outils, il faut développer un *tokenizer* (pour reconnaître les mots) ; pour les ressources, il faut des corpus, des descriptions de la morphologie et des grammaires, et des bases de données lexicales.

Dans la phase 2 on développe les premiers outils et applications basiques tels que le correcteur orthographique, le lemmatiseur, un analyseur morphologique, des outils statistiques de traitement de corpus ; pour les fondations et ressources, on développe des bases de données lexicales enrichies, des corpus étiquetés morphologiquement, des descriptions computationnelles de la morphologie, des dictionnaires.

La phase 3 concerne des outils et des applications plus avancées, comme l'enseignement des langues assisté par ordinateur, les outils de vérification et correction de la grammaire, les correcteurs orthographiques avec reconnaissance d'entités nommées, les analyseurs syntaxiques, les *crawlers*, les *parsers*, les techniques de désambiguïsation du sens des mots.
Enfin, la phase 4 aboutit à des applications multilingues et générales comme la traduction assistée par ordinateur et l'extraction de l'information, et, pour les ressources, la constitution de corpus étiquetés sémantiquement, par exemple.

Dans notre groupe, chacune de ces étapes a duré environ six ans.

Les bases de données lexicales, les lemmatiseurs/étiqueteurs et les corpus sont indispensables pour traiter une langue minoritaire. Les outils d'analyse sémantique et syntaxique, l'identification des entités nommées ne viennent que par la suite.

Actuellement, Ixa participe à plusieurs projets, par exemple : un projet européen sur la qualité de la traduction (*Quality traduction by deep language engineering approaches*), un projet du gouvernement espagnol intitulé *Traducción automática en contexto y aumentada con recursos dinámicos de internet (*TACARDI), et un projet de recherche stratégique au niveau du Pays basque, Berbatek.

État des lieux et des besoins pour quelques langues régionales en France

Président de séance : **Jeremy Evas**

Langue bretonne et nouvelles technologies : une vitalité à soutenir

Olier Ar Mogn, Office public de la langue bretonne / Ofis publik ar Brezhoneg

Demat d'an holl. Laouen on o kinnig stad an teknologiezhoù nevez e brezhoneg dirazoc'h hiziv e Meudon. Va gerioù kentañ a lavaran e brezhoneg dre ma soñj din ez eo mat lakaat klevet un tamm ar yezhoù a gomzomp kement diwar o fenn.[1]

Après les trois interventions que l'on vient d'entendre sur le gallois, le basque et le catalan, j'ai l'impression que l'on va redescendre d'un cran. Ma présentation va faire un état des lieux de tout ce qui existe en langue bretonne.

Tout d'abord, quelques mots sur la langue bretonne

On estime que le breton est parlé par 200 000 personnes, avec en outre 30 000 apprenants. Les foyers actuels de développement de la langue sont des foyers urbains, à savoir les trois principales villes de Bretagne : Rennes, Brest, et Nantes. Plus de 15 000 enfants sont scolarisés dans

1 Bonjour à tous. Je suis heureux de présenter un état des nouvelles technologies en langue bretonne devant vous aujourd'hui à Meudon. Je prononce mes premiers mots en breton, car je pense qu'il est bon de faire entendre un peu les langues dont nous parlons tant.

l'enseignement bilingue. Le Conseil régional de Bretagne a adopté un plan de politique linguistique en 2004, réactualisé et approfondi en 2012 avec des objectifs très précis, notamment en ce qui concerne les nouvelles technologies. De plus, il est à signaler que nous avons la chance que le breton soit une langue normée, avec un standard. La question des versions différentes ne se pose pas pour ce qui relève des nouvelles technologies.

Quelques particularités du breton

Le breton présente un certain nombre de digraphes voir trigraphes insécables. Il comporte également un système de mutations consonantiques : l'initiale d'un mot peut changer de multiples façons, suivant son rôle dans la phrase ou le mot qui précède. Ces particularités représentent des difficultés supplémentaires pour le traitement de la langue, notamment en ce qui concerne les moteurs de recherche. Enfin, le breton utilise le « ñ », ce qui a amené certaines personnes à créer un clavier spécifique prenant en compte les spécificités de la langue bretonne.

Les principaux acteurs des nouvelles technologies sont les suivants :

• L'Office Public de la langue bretonne, établissement public de coopération culturelle (EPCC), qui travaille directement à la réalisation d'outils et réalise un travail de lobbying pour développer la place de la langue bretonne dans les nouvelles technologies.

• Les associations, puisque le monde associatif est très actif dans le domaine des nouvelles technologies en Bretagne. Certains y voient une faiblesse, j'y verrais plutôt une grande force, car cela signifie d'abord que la société civile est demandeuse, et ensuite que les Bretons se prennent en main pour créer les outils dont ils manquent, sans attendre forcément que les choses viennent d'en haut. Parmi ces associations, l'une d'entre elles a obtenu la création de l'extension *.bzh*, qui est une nouveauté. Nous avons également une myriade d'informaticiens bénévoles, qui traduisent et amènent la langue bretonne dans les nouvelles technologies.

Il nous faut cependant remarquer l'absence des universités sur ce terrain en Bretagne.

Comme je l'évoquais tout de suite, nous avons ce *.bzh*, « *pik bzh* » comme on dit en breton. Nous l'avons obtenu après plus de dix ans de gestation et le soutien très fort du Conseil Régional. La Bretagne est la première région de France disposant de sa propre extension, et nous pensons que cela va contribuer à donner une meilleure visibilité aux ressources linguistiques en ligne.

Une autre grande avancée (fin 2014) est l'ouverture à la traduction de l'interface de Facebook. Cela n'a pas été chose facile, mais l'Office Public de la langue bretonne là aussi a joué son rôle de prise de contact, épaulé par une vraie demande sociale. Il est donc désormais possible utiliser l'interface de Facebook en breton. La traduction a été très rapide et atteint un taux de 91 % aujourd'hui. Sachant que le gallois est à 94 % et le catalan à 98 %, nous constatons qu'en trois mois un bon travail a été fait. Environ 500 personnes ont contribué à cette traduction. Cela montre qu'il y avait du monde derrière prêt à mettre la main à la pâte.

Voilà deux bonnes nouvelles de la fin 2014 montrant la vitalité de la langue bretonne.

D'autre part, l'Office public de la langue bretonne est l'organisme référent pour le *Common Locale Data Repository* (CLDR) du Consortium Unicode, et nous assurons le suivi afin que ce répertoire de données soit renseigné régulièrement. Pour ce qui concerne la langue bretonne, nous sommes en ce moment au plus haut niveau, dit « *comprehensive* ».

En ce qui concerne les ressources terminologiques, nous ne sommes pas dépourvus non plus. Nous avons au sein de l'office public de la langue bretonne un service de terminologie qui s'appelle TermBret. Ce centre a une base de données en ligne, TermOfis, dont la consultation en ligne est gratuite et qui présente actuellement plus de 63 000 termes, tous domaines confondus. C'est une base de données très fréquemment consultée, qui a eu une grande influence sur la normalisation de la langue, notamment dans la presse et les médias. On peut également citer deux autres sources en matière de terminologie :

• La base Brezhoneg21, qui est un dictionnaire des sciences et techniques émanant du mouvement des écoles bilingues Diwan. Il est intéressant

de noter que l'on assiste actuellement à un mouvement de convergence des différents centres produisant de la terminologie pour aller vers plus de coordination, notamment entre l'office public de la langue bretonne et Brezhoneg21.

• Et une autre base en ligne de vocabulaire informatique de presque 5 000 termes : *Geriadur Preder ar stlenneg*.

Pour ce qui concerne les dictionnaires, je citerai d'abord un travail important lancé officiellement il y a à peine deux ans, Meurgorf. C'est un dictionnaire historique numérique présentant déjà plus de 54 000 entrées, 82 000 formes historiques attestées et 650 000 formes fléchies. Des ressources lexicales en ligne sont disponibles.

Nous disposons également de correcteurs de textes, aussi bien orthographiques, grammaticaux, que les deux en même temps. Ce sont des produits qui sont en perpétuelle amélioration mais qui sont déjà utilisables. Dans le domaine de la traduction automatique, on trouve un outil développé par l'office public de la langue bretonne, en collaboration avec l'université d'Alacant. C'est un produit en constante amélioration, qui pour le moment ne fonctionne que dans le sens breton-français. C'est un parti pris, dans la mesure où si le grand public est demandeur du sens français-breton, il ne serait pas forcément en mesure d'appréhender la correction de la langue produite, et nous ne souhaitons pas courir le risque de voir apparaître, sur des dépliants touristiques par exemple, des textes incorrects qui seraient attribués au traducteur automatique de l'office public de la langue bretonne.

Au sujet des corpus de textes, nous disposons actuellement d'un concordancier de 43 000 phrases français-breton. Pour l'oral, nous avons deux sites présentant des enregistrements de brittophones de naissance, retranscrits sous leur forme dialectale, standard et phonétique, et traduits en français. Les quatre séquences sont disponibles côte à côte.

Pour la grammaire de la langue, nous avons un site qui s'appelle ARBRES, qui étudie la morphologie, les constituants, la syntaxe de la phrase, etc.

En ce qui concerne les logiciels et applications, beaucoup de travail a été

fait, souvent bénévole, pour des logiciels comme LibreOffice, OpenOffice, Mozilla, etc.

Quant à la matière en ligne, voici quelques exemples qui font sentir la dynamique et la vitalité de la langue : EduBreizh, un site d'autoformation gratuit, la place du breton sur Wikipedia qui est plutôt bonne, les sites de livres électroniques, les vidéos, etc.

Parmi les développements que nous considérons comme vitaux pour la langue bretonne, nous comptons l'accession à la synthèse vocale. En effet, si la langue bretonne a commencé à trouver sa place dans la société à l'écrit, il faut maintenant l'entendre dans les bus, dans les trains, dans les services, sur les smartphones et les GPS. C'est un pas décisif à franchir. L'obstacle est essentiellement financier, sachant que lorsqu'on pense aux enjeux, les besoins financiers ne devraient pas être un obstacle insurmontable.

Je conclurai en citant les obstacles à surmonter :

• La question du suivi : traduire un logiciel une fois c'est bien, mais il est nécessaire de le suivre, sinon cela ne sert à rien. Pour cela, des structures pérennes sont nécessaires.

• Nous avons des soucis concernant la correction de la langue, notamment dans le cadre des initiatives privées qui peuvent donner des produits de qualité inégale.

• L'absence d'interfaces, comme dans le cas des logiciels Microsoft. Il y a une grande attente d'un soutien de la part de l'État pour la localisation de produits tels que ceux de Microsoft, pour mettre le breton sur les smartphones par exemple. On sait qu'en Espagne il y a eu des démarches communes pour le basque, le galicien, le catalan, pour faire en sorte que les opérateurs soient incités à proposer des services dans ces langues. Nous pensons que l'État pourrait jouer un rôle dans le cas des langues de France.

Pour terminer sur des points positifs, l'office public de la langue bretonne assure ce suivi des traductions, tant que possible et tant

que les ressources humaines le permettent. De nombreuses initiatives spontanées continuent, avec une nouvelle génération de brittophones de plus en plus demandeuse et encline à utiliser ces nouvelles technologies. Il y a également le travail de stabilisation de la langue des nouvelles technologies, le chemin parcouru dans les dix dernières années est phénoménal. Il faut bien sûr aussi penser aux collaborations internationales, que l'on peut évoquer dans le cadre de ce colloque, qui seront à nouveau évoquées à Bruxelles au mois de mars et que nous continuerons à suivre au plus près.

La langue corse numérique, un chantier

Sébastien Quenot, Collectivité territoriale de Corse /
Capiserviziu di u Cunsigliu linguisticu

Après les interventions portant sur le breton, le basque, le catalan, le gallois, qui ne posent plus la question de l'élaboration des outils mais de leurs usages, on voit qu'en Corse on se croit assez avancé mais qu'en fait on joue dans une catégorie inférieure, et qu'il s'agit véritablement d'un chantier à ouvrir.

Je vais vous présenter dans un premier temps la politique linguistique de la Collectivité Territoriale de Corse (CTC). Nous disposons depuis plus d'une trentaine d'années d'un statut particulier : nous sommes une collectivité territoriale disposant d'une compétence relativement pleine et entière en matière de politique linguistique, néanmoins contrainte par le cadre constitutionnel. Ensuite, je vous présenterai les équipements numériques concernant la langue corse, les ressources, les acteurs et les moyens, et pour finir, en troisième partie, nos projets et intentions en matière de planification linguistique, notamment le plan qui devrait être voté dans les jours à venir concernant le plan Lingua 20201, une planification totale concernant la langue corse, au-delà des nouvelles technologies.

Depuis 1991, la CTC met en œuvre la politique culturelle. Le premier plan pour la langue corse, axé principalement sur l'éducation et sur les médias a été adopté en 1997. La loi du 22 janvier 2002 a intégré l'enseignement de la langue corse dans le cadre de l'horaire normal des classes maternelles et élémentaires, ce qui est une avancée considérable : l'enseignement du corse est offert à tous les élèves, obligatoire pour l'État mais pas pour les familles, et l'assemblée de Corse a la responsabilité d'adopter un plan de promotion de la langue corse. Ce n'est qu'à partir de 2005 qu'il y a véritablement une orientation sociétale dépassant les secteurs de l'éducation et des médias, avec l'adoption à l'unanimité de l'Assemblée des élus de Corse en 2007 du plan stratégique d'aménagement et de développement pour la langue corse. En mai 2013 a également été

1 Planification adoptée par l'Assemblée de Corse le 14 avril 2015 disponible sur : www.corse.fr/file/161442/.

adoptée une proposition de statut pour la coofficialité de la langue corse, de façon à lui donner des droits linguistiques, puisque nous pensons qu'il faut à la fois disposer d'un socle juridique et élaborer différents outils pour mener une politique de revitalisation, et plus encore, une politique de normalisation linguistique.

Concernant quelques éléments relatifs à la vitalité de la langue corse, nous avons mené une enquête en 2013. Sur 320 000 habitants nous compterions 100 000 locuteurs corsophones, tous bilingues avec le français. 58 % déclarent comprendre le corse bien ou assez bien, 28 % le parleraient bien. 26 % des 18-24 ans, la tranche d'âge la plus jeune que nous ayons prise en compte, seraient plus à même de parler le corse que la tranche d'âge supérieure des 25-34 ans. Est-ce dû à l'enseignement, à la présence du corse dans les médias, aux effets de l'évolution de certaines représentations linguistiques, ou bien aux effets de l'importance de l'immigration sur la tranche d'âge 18-24 ans ? En tout en état de cause, nous observons un tassement chez les plus jeunes, ce qui est déjà plutôt positif même si le pourcentage est toujours beaucoup trop faible.

Au niveau des pratiques, 93 % des personnes sondées écoutent de la musique en langue corse, 50 % en chantent, 18 % envoient des SMS en langue corse. Enfin une donnée importante, en famille l'usage du français est majoritaire, mais le monolinguisme est minoritaire (il y a également d'autres langues parlées en Corse). 90 % des sondés pensent qu'à l'avenir il faudrait parler corse et français, il y a donc un fort désir de langue. D'autre part concernant la structuration de la politique linguistique, depuis 2010, le budget consacré à la langue a fortement augmenté : il a été multiplié par trois et demi, ce qui dans un contexte de restriction budgétaire est quand même important, même si nous sommes à un niveau de dépense par habitant d'environ 7€, en comparaison avec 20€ pour la Catalogne.

Au niveau de l'équipement numérique du corse, voici un bilan très succinct au niveau des responsabilités et des actions qui ont été menées.

Nous ne sommes pas à zéro. Actuellement, la CTC pilote et finance principalement la politique de création d'outils et de nouvelles technologies. Nous projetons à travers le *Cunsigliu de la lingua* de créer un dictionnaire général de la langue corse ainsi que plusieurs déclinaisons, de la traduction

automatique, le recensement général de la toponymie corse de façon à alimenter les GPS, etc. L'université, qui est un partenaire très important en Corse, a également développé différents outils : un atlas avec la banque de données linguistiques (BDLC), la publication en novembre dernier de la M3C, la médiathèque culturelle de la Corse et des Corses, sur laquelle on peut trouver de très nombreux documents (à la fois en corse et en français mais aussi dans d'autres langues portant sur la Corse), ainsi que des méthodes d'apprentissage du corse, sur CD ou DVD et qui sont maintenant en cours de mise en ligne sur le site InterRomania pour développer le e-learning, qui est aussi une voie très intéressante. Le Centre Régional de Documentation Pédagogique (CRDP), que nous appelons maintenant le réseau Canopé, a mis en ligne tous les documents qu'il a pu produire.

S'il y a une quinzaine d'années nous disposions de peu de documents, réalisés de bric et de broc par les différents enseignants, aujourd'hui nous pouvons dire que quasiment toutes les disciplines sont couvertes, de la maternelle au secondaire. Tous ces documents sont accessibles gratuitement en ligne, ce qui est un acquis très important. Ils sont également mis à jour, au fur et à mesure de la réactualisation des programmes. Récemment, nous avons même assisté à une polémique que l'on peut voir comme positive : des livres en langue corse ont été passés au pilon. Finalement, le CRDP a réagi en annonçant la publication de nouveaux manuels mis à jour. Paradoxalement, la destruction de manuels papier en langue corse est plutôt bon signe, il y a quelques années cela aurait été totalement inespéré.

Au niveau de la société civile, nous avons des blogues en langue corse qui marchent très bien, tenus par des personnes qui ne sont pas forcément corsophones, avec des textes faciles d'accès et un peu décalés quant à l'actualité. Nous avons aussi emboîté le pas au breton pour traduire Facebook en corse, ce qui est compliqué à cause des nouvelles terminologies qui ne sont utilisées pratiquement que sur Facebook : il faut donc veiller à l'inter-compréhension entre le traducteur et l'utilisateur. Nous disposons aussi d'un dictionnaire en ligne réalisé par l'Association pour le Développement des Études Archéologiques, Historiques, Linguistiques et Naturalistes du Centre-Est de la Corse (ADECEC), doyenne des associations de promotion de la langue corse.

Cela dit, j'avais réalisé une étude en 2008-2009 sur la corsophonie, internet et les nouvelles technologies, la place du corse dans la cyberguerre mondiale des langues[1], et si l'on fait un état des lieux depuis, je crois que nous n'avons pas tellement avancé. Les initiatives importantes relèvent plutôt du niveau individuel, ce qui me semble être un signe de faiblesse, d'autant plus que lorsqu'on est 300 000, dont 100 000 locuteurs corsophones, la masse est moins importante que pour les Bretons par exemple, qui peuvent avancer davantage. Lorsqu'on compare les wikis des deux langues, la forte distorsion est visible. Il ne s'agit pas seulement d'un effet de volonté ou de compétences mais vraiment d'un effet de masse. Les institutions n'ont pas suffisamment avancé, malgré les dispositifs votés qui sont encore aujourd'hui mis en œuvre, comme la charte de la langue corse signée par plusieurs communes.

L'un des axes principaux dans les années à venir va donc être la mise en œuvre de cette planification adoptée par le conseil exécutif de l'assemblée de Corse, et qui devrait être adoptée par l'assemblée de Corse en avril prochain. Elle devra s'appuyer sur le moyen stratégique que représente la coofficialité même si celle-ci nécessite de nouveaux aménagements notamment constitutionnels.

Notre planification s'appuiera sur cinq objectifs opérationnels : permettre à chacun d'apprendre le corse, quel que soit son âge, sa situation professionnelle ou ses origines, offrir à chaque locuteur un maximum d'opportunités d'usage de la langue corse, créer les conditions de l'offre de service bilingue par les organismes publics et privés, veiller à la qualité de l'équipement de la langue, et veiller au rayonnement interne et externe de la langue corse.

Les nouvelles technologies de la langue sont concernées par l'ensemble de ces cinq objectifs opérationnels.

Concernant le plan pluriannuel que nous comptons mettre en œuvre pour le développement d'outils numériques, il se compose de trois volets :

- Le développement de programmes pour les applications numériques

1 Disponible sur : https://halshs.archives-ouvertes.fr/docs/00/40/81/26/DOC/Le_corse_dans_la_cyberguerre_mondiale_des_langues_quenot_teramo.doc.

pour la langue corse, comme ce qui s'est déjà fait pour le catalan, le basque ou le gallois par exemple ;

• L'encouragement des grandes entreprises à localiser leurs applications en langue corse ;

• Le développement des représentations positives de ses propres compétences, comme ce qu'évoquait Jeremy tout à l'heure, en encourageant l'usage des outils numériques en langue corse.

Les deux obstacles les plus importants pour la Corse sont les suivants :

• L'obstacle financier : nous sommes 300 000, mais que l'on soit 60 millions ou 300 000 les coûts de traduction et d'élaboration linguistique sont les mêmes. C'est pour cela qu'il faudrait peut-être que le ministère reconsidère vraiment sa politique financière en matière de langues de France, de façon à y apporter une contribution. Même si la collectivité définit la politique linguistique, les financements doivent pouvoir affluer non seulement du ministère mais également de l'Europe. Actuellement, le fait que le corse ne soit pas une langue officielle le prive de financements européens.

• Un problème de formation : à l'université, nous avons d'un côté une filière « études corses », orientée sur l'anthropologie, la sociolinguistique, et de l'autre côté de l'université sur un autre campus, nous avons la filière informatique. Ce sont deux mondes très distincts, avec des objectifs très différents, très avancés chacun dans leurs domaines mais cloisonnés. Il faut essayer de mutualiser, non seulement sur notre campus mais plus largement, certains projets entre différentes langues. Je crois que c'est l'objet de ce colloque et c'est une première marche à franchir avant de poursuivre ce chantier que nous entendons poursuivre dans les mois à venir.

Je vous remercie.

Le numérique au service de la transmission de la langue occitane : situation et perspectives de développement

Gilbert Mercadier et Aure Séguier
Lo Congrès permanent de la lenga occitana

Gilbert Mercadier
Tout le monde ne comprend pas encore l'importance du numérique pour notre langue. Récemment, je marchandais avec un auteur de dictionnaire qui m'a dit, pour me refuser gentiment d'entrer sur notre site et dans notre base de données, que si les étudiants voulaient consulter son dictionnaire, il fallait qu'ils aillent à la bibliothèque. Nous avons donc du chemin à faire. Nous sommes un peu en retard et ça nous fait un drôle d'effet d'arriver à la fin de ce beau panorama de langues qui ont plus de chances, et de moyens aussi.

Le Congrès que l'on représente, tout jeune, essaie depuis sa création de faire un gros effort et se concentre essentiellement sur le numérique.

Le Congrès, qu'es aquò ?

C'est une fédération des associations et des institutions qui se préoccupent de la langue occitane. Il bénéficie du soutien des collectivités territoriales, dont certaines sont représentées ici, et aussi de la DGLFLF, que je tiens à remercier pour son soutien.

Ce Congrès est constitué d'un conseil scientifique, présidé par un professeur de l'université Toulouse Jean Jaurès, et d'un conseil des usagers représentant la demande sociale. Il doit agir en fonction de quelques grands principes, notamment, chez nous c'est important, le respect de l'unité, mais aussi de la diversité, car si nous ne respections pas suffisamment la diversité, les gens qui doivent adhérer à la langue et à la culture occitane ne s'y reconnaîtraient pas et n'y adhéreraient pas.

Pour répondre aux besoins des usagers, le Congrès a développé une plate-forme numérique. Je tiens à dire que le Congrès a trois ans. Sur notre site, votre trouverez un multi-dictionnaire occitan. Jusqu'en 2011, il n'y avait pas de dictionnaire sérieux occitan en ligne. Nous avons aussi un dictionnaire historique, un conjugueur, une base terminologique, une base toponymique et une rubrique pour les normes et les œuvres normatives. Nous avons aussi en chantier bien avancé un lexique inter-dialectal, mettant en valeur ce qui est commun à la langue occitane, qui est bien une langue, ainsi qu'un grand dictionnaire informatisé. Il faut bien dire que depuis le grand Mistral et son *Trésor du Félibrige*, nous n'avons pas de dictionnaire de l'ensemble de la langue moderne. Je dirais que le site du Congrès, modestement, a été pensé comme un service public pour la langue occitane, et que finalement au bout de peu de temps avec plus de 140 000 visites dans l'année 2014, il est bien devenu ce service public. Dans les collèges et les lycées où l'on enseigne l'occitan, on trouve au CDI une application directe pour joindre le site du Congrès.

Bien sûr, tout cela n'est pas suffisant. Nous continuons d'être en retard, même si d'autres initiatives devraient être citées. Je pense notamment à Wikipedia où vous pouvez trouver une bonne place pour l'occitan. Il y a aussi des démarches scientifiques comme la base de données textuelles en occitan (BaTelÒc) à Toulouse, ou le Thesoc à Nice. Il faut bien dire qu'il y a une dynamique, certes insuffisante, mais qui existe. Sans le numérique, la langue occitane, déjà minorée et marginalisée, le serait encore plus.

L'occitan s'étale sur un ensemble linguistique de 14 millions de personnes sur 3 États de l'Union Européenne (France, Espagne, Italie : 32 départements français, Val d'Aran en Espagne et quelques vallées italiennes du Piémont). Comme les autres langues régionales, l'occitan a subi l'érosion de l'histoire. Sur les 14 millions d'habitants qui vivent dans cet espace, on peut considérer qu'un certain nombre (10 % environ mais les estimations varient beaucoup) le parlent encore, souvent des personnes âgées, même si l'enseignement a progressé. Nous comptons 80 000 enfants scolarisés bénéficiant d'un enseignement bilingue ou, pour une grande majorité, d'une initiation à la langue occitane. Gardons l'espoir d'obtenir demain très vite les moyens qui nous permettront de réaliser tous les outils que nous avons à réaliser. C'est bien pour cela que le Congrès a décidé de s'attaquer à ce retard numérique, pour essayer de contribuer, dans ce domaine au moins, à un

véritable avenir pour notre langue.

Je dois remercier l'association de développement des Pyrénées pour la formation (ADEPFO) ainsi que tous les membres du comité de pilotage qui nous ont aidé et qui ont soutenu notre initiative qui a abouti à un diagnostic et à une feuille de route pour le développement du numérique occitan. Vous pourrez voir que c'est le résultat d'une recherche-action, c'est-à-dire que les formés ont été les acteurs de leur formation avec des experts et ont produit le diagnostic et la feuille de route. J'en profite également pour remercier la fondation Elhuyar. Quand on est pauvre et un peu en retard comme nous, on va voir ceux qui sont à côté, les basques par exemple.

Ce document qui va vous être présenté est véritablement une première pour l'occitan. Mais nous savons bien qu'en lui-même il ne suffira pas, parce que si en pays d'oc on dit traditionnellement « *la fe sens òbra, mòrta es* », la foi sans les œuvres est morte, aujourd'hui on devrait ajouter « *la fe sens moneda, sens mejans, mòrta es tanben* » : vous avez compris, la foi sans moyens est morte. Il faudra que tous ceux qui y sont attachés ou devraient l'être un peu plus, y compris notre État, qui pour le moment nous traite parfois un peu avec des soins palliatifs plutôt qu'avec une véritable politique linguistique, nous aident véritablement. Finalement, la clé de la réussite de ce programme, ce sera bien les crédits qu'il nous faudra trouver ensemble.

Je passe donc la parole à Aure Séguier pour vous présenter ce travail dont elle est une des chevilles ouvrières, et j'en profite aussi pour la remercier et remercier notre directeur, Benaset Dazeàs.

Aure Séguier

Je suis Aure Séguier, *webmaster* du Congrès, qui est à l'origine de l'idée de cette feuille de route, dont le but était de définir des objectifs spécifiques et précis pour développer les technologies du langage pour l'occitan d'ici 2019. Pour ce faire, nous avons réuni différents acteurs de la transmission de la langue occitane et nous avons procédé en trois étapes.

La première étape était de réaliser un inventaire de tout ce qui existe pour l'occitan en matière de ressources et outils linguistiques.

Ensuite, nous avons fait intervenir des experts qui nous ont expliqué l'expérience de quelques langues. À partir de ces expériences et de notre inventaire, nous avons établi un diagnostic des besoins et rédigé le calendrier pour essayer d'y répondre.

Le diagnostic portait sur les outils et les ressources linguistiques, et l'inventaire a été réalisé par les intervenants : chacun a essayé de recenser tout ce qu'il connaissait. Ce n'était sans doute pas exhaustif mais nous avons pu obtenir un bon aperçu de ce qui existe.

Nous avons voulu différencier les ressources linguistiques brutes nécessitant un prétraitement de celles qui sont utilisables directement en informatique : corpus de textes bruts ou corpus de textes déjà annotés. Cela nous a permis de constater qu'il y a beaucoup de ressources potentielles, mais que très peu sont utilisables en informatique. Les outils sont également peu nombreux. Il y a donc du travail.

Quelques exemples : corpus oraux non transcrits, corpus textuels non annotés, dictionnaires numérisés pas toujours *OCRisés*[1] et jamais balisés, grammaires à visée pédagogique non-transcrites pour être utilisées en algorithme informatique, et questions de la validité et de la qualité linguistique. Il faut créer des ressources de base avant de développer des outils. Nous avons donc fait intervenir quatre experts, pour les langues catalane, basque, bretonne et galloise. Chacun a donné ses impressions, ses expériences et des choses ont convergé, notamment l'importance de bien planifier pour ne pas se disperser et perdre de l'énergie, développer autant les ressources que les outils, et surtout l'intérêt de la collaboration, pour laquelle les licences libres sont primordiales afin de pouvoir échanger les ressources et les outils déjà construits.

Nous sommes ensuite passés à la feuille de route elle-même. Il a fallu définir des besoins, les prioriser, rédiger la feuille de route, et estimer l'ordre de réalisation sous la forme d'un calendrier.

Pour cela, une fois que nous avions identifié les besoins, nous avons identifié les dépendances : quels outils ou ressources sont nécessaires

1 *OCRiser* ou *océriser* : transformer automatiquement (un fichier contenant l'image d'un document) en fichier texte, grâce à un logiciel OCR.

pour construire quels autres outils. Certains outils, comme la base lexicale monolingue pour le clavier prédictif et l'auto-correction par exemple, sont absolument nécessaires. En revanche, on peut se passer des corpus au début, mais ils deviendront nécessaires pour perfectionner l'outil à partir de données statistiques.

Nous avons ensuite dressé le calendrier.

Pour les corpus, nous avons décidé de faire un corpus parallèle occitan-français et deux corpus monolingues :
• un premier qui serait un corpus assez cadré d'une grande validité linguistique pour lequel on peut demander un certain nombre de textes littéraires, un certain nombre de textes d'actualités, etc., pour des utilisations bien avancées ;
• un corpus web monolingue qui pourra être constitué automatiquement grâce au détecteur d'occitan.

Pour les ressources lexicales :

• une base lexicale monolingue, avec une première version rapide et une deuxième version plus poussée en 2017 ;
• une base lexicale bilingue.

La base grammaticale n'est pas vraiment une base, il s'agit plutôt de définir des règles de grammaires aisément transposables en algorithme informatique et langage de programmation.

Les technologies de traitement de la parole sont assez difficiles à réaliser. Nous n'avons pas souhaité mettre la priorité sur la reconnaissance et la synthèse de la parole d'ici 2019, car cela risquerait d'empêcher beaucoup d'autres réalisations. Nous commençons en revanche à identifier les ressources nécessaires à la reconnaissance vocale et à les créer. La synthèse vocale, elle, sera développée seulement en 2019.

Pour le détecteur d'occitan et le détecteur de variantes, des travaux sont déjà en cours.

Il existe plusieurs correcteurs orthographiques mais nous cherchons à en faire un seul, proposant une utilisation pour toutes les variantes et

pour tous les logiciels les plus courants (traitement de texte, messagerie, navigateur).

En ce qui concerne le clavier prédictif et l'autocorrection, cela sera utile surtout pour les mobiles.

Enfin, les deux outils primordiaux pour réaliser d'autres outils sont l'analyseur morphologique et l'analyseur syntaxique, qui sont indispensables pour perfectionner la traduction automatique ou annoter des corpus. La base de connaissance lexicale sera une base reliant les mots entre eux avec des relations d'hyperonymie, de catégorisation, de champs lexicaux, pour permettre notamment la désambiguïsation de mots en traduction automatique.

Enfin, nous arrivons au traducteur automatique occitan-français à partir de toutes les variantes. En effet, si des projets existent déjà, ils ne concernent que certaines variantes. L'idée est d'en avoir un qui fonctionne avec toutes et dans le sens français-occitan. La question des variantes s'est posée : est-ce qu'il faut pour chaque outil développer une version pour chaque variante ? Est-ce que l'on préfère choisir une variante, mais dans ce cas-là laquelle et sur quels critères ? La solution est venue avec le transcripteur automatique entre les variantes, qui permettrait de passer un texte d'une variante de l'occitan à une autre. Nous développerions ainsi uniquement un traducteur français-occitan languedocien par exemple, auquel on appliquerait ensuite le transcripteur occitan languedocien-occitan gascon : on obtiendrait un traducteur automatique français-occitan gascon, et cela résoudrait le problème de multiplicité des outils.

Enfin, nous voulions proposer aux utilisateurs un pack constitué d'un système d'exploitation et des principaux logiciels dans la langue, très simple d'installation, pour 2017.

En conclusion, la feuille de route est une stratégie harmonisée pour le développement, et un cadre cohérent duquel peuvent déboucher des actions concrètes. Il ne faut pas partir de rien, du travail a déjà été fait et il faut le continuer tout en maximisant la coopération et l'effort collectif.

Débat avec le public

Philippe Boula de Mareüil (depuis la salle)
Vous avez été plusieurs à évoquer la synthèse vocale, notamment en breton. Or, il y a tout un vivier autour de Lannion avec le logiciel Voxygen. Leur avez-vous demandé de chiffrer le coût de développement d'un système pour le breton ? J'adresse cette question également concernant le corse et l'occitan.

Olier Ar Mogn
Oui tout à fait, nous connaissons bien les gens de Voxygen, nous sommes entrés en contact avec eux. Pour vous répondre très concrètement, il nous faut trouver 200 000 euros.

Gilbert Mercadier
Le tarif doit être le même pour les autres.

Aure Séguier

En ce qui nous concerne, nous avons réalisé la feuille de route cette année. Nous sommes donc en train de passer à la mise en œuvre pour les financements. Pour le moment, la synthèse vocale est tout à la fin.

Joseph Mariani (depuis la salle)
On a vu très rapidement dans les transparents de Sébastien des chiffres défiler en face des actions à mener. J'aimerais savoir si ces chiffres vont être proposés au vote pour la mise en place du programme, en demandant donc des financements qui viendraient de la collectivité de Corse. De la même manière, pour l'occitan, est-ce qu'il y a des chiffres à mettre en face du programme que vous nous proposez, et qui les finance ?

Aure Séguier
Le chiffrement est la seconde étape de la feuille de route. Il fallait d'abord définir les besoins priorisés. Nous sommes en train de définir les acteurs, en fonction desquels les financements seront cherchés selon les soutiens possibles. Nous travaillons aussi en collaboration avec d'autres langues, qui pour certaines ont déjà développé des compétences, donc tout dépend vraiment des acteurs.

Gilbert Mercadier
Le président vous dira qu'il faut aller taper à toutes les portes, y compris celles des collectivités territoriales qui ont voulu le Congrès et qui le soutiennent : celles du ministère de la Culture, celles aussi de l'Europe probablement, même si comme vous l'avez évoqué, il est aujourd'hui beaucoup plus difficile d'obtenir des financements de la part de l'Europe pour les langues et cultures régionales qu'autrefois.

Benaset Dazéas (depuis la salle)
Pour répondre à M. Mariani, la feuille de route étant assez récente, on ne connaît pas encore le budget total. Je prends bonne note de ce qui concerne la synthèse vocale. Pour le moment nous sommes en train de faire un résumé de cette feuille de route, nous cherchons à voir quels partenaires potentiels peuvent intervenir sur les différents développements. Deux développements par exemple vont être lancés très vite : un sur le traducteur automatique et un autre sur une base de données lexicale. Nous sommes en train de réfléchir sur un dossier de Fonds européen de développement régional (FEDER) pluri-annuel, sur les crédits-recherche, puis il y a tout ce qui concerne la coopération trans-frontalière, puisque l'occitan, comme on le disait tout à l'heure, est une langue qui est parlée sur trois États et qui en plus est voisine de deux autres langues dont il y avait des représentants tout à l'heure, le basque et le catalan. Cette proximité nous a déjà permis de développer certaines ressources avec la fondation basque Elhuyar. Ce sont les trois grandes pistes que l'on voit pour le moment. C'est à creuser, et j'espère que d'ici un mois nous aurons quelques chiffres un peu plus avancés.

Joseph Mariani
Vous parlez de synthèse vocale avec un prix de 200 000 euros, et vous avez présenté une façon de gérer la traduction en passant par une transcription de différentes variantes, est-ce que c'est aussi quelque chose sur lequel vous allez réfléchir ? Par exemple, pour la synthèse vocale, si vous voulez faire de l'occitan, tout en traitant aussi le gascon, est-ce que vous aurez besoin de deux fois 200 000 euros, ou est-ce que vous trouverez une façon plus intelligente et moins chère de traiter les deux ?

Aure Séguier
Je pense que là le transcripteur de dialectes n'est pas directement utilisable

tel quel, mais il est certain que nous ne développerons pas cinq ou six outils en partant de zéro. L'outil sera le même, seules les règles syntaxiques changeront. Théoriquement, nous devrions obtenir de meilleurs résultats que pour de la traduction automatique de langue à langue, puisqu'il s'agit de la même langue, donc le taux d'erreur sera beaucoup moins élevé. On peut déjà le constater entre le catalan et l'occitan, le taux d'erreur est bien moindre par rapport à une paire de langue « classique ».

Gilbert Mercadier
Il existe déjà un traducteur pour passer du catalan à l'occitan.

Approches scientifiques : traitement linguistique et automatique

Président de séance : **Olivier Baude**

Variation et norme : des transferts linguistiques aux transferts technologiques

Philippe Boula de Mareüil, laboratoire de recherche en informatique pluridisciplinaire (LIMSI) – Centre national de la recherche scientifique (CNRS)

Le langage permet de communiquer mais aussi de refléter notre identité. L'argument selon lequel, si tout le monde parlait français (ou anglais), la compréhension serait plus facile, est difficile à contrecarrer. Il était déjà en place chez l'abbé Grégoire, dans son *Rapport sur la nécessité et les moyens d'anéantir les patois*, dont j'ai choisi un extrait en guise d'exergue :

> Proposerez-vous [...] des traductions ? Alors vous multipliez les dépenses [...]. Ajoutons que la majeure partie des dialectes vulgaires résistent à la traduction, [...] les uns [...] sont absolument dénués de termes relatifs à la politique ; les autres sont des jargons lourds et grossiers, sans syntaxe déterminée, parce que la langue est toujours la mesure du génie d'un peuple.

Aujourd'hui, fort heureusement, on ne s'exprime plus en ces termes ; on s'attriste plutôt de la mort des langues, de même que de la disparition des espèces animales. Je ne pousserai pas trop loin ce parallèle qu'a fait Claude Hagège avec les espèces vivantes, car la langue est avant tout une

construction sociale. Toujours est-il que ce rapport de l'abbé Grégoire s'inscrit dans une lignée d'ouvrages normatifs, au XVIII[e] siècle, comme *Les gasconismes corrigés*, qui avaient pour but de gommer les particularités régionales. Je pense qu'aujourd'hui ici nous sommes tous attachés à la diversité des langues, qui toutes nous apprennent quelque chose sur l'homme.

Voici le plan de mon exposé :

La question de la norme linguistique se pose rapidement, dès lors qu'on s'intéresse à la variation à l'intérieur des langues. C'est donc par là que je commencerai.

Je prendrai ensuite l'exemple du corse, de l'occitan et du catalan, que j'ai eu l'occasion d'étudier, et qui montrent entre autres différentes stratégies possibles dans la façon de poser des questions, au niveau intonatif, avec de possibles transferts vers le français.

Pour finir, j'élargirai la problématique des transferts linguistiques aux transferts technologiques.

On peut distinguer au moins deux types de normes : une norme prescriptive, qui nous indique un modèle promu comme correct, et une norme statistique, définie par des usages quantifiables. Pour les langues régionales de France minorées, qui nous occupent, en l'absence de norme prescriptive acceptée par tous, une grande variation entre en ligne de compte. Au niveau lexical, la chose est bien documentée, mais c'est une difficulté à laquelle on se heurte vite lorsqu'on va mener des enquêtes où l'on demande de traduire par exemple la fable *La bise et le soleil*, un texte qui a été étudié dans nombre de langues par l'Association Phonétique Internationale, à des fins d'analyse acoustique. Est-ce que, pour traduire ce mot, on a recours à un équivalent de *tramontane*, ou à d'autres formes, comme, en catalan, « *bispa* » ou « *vent geliu* », en occitan, « *cisampa* » ou « *(aura de) bísia* », en corse, « *zilefra* » ou « *ventulellu* » ?

On a également affaire à la variation orthographique : est-ce que, pour transcrire l'occitan, on adopte plutôt la graphie alibertine, également dite « classique », ou la graphie mistralienne, que certains provençaux préfèrent, comme c'était le cas de mes informateurs lorsque je suis allé faire des

enregistrements autour d'Avignon ? Il y a aussi évidemment la variation phonétique, notamment prosodique. Est-ce que, par exemple, un mot comme « *boulgara* » est accentué sur l'antépénultième syllabe, ou bien comme c'est le cas chez certains locuteurs corses ou catalans roussillonnais, sur la pénultième ? Ces problèmes de la variation seront développés tout à l'heure par Pascal Vaillant, dans le cas des langues créoles.

Pour les langues comme le corse, le statut de langue polynomique, c'est-à-dire relativement tolérante vis-à-vis de la variation, est généralement admis. C'est du reste pour cette langue que ce concept a été élaboré par le sociolinguiste Jean-Baptiste Marcellesi, mais une orthographe phonétique trouve vite ses limites. En corse, « nous montons » se dira « *cullemu* » dans le nord, « *cuddemu* » dans le sud, tandis qu'à Corte on peut entendre « *cullimu* ». Faut-il le transcrire orthographiquement ? Je crois savoir que ceci est relativement bien accepté. En revanche, il semble que la chute du 'l' en Balagne ou dans d'autres régions ne se transcrit pas en général. Dans un cas comme dans un autre, les problèmes restent entiers pour le traitement automatique, qui nécessite un minimum de standardisation.

Cependant, on peut s'interroger sur la faisabilité et l'acceptabilité par les acteurs eux-mêmes de cette entreprise de standardisation des langues de France. À trop vouloir standardiser, en effet, on court le risque d'une double diglossie entre la langue régionale, qui peut être ressentie comme une construction savante par certains, et d'une part le français ; d'autre part la langue de la famille, du quotidien, de la connivence, de l'immédiateté. Je crois que ce n'est pas le lieu ici d'essayer de résoudre ces problèmes d'aménagement linguistique et d'action glottopolitique.

Étant chercheur, on m'a demandé d'introduire la problématique scientifique. Je vais donc le faire en présentant un travail dont l'objectif était d'enrichir un atlas dialectologique, en appliquant un protocole commun, centré sur l'intonation notamment des questions. J'ai pu mener des enquêtes auprès de locuteurs corses, occitans, catalans, à qui on demandait de dire des phrases relativement similaires dans différentes langues romanes, comme, en français « la touriste malade trouve la caserne », corse « *a turista malata trova a caserna* », occitan « *la torista malauta troba la caserna* », catalan « *la turista malalta troba la caserna* ». De telles phrases ont fait l'objet d'analyses acoustiques, qui ont donné des résultats que je résume rapidement. En

Corse, on a majoritairement un ton haut initial et une descente mélodique finale. En occitan, on a majoritairement, dans les questions terminées par un paroxyton, c'est-à-dire un mot accentué sur l'avant-dernière syllabe, des patrons mélodiques montant-montants (sur les deux dernières syllabes), et de même en français en contact avec l'occitan. Et en catalan, dans ces questions terminées par un paroxyton, on a majoritairement un patron descendant-montant (sur les deux dernières syllabes), qui se trouve parfois, mais pas systématiquement, transféré au français régional.

Cette étude à base de phrases transparentes dans différentes langues romanes a permis de mettre en lumière différentes stratégies cognitives pour poser des questions. Nous avons mené des expériences perceptives notamment à base de modification et de re-synthèse de la parole, qui ont donné des résultats intéressants, en termes de transfert vers le français régional également. L'interprétation linguistique et les comparaisons sont à poursuivre. On peut dire qu'une double force s'exerce sur le langage, unificatrice et séparatrice, dont les effets ne sont pas nécessairement les mêmes sur tel ou tel trait linguistique. Pour poursuivre ce travail de comparaison avec d'autres variétés d'occitan, des enregistrements ont récemment été collectés en gascon, en limousin et en auvergnat.

Pour passer des transferts prosodiques aux transferts technologiques, j'aimerais dire deux mots de la synthèse et de la reconnaissance de la parole. L'une et l'autre font appel à des traitements linguistiques et à des traitements phonético-acoustiques. Comment mutualiser les efforts pour les langues régionales ? En recueillant des textes dans différentes langues, à partir de la presse, de Wikipedia, à travers une gestion commune des exceptions en conversion orthographique-phonétique pour les noms propres et les emprunts, et également, en synthèse, en enregistrant des locuteurs bilingues (langue régionale/langue mieux dotée) ou, en reconnaissance de la parole, en amorçant des modèles acoustiques à partir du français : j'ai pu aligner en phonèmes, c'est-à-dire segmenter le flux sonore en unités acoustiques, à l'aide de modèles français, des données en corse, occitan et catalan. Il y a donc un moyen de réduire les coûts.

Pour conclure, les langues territoriales de France ont trop longtemps été délaissées pour des raisons politico-économiques. Aujourd'hui, pour le traitement automatique, il est nécessaire de collecter des enregistrements

de qualité et des textes en quantité, car l'automatisation du processus se montre assez exigeante en outillage. Dans ce travail, j'ai présenté différentes stratégies possibles pour poser des questions en corse, en occitan et en catalan. J'espère que cette étude a pu contribuer à la modélisation de ces langues, qu'elle a permis de quantifier des tendances connues ou moins connues, et des transferts prosodiques vers le français.

Pour aller plus loin, il faut prendre en compte la dimension sociale qui a été quelque peu négligée ici, dans des applications à d'autres langues régionales, et également à la synthèse ou à la reconnaissance de la parole, qui, espérons-le, peuvent bénéficier de ressources dans des langues mieux dotées comme l'est le français.

Le projet RESTAURE

Delphine Bernhard, Laboratoire linguistique, langues, parole (LiLPa) – université de Strasbourg
Marianne Vergez-Couret, Laboratoire Cognition, Langues, Langages, Ergonomie (CLLE) - Équipe de recherche en synthaxe et en sémantique (ERSS) – université de Toulouse II Jean Jaurès

Transcription revue et complétée avec la participation de Myriam Bras, CLLE-ERSS – université de Toulouse II Jean Jaurès et de Christophe Rey, Linguistique Et Sociolinguistique : Contacts, Lexique, Appropriations, Politiques (LESCLAP) (CERCLL-EA 4283), université de Picardie Jules Verne

Le projet Ressources Informatisées et traitement automatique pour les langues régionales (RESTAURE) est un projet financé par l'ANR, entamé au mois de janvier 2015 pour une durée de 42 mois. Il comporte trois objectifs principaux :

- acquisition et normalisation de ressources (corpus et lexiques) ;
- développement d'outils pour l'acquisition et l'analyse de corpus ;
- diffusion des résultats auprès du grand public.

Les langues régionales de France concernées par le projet sont au nombre de trois : le picard, l'alsacien et l'occitan. Chacune de ces langues est représentée par un laboratoire partenaire : LESCLAP à Amiens pour le picard, LiLPa à Strasbourg pour l'alsacien, et CLLE-ERSS à Toulouse pour l'occitan. À cela s'ajoute un laboratoire en région parisienne, le LIMSI-CNRS, qui travaille sur les aspects de traitement automatique des langues.

La motivation principale du projet est le manque de ressources informatisées pour les langues régionales de France, en particulier pour les trois langues concernées par le projet.

1. État des lieux des ressources et outils existants

Pour ce qui est des corpus, la langue la plus avancée des trois est le picard, car il existe une base textuelle appelée Picartext, que nous présenterons plus

en détails par la suite. L'occitan est aussi relativement bien avancé, grâce à plusieurs projets en cours, dont la construction de la base textuelle BaTelÒc, que nous présenterons également, alors qu'il n'existe aucun corpus à l'heure actuelle pour l'alsacien : il s'agit là d'une lacune que nous souhaiterions combler.

Pour ce qui est des lexiques pour le Traitement Automatique des Langues et l'étiquetage morpho-syntaxique, quelques travaux ont été réalisés pour l'alsacien et l'occitan (Bernhard, 2014 ; Bernhard et Ligozat, 2013 ; Vergez-Couret et Urieli, 2014), mais il faut reconnaître que l'on n'atteint pas encore des niveaux de performance similaires à ceux du français. Les lacunes sont encore plus importantes pour l'analyse syntaxique et la lemmatisation.

Différentes raisons expliquent ce manque de ressources, mais le défi majeur est celui de la variation graphique, qui peut poser des problèmes aux outils automatiques.

En alsacien, l'exemple du mot « lundi » est un cas intéressant : on peut le trouver sous les formes « Mantig », « Mandig », « Mandi », « Mändàach », « Mändàà », « Mondàà ». Ces diverses formes ont peu de caractères en commun, mais constituent tout de même des variantes d'une même unité lexicale et devraient donc être reconnues comme telles. Il s'agit là du défi scientifique majeur auquel le projet RESTAURE va tenter d'apporter des solutions.

Partant de ces constats, nous avons décidé pour le projet de mutualiser nos connaissances et nos compétences, notamment en nous inspirant de l'existant, même s'il n'est pas encore très développé.
Pour ce qui est des corpus, nous allons nous appuyer sur les méthodologies employées pour les projets Picartext pour la langue picarde et BaTelÒc pour la langue occitane. Ces deux projets ont parallèlement débuté en 2006 (sans alors avoir connaissance l'un de l'autre) et visaient un même objectif : doter la langue picarde et la langue occitane d'une base de textes consultables en ligne.

La base de texte picarde, PicarText[1] (*Eloy et al*, 2015), a été réalisée au LESCLAP à Amiens sous la direction de Jean-Michel Eloy et Christophe Rey avec le soutien financier du conseil régional de Picardie. Elle comprend à l'heure actuelle environ dix millions de mots, de textes allant du XVIII^e^ au XXI^e^ siècle et de genres très variés (dictionnaires, contes, recueils de

1 https://www.u-picardie.fr/LESCLaP/PICARTEXT/Public/

poésie, romans, chansons). Les méthodes de recherche dans la base sont particulièrement intéressantes, car elles prennent en compte la variation graphique : on peut faire des recherches sous forme littérale ou avec des expressions régulières, mais il y a également des fonctionnalités qui intègrent la correspondance phonétique, ce qui permet de retrouver différentes formes orthographiques utilisées par les auteurs à condition que la prononciation soit identique. Il est également possible de prendre en compte la correspondance dialectale, c'est-à-dire retrouver les formes théoriquement possibles en picard, y compris avec d'autres prononciations que celle fournie. À cela s'ajoutent d'autres fonctionnalités de recherche : empan temporel, zone géographique, genre textuel.

La base BaTelÒc (Bras et Thomas, 2011) est réalisée au laboratoire CLLE-ERSS, Université de Toulouse 2 Jean Jaurès sous la direction de Myriam Bras. La première version de la base est en cours de finalisation. Elle contient environ trois millions de mots (85 œuvres d'une quarantaine d'auteurs), sur une période allant du XIX^e^ au XXI^e^ siècle, représentant des genres variés (contes, poésies, romans, nouvelles, mémoires) et relevant de plusieurs dialectes et de plusieurs graphies. À ce jour, tous les textes de la base ont été acquis au format numérique et encodés au format XML (TEI P5) avec le souci de rester le plus fidèle possible à l'édition papier (mise en forme). Les textes sont intégrés dans une base dotée d'une interface qui propose plusieurs outils de consultation. Le premier outil est une interface de sélection des textes qui permet de se constituer un corpus de travail (selon le titre, l'auteur, sa date de naissance, l'année de création ou d'édition de l'œuvre, le dialecte, la graphie, le genre…). Les autres outils sont des concordanciers qui permettent de rechercher les contextes d'emploi d'une forme ou de plusieurs formes avec des fonctionnalités telles que la forme « est », « contient », « commence par », « finit par » ou en exploitant le langage des expressions régulières.

2. Objectifs du projet RESTAURE

Le projet RESTAURE s'inscrit dans la complémentarité des projets PicarText et BaTelÒc et vise en premier lieu pour l'alsacien, l'occitan et le picard le développement de ressources et d'outils linguistiques. Il est prévu selon les états d'avancement pour chaque langue de constituer ou d'enrichir les corpus

de textes. La matière pour les trois langues concernées est abondante mais pas toujours disponible au format numérique. Un des objectifs visés par le projet est donc de numériser et *OCRiser*[1] une partie de cette matière. Nous souhaitons développer des outils *OCR* spécifiques à chaque langue et adaptés au traitement de la variation. Ces outils permettront de constituer et compléter nos corpus et seront également diffusés avec des licences libres à la fin du projet.

Le projet vise ensuite l'enrichissement de ces corpus avec des annotations linguistiques, en l'occurrence des annotations morphosyntaxiques. Cette annotation vise à associer à chaque forme des corpus, le lemme ou forme de citation (par exemple le verbe à l'infinitif, l'adjectif au masculin singulier, le nom au singulier), une catégorie grammaticale (nom, verbe, adjectif, adverbe…) et des informations morphosyntaxiques (personne, nombre, genre…). Ces annotations, dans le dispositif des bases de textes présentées ci-dessus, permettront de nouveaux modes de consultation des contextes d'emploi, par exemple la recherche de toutes les formes fléchies d'un verbe à partir de son lemme.

Nous avons fait le choix, dans le projet, d'utiliser des algorithmes par apprentissage, c'est-à-dire qui cherchent à apprendre des règles générales à partir d'exemples particuliers. Cela nécessite un travail pointu d'annotation des données. Dans le projet, les compétences linguistiques particulières à chaque langue régionale se trouvent dans les trois laboratoires qui développeront les annotations nécessaires. Néanmoins, nous souhaitons mettre en place un soutien commun sur toutes les compétences techniques et sur tous les aspects méthodologiques. Nous souhaitons nous doter d'une méthodologie commune qui pourra être également appliquée à d'autres langues qui souhaiteraient avoir le même parcours que le nôtre.

La recherche en traitement automatique des langues pour l'alsacien, l'occitan et le picard soulève des questions nouvelles sur le traitement de la variation qui sont incontournables pour doter chacune des langues d'une boîte à outils minimale en TAL. Nous œuvrons à la constitution de bases solides pour qu'ensuite de nombreuses applications puissent être développées (moteurs de recherche, correcteurs orthographiques, outils d'aide à la rédaction et à la traduction, synthèse vocale…).

1 *OCRiser* ou *océriser* : cf. supra, note 1, page 85.

Indications bibliographiques communiquées par les auteurs

Bernhard, D. (2014). Adding Dialectal Lexicalisations to Linked Open Data Resources: the Example of Alsatian, in *Proceedings of the Workshop on Collaboration and Computing for Under Resourced Languages in the Linked Open Data Era (CCURL 2014)*, May 2014, Reykjavík, Iceland. pp.23-29, 2014

Bernhard, D. et Ligozat, A.-L. (2013). Es esch fàscht wie Ditsch, oder net ? Étiquetage morphosyntaxique de l'alsacien en passant par l'allemand. In Actes de TALARE 2013 : Traitement Automatique des Langues Régionales de France et d'Europe, p. 209–220.

Bras, M. et Thomas, J. (2011). Batelòc : cap a una basa informatisada de tèxtes occitans. In A. Rieger (ed.) L'Occitanie invitée de l'Euregio. Liège 1981-Aix-la-Chapelle 2008 Bilan et perspectives, Actes du IX^{e} Congrès International de l'AIEO, Aache, Shaker.

Eloy, J.-M., Rey, C., Martin, F. (2015). PICARTEXT : Une ressource informatisée pour la langue picarde, Actes de *TALaRE 2015 - Traitement Automatique des Langues Régionales de France et d'Europe*, Juin 2015, Caen, France.

Vergez-Couret, M., Urieli, A. (2014). 'POS-tagging different varieties of Occitan with single-dialect ressources', Eds M. Zampieri, L. Tan, N. Ljubešić, J. Tiedemann, *Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects*, Eds. (Dublin : Association for Computational Linguistics and Dublin City University), 21-29.

Le cas des créoles français : mutualisation des ressources pour des dialectes apparentés

Pascal Vaillant, Université Paris XIII

Je m'appelle Pascal Vaillant, je travaille à l'université Paris 13, et je vais vous parler des langues créoles, sur lesquelles j'ai commencé à travailler lorsque j'étais enseignant-chercheur à l'université des Antilles et de la Guyane, d'abord en Martinique, puis en Guyane, dans l'équipe de recherche « Structure et Dynamiques des Langues » du CNRS, basée à Villejuif.

Je vais commencer par vous présenter quelques informations sur la nature de ces langues un peu particulières, en ce qu'elles forment un groupe de langues qui sont à la fois diverses mais comportent également de nombreux éléments communs, ce qui justifie une réflexion sur la nature de leur norme et sur l'intérêt de réfléchir à la mutualisation de ressources.

Histoire et genèse

Pour commencer, je voudrais vous présenter quelques éléments de compréhension historique et linguistique.

La situation de ces langues est assez particulière, car contrairement à beaucoup d'autres langues régionales de France métropolitaine, qu'elles soient de l'aire gallo-romane ou non, les créoles ne résultent pas du développement naturel lent, « écologique », de langues au fil des générations sans rupture brutale.

Les créoles se répartissent sur deux aires : l'aire Amériques-Caraïbes, et l'aire de l'océan Indien avec la Réunion, les îles Mascareignes, Maurice. Je vais, dans la suite de cet exposé, parler surtout de l'aire Amérique-Caraïbes, car c'est celle que je connais.

L'émergence des langues créoles dans ces régions est liée à une période historique particulière. Après la découverte du Nouveau Monde, l'Espagne et le Portugal se précipitent dans une ruée vers les nouvelles richesses. Le traité de Tordesillas, signé entre les souverains de l'Espagne et du Portugal en 1494, suite à l'initiative du pape Alexandre VI, répartissait le monde entre ces deux grandes puissances catholiques alliées à la papauté. L'Espagne s'est précipitée sur les mines d'argent et les grandes civilisations à piller (les empires inca et aztèque), et le Portugal sur les comptoirs côtiers où l'on pouvait installer des ports pour faire du commerce. Les pays européens qui étaient des puissances maritimes émergentes mais un peu en retard par rapport à l'Espagne et au Portugal, c'est-à-dire la France, l'Angleterre et les Pays-Bas, sont venus s'intercaler dans un petit coin, qui n'intéressait à l'époque ni les Portugais ni les Espagnols, sur ce qui est devenu les petites Antilles et la côte nord de l'Amérique du Sud (les trois Guyanes). C'est la raison pour laquelle aujourd'hui toute l'Amérique du Sud parle portugais ou espagnol, sauf ces trois petits pays qu'on appelle les Guyanes, et les îles Caraïbes.

Au départ, ces trois puissances maritimes du nord ont donc pris pied dans ce petit coin du monde, qu'elles ont commencé par arracher à l'emprise de l'Espagne, puis qu'elles se sont pendant un certain temps disputé entre elles. Une période d'économie agricole de plus en plus intensive a commencé, amenant à l'importation d'esclaves africains, une importation qui est devenue de plus en plus massive et qui a connu un siècle extrêmement intense entre 1750 et 1850 environ, au cours duquel beaucoup de gens étaient incorporés à la population à chaque génération sans avoir le temps de s'adapter. Quand des minorités arrivent dans un pays et apprennent la langue, il y a généralement une dose filée de personnes qui ont le temps de s'incorporer et d'apprendre une langue cible. Dans cette situation-là, il n'y avait plus d'équilibre. D'après une étude, réalisée par Jaques Arends, de ce qui s'est passé sur le plan démographique pendant ce siècle intense d'économie esclavagiste sur la colonie du Surinam, la moitié des gens présents à un instant T ne l'était pas dix ans auparavant. Le cas du Surinam est assez représentatif de l'évolution des autres colonies esclavagistes à la même époque. Il s'agit donc d'un arrivage massif de gens qui n'ont pas le temps de s'adapter, générant un laboratoire in vivo d'évolution linguistique assez original.

Comment comprendre ce qui se passe dans une telle situation d'évolution linguistique « chaotique » ?

La description proposée par Robert Chaudenson pour le créole réunionnais s'adapte à notre cas. La langue cible est le français. Il ne s'agit pas du français d'aujourd'hui tel que standardisé par l'Académie Française, mais sûrement une *koïnè* de français oraux de différentes régions – à l'origine surtout des dialectes de l'ouest de la France, mais on a vu au fil du 18e siècle beaucoup d'arrivants de région parisienne et de Picardie notamment. Puis est venue cette période d'intense traite esclavagiste où la langue a perdu pied, puisque les gens qui arrivaient n'avaient pas le temps d'apprendre. Lorsqu'un esclave arrive sur une plantation, il ne suit pas de cours de langue, il est mis directement au travail, et les gens avec lesquels il est en contact ne sont pas des francophones : ce sont des gens qui eux-mêmes sont arrivés quelques années avant lui et qui n'ont pas eu le temps d'apprendre la langue. Après un siècle de ce cycle-là, la langue perd pied sur les bases de l'évolution éco-linguistique « naturelle ». Ce dérapage rapide et massif fait que si le créole martiniquais d'aujourd'hui est, en un certain sens, apparenté au français, de la même manière que l'on peut dire que le français est apparenté au latin, il l'est à la suite d'une dérive dans l'évolution du système et des caractéristiques typologiques qui les rendent extrêmement différents. Pourtant un siècle d'évolution seulement sépare le créole et le français, alors qu'il a fallu dix-sept siècles pour que le latin devienne du français.

Aire linguistique

Le contexte historique est intéressant pour comprendre la genèse de ces langues et leurs caractéristiques. Par ailleurs, il ne faut pas perdre de vue le fait qu'elles sont apparentées, parce que le contact entre les différentes îles et colonies n'a pas été rompu (des colons allaient d'une île à l'autre avec leurs esclaves), mais que la situation était quand même celle d'un archipel. Ce n'est pas un continuum dialectal comme sur un continent, où les gens gardent contact d'un village à l'autre. Nous sommes dans le cas d'îles, des communautés se forment, convergent à certains moments, puis divergent et restent longtemps dans des situations divergentes. La question de l'existence d'une véritable famille linguistique fait elle-même

débat : certains parlent des créoles comme des différents dialectes d'une même langue, alors que d'autres, à l'extrême, affirment que même au sein de la zone Amérique-Caraïbes, les différents créoles résultent d'évolutions entièrement parallèles. Une position raisonnable consiste à considérer que l'on est en présence d'une aire dialectale, qui n'est certes pas comparable à ce que l'on peut appeler un continuum dialectal comme dans le cas des aires des langues d'oc, mais qui constitue malgré tout une répartition géographique d'une famille de dialectes apparentés. C'est par exemple l'avis du linguiste allemand Stefan Pfänder, qui a étudié cette aire linguistique qu'il a baptisé la *Creolia* – en référence à la *Romania*, comme les linguistes appellent l'aire des langues romanes. Je précise au passage que pour des linguistes, le terme « dialecte » n'a pas la moindre connotation péjorative : il désigne simplement une langue, considérée comparativement à d'autres langues d'un groupe, et n'implique aucun jugement de valeur ou hiérarchisation.

En ce qui concerne l'aire que je connais, la zone Amériques-Caraïbes, des créoles sont parlés dans quatre ex-colonies françaises, cinq si on compte l'île de la Trinité, où le créole est en voie de disparition : Haïti – qui bien qu'indépendant depuis 1804, continue à avoir le créole et le français comme langues officielles ; et trois territoires qui sont encore aujourd'hui dans le giron de la République, et qui sont la Guyane Française, la Martinique et la Guadeloupe. Le créole à base française des Petites Antilles est par ailleurs parlé dans des îles voisines de la Martinique et de la Guadeloupe, qui sont la Dominique et Sainte-Lucie (un temps des colonies anglaises, aujourd'hui des pays indépendants). Pour simplifier le tableau, nous en resterons à quatre variantes principales : Haïti, Guadeloupe, Martinique, Guyane.

Traits caractéristiques des créoles : similitudes et divergences

Nous sommes dans le cadre de langues apparentées et dans une certaine mesure inter-compréhensibles, malgré des différences phonologiques et syntaxiques que je vais un peu expliciter.

En ce qui concerne les caractéristiques de ces langues, le lexique est à base française à environ 90 %, selon les estimations de Marie-Christine Hazaël-Massieux. En revanche, au niveau typologique et au niveau grammatical,

ce n'est plus du français. L'ordre y est certes aussi Sujet-Verbe-Objet, mais ce sont des langues isolantes, avec des mots invariables, une grammaire très analytique, pas de morphologie flexionnelle, et pour ainsi dire pas de morphologie dérivationnelle à part quelques réintroductions récentes. Par ailleurs, on a une caractéristique très intéressante que le créole partage en tant que langue isolante avec d'autres langues comme le chinois par exemple, qui est la très grande plasticité des catégories morpho-syntaxiques : un mot peut sans problème être utilisé en tant que verbe dans une phrase, en tant que nom entre un déterminant et un adjectif dans une autre, etc.

Pour illustrer des caractéristiques communes et des caractéristiques divergentes de ces langues, je vais vous présenter un exemple de modélisation des parties communes et des parties divergentes, d'abord avec le système du marquage du temps et de l'aspect pour l'expression du verbe, puis avec le système de détermination dans le groupe nominal.

Groupe verbal

Le groupe verbal est organisé autour du prédicat de la phrase (appelons-le ici, pour simplifier, le verbe, bien qu'il puisse parfois s'agir d'une autre catégorie de mots). La phrase typique, dans sa forme minimale, comprend donc un sujet et un verbe, par exemple « nou dansé ».

• Par défaut, un pronom et un verbe, sans autre mot, expriment l'aspect accompli : *nou dansé* = « nous avons dansé ».
• La particule *ka* permet d'exprimer l'imperfectif (action fréquente ou répétée) ou le progressif (action vue dans son développement) : *nou ka dansé* = « nous sommes en train de danser ».
• La particule *ké* exprime l'aspect prospectif, le futur : *nou ké dansé* = « nous danserons ».
• La particule *té* sert à exprimer le passé : *nou té dansé* = « nous avions dansé ».
• Il est possible de combiner *té* et *ka* pour exprimer l'imperfectif dans le passé : *nou té ka dansé* = « nous étions en train de danser ».
• La combinaison de *té* et *ké* donne le sens d'un conditionnel : « Nou té ké dansé », « nous danserions » (le conditionnel est ici exprimé comme une possibilité future vue depuis un point situé dans le passé ; étymologiquement, c'est d'ailleurs la même image qui a donné naissance au conditionnel des langues romanes).

• On peut enfin même combiner les trois pour exprimer un conditionnel imperfectif : *nou té ké ka dansé* — comme on l'entend dans une chanson de zouk très célèbre : *kolé séré nou té ké ka dansé* —: « nous serions en train de danser aujourd'hui » (sous-entendu : si nous n'avions pas rompu il y a vingt ans).

On a donc là toutes les combinaisons possibles d'un système à trois unités, dont chacune peut alterner entre l'aspect zéro (non-exprimé) et l'aspect particule invariable (exprimée). Ces trois unités peuvent se combiner pour donner tous les effets de sens possibles.

nou	*Ø*	*Ø*	*Ø*	*dansé*
	té	*ké*	*ka*	
	Passé	Prospectif	Imperfectif	

Groupe nominal

Pour le groupe nominal, il y a comme en ancien français la possibilité d'avoir un nom sans article avec une valeur générique : *moun* = « la personne », « les personnes », « les gens ». Il y a un article indéfini antéposé comme en français, et un article défini qui, dans tous ces créoles, est postposé. Il se présente sous une forme de base *la*, et vient probablement étymologiquement d'un déictique (« là »).

Au-delà de ces quelques points communs, la construction des groupes nominaux montre plusieurs divergences dans les différents créoles. L'une d'entre elles réside dans le démonstratif : il est antéposé et se combine avec le défini en guyanais, alors qu'en martiniquais il est postposé, et se combine également avec l'article défini, mais en s'amalgamant avec lui.

Pour le pluriel, on a un article défini postposé pluriel en guyanais, *moun yan* (= « les gens »), mais on a une particule pré-nominale « sé » en martiniquais et en guadeloupéen. Il s'agit d'un morphème « transparent », c'est-à-dire : qui ne sert qu'à exprimer une seule valeur, le pluriel. Il se combine donc avec d'autres articles ou des démonstratifs.

Les différentes possibilités pour la construction du groupe nominal sont illustrées dans le tableau ci-dessous.

		haït.	guad.	mart.	guya.	français
Degré générique		*moun*	*moun*	*moun*	*moun*	personne (humain)
Singulier	indéfini	*yon moun*	*on moun*	*an moun*	*roun moun*	une personne
	défini	*moun nan*	*moun la*	*moun lan*	*moun an*	la personne
	démonstratif	*moun sa a*	*moun lasa*	*moun tala*	*sa moun an*	cette personne
Pluriel	indéfini	*moun*	*moun*	*moun*	*moun*	des personnes
	défini	*moun yo*	*sé moun la*	*sé moun lan*	*moun yan*	les personnes
	démonstratif	*moun sa yo*	*sé moun lasa*	*sé moun tala*	*sa moun yan*	ces personnes

Modélisation de la langue et mutualisation des ressources

Pour faire entrer ces langues dans l'ère des technologies de l'information, il faut les « mettre en boîte » dans un ordinateur. Il s'agit concrètement d'un travail de développement de ressources numériques (grammaires, lexiques) qui est très coûteux en temps et en ressources. Le point principal que je souhaite défendre dans cet exposé est que tout l'enjeu, dans le cas de langues apparentées, est de savoir mettre en commun une partie de ce travail.

En ce qui concerne la grammaire, ce que je propose est une modélisation de la langue qui permette de ne pas refaire ce travail autant de fois qu'il y a de dialectes à représenter. On l'a vu, les différents créoles ont des structures communes, et des structures spécifiques. Il s'agit de pouvoir les considérer comme une seule langue lorsque l'on représente leurs structures communes, et comme des langues différentes lorsque l'on représente leurs structures spécifiques. Comment faire ? Une approche qui a déjà été testée, avec succès, dans le domaine des technologies

de la langue, est celle des « méta-grammaires » : elle consiste à créer un étage abstrait de description des structures communes, qui engendre ensuite, dans un processus automatique, des grammaires partiellement incomplètes de chacune des langues ; on complète ensuite chacune de ces grammaires en y développant les parties spécifiques. Cependant, cette approche aboutit encore, au final, à un résultat consistant en quatre grammaires distinctes et étanches. Je pense que nous pouvons encore faire mieux, en développant une grammaire *modulaire*, c'est-à-dire une grammaire en « cercles concentriques », dont certaines parties seraient communes à plusieurs langues et d'autres spécifiques, et qui coexisteraient dans le même processus informatique.

J'ai proposé un prototype de ce type de grammaire modulaire. J'y utilise un formalisme bien connu (appelé TAG) qui représente chaque construction élémentaire d'un mot (par exemple la manière d'utiliser un article, ou de construire la phrase autour d'un verbe) comme un petit « arbre syntaxique » élémentaire, ces arbres pouvant ensuite se combiner entre eux. En général, les paramètres permettant de combiner ces petits arbres sont des paramètres internes à une langue (la catégorie grammaticale d'un mot, le genre, le nombre, la personne, etc.). La nouveauté dans le modèle proposé est que l'on ajoute, à ces paramètres, un paramètre qui représente tout simplement la langue en cours d'utilisation. On a donc une grammaire avec des arbres qui modélisent des aspects communs de ces différentes langues, et quelques arbres grammaticaux qui vont nous permettre d'exprimer des choses en spécifiant le paramètre de langue lorsqu'il s'agit de structures spécifiques à une langue ou à un sous-ensemble de langues. Tant que les structures sont communes, on laisse le paramètre de langue non-instancié. Ainsi, les règles de production du passé en utilisant la particule *té*, par exemple, font partie du noyau commun aux quatre variétés créoles. Lorsqu'il s'agit en revanche de faire comprendre à l'ordinateur que tel mot, ou telle construction, n'est valable que pour l'un des dialectes (par exemple l'article défini pluriel du guyanais), on indique la valeur obligatoire de ce paramètre de langue. À tout instant, ces différents petits arbres coexistent dans l'ordinateur : c'est le processus de construction qui va déterminer, en fonction de la langue utilisée, lesquels sont utilisables et combinables entre eux, et lesquels ne le sont pas.

Ce qui est amusant, pour me diriger vers la conclusion, c'est que le même problème de variation se pose au niveau du lexique. Dans le lexique, le vocabulaire commun est important, ne serait-ce que parce que nous avons affaire à des langues basées sur du vocabulaire français, qui ont subi des processus d'érosion et d'évolution comparables dans des zones géographiques comparables. Mais en même temps, tous les dictionnaires actuels de créoles – je me limite à parler de ceux que je connais, de l'aire Amérique-Caraïbes – résultent d'initiatives locales et n'ont pas été construits en commun : le dictionnaire de Barthélémy pour le guyanais ; celui de Frank pour le saint-lucien ; deux dictionnaires pour le guadeloupéen : celui de Poullet, Telchid et Ludwig et celui de Tourneux et Barbotin ; et Confiant pour le martiniquais. Peut-on les rassembler en essayant de construire une norme sur la base de méta-phonèmes, comme on l'a fait pour d'autres aires dialectales, pour le breton avec l'étude des variations qui ont servi de base à l'élaboration de la norme *Kerne-Leon-Treger* (KLT) ? Ce n'est pas si évident que cela. Il y a quelques méta-phonèmes systématiques, mais surtout beaucoup de variation. La publication très récente d'un atlas linguistique des petites Antilles, résultant d'une enquête réalisée sous la direction de Guylaine Brun-Trigaud et de Jean Le Dû, montre qu'en réalité beaucoup de mots sont acceptés en compréhension dans plusieurs aires dialectales, donc dans plusieurs sites de ces territoires. J'ai simplifié en disant Guadeloupe-Martinique-Guyane, mais on ne parle pas de la même manière à l'ouest et au sud de la Martinique, et on voit apparaître parfois des répartitions de fréquences qui ne sont pas exactement les mêmes. Mais il y a surtout beaucoup de variation, au sein même des différentes communautés.

Cela pose la question, pour conclure sur les questions et les perspectives, d'abord de la difficulté de la détermination d'une norme, et ensuite de la pertinence de la mutualisation, qui est un peu je crois ce que les technologies de la langue pourraient apporter à ce type de groupe de langues apparentées.

La difficulté de la norme est de savoir de quelle norme on parle, et qui la fixe. A-t-on affaire à un ou plusieurs créoles ? Pour se limiter à l'aire Amériques-Caraïbes pour laquelle on est sûr qu'il s'agit de langues apparentées, certaines personnes, à un moment donné, ont eu envie de dire « le créole ». Il y a eu un débat autour des années 2000-2001, car le ministre de l'Éducation avait décidé de créer un CAPES de créole afin

de l'introduire comme enseignement de langue régionale au lycée. Il y a des gens qui ont insisté pour que l'on dise « un » créole. En même temps, d'autres arguments pourraient tout aussi bien démontrer que l'on devrait parler « des » créoles. « Le créole » a même été utilisé pour englober aussi les créoles de l'océan indien, ce qui linguistiquement est une absurdité, mais vous savez tous que dans la gestion des termes liés à la dénomination des langues il y a au moins autant de politique que de linguistique. Ce qui un jour est du serbo-croate devient vingt ans plus tard une langue serbe et une langue croate qui n'ont officiellement plus rien à voir. Alors, a-t-on affaire à une langue unique avec des variantes ? À plusieurs langues apparentées ? Y a-t-il a un continuum ?

Une autre question est celle de la position de la norme au sein d'un continuum post-créole. C'est une notion qui existe dans les études créoles depuis 1971 : on distingue un *acrolecte*, qui est la langue vers laquelle les gens tendent quand ils font partie des couches sociales favorisées, qui est en général la langue de base, l'anglais dans les pays anglophones, le français dans les pays francophones, et un *basilecte* qui est le créole le plus « rustique », celui que l'on parle dans les campagnes. C'est une conception certes un peu simplificatrice, que j'énonce comme cela pour faire vite ; il y a eu différents modèles. Mais dans tous les cas une question centrale demeure : si l'on doit fixer une norme, où la fixe-t-on ? On ne peut pas fixer le français, car cela n'aurait aucun intérêt. Certaines personnes essaient de la fixer au basilecte le plus rustique possible, mais souvent les locuteurs de la langue elle-même ne s'y reconnaissent plus. Il y a donc des enjeux politiques et idéologiques à cette normalisation, ne serait-ce qu'au niveau de l'orthographe : est-ce qu'on veut favoriser l'autonomie de la langue par rapport au français, en faisant un choix de déviance, ou est-ce qu'on veut favoriser parfois un critère de lisibilité ? La question suscite un véritable débat, qui draine des enjeux à la fois idéologiques, linguistiques et didactiques. Fixer une norme proche de celle du français tend à favoriser la confusion, à ne pas aider ceux qui veulent prendre à bras le corps le problème de l'écriture en créole à réussir à franchir le pas. En outre, elle entretient, auprès d'un certain public, l'idée erronée que le créole est du mauvais français. Mais par ailleurs, dans des pays comme la Martinique, la Guadeloupe ou la Guyane, où les gens ont tous appris à lire et à écrire en français avant d'apprendre à lire et à écrire en créole, même si le créole est la langue première, la lisibilité n'est parfois

pas améliorée si l'on fait un choix de représentation orthographique qui est trop éloigné du français.

Pour terminer, la pertinence de la mutualisation des ressources, à mon sens, est importante, et je crois qu'on va en reparler dans les exposés suivants. Lorsqu'on a des langues qui se ressemblent, pourquoi refaire 100 % du travail quand une partie du lexique et de la grammaire sont communes ? La mutualisation permet la représentation d'un système hétérogène – on rend compte non pas d'une seule langue, mais d'un ensemble de dialectes, ce qui n'est pas inintéressant en soi –; non seulement elle permet un gain en termes de taille de description et de ressources de travail, mais en outre elle permet de prendre en compte la question du plurilinguisme, ce qui m'amène à mon mot de conclusion.

Où sont les gens qui parlent créole aujourd'hui ? Nombre d'entre eux sont plurilingues, pas seulement créole-français, mais de plusieurs créoles, et ils ne sont plus seulement dans les îles des Antilles, ils sont aussi dans l'agglomération parisienne, lyonnaise, bordelaise, toulousaine, nantaise, rouennaise, etc. Il y a une nouvelle *koïnè* créole : si vous écoutez dans le métro, les gens parlent un créole qui n'est plus du créole martiniquais, mais du créole de seconde génération, qui est presque plus vivant que celui qui est promu dans les milieux universitaires en Martinique. C'est un véritable enjeu de recherche qui est très intéressant.

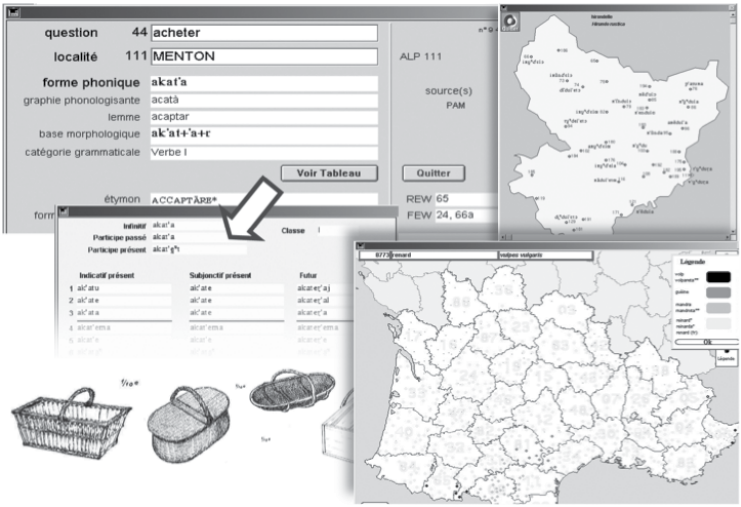
Pour conclure, je souhaite citer Amin Maalouf, sur le fait que les technologies de la langue peuvent être à la fois une menace mais aussi une opportunité pour les langues minoritaires, dans *Les Identités Meurtrières* (1998) :
« *Je ne doute pas que la mondialisation menace la diversité culturelle, en particulier la diversité des langues et des modes de vie ; je suis même persuadé que cette menace est infiniment plus grave que par le passé [...]. Seulement, le monde d'aujourd'hui donne aussi à ceux qui veulent préserver les cultures menacées les moyens de se défendre. Au lieu de décliner et de disparaître dans l'indifférence comme ce fut le cas depuis des siècles, ces cultures ont désormais la possibilité de se battre pour leur survie ; ne serait-il pas absurde de ne pas en user ?* »

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.

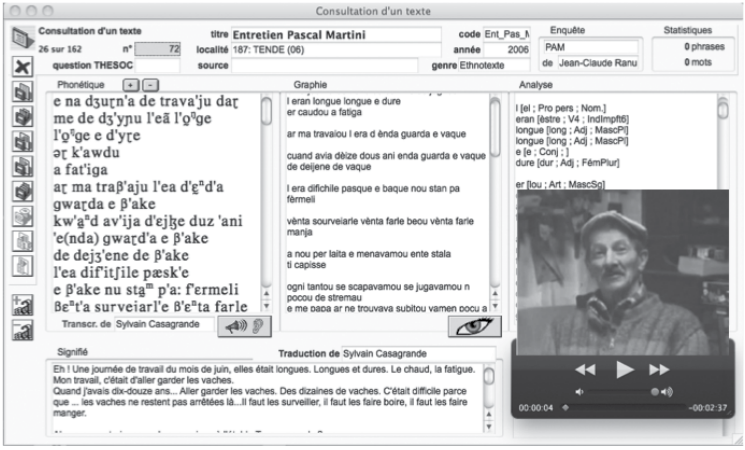
Le Thesaurus Occitan (ou Thesoc, en abrégé) est probablement plus connu pour sa base lexicale, développée au sein du laboratoire BCL[1] depuis 1992, et qui est d'ailleurs mentionnée dans l'inventaire des ressources linguistiques des langues de France (réalisé par l'ELDA, de mars 2013 à novembre 2014).



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).

Développement de ressources syntaxiques : retour d'expérience et méthodologies

Eric de la Clergerie, Inria, équipe Alpage

J'ai en fait peu à dire pour les langues régionales, n'étant pas spécialiste de ce domaine. Mais comme mentionné plus tôt dans la journée, même une langue majeure comme le français souffre de la comparaison au niveau international, et ainsi on a souvent l'impression de courir après l'anglais en termes d'outils et de ressources à notre disposition.

Sur environ 6 000 langues existantes, on en compte peut-être quelques dizaines seulement qui sont outillées, et encore, à des niveaux très divers. Il faut sûrement chercher dans la complexité des langues la raison de cette situation. Il y a en effet beaucoup de niveaux d'attaque (sémantique, syntaxe, morphologie) pour un traitement complet d'une langue et chaque niveau se découpe lui-même en une succession de problèmes complexes.

Je travaille essentiellement sur les aspects syntaxiques, sur lesquels je vais donc focaliser le reste de cet exposé. Mais je pense que ce que je peux dire pour le niveau syntaxique a un écho plus large sur l'ensemble du TAL.

Alors que je développais un analyseur syntaxique vers 2004, je me suis rendu compte qu'il me manquait des ressources pour le français, essentiellement une grammaire avec une bonne couverture des phénomènes syntaxiques. Ce travail s'inscrivait dans une approche que l'on peut qualifier de traditionnelle du TAL, reposant sur le développement manuel de ressources linguistiques.

J'ai donc décidé de développer une telle grammaire, en fait plus précisément une méta-grammaire : une description modulaire à base de contraintes des phénomènes syntaxiques. Pour donner quelques ordres de grandeur, je me suis inspiré de travaux qui existaient bien sûr, mais le développement initial de cette grammaire a pris quelques mois, ce qui est finalement très rapide. Mais ensuite, l'effort nécessaire pour améliorer progressivement les performances, la qualité, la rapidité, s'est échelonné sur environ dix ans,

avec cependant de nombreuses phases de sommeil. Cet effort s'appuie sur un certain nombre de techniques que j'explore : traitement de corpus, participation à des campagnes d'Évaluation d'Analyse Syntaxique (EASy) et Produire des Annotations Syntaxiques à Grande Échelle (PASSAGE), qui sont très importantes pour améliorer les ressources), techniques de fouille d'erreurs, etc.

Cependant en tant que telle, une grammaire n'est pas suffisante. D'autres ressources sont nécessaires pour former un écosystème, avec par exemple le lexique Lefff et la base d'entités nommées ALEDA, tous deux développés par l'équipe d'Analyse linguistique profonde à grande échelle (Alpage). Nous commençons aussi à développer des ressources plus sémantiques, comme des versions françaises de WordNet, FrameNet et VerbNet.

J'aimerais vous offrir un aperçu des points positifs et négatifs de ces approches fondées sur le développement de ressources linguistiques. Parmi les points négatifs, je compte :

- la nécessité d'une expertise linguistique assez forte ;
- les ressources sont d'assez grande taille, et assez complexes à mettre au point ;
- on estime que la couverture reste souvent faible, que les experts qui développent les ressources sont trop focalisés sur les détails, et que la ressource n'est pas utilisable en pratique ;
- enfin, ces ressources sont généralement non-probabilisées. Par exemple dans le Lexique des Formes Fléchies du Français (LEFF), lorsqu'on trouve « pomme », on pense au fruit, mais cela correspond aussi à une forme du verbe « pommer », comme dans « pommer un chou ». Cela peut interférer dans les analyses, alors que la seconde lecture est très rare.

Dans les points positifs en revanche, ces ressources s'appuient normalement sur des théories linguistiques assez bien construites, elles sont compréhensibles par des humains et, surtout, on peut les faire évoluer dans le temps, si on prend le soin de suivre des bonnes pratiques de maintenance. On peut ainsi corriger des erreurs, les enrichir, et les modifier. C'est un cycle assez proche de celui qui se trouve dans le monde du développement logiciel. Pour aller plus loin dans ce processus, comme j'essaie actuellement de le faire, on peut estimer que développer ce genre de ressources est un effort important pour une personne et même une

équipe, et qu'il faut passer à un niveau collaboratif pour demander l'avis d'une communauté et récupérer des informations attestées sur la langue. C'est ce qui est maintenant tenté avec FRMG Wiki, un wiki linguistique permettant de présenter et discuter les phénomènes syntaxiques pour le français, de proposer des phrases d'exemple, de tester les sorties de l'analyseur sur ces phrases d'exemple et même d'examiner les sources de la grammaire pour éventuellement faire des suggestions.

L'approche actuellement majoritaire en TAL ne réside plus cependant dans le développement manuel de ressources comme des lexiques ou des grammaires. Elle s'appuie plutôt sur l'annotation manuelle ou semi-manuelle de corpus textuels (ou oraux) de taille conséquente (plusieurs centaines de milliers à quelques millions de mots). On applique ensuite des techniques d'apprentissage dites supervisées sur ces corpus annotés qui fournissent beaucoup d'informations pour guider l'apprentissage. Ces techniques d'apprentissage sont devenues au fil du temps extrêmement sophistiquées et fonctionnent très bien, en particulier pour l'analyse syntaxique. On voit apparaître ici un découplage entre, d'un côté, un effort linguistique d'annotation et, de l'autre côté, un effort sur les outils et les techniques d'apprentissage.

Le corpus de référence pour le français est le *French Tree Bank*. Il est constitué d'un peu plus de 10 000 phrases de contenu journalistique, sur lequel j'ai entraîné Dyalog-SR, un analyseur statistique par transition que je développe. Le résultat de l'apprentissage n'est pas du tout une grammaire mais un modèle statistique qui permet de prendre des décisions à chaque mot en fonction d'un certain nombre de traits discriminants. Pour illustrer ce découplage, Dyalog-SR peut être entraîné sur le *French Tree Bank*, mais aussi sur beaucoup d'autres langues, comme je l'ai fait lors de la compétition SPMRL'13, sur des langues que je ne comprends absolument pas. Nous l'avons même utilisé cette année pour des tâches d'analyse sémantique (SemEval 2014, Tâche 8), on peut donc finalement l'utiliser dans un cadre assez large.

Le problème avec un corpus annoté comme le *French Tree Bank* vient de son contenu spécifiquement journalistique, alors que le monde n'est pas composé uniquement de textes journalistiques. Nous avons donc commencé dans Alpage à produire des annotations supplémentaires

au niveau syntaxique : par exemple pour le *Sequoia Bank*, un corpus de textes un peu plus hétérogène (médical, Wikipedia, etc), pour *Question bank*, un corpus de questions, pour *Social Media Bank*, un corpus basé sur les réseaux (forums, blogs, etc.) avec des textes assez dégradés et beaucoup plus de variations dans les formulations, etc.

Il existe aussi des corpus annotés (ou en cours d'annotation) pour les annotations temporelles (*French Time Bank*), pour le discours et l'argumentation (*French Discourse Tree Bank*), pour l'oral, la syntaxe et la prosodie (*Rhapsody Tree Bank*), et ainsi de suite.

Les corpus annotés se multiplient progressivement, mais il faut bien prendre en compte le coût humain d'annotation, qui est important. Constituer ces corpus reste un travail très fastidieux. En conséquence, ces corpus annotés demeurent relativement figés, malgré leurs défauts (erreurs, taille insuffisante, choix d'annotations). Surtout, en ce qui concerne l'apprentissage, les modèles appris sont sensibles au domaine du corpus d'entraînement.

De plus, ces modèles statistiques sont des « boîtes noires », des ressources qu'on ne peut pas vraiment comprendre et interpréter en bout de course. On ne peut donc pas les modifier facilement pour les adapter à un nouveau domaine. Concernant les points forts, l'approche par annotation est par contre très efficace et relativement robuste. Alors que certaines phrases peuvent ne pas être analysables par un analyseur syntaxique traditionnel (par manque de couverture), globalement un analyseur statistique peut traiter toutes les phrases (sans garantie néanmoins sur la qualité du résultat).

Il est également possible de conduire des évaluations directement sur le corpus qui sert aussi pour l'entraînement (en le divisant en plusieurs parties). Il y a ainsi un côté autonome, c'est-à-dire que le corpus annoté fournit souvent toutes les informations nécessaires pour apprendre et faire fonctionner un outil, et peu de ressources ou outils supplémentaires sont nécessaires (à la différence des écosystèmes de ressources évoqués précédemment).
J'aimerais aussi mentionner l'émergence d'une troisième approche, fondée sur de l'apprentissage non ou faiblement supervisé. Nous avons vu que les annotations sont rares, car le processus de construction demeure long, fastidieux et coûteux. Par contre, de nombreux corpus (textuels, oraux,

vidéos) existent, au moins pour un certain nombre de langues même si la situation est peut être plus compliquée pour les langues régionales. L'idée est alors d'essayer d'en tirer parti sous leur forme « brute », en exploitant des propriétés génériques des langues, comme des distributions sur les fréquences. Ainsi on peut tenir compte de la loi de Zipf qui indique qu'il y a très peu de phénomènes très fréquents avec malheureusement une très longue traîne de phénomènes de plus en plus rares. On a aussi des hypothèses distributionnelles au niveau sémantique, stipulant qu'on appréhende le sens d'un mot par la manière dont il est utilisé, par ses contextes, et ses voisins. On a aussi des informations sur la typologie et la morphologie des langues : isolée ou pas, agglutinante, avec un ordre préférentiel sujet-verbe-objet (SVO), etc.

Tout ceci peut être exploité par des méthodes d'apprentissage non ou faiblement supervisées pour induire des ressources. Par exemple, la partie morphologique du lexique Lefff a été induite à partir de corpus bruts, en exploitant des informations sur les paradigmes de flexion morphologique du français. On peut faire ça pour la morphologie, l'étiquetage morphosyntaxique, la segmentation en mots pour certaines langues, pour l'extraction terminologique, pour la construction de ressources plus sémantiques (construction « d'ontologies »).

Il se trouve néanmoins que ces approches ne sont pas encore très efficaces pour la syntaxe, car même si les résultats sont intéressants, ils restent loin de ceux obtenus avec des méthodes supervisées. Un autre avantage de ces méthodes non-supervisées est qu'elles permettent d'améliorer des ressources existantes : on peut ainsi améliorer la couverture en termes de mots ou de phénomènes manquants dans une ressource existante par des techniques de fouille d'erreurs, ce qui a été fait pour Lefff en exploitant les échecs de l'analyseur syntaxique sur des corpus. On peut aussi récupérer des informations de fréquence qui peuvent être intéressantes. En général, il faut essayer de compléter ces acquisitions au travers de validation humaines collaboratives.

On peut observer une augmentation phénoménale du volume de données utilisées pour ces tâches d'apprentissages. Google utilise ainsi des volumes textuels de l'ordre de plusieurs centaines de milliards de mots. C'est une course à la taille des ressources qui n'est peut-être pas très réaliste.

J'ai conduit quelques expériences d'acquisition de connaissances en cherchant à rapprocher des mots : par exemple, des mots comme chaise, tabouret, banquette, etc, sont rapprochés simplement en examinant leurs contextes d'utilisation et les voisins de ces mots-là. On peut ensuite utiliser ces informations acquises pour améliorer le traitement syntaxique, en faisant des inférences logiques ne résultant que d'informations statistiques apprises sur corpus.

Parmi les approches faiblement supervisées, le transfert me paraît particulièrement intéressant, c'est-à-dire le fait de transférer tout ou partie d'une ressource d'une langue source vers une langue cible. Ce transfert peut se faire directement et manuellement, par exemple en utilisant la méta-grammaire du français comme méta-grammaire de l'espagnol (ce que nous avons fait en partie). Comme la méta-grammaire comporte des aspects modulaires, on en garde certaines parties et on en modifie d'autres progressivement.

Mais il est aussi intéressant de transférer de façon semi-automatique comme nous l'avons fait pour WOLF, notre version française de WordNet, en partant du WordNet anglais de référence, de plusieurs autres WordNet existants, de dictionnaires bilingues, et de corpus alignés, pour pouvoir remplir progressivement WOLF. Pour le transfert, il existe des ressources bilingues ou multilingues (lexiques, dictionnaires) qui sont intéressantes, des corpus parallèles ou comparables (traitant des mêmes sujets) et il existe également des ressources qui ont des liens multilingues, comme Wikipedia et Wiktionary. Ce sont des sources d'information très riches qui peuvent et doivent être exploitées.

En conclusion, je dirais que si vous avez des financements et qu'il faut des résultats rapides, il vaut sans doute mieux choisir l'approche la plus directe reposant sur l'annotation de corpus. Si les moyens sont plus limités et qu'il faut plutôt s'inscrire dans une stratégie à long terme, il vaut mieux favoriser le travail collaboratif, qui peut être ludique (comme le sont par exemple les projets *ZombiLingo*[1] de Karën Fort ou Jeux de Mots de Matthieu Lafourcade). Un avantage pour les langues régionales réside dans l'existence de communautés de gens a priori passionnés par leur langue et qui peuvent facilement intervenir si on leur offre la possibilité de le faire.

1 http://zombilingo.org/

De plus ces lieux de collaboration créent aussi des espaces de visibilité pour la langue, et assurent une meilleure pérennité des ressources constituées progressivement.

Ces ressources sont aussi, à mon avis, plus dynamiques et vivantes : elles entrent dans un cycle d'évolution, en s'inspirant des modèles de développement issus du logiciel libre. Je pense qu'il faut essayer de minimiser les besoins en annotation de corpus. Ils restent nécessaires ne serait-ce que pour des besoins d'évaluation et pour calibrer les outils, mais on peut essayer de réduire le volume en recherchant les données qui sont les plus utiles à annoter. À mon avis, il faut favoriser les recherches sur les mécanismes de transfert, en particulier pour aller des langues assez bien dotées (comme le français, l'espagnol, l'allemand) pour aller vers des langues régionales moins dotées, tout en examinant plus sérieusement les approches non ou faiblement supervisées.

Traitement de la parole et traduction pour les langues sous-dotées

Fethi Bougares, université du Maine

Avant de continuer sur les questions d'apprentissage supervisé et non-supervisé, je souhaite rappeler un passage de la Déclaration universelle des Droits Linguistiques de l'UNESCO (juin 1996) :

Toute communauté linguistique a le droit de disposer, dans le domaine de l'informatique, d'équipements adaptés à son système linguistique et d'outils de production dans sa langue, afin de profiter pleinement du potentiel qu'offrent ces technologies pour l'auto-expression, l'éducation, la communication, l'édition, la traduction et en général le traitement de l'information et de la diffusion culturelle.

Environ 6 800 langues sont répertoriées à travers le monde. Certains pays en comptent plusieurs, et généralement plus d'une langue est utilisée dans la vie quotidienne de chacun d'entre nous. L'anglais, le chinois, l'espagnol et l'arabe sont les langues les plus parlées dans le monde. Les autres langues représentent uniquement 36 % de l'ensemble de locuteur. De plus, 70 % des langues sont concentrées dans vingt nations.

La langue française est la langue de la République. Il y a 80 millions de locuteurs francophones natifs, spécialement en France, en Belgique et en Suisse. C'est une langue parlée dans 36 pays comme langue seconde, et on trouve aussi des langues régionales dont nous avons déjà discuté au cours de la journée.

L'architecture d'un système de transcription automatique de la parole est la suivante : à partir de données issues de ressources linguistiques, on essaye d'apprendre statistiquement des modèles qui servent ensuite à reconnaître ou transcrire la parole.

Dans un système de transcription de la parole il y a trois modules : le modèle de langage, qui vérifie si la suite de mots produite ou reconnue

par le système correspond bien à une phrase correcte dans la langue traitée, le dictionnaire de phonétisation qui fait le lien entre les mots du langage et l'aspect phonétique, et le modèle acoustique, qui modélise la correspondance entre les phonèmes et les morphèmes. À partir d'un signal de parole, on extrait des paramètres acoustiques utilisé par le décodeur pour produire une hypothèse de transcription. Cette hypothèse correspond à la meilleure solution selon les modèles.

Nous faisons face à plusieurs difficultés pour créer le système de transcription parfait. La variabilité intra et inter-locuteurs fait qu'on ne peut pas reproduire la même phrase deux fois de suite de la même façon : tous les locuteurs ne parlent pas de la même manière.

Ensuite, l'absence de séparateurs dans le flux de parole, les disfluences, l'homophonie entre les mots, le bruit, la parole superposée, le « *code switching* » (le fait de s'exprimer dans une langue mais de glisser quelques mots dans une langue étrangère), ainsi que les aspects de la communication humaine multimodale nous compliquent la tâche. En effet, la communication entre humains implique aussi des gestes, des modalités qui ne relèvent pas seulement de la parole. Or un système de transcription ne peut pas prendre en compte ces aspects-là. L'origine, l'action, le dialecte, la situation sociale, le type de prononciation de la langue peuvent aussi rendre la tâche de transcription plus difficile. On peut considérer donc que le développement d'un système de transcription de la parole pour un dialecte revient à construire un nouveau système de transcription et pas uniquement à l'adapter.

Au Laboratoire d'informatique de l'université du Maine (LIUM), nous travaillons sur ce sujet depuis longtemps. Pour construire nos systèmes nous utilisons des outils open-source, principalement *Carnegie Mellon University* (CMU) Sphinx et Kaldi. Le développement de nos propres outils autours de ces outils nous permet d'améliorer la qualité de la transcription. Lors de notre participation à l'*International Workshop on Spoken Language Translation* (IWSLT) en 2014, une campagne d'évaluation internationale, nous avons développé un système de transcription pour l'italien qui a été classé numéro 1, sans avoir des connaissances linguistique dans cette langue. Nous traitons aussi l'arabe, l'anglais, l'allemand, l'italien, le français, etc, ainsi que des langues sous-dotées. Enfin, des travaux, notamment de thèse, sont en cours sur les dialectes arabes et principalement le dialecte tunisiens.

La traduction automatique relève d'une approche similaire à la transcription automatique : elles sont basées toutes les deux sur un processus d'apprentissage automatique similaire. Traduire une phrase source consiste à retrouver à l'aide d'un modèle probabiliste la phrase correspondante la plus probable dans la langue cible. Le modèle maximise la probabilité de la phrase cible connaissant la phrase source. On peut distinguer deux modèles d'après cette formule : le modèle de traduction et le modèle de langage. L'algorithme de décodage utilise ces deux modèles pour chercher la traduction la plus probable. La qualité de la traduction dépend du système utilisé, du domaine et de la langue traitée. Le modèle de traduction est comme un dictionnaire bilingue de phrases et le modèle de langage modélise les phrases côté cible.

Pour la traduction de la parole, il est nécessaire de coupler les deux systèmes. On utilise le premier pour passer du signal en langue source, c'est-à-dire pour créer le texte puis le traduire avec un système de traduction. Le couplage des deux systèmes n'est pas direct, il faut passer par des étapes de pré-traitement des données d'apprentissage pour les rapprocher, afin que le système de traduction reconnaisse bien la sortie du système de transcription. Ensuite on traduit la meilleure sortie du système de transcription, ou on cherche dans les autres hypothèses, car une ou plusieurs sorties peuvent être renvoyées. Beaucoup d'avancées ont été réalisées dans ces systèmes grâce à des campagnes d'évaluation, qui sont, comme cela a déjà été dit, très importantes pour obtenir des données et comparer les techniques utilisées par les différents participants.

En ce qui concerne les langues sous-dotées, leur particularité est qu'on ne dispose pas de beaucoup de données pour créer les modèles afin de créer un système de transcription ou de traduction assez puissant. Lorsque peu de données sont disponibles, la traduction est très coûteuse. Il faut de nouvelles techniques réduisant les coûts de création des systèmes, par exemple en facilitant la collecte avec la création de corpus de façon semi-supervisée ou avec des techniques d'adaptation pour traiter de nouveaux domaines.

Dans le cadre d'un projet avec différents partenaires (le projet PEA-TRAD (*Post-Editing Actions* - Traduction pour l'aide à l'analyse documentaire)) nous avons développé un système de transcription et de traduction pour une langue peu dotée qui est le pashto. Je vais vous présenter ce qu'on

a fait dans le cadre de ce projet pour rassembler le plus de données possibles. Un système de traduction nécessite des corpus bilingues et monolingues, mais on ne les a pas toujours et ce n'est pas facile d'en obtenir (ils sont soit chers, soit indisponibles). L'une des solutions est d'avoir recours à une langue richement dotée pour créer de façon non-supervisée, en utilisant une langue pivot, un corpus bilingue. On peut par exemple utiliser le couple anglais-français pour créer un corpus bilingue pour l'arabe-français. Pour cela, nous prenons la partie anglaise dans le système arabe-anglais, et nous utilisons un système de traduction automatique anglais-français pour réaliser la traduction. Nous obtenons en sortie la traduction française et cela permet de prendre la partie arabe et de créer un nouveau corpus à moindre coût. On parle alors de synthèse des données bilingues. Nous utilisons aussi des techniques pour vérifier la qualité de la sortie du système de transcription, car parfois ce n'est pas évident, même si la paire de langues est assez riche et assez dotée. Lorsque la technique n'est pas suffisamment efficace, nous avons recours a des techniques sélectionnant une sous-partie du corpus, pour créer un premier corpus arabe-français qui est assez correct pour entraîner un système de traduction.

Pour l'apprentissage semi-supervisé, la chaîne de traitement est la suivante :

Tout d'abord, il faut créer un premier système (avec la transcription et la traduction) avec une grande quantité de données. Ce système est ensuite corrigé, soit manuellement, ce qui a un faible coût puisque les données sont corrigées et non pas produites à partir de rien, ce qui prend en général beaucoup moins de temps, soit automatiquement en sélectionnant un sous-ensemble comme je viens de le dire. Ces nouvelles données sont ensuite réinjectées pour produire un nouveau système, et l'on revient à la deuxième étape en traduisant de nouveau. Le système est amélioré petit à petit à travers cette boucle.

Nous avons aussi utilisé dans le cadre de ce projet un système de traduction automatique pour créer des phonétisations. On remplace la langue source par le français et la langue cible par la phonétisation. Avec peu de données nous parvenons donc à créer un système statistique qui permet de phonétiser automatiquement et de créer des dictionnaires de phonétisation. Pour le pashto, nous n'avons que 9 000 mots comme

données d'apprentissage (9 000 mots manuellement phonétisés). Nous nous sommes servis de ce corpus pour entraîner un système de traduction pour la phonétisation. Avec ce système nous avons créé un dictionnaire de 32 000 mots, que nous avons pu utiliser ensuite pour générer un système de transcription de la parole.

Le problème du manque de ressources ne concerne pas que les langues peu dotées : on a aujourd'hui beaucoup de données pour les actualités mais peu pour les domaines de spécialité. En apprentissage automatique on considère que plus l'on dispose de données, plus la qualité du système augmente. Mais ce n'est pas toujours le cas. J'ai pu constater une amélioration des résultats dans le cas d'un système anglais-français, si je réduis et je sélectionne intelligemment une sous-partie des données. Avec cette sélection les phrases pertinentes ne sont plus noyées dans la masse. On essaie donc d'avoir des données « *in-domain* ».

Wikipedia comme ressource

Rémi Mathis, association Wikimedia France

Wikipedia est un des sites de la galaxie Wikimedia, avec d'autres sites de grand intérêt pour l'usage et le traitement des langues : le Wiktionnaire, Wikisource, en lien avec des communautés structurées par des associations. Je suis là pour représenter Wikimedia France, la structure qui en France essaie de promouvoir et monter des projets ainsi que des partenariats de manière générale. J'ai été président de Wikimédia pendant presque quatre ans, de 2011 à l'automne 2014, et cette période a été celle d'un accent sur les langues. Nous avons travaillé sur la langue française et le développement de la francophonie, en particulier avec des pays africains, mais également les langues régionales. Comme vous le savez, Wikipedia fonctionne par langue : on parle de Wikipedia francophone, anglophone, occitane, bretonne, etc. Mais Wikimedia France et toutes les structures fonctionnent par pays. Wikimedia France est donc amenée à travailler sur l'ensemble des langues présentes sur le territoire français, en particulier les langues régionales.

Quand on parle de Wikipedia dans le grand public, on parle souvent de la Wikipedia francophone ou anglophone. Les gens ne sont pas toujours clairs. Toujours est-il que le principe des Wikipedias est toujours le même : c'est une encyclopédie, collaborative, libre, à laquelle chacun peut participer sans avoir à dire qui il est, en se créant un compte ou pas. Par son fonctionnement, c'est quelque chose qui a pris une part importante dans nos vies, constituant des corpus qui sont très importants.

Le site est devenu incontournable, c'est le 5e site internet en France, très utilisé et connu, ce qui veut dire qu'il y a peut être moyen de profiter de cette popularité pour mettre en valeur un certain nombre de choses qui sont moins connues et en particulier les langues régionales. Wikipedia n'est pas *un* site internet, ce sont 287 sites, chacun correspondant à une langue, qui sont parfois des « grandes » langues, pour ce que cela veut dire, des langues très parlées, internationales, et pour un certain nombre des langues plus confidentielles, parlées par moins de personnes. Lorsqu'on qu'on s'intéresse aux langues régionales de France on voit qu'il y a la

Wikipedia en alémanique, basque, breton, catalan, corse, franco-provençal ou arpitan, flamand occidental, luxembourgeois en Lorraine, normand (Jersey/Guernesey), dialecte du Cotentin, occitan, picard, tahitien, wallon (Ardennes). On se rend compte en regardant les dates de création que ce sont des créations qui sont extrêmement précoces pour la plupart. On voit qu'un certain nombre de personnes trouvent un intérêt aux langues régionales et profitent des nouvelles technologies pour les mettre en valeur, à tel point que la Wikipedia en catalan est la deuxième Wikipedia à avoir été créée après l'anglophone, et qu'elle existait une semaine avant celle en français.

Il y a des règles pour créer ces Wikipedias. Si elles existent dans ces langues-là et pas dans d'autres, c'est en fonction d'un choix reposant sur la communauté d'une manière générale au niveau international, comme pour toutes les décisions qui sont prises. Une Wikipedia peut être créée à partir du moment où une langue possède un code ISO international qui permette de l'identifier, et de s'assurer après quelques essais qu'il y a une communauté minimale qui permettra qu'elle ne tombe pas dans les limbes mais que des articles soient produits. Si l'on considère le nombre d'articles, on peut voir qu'ils sont assez divers. On va de 1 000 ou 2 000 articles dans les petites Wikipedias dans les langues peu parlées actuellement (franco-provençal, picard, tahitien), mais qu'il ne faut pas sous-estimer, car les grandes langues internationales dans Wikipedia représentent environ 50 % des articles, alors que toutes les petites langues représentent les 50 autres pour cents. Ce n'est donc pas du tout anecdotique et cela fait vraiment partie de Wikipedia et de son fonctionnement de manière générale. Selon le même principe, il existe d'autres sites de la galaxie Wikimedia : le Wiktionnaire est particulièrement important, notamment pour les langues régionales, et pour tout ce qui est travail sur les langues puisque cela permet l'accès à des lexiques dans les différentes langues de manière structurée, avec des liens entre ces langues. Le Wiktionnaire décrit dans une langue les mots d'autres langues : le Wiktionnaire français ne comporte pas uniquement des mots français, mais est écrit en français, pour décrire l'ensemble des formes de toutes les langues. On peut donc trouver plus de 17 000 mots en occitan sur le Wiktionnaire francophone, et il existe aussi dans les langues régionales. Il faut également citer Wikisource, qui est très important pour la création de corpus qui peuvent être réutilisés, puisque c'est une bibliothèque numérique de textes, de livres

dans le domaine public (existant aussi dans un grand nombre de langues).

Il faut également être conscient qu'une Wikipedia fonctionne à partir du moment où il y a une communauté pour la faire vivre. Les langues régionales ont généralement un petit noyau de contributeurs passionnés, formant une communauté en général très bien identifiée qui permet d'intervenir et de faire des choses, en s'appuyant éventuellement sur des institutions spécialisées, des associations, les pouvoirs publics locaux, etc., pouvant permettre de monter un certain nombre de pratiques, sachant évidemment que plus la communauté est grande meilleur est le travail, comme on peut le constater sur Wikipedia en général. La Wikipedia anglophone est meilleure que la Wikipedia francophone, car les contributeurs sont plus nombreux. On risque donc, si la communauté n'est pas suffisamment grande, de tomber dans l'anecdotique ou pire, dans un certain côté militant comme cela peut arriver dans le cas de certaines langues régionales actuellement, où le côté « régionaliste », dans le mauvais sens du terme, prend parfois le pas sur la simple volonté de mettre une langue en valeur. Nous avons un certain nombre de cas de personnes qui veulent imposer leur point de vue au détriment des règles de Wikipedia. Il est donc important que les communautés soient suffisamment grandes et représentatives pour éviter cela.

Il faut noter également que le but est de faire fusionner les communautés qui pré-existent au niveau des langues avec les communautés de Wikipédiens qui existent également et qui sont souvent eux-mêmes très preneurs et organisés régionalement. Il y a en particulier un groupe de Wikipédiens bretons, avec des bretonnants déjà présents réalisant un certain nombre de choses. Il suffirait alors de faire se rencontrer les gens pour que cela fonctionne et que les langues régionales soient plus présentes, y compris à travers un certain nombre d'événements qu'il serait très facile d'organiser. On travaille en particulier sur des « *edit-athons* », c'est-à-dire l'organisation d'événements sur une thématique. Ces *edit-athons* pourraient très facilement être organisés par des communautés régionales sur les langues régionales.

Dernier enjeu quand on parle des communautés : on parle des personnes qui contribuent, mais il faut aussi mentionner les lecteurs. Souvent, le problème avec les langues qui touchent un faible nombre de personnes, c'est que l'on risque de rester dans des usages qui cherchent à mettre

en valeur ces langues sans qu'il y ait un véritable usage de l'outil qui soit fait. On va écrire en breton pour le faire exister, mais dès que l'on a besoin de renseignements plus précis, on va lire en français et en anglais. Il est donc important de faire en sorte que cela évolue, pour que cela ne reste pas dans l'anecdotique mais que la Wikipedia en breton soit un véritable outil, une véritable encyclopédie qui soit aussi utile que la Wikipedia francophone ou anglophone.

C'est d'autant plus utile et pratique que sur Wikipedia tout est fait sous licence libre et est donc réutilisable. Il ne s'agit pas seulement de faire des choses pour Wikipedia, mais de faire un corpus, des travaux, qui soient ensuite utiles pour l'ensemble de la communauté parlant ces langues, pour réutiliser les corpus comme vous le faites dans vos projets de recherche.

Au-delà des textes eux-mêmes et de ces données, l'avantage principal de Wikipedia est d'être complètement prise dans un paysage de l'internet qui permet de faire des liens avec d'autres sites, d'autres ressources, et permettant vraiment de travailler sur des corpus beaucoup plus larges sans s'enfermer dans Wikipedia. En bas des articles se trouvent des catégories structurées qui permettent de lier les articles entre eux et de faire des liens vers l'extérieur et les Wikipédias en différentes langues. Cela permet de travailler à des sites multilingues qui font que les langues régionales ne sont pas isolées, mais intégrées dans ce panorama de l'internet dans toutes les langues, et de travailler à des traductions automatiques. Par exemple, le site expérimental du ministère de la Culture sur l'histoire des arts, HDA-Lab[1] utilise le fait que Wikipedia soit multilingue pour traduire le site automatiquement.

Tant que nous sommes dans les questions techniques, Wikipedia se trouve au centre du web sémantique, un aspect important du traitement automatique des données et des langues. Le projet de sémantisation de Wikimedia France, comme cela a déjà été fait en langue anglaise et allemande, a été réalisé avec le Ministère de la culture, la DGLFLF et l'Inria, avec la possibilité de sémantiser les autres sites (Wiktionnaire, etc.) et donc de disposer de données structurées, sémantisées, réutilisables, qui permettent de travailler sur des données directement traitables d'un point de vue automatisé. Au-delà de la sémantisation de Wikipedia, on

1 http://hdalab.iri-research.org/hdalab/

peut travailler directement sur les données structurées à travers un autre site frère de Wikipedia, Wikidata, un site international mais encore une fois multilingue, où il est également possible de lier toutes les données dans les langues régionales avec le paysage anglophone.

En conclusion, j'ai essayé de synthétiser tout ce qui est faisable, tout ce qui existe et en quoi les projets Wikimedia peuvent être utiles pour la mise en valeur des langues régionales et à travers elles des cultures régionales, en soulignant qu'il reste bien évidemment beaucoup de travail à faire pour des communautés qui existent, qui ont certainement parfois envie de travailler et qu'il suffirait de faire surgir. Il a été mentionné un certain nombre de fois au cours de cet après-midi qu'il y a des manques de moyens. L'avantage de Wikipedia est que peu de moyens sont nécessaires, et qu'il y a des structures qui peuvent mettre à disposition ces moyens. Tout ce qu'il faut ce sont des gens qui travaillent et qui ont de la matière grise, et comme il y en a ici, nous pouvons travailler ensemble.

Les projets structurants

Président de séance : **Gilles Adda**

Ortolang : un équipement d'excellence pour la mutualisation et la valorisation des ressources sur le français et les langues de France

Jean-Marie Pierrel, Analyse et traitement informatique de la langue française (ATILF) – université de Lorraine et Centre national de la recherche scientifique (CNRS)

Vous savez tous, et nous en avons suffisamment parlé, que les ressources linguistiques sont fondamentales pour pouvoir développer un certain nombre de travaux de recherche en traitement automatique des langues, à la fois pour faire émerger des modèles (approche stochastique ou symbolique) ainsi que pour les valider. Il est également indispensable d'avoir des ressources partagées et pérennes pour pouvoir comparer les résultats obtenus par différents systèmes de traitement.

Enfin, il est important de pouvoir valoriser nos langues et cet important travail de production de ressources. Pendant de trop nombreuses années, un nombre très important de ressources dont la constitution a pourtant coûté très cher, ont été perdues. Dans les années 1990-2000, beaucoup de travaux de recherche et de thèses ont commencé par construire une ressource spécifique, un corpus, un lexique ou un outil de traitement. Une fois la thèse ou le projet terminés, le thésard s'en va, le corpus reste sur un ordinateur, et cinq ans après plus personne n'est capable de savoir où il est ni comment cela fonctionne.

Le premier objectif est donc la pérennisation de toutes les ressources

produites par nos laboratoires sur les langues, et, au travers de leur mutualisation, permettre de conforter nos recherches. Par ailleurs il est évident qu'une équipe de recherche ne peut pas être performante dans tous les domaines et sans une véritable mutualisation de ressources chaque équipe de recherche se verrait dans l'obligation de tout réinventer. La constitution de ressources est tellement chère qu'il était nécessaire de faire quelque chose pour faciliter cette mutualisation et cette valorisation de ressources linguistiques et c'est précisément l'objectif d'ORTOLANG.

En termes de positionnement institutionnel, ORTOLANG est un équipement d'excellence, validé dans le cadre du programme d'Investissement d'avenir lancé par le gouvernement. Géré par le CNRS, il regroupe un certain nombre d'équipes, l'ATILF avec son Centre National de Ressources Textuelles et Lexicales (CNRTL), l'Institut de l'information scientifique et technique (INIST) et le Laboratoire lorrain de recherche en informatique et ses applications (LORIA), le Laboratoire parole et langage (LPL), qui gérait le *Speech and Language Data Repository* (SLDR) centre de ressources sur l'oral, Modèles, Dynamiques, Corpus (MoDyCo) et le LLL, et implique quatre universités (université de Lorraine, Aix Marseille université, université Paris Ouest Nanterre, université d'Orléans), le CNRS et l'Inria.

Les objectifs d'ORTOLANG sont essentiellement d'outiller et valoriser les travaux qui se font sur le français et les langues de France, avec en particulier une commande de l'État pour proposer un certain nombre de ressources et d'outils de base pour le français. Nous aurons une présentation d'Huma-Num tout à l'heure, qui couvre tous les aspects des sciences humaines et sociales, ORTOLANG est un service spécialisé dans la langue, complémentaire de l'offre générale d'Huma-Num.

Comme pour tous les équipements d'excellence, deux phases ont été définies par le programme d'investissement d'avenir, une phase dite d'investissement, qui s'étale jusqu'à la fin de l'année 2016 et durant laquelle nous devons définir et stabiliser la plate-forme que nous proposons, puis une phase de fonctionnement dont les financements sont assurés jusqu'en 2020. La plate-forme ORTOLANG actuelle s'appuie sur une grappe de trois serveurs biprocesseurs avec 128Go de RAM chacun, offrant ainsi une puissance de calcul et de stockage suffisamment importante. Nous disposons en effet actuellement de 40To de disque utile partagés et d'un espace de sauvegarde

allant jusqu'à 312To. Cela nous permet d'assurer, au fur et à mesure que des ressources sont déposées sur la plate-forme, des sauvegardes journalières et incrémentales. Pour permettre une adaptation maximale aux besoins des chercheurs, nous avons développé une plate-forme logicielle spécifique qui peut être schématisée par le flux de travail présenté dans la figure 1 et qui structure les grandes fonctionnalités que nous proposons dans ORTOLANG.

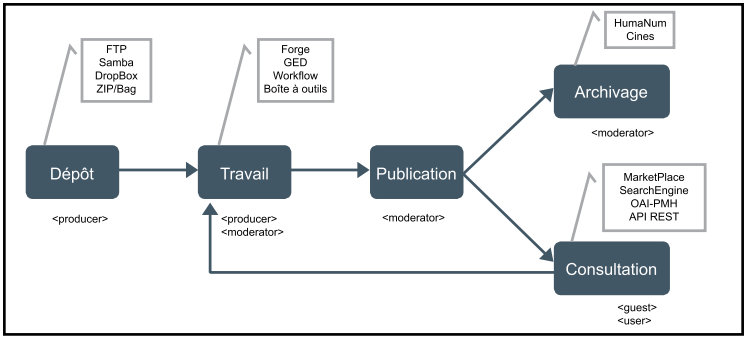


Figure 1 : Flux de travail au sein d'ORTOLANG

Chacun sait qu'héberger, stocker, archiver des ressources est une chose, mais que la plupart du temps quand des chercheurs ou un groupe de passionnés de la langue construisent une ressource, elle est difficilement publiable en l'état, car elle ne respecte pas forcément un certain nombre de standards indispensables pour qu'elle puisse être réutilisée par d'autres.

Le schéma classique, auquel on se trouve donc confronté, est que les utilisateurs souhaitent pouvoir déposer une ressource, mais ces derniers n'étant pas informaticiens la plupart du temps, il faut leur proposer une solution simple pour qu'ils puissent le faire. Cependant lors du dépôt, les ressources ne sont pas forcément prêtes à être publiées. Il est donc nécessaire d'accompagner les utilisateurs pour la standardisation et la documentation de leurs ressources en termes de métadonnées facilement interopérables avec d'autres systèmes. Il faut également mettre en place un processus de validation technique de la ressource avant de prendre la décision de la diffuser. À ce moment, la pérennité doit être assurée, c'est-à-dire que dès qu'une ressource est publiée, il faut pouvoir la retrouver sans aucun problème plusieurs années plus tard. Une fois publiée, elle pourra éventuellement être

archivée ou exploitée par d'autres pour enrichir de nouvelles ressources. C'est ce qui nous a conduits à proposer un flux de travail qui distingue cinq étapes. La première étape est la plus simple, il s'agit du dépôt. Nous avons choisi de proposer des choses simples pour l'utilisateur. Si la ressource n'est pas trop importante (ressource textuelle, lexique morphosyntaxique par exemple), il suffit de la glisser dans l'espace de travail sur la plate-forme, et le téléchargement se fait automatiquement. Cela est impossible s'il s'agit d'une ressource vidéo, car le débit du réseau ne permet pas ce genre de choses. Nous proposons donc aussi des systèmes de réseaux partagés asynchrones qui permettent de déposer facilement une ressource. Le dépôt peut se faire dans différents formats, y compris dans des formats compressés qui seront décompressés automatiquement par la plate-forme.

Dès qu'elle est déposée, la ressource est prise en charge par ORTOLANG, ce qui signifie qu'elle va être sauvegardée systématiquement. Il s'agit ici d'une sauvegarde sécurisée : un système informatique assure une sauvegarde dans deux lieux différents, sur des machines différentes, pour être sûr de ne pas avoir de problème. Un nombre important de ressources ont en effet été perdues dans le passé à cause d'un crash du disque du chercheur. Dès que le dépôt est effectué, les déposants sont de plus accompagnés par les centres de compétences d'ORTOLANG : le LPL d'Aix-en-Provence et MoDyCo pour les ressources orales et multimodales, l'ATILF pour les ressources sur l'écrit. Cela implique des interactions directes entre les déposants et les membres des centres de compétences qui les accompagnent pour aboutir à une ressource publiable. Pour permettre à l'utilisateur de suivre l'évolution de sa ressource et être informé du stade où elle se trouve, des informations régulières lui sont alors transmises par les équipes d'ORTOLANG. Ainsi le déposant se voit ouvrir sur la plate-forme un espace de travail sécurisé où il pourra à la fois normaliser, standardiser sa ressource, enrichir ses métadonnées grâce à des systèmes interactifs que nous avons mis au point, et définir les accès qu'il souhaite pour ses données ainsi que la licence attachée à sa ressource : à ce stade, la ressource n'est visible que par les déposants ou par des ayants droit qu'il convient de définir.

Ce n'est que lorsque ce travail est terminé que l'on passe à la phase suivante dite de publication. Là, l'utilisateur peut faire des choix sur l'ouverture et la visibilité de sa ressource. Ce que nous souhaitons, c'est que les

ressources soient le plus possible entièrement libres. Néanmoins, un certain nombre de ressources ne peuvent être ouvertes qu'à la communauté de l'enseignement supérieur et de la recherche, voire limitées à une liste précise de personnes. Ainsi, si vous avez un corpus vidéo de suivi longitudinal de la constitution de lexiques et de l'apprentissage de la langue chez un enfant autiste en famille, il n'est pas possible de partager librement ces vidéos ne serait-ce qu'à l'ensemble des chercheurs de l'enseignement supérieur et de la recherche. Le déposant a alors la possibilité de n'ouvrir la publication qu'à un nombre limité de personnes clairement définies.

Une fois la ressource publiée, elle devient accessible via le site d'ORTOLANG. À ce moment-là, la ressource peut être archivée et consultée.

• Pour l'archivage, si le déposant le souhaite, une commission commune à ORTOLANG et à la TGIR Huma-Num décide si la ressource doit être archivée à long terme (sur trente ans ou plus). Le coût de l'archivage étant non-négligeable, si ce n'est qu'une version numérisée d'un texte imprimé la meilleure archive est sans doute le papier. Ce n'est peut-être pas nécessaire d'en faire une archive informatique qui coûtera plus cher que sa renumérisation. Par contre dans certains cas, comme pour la ressource ESLO, une enquête sociologique sur le parler français à Orléans dans les années 60, il est indispensable de l'archiver, car il ne sera jamais plus possible de la reconstituer.

• La ressource part donc côté archivage et peut être consultée. La consultation, quant à elle, peut se faire selon plusieurs modes : en téléchargement simple, avec une visionneuse spécifique présente sur la plate-forme ou sous forme de projet intégré, c'est-à-dire avec tout un site web spécifique pour pouvoir exploiter ou parcourir une ressource particulière, avec des possibilités de navigation, de recherche dans l'ensemble des ressources proposées.

Grâce à ce flux de travail, qui peut paraître un peu complexe, nous sommes certains que lorsqu'une ressource est proposée en consultation les données sont propres et homogènes : tous les tests ont été faits préalablement pour s'assurer de sa robustesse. Une fois qu'elle est publiée et consultable, la ressource peut être réutilisée dans une autre espace de travail ou projet pour pouvoir l'enrichir, en respectant bien entendu la licence qui lui est propre.

Pour finir, précisons comment ORTOLANG s'insère dans le dispositif national et international.

Au niveau national, ORTOLANG est un service spécialisé pour la langue, complémentaire de l'offre générale proposée par la TGIR Huma-Num. L'archivage de nos ressources se fait via la solution proposée par Huma-Num. Notons par ailleurs que dans le cadre du partenariat que nous avons la TGIR nous avons lancé des appels communs avec les consortiums Ecrit et IRCOM pour la standardisation et la finalisation de corpus ou de ressources existantes dans les laboratoires, mais qui étaient jusqu'ici non diffusées pour la plupart.

Au niveau international, nous participons aux infrastructures européennes de recherche, dont Dariah qui sera présenté tout à l'heure. Faisons donc un petit focus sur l'autre infrastructure de recherche européenne, CLARIN, spécialisée sur les langues. La position de la France actuellement vis-à-vis de CLARIN n'est pas complètement claire, parce que la France n'a pas encore pris la décision de la rejoindre, mais nous sommes prêts à contribuer à la création d'une CLARIN-France dès que la France donnera son feu vert officiellement[1]. Pour mémoire, notre laboratoire, l'ATILF, était contractant du projet européen qui a conduit à cette infrastructure.

Quels sont les objectifs de CLARIN ?

• CLARIN veut proposer des ressources dans diverses langues : nous sommes prêts à gérer les ressources pour le français et les langues de France.

• Pour la diffusion, CLARIN veut qu'il n'y ait pas d'autre restriction que celle découlant de considérations éthiques ou juridiques : nous plaidons pour des ressources entièrement ouvertes aussi.

• CLARIN veut une intégration maximale des données, c'est-à-dire avoir des métadonnées de recherche des contenus qui permettent aux chercheurs, où qu'ils soient, de parcourir les ressources qui sont dans le réseau CLARIN. Nous respectons au niveau d'ORTOLANG les normes de

1 Depuis la conférence, la position de la France vis-à-vis de CLARIN s'est clarifiée mi-juin 2015 par une décision d'être observateur au sein de cette infrastructure européenne.

représentation, en particulier les métadonnées compatibles avec celles de CLARIN.

• CLARIN souhaite qu'il y ait une intégration de service, pour toutes les modalités (texte, parole, geste), sous forme en particulier de web services. Nous allons proposer de tels web services d'exploitation au niveau d'ORTOLANG sur les ressources que nous gérons.

• CLARIN souhaite que les ressources accessibles via son réseau soient pérennisées, nous avons fait ce choix aussi. Cela signifie que chaque ressource peut être retrouvée à travers son identifiant pérenne n'importe quand, avec bien entendu un versionnage, puisque les ressources peuvent évoluer.

• CLARIN plaide pour une durabilité dans le temps. Dans notre système français ce point peut apparaître comme problématique : vous savez que même les plus grands laboratoires en France ont une durée de vie limitée, leur renouvellement étant discuté tous les cinq ans. Pour le moment notre existence est assurée jusque 2020, et c'est la réussite du projet qui fait qu'ORTOLANG sera pérennisé.

Aujourd'hui la première version de la plate-forme est accessible à l'adresse www.ortolang.fr, nous y intégrons toutes les ressources que nous avons déjà récupérées, et celles existant au sein des centres de ressources que le CNRS avait lancés en 2006 (Centre national de ressources textuelles et lexicales (CNRTL) et Centre de ressources sur l'oral *Speech and Language Data Repository* (SLDR) à Aix) ainsi au cours de cette année ce sont plusieurs centaines de ressources (corpus, lexiques et outils de traitement) qui seront directement accessibles sur notre plate-forme.

Présentation des infrastructures européennes pour les ressources linguistiques

Khalid Choukri, agence pour l'évaluation et la distribution des ressources linguistiques (ELDA) - association européenne pour les ressources linguistiques (ELRA)

Je vais essayer de vous brosser un panorama de quelques infrastructures européennes assez importantes, sur une vingtaine d'années.

La première grande initiative est une initiative européenne intitulée Relator, datant de 1994. La commission européenne a demandé à l'ensemble de la communauté scientifique européenne de réfléchir sur les infrastructures nécessaires pour pérenniser les ressources financées par la commission européenne. Les ressources étaient produites, puis perdues au bout de deux ou trois ans, l'industrie demandait des financements pour constituer le même genre de ressources plusieurs fois. La commission a donc demandé aux industriels de se rassembler pour travailler ensemble afin de mettre en place une infrastructure qui permette d'archiver et de rendre ce genre de ressources disponibles pour tout le monde. C'est de là que découle la création d'ELRA en 1995, l'Association européenne pour les ressources linguistiques, avec la participation d'un certain nombre de partenaires européens, qui a lancé d'autres initiatives listées ici et que je vais essayer de décrire sommairement.

On se demande si les données ont vraiment une certaine importance. J'ai pris l'exemple d'un cas de traduction automatique. Peu importe la technologie, en partant d'un corpus d'apprentissage de 100 000 mots, on voit que les performances s'améliorent au fur et à mesure que l'on injecte des mots supplémentaires pour atteindre le meilleur score avec 95 millions de mots, sans rien faire de plus. La notion de données est donc quand même assez importante, les infrastructures dont on parle aujourd'hui ont leur raison d'être. Lorsque ELRA a été créée comme je le disais en 1995, il y a vingt ans cette année, l'idée était d'identifier, de collecter, d'archiver toutes les ressources que les laboratoires de recherche

et les industriels utilisaient et de les rendre disponibles. Cela supposait de résoudre des difficultés, notamment techniques. Rapidement, les gens ont souhaité mettre en place des campagnes d'évaluation, afin de mesurer les performances qu'ils pouvaient obtenir avec les données. Nous sommes essentiellement une association d'utilisateurs de ressources, tous les gens qui en utilisent sont les bienvenus dans l'association. Les différents aspects que l'on gère vont des aspects techniques comme ceux que l'on évoque depuis hier, à l'aspect catalogage et listes de ressources.

Vous pouvez donc voir que ELRA est en quelque sorte l'intermédiaire entre des fournisseurs de ressources linguistiques et un certain nombre d'utilisateurs, et vous avez un aperçu de l'ensemble des tâches que l'on effectue (production de ressources, catalogage, description, validation, etc). La typologie de ressources couvre tout ce qui entre dans une interaction Homme-Machine, c'est-à-dire des données audio, vidéo, textuelles, OCR, multimédia/multimodale, etc.

Nous disposons essentiellement de quatre catalogues assez importants. Le premier est le « vrai » catalogue ELRA, c'est-à-dire le catalogue des produits qui sont disponibles. Aux dernières nouvelles, entre 30 et 35% de ces ressources sont gratuites (contre la signature d'une licence et sans contrepartie financière). Nous avons ensuite un catalogue universel, où on stocke tout ce qu'on a identifié. Il y a le *LRE-MAP*, un catalogue introduit il y a environ cinq ans : nous avons demandé à tous les chercheurs qui soumettaient des publications à des congrès comme *Language Resources and Evaluation Conference* (LREC), dont je parlerai tout à l'heure, d'ajouter à leurs articles scientifiques une description des ressources qui sont évoquées dedans ou sur lesquelles ils travaillent. Cela permet d'avoir à la fois le côté scientifique et le côté ressources. Nous disposons enfin de l'inventaire *Meta-share* dont je parlerai plus tard.

LREC est la grande conférence que l'on organise pour débattre de ce genre de sujet, avec plus de 1 200 participants. Un certain nombre de travaux sur les langues régionales y sont présentés.

Sur les autres grandes initiatives qui ont été soutenues par l'union européenne, la première vraiment importante est une action qui s'appelle FlaReNet, coordonnée par une équipe du CNRS italien à Pise avec

l'implication du CNRS français, d'ELDA et d'un certain nombre de partenaires. L'objet essentiel était de mettre en place un premier forum international qui essaye de canaliser toutes les initiatives qui avaient lieu en Europe, dans le but de produire à la fois une sorte de réseau d'experts impliqués dans le domaine mais aussi de faire des recommandations à la commission européenne sur ce qu'elle devait soutenir. L'action s'est déroulée entre 2008 et 2011. Le rapport final parle des ressources linguistiques du futur et du futur des ressources linguistiques dans le cadre européen. Le FlaReNet Book évoque ce que les chercheurs européens souhaitent voir entrepris par la commission européenne (disponible en ligne). Ensuite, en 2010, deux grandes initiatives ont été lancées par la commission européenne, la première s'appelait officiellement *Technologies for the Multilingual European Information Society* (*T4ME)*, qui a donné naissance au programme META-NET, un grand projet européen, et le deuxième projet similaire s'appelait *Platform for Automatic, Normalized Annotation and Cost-Effective Acquisition of Language Resources for Human Language Technologies* (PANACEA).

T4ME est le premier grand réseau d'excellence européen fédérant environ 800 partenaires européens. L'idée était de mettre en place un consortium de recherche sur la thématique de la traduction automatique, mais aussi une infrastructure ouverte pour la distribution et le partage de ressources linguistiques. Il y avait un troisième pilier qu'on appelait META-VISION, avec la production d'un agenda stratégique pour la recherche européenne, dans le cadre de quelques thématiques dont la traduction automatique, l'interaction homme-machine, etc.

Le projet PANACEA, celui qui devrait peut-être vous intéresser le plus même si j'en parle très sommairement, consiste à mettre en place une usine de production de ressources pour un certain nombre de langues, pour éviter que toutes les productions soient faites de façon manuelle et donc coûteuse. C'est vraiment l'idée d'une « *factory* ».

Meta-share était le deuxième pilier de Meta-net, dont l'objectif était de mettre en place un réseau distribué de centres de dépôts. C'était en 2010, dans la grande mouvance de la distribution, on voulait passer d'un groupe d'organismes avec des ressources et des archivages centralisés à quelque chose d'ouvert. L'objectif de Meta-share était de mettre en place une infrastructure d'échange de ressources linguistiques ouverte

en termes de logiciel gérant les centres de dépôts, mais aussi au niveau des ressources qui y seraient stockées, tout en assurant la sécurisation (données de valeur scientifique, commerciale, éthique, traitant de la vie privée parfois) et l'interopérabilité pour permettre de passer d'un centre de dépôt à un autre. De plus, l'idée de la commission européenne était vraiment de mettre en place un marché où les industriels puissent avoir accès à des données et des outils qui soient bien documentés, bien catalogués.

Dans Meta-share trois points assez importants ont été réalisés essentiellement. Le premier est la constitution du réseau lui-même. Le deuxième est le travail sur le modèle de « métadonnées », c'est-à-dire les éléments de description des ressources, pour pouvoir décrire de façon assez exhaustive aussi bien un lexique textuel qu'un lexique phonétique, un enregistrement audio, vidéo, avec de l'image, du texte qui s'affiche, et en troisième lieu l'établissement d'un cadre légal commun. Nous avons essayé de revisiter les licences qui sont utilisées par les uns et les autres pour essayer de les harmoniser. Nous nous sommes inspirés essentiellement de ce que ELRA faisait mais aussi de Creative Commons. Aujourd'hui existent les Meta-share Licences, qui sont à la fois spécifiques, car dans notre communauté tout n'est pas ouvert et tout n'est pas distribué avec des licences permissives, mais aussi des répliques de Creative Commons adaptées aux ressources linguistiques.

La base de l'architecture décentralisée et distribuée de Meta-share est un ensemble d'inventaires ou de centres de dépôts qui sont dans des labos de recherche, dans des agences linguistiques, chez des organismes comme ELRA.

Nous essayons de collecter toutes les métadonnées des ressources qui sont dans les différents dépôts pour avoir un inventaire commun cohérent, l'inventaire Meta-share, afin de permettre de réaliser un certain nombre d'opérations comme la recherche et la navigation mais aussi le téléchargement des données selon les licences. Le site présente un ensemble d'organismes et de partenaires européens, nationaux et régionaux qui permettent d'arriver directement sur le répertoire Meta-share. Ces organismes et partenaires permettent de collecter toutes leurs ressources et de les rendre visibles à travers le premier niveau du réseau.

Le deuxième projet dont j'ai parlé, et que je pense vraiment utile pour nous aujourd'hui, est PANACEA. Il s'agit d'une plate-forme de production

de ressources. PANACEA constitue par exemple des corpus parallèles avec un moteur de gestion de flux de données. L'utilisateur définit un certain nombre d'étapes, par exemple naviguer sur le web et crawler des données, en identifiant éventuellement une langue et un domaine. Ensuite on peut envoyer les résultats dans un deuxième service web pour réaliser l'alignement de ce qui a été *crawlé*. Cette usine de production de ressources est disponible et open source.

La dernière infrastructure est encore un projet européen, intitulé *MultiLingual* (MLi), lancé par la commission européenne fin 2013, et l'idée de ce projet est de définir auprès de la commission européenne l'infrastructure nécessaire aujourd'hui, compte tenu du contexte que j'ai évoqué : quelle serait la prochaine génération d'infrastructures pour gérer les ressources linguistiques en Europe ? Le résultat doit être disponible fin 2015. Cela porterait sur le genre de ressources linguistiques nécessaires pour demain, faut-il continuer à faire les choses manuellement ou bien instaurer l'automatisation et si oui de quelle manière, est-on capable aujourd'hui de mettre en place des services web qui permettent de prendre un corpus et un outil de traduction automatique et de les faire travailler ensemble, et est-ce qu'on peut voir ça dans le contexte de la R&D aussi bien publique que privée, mais aussi de l'industrie et de l'innovation : est-ce que ce hub va permettre à un certain nombre de sociétés et de *start-ups* d'être les Google de demain ? Est-ce qu'il est possible de se servir de l'existant comme blocs de construction, et de constituer un système de traduction parce que les technologies et les ressources existent ?

Enfin, les deux dernières initiatives financées il y a moins d'un mois par la commission sont *Cracker* et *LT-Observatory*. L'essentiel du rôle demandé à ces deux initiatives est de produire un agenda stratégique pour la recherche et l'innovation, une tâche confiée à un groupe de chercheurs européens. Il m'a semblé important de vous en parler, car on le considère parfois comme un travail académique à l'influence limitée. Il se trouve que de temps en temps la commission européenne émet des appels d'offre sur ces bases-là.

Évaluations des technologies de la langue

Juliette Kahn, laboratoire national de métrologie et d'essais (LNE)

Le LNE est un établissement public à caractère industriel et commercial, dont les principales missions de service public consistent à sécuriser les produits de consommation courante et piloter la métrologie française. Petit à petit, on se rend compte que même si la métrologie concerne principalement les questions du domaine physique, les technologies de l'information entrent de plus en plus dans ces nécessités de posséder des étalons, des référentiels communs, et le LNE travaille à cet objectif.

En termes de ressources, nous sommes composés de 800 collaborateurs implantés dans le monde avec environ 55 000 mètres carrés de laboratoire, ce qui s'explique par le fait que la physique nécessite de l'espace. Au niveau de nos missions, nous sommes tiers de confiance pour un certain nombre d'organismes publics, de façon à évaluer les progrès des systèmes de traitement de l'information multimédia. Cela passe par la mise en place de métriques, par l'organisation de campagnes d'évaluation et par l'analyse des résultats obtenus. Finalement, l'objectif est d'établir des référentiels communs, qui se composent à la fois de données (construites le plus souvent en collaboration avec des producteurs de données comme ELDA par exemple), mais aussi de proposer des outils d'évaluation qui permettent de mesurer la même chose et d'établir un certain nombre de règles communes. Au niveau des technologies de la langue, comme cela a été largement abordé au cours de ce colloque, nous travaillons aussi bien sur des données audio, de façon à répondre à des questions comme « Qui parle ? », « Quelle langue parle-t-il ? », « Que dit-il ? », que sur des données textuelles, avec des questions du type « Quel document est pertinent ? » pour répondre à une question, ou « Quelles personnes sont citées ? » dans un texte. Tout cela donne lieu à des tâches très précises que je vous ai listées, car la question en elle-même ne suffit pas : elle se décline ensuite de manière très précise en termes de tâches. Nous travaillons également sur l'évaluation de la traduction de documents textuels ou de transcription de parole, ainsi que sur le traitement d'image, en particulier sur la reconnaissance optique de caractères (OCR) pour répondre à la question « Qu'est-il écrit dans ces images ? ».

La question importante est : pourquoi créer et évaluer des systèmes ? Nous évoquons souvent trois raisons principales. La première consiste à aider l'humain à réaliser une tâche. L'évaluation cherche donc à voir si le temps de réalisation a été accéléré et si la précision dans la réalisation de la tâche a été accrue. Un autre objectif peut être de suppléer l'humain dans la réalisation d'une tâche. Dans ce cas, un bon critère d'évaluation est la comparaison des différences de réalisation entre la production humaine et celle du ou des systèmes. Enfin, la dernière question est de comprendre comment l'humain réalise la tâche et d'essayer de reproduire le processus.

Le terme « évaluation » recoupe finalement plusieurs types d'évaluation. On peut se positionner en termes de test d'usage, où l'on va mettre les utilisateurs face à un certain nombre de systèmes, en cherchant à savoir si les technologies développées sont un outil pertinent pour l'humain, si le temps de réalisation progresse. Nous avons un deuxième outil à notre disposition : les campagnes d'évaluation ouvertes ou techniques. Dans ce cas, les questions vont plutôt évaluer la pertinence des modèles développés et la fiabilité des systèmes. Ces campagnes sont des outils de comparaison et de développement de systèmes qui sont extrêmement précieux. Il y a bien entendu une complémentarité entre ces approches, il est évidemment intéressant que les technologies soient évaluées à la fois en test d'usage et en campagne d'évaluation. Ici, je n'aborde absolument pas la question de la performance dans le sens du temps de traitement des données, ou du suivi de la qualité de développement des technologies. Le LNE se focalise principalement sur ces questions de campagnes d'évaluation ouvertes. Finalement, on est vraiment face à un cycle vertueux grâce à ces évaluations. À partir de théories et d'applications définissant un certain nombre de besoins, une tâche très précise est délimitée, à laquelle différents systèmes essayent de répondre concomitamment en développant différents modèles. En somme, on entre dans un cycle où l'évaluation permet de statuer sur la pertinence des modèles, et donc de redéfinir une nouvelle tâche, qui permet de réévaluer à nouveau les modèles.

Au niveau des évaluations, on part de données de test, au sujet desquelles les humains donnent un certain nombre d'indications concernant ce qui est attendu : on créée des références. Les systèmes ont exactement les mêmes données à traiter et soumettent leurs hypothèses. Tout le travail consiste à comparer les sorties des systèmes avec ces références de façon à estimer leur fiabilité et leur performance.

Il est quand même possible de travailler avec deux approches différentes. La première consiste à travailler sur corpus a priori. L'idée est de comparer les sorties de systèmes avec un corpus de référence annoté avant la mise en place du test. C'est souvent ce qu'on utilise dans les tâches de transcription de la parole, de repérage d'entités nommées, voire de traduction. L'avantage c'est que la reproductibilité de l'évaluation est garantie très facilement puisque les corpus sont créés à l'avance, il suffit donc de les utiliser pour évaluer les éléments. Évidemment, on peut quantifier aisément les progrès des systèmes, on fournit aux développeurs une fonction qu'ils cherchent à optimiser, et ils peuvent travailler très facilement à partir de ces corpus a priori. L'inconvénient c'est qu'il n'est pas toujours possible d'appliquer ces types de méthodes quand la réponse est équivoque. Par exemple, je mentionne le fait que pour la traduction on peut utiliser ce type d'évaluation, mais ce n'est pas toujours ce qui est privilégié, parce qu'il peut y avoir une attente de plusieurs références. Une façon de traduire une phrase vers une autre peut attendre plusieurs réponses, et il est évidemment impossible, a priori, de proposer toutes les solutions possibles pour une traduction donnée. C'est pour cela qu'on travaille aussi sur des campagnes d'évaluation avec des corpus a posteriori. L'idée est plutôt que le système traite un corpus qui n'est pas annoté au départ, la réponse est jugée par des annotateurs et les jugements sont comptabilisés de façon à estimer la qualité de ces sorties. Ce sont des solutions qui sont plutôt utilisées en traduction, en recherche d'information. On retrouve l'avantage de quantifier les progrès des systèmes, on peut traiter des tâches qui sont plus complexes et on fournit en fin d'évaluation des sorties qui peuvent être réutilisées, notamment par les développeurs pour améliorer leurs systèmes. Un autre avantage est que cela fournit une évaluation proche des sorties de système et qui s'adapte à l'ambiguïté linguistique. L'inconvénient est que la reproductibilité est partielle, puisqu'on ne dispose pas de réponse pour toutes les sorties de système possible et qu'à chaque nouvelle sortie de système, si celle-ci n'a jamais été traitée, il faut demander à un humain de recommencer son jugement.

Au niveau du déroulement d'une campagne d'évaluation, le premier point très important est la définition de la tâche à évaluer en tant que telle. Il faut pouvoir formaliser les objectifs, la méthode de comparaison et surtout les données sur lesquelles on travaille. Ce n'est pas du tout la même chose de faire de la transcription de parole en choisissant de ne travailler que

lorsqu'un seul locuteur parle, ou en travaillant avec de la parole superposée. Il faut se mettre d'accord sur un certain nombre de formalismes, de façon à ce que tout le monde puisse répondre à la tâche proprement, et se mettre d'accord sur la comptabilisation des erreurs et la mesure de la performance du système. La deuxième étape consiste en la préparation de la campagne. Il s'agit d'implémenter les métriques de façon à fournir des outils permettant la comparaison, produire les données, les répartir entre celles qu'on donne au développeur (données d'entraînement), les données qui permettent d'optimiser leurs systèmes (données de développement), et les données que l'évaluateur conserve et qui sont les données de test permettant de rendre compte de la performance des systèmes.

La troisième étape consiste en l'exécution de la campagne, où l'on applique les métriques, puis l'étape très importante d'adjudication. Cette étape consiste à faire un point une fois qu'une première phase de résultats sur le système est donnée : on revisite le corpus de référence de façon à l'améliorer au cas où il y aurait eu un certain nombre d'erreurs. Une troisième étape très importante est l'analyse des résultats, puisqu'il ne suffit pas de donner un score. Le score permet de comparer les solutions des systèmes les uns par rapport aux autres, mais le but est surtout de progresser par l'évaluation, et de comprendre les erreurs que les systèmes produisent actuellement, de façon à toujours pouvoir être dans ce cercle vertueux d'amélioration des modèles à partir de ces éléments-là. Évidemment une dernière tâche consiste à animer la campagne en tant que telle. Une campagne d'évaluation ce n'est pas seulement confronter les chercheurs les uns par rapport aux autres, c'est au contraire une saine émulation, leur permettre d'échanger les uns avec les autres, d'organiser des colloques et des rencontres autour de ces événements.

Bien entendu pour nous l'enjeu d'évaluation autour des corpus est très important. Il faut avoir un corpus qui soit représentatif de la tâche, duquel on puisse caractériser les difficultés, comprendre si finalement les données fournies vont être complexes ou simples à traiter par le système. Il faut que les documents utilisés dans le corpus soient ceux que l'on veut réellement traiter par la suite. En termes d'annotation, nous en avons de deux types : ce qu'on attend que le système réalise, identifier le phénomène à trouver, et ce qui permet de révéler des facteurs explicatifs des potentielles erreurs (bien connaître le corpus de façon à pouvoir comprendre où se situent les

erreurs et pourquoi elles sont observées). Le dernier outil que je voudrais évoquer est l'accord inter-annotateur. Nos corpus d'évaluation sont soumis à plusieurs annotateurs, ils sont annotés plusieurs fois de façon à vérifier que les références utilisées font consensus, que le phénomène a bien été défini et qu'il existe une homogénéité entre les corpus.

En ce qui concerne les métriques, leur objectif premier est de comparer les systèmes les uns par rapport aux autres. Il faut donc qu'elles soient capables d'estimer la significativité des différences de performance, ce qui n'est pas forcément évident. Khalid Choukri montrait tout à l'heure du BLEU. Aujourd'hui il est très difficile de dire si un BLEU de 40 c'est mieux qu'un BLEU de 45, ou même de 38. Statistiquement parlant, la communauté a un peu de mal à dire quelles sont les différences entre ces chiffres. Elle doit nous permettre aussi cette analyse d'erreur, de façon à ce que les effets de certains facteurs soient identifiés. De plus, c'est quand même intéressant si elle correspond au jugement des humains. S'il n'y a pas du tout d'adéquation entre ce que les développeurs cherchent à optimiser et ce que les utilisateurs jugent comme pertinent, c'est qu'il y a un souci.

Pour finir cette présentation, je vais vous donner quelques exemples de campagnes qui ont été organisées ces dernières années. Toutes ces campagnes sont référencées sur le site du LNE. La première campagne est la campagne Évaluations en Traitement Automatique de la Parole (ÉTAPE), financée par l'ANR, coorganisée par l'Association Francophone de Communication Parlée, la Direction Générale de l'Armement et ELDA. Elle rassemblait seize participants académiques et industriels. Les tâches principales étaient du suivi de locuteur (qui parle quand) et la transcription de parole télévisuelle et radiophonique, la particularité étant que l'on travaillait aussi bien en parole superposée que non-superposée. Lorsque deux locuteurs parlaient en même temps, il fallait que les systèmes soient capables de transcrire ce qui était dit, ce qui est une tâche très difficile pour les participants. Il y avait aussi une question autour de la détection d'entités nommées.

La seconde est la campagne de moyens automatisés de reconnaissances de documents écrits (MAURDOR), financée par la direction générale de l'armement (DGA) et coorganisée par le LNE et Cassidian. Le corpus a été entièrement créé par ELDA. C'est une campagne composée de six

participants académiques et industriels. Il faut savoir que ce sont des tâches très nouvelles en matière de suivi de séquence, nous étions donc assez contents de ce niveau de participation. L'objectif était d'évaluer des chaînes de traitement complètes de documents numérisés, c'est-à-dire qu'il s'agissait d'abord de détecter les zones qui composaient les documents : séparer les images des zones de textes, des zones de tableaux, des ratures, des signatures, etc. Pour les zones d'écriture on cherchait à identifier si l'écriture était manuscrite ou tapuscrite, les deux étant représentées. Nous cherchions également à identifier la langue des documents, puisque le français, l'anglais et l'arabe étaient représentées, d'abord en reconnaissant les caractères et ensuite en identifiant les liens logiques existant entre différentes zones (repérer la légende d'un graphique, un titre par rapport à un corps de texte).

La troisième est la campagne de traduction pour l'aide à l'analyse documentaire (TRAD), également financée par la direction générale de l'armement, coorganisée par le LNE et Cassidian, dont les corpus ont été constitués par ELDA. La particularité de cette campagne était de travailler sur le couple de traduction arabe-français, sur différents types de sources, aussi bien des textes journalistiques que de la parole, du blog, du mail. Le but était toujours d'aller vers le français, mais sur différents types de textes ou de parole.

Enfin, la dernière campagne est la campagne de Reconnaissance de PERsonnes dans des Émissions audiovisuelles (REPERE), financée par l'ANR et la DGA, coorganisée par le LNE et ELDA, où 16 participants académiques et industriels ont participé en étant regroupés sur trois consortiums. L'objectif de ce défi était d'indiquer qui était présent, soit visuellement soit oralement, dans des vidéos issues de la télévision. Cela a donné lieu à de nombreuses tâches, notamment le suivi du locuteur, son identification, la transcription de parole, le repérage d'entités nommées, la reconnaissance optique de caractères. L'avantage de toutes ces campagnes est qu'elles ont déjà eu lieu. Les corpus et les soumissions sont disponibles et distribués principalement par ELDA. Par conséquent quelqu'un qui voudrait entrer dans le milieu et connaître l'état de l'art pour ces types de tâches peut se servir de toutes ces évaluations pour situer son système par rapport à cet ensemble.

Au niveau des résultats, les systèmes de la campagne REPERE ont été évalués chaque année, et l'on peut voir qu'il y a une grande progression des consortiums sur la réalisation des tâches, en particulier pour le consortium 1. Ce qu'il faut voir par rapport aux deux autres consortiums c'est que le corpus qui a été utilisé lors de la deuxième campagne était beaucoup plus complexe que pour la première, dans le sens où on a rajouté beaucoup de données et qu'il y avait une émission surprise à traiter que les systèmes n'avaient jamais vue. Il faut toujours bien insister sur le fait qu'il faut comparer ce qui est comparable, et qu'il est nécessaire de rejouer les systèmes sur les mêmes corpus. À chaque fois que le corpus d'évaluation est changé, le référentiel est changé aussi, donc même si l'on utilise la même métrique on ne peut pas comparer la progression des systèmes sur deux corpus différents, puisque la difficulté du corpus joue énormément sur les résultats. Ceci dit on voit quand même une progression spectaculaire des consortiums. On ne note pas de différence significative entre les systèmes, malgré les différences de points. Concernant l'émission surprise qui n'était pas dans les corpus d'apprentissage, elle se situe dans la classe des émissions qui étaient plus difficiles que d'autres, mais ce n'était pas forcément la plus difficile. Les difficultés se retrouvent vraiment selon le type d'émission. Celles qui sont globalement les mieux traitées sont les débats, puis les journaux, et de façon beaucoup plus difficile pour tous les systèmes on trouve les questions au gouvernement et un magazine *people* où le langage n'est pas du tout le même que dans les journaux télévisés, ce qui fait que les systèmes ont du mal à repérer qui parle.

En conclusion, les évaluations sont vraiment un outil d'estimation de la fiabilité et de la pertinence des systèmes. Elles permettent de situer des outils ou des approches par rapport à un état de l'art, et de faire progresser les systèmes en proposant un cadre commun permettant à toute la communauté d'avancer sur les problématiques de traitement du langage.

L'ERIC DARIAH : Une infrastructure européenne pour les sciences humaines et sociales

Jean-Luc Minel, Modèles, Dynamiques, Corpus (MoDyCo) – université Paris X

Je vais vous parler de *Digital Research Infrastructure for the Arts and Humanities* (DARIAH), une infrastructure européenne pour les sciences humaines et sociales. Théoriquement c'est Laurent Romary, l'un des directeurs de DARIAH, ou Sophie David, coordinatrice de DARIAH France, qui auraient dû être présents, mais ils sont tous les deux absents et je les remplace aujourd'hui. Je suis impliqué dans DARIAH mais n'en suis pas un membre statutaire.

Vous pouvez trouver des informations sur le site http://www.dariah.eu, site européen de DARIAH. Il existe également un site français[1].

Je vais essayer de vous faire comprendre ce qu'est DARIAH, ce que cela peut vous apporter, et j'évoquerai l'infrastructure institutionnelle sur laquelle s'appuie DARIAH (rôles, budgets, implication de la France, possibilités de contribution et d'utilisation de DARIAH).

Qu'est-ce que DARIAH ?

Je préfère le formuler négativement : DARIAH n'est pas une agence de moyens, DARIAH ne finance rien. Au contraire, c'est vous en tant qu'unité de recherche qui allez apporter des éléments dans DARIAH. J'insiste, car c'est souvent la source de malentendus avec les enseignants-chercheurs, et étant moi-même directeur d'un laboratoire, on me demande parfois l'intérêt, s'il n'y a pas d'argent, d'aller dans DARIAH. Ma réponse en général est qu'en tant qu'enseignant-chercheur, il faut contribuer à la diffusion des connaissances, ce que DARIAH permet de faire. Le point de départ, en 2006-2007, est une opération politique initiée essentiellement par la France et l'Allemagne décidant de mettre en place une infrastructure (terme très ambigu à mon sens) dédiée à toutes les disciplines des sciences

1 http://www.dariah.fr/

humaines et sociales pour mettre en place un réseau d'expertise, d'outils, de services, d'acteurs (chercheurs) qui s'intéressent à un certain type d'objets, numériques essentiellement (texte, son, image, vidéo). J'insiste sur la notion de réseau. Peut-être que les initiateurs de DARIAH avaient l'idée de mettre en place une infrastructure physique européenne dans laquelle il y aurait des données. Je pense qu'on a beaucoup évolué, et qu'on est maintenant sur l'idée de fédérer des infrastructures existantes dans des pays pour les rendre éventuellement interopérables et pour être un lieu d'échange. On s'oriente plutôt maintenant vers un DARIAH qui serait un réseau. Enfin, il n'y a dans DARIAH que des chercheurs. C'est un réseau pour les chercheurs et par les chercheurs. On peut éventuellement le regretter, mais il y a très peu d'infrastructure administrative et technique dans DARIAH. Il n'y a pas d'espace dans lequel il y aurait un ensemble d'ingénieurs, de techniciens et d'administratifs qui s'occuperaient de la structure. Cela la rend fragile tout en lui donnant une certaine flexibilité.

DARIAH permet d'accéder à un certain nombre d'événements, d'articles scientifiques, d'archives ouvertes, de données. On pourrait me rétorquer qu'on a déjà ça en France : le calendrier des lettres et sciences humaines et sociales (Calenda), les archives ouvertes. Pour hypothèses.org, vous remarquerez que le blog n'est pas en français mais en anglais, et que les données viennent d'Italie. Je cherche à vous montrer que c'est bien un réseau : des contributeurs, des Français, des Italiens, des Allemands, des Hollandais, vont placer dans ce nuage DARIAH des ressources déjà existantes, et les mettent à la disposition des partenaires de DARIAH. Par exemple, *Open Edition* a mis à la disposition de DARIAH l'espace des blogs : il y a des blogs en allemand, en italien, en irlandais, etc. Il faut le voir sous la forme d'un réseau de plate-forme dans lequel les contributeurs amènent la ressource ou des outils. C'est aussi un réseau de formation, selon une volonté de l'Europe : fournir un panorama des formations en « *art et humanities* », « *digital humanities* » même si je participe pas à ce vocabulaire. C'est l'idée que l'ensemble des étudiants, des chercheurs et des enseignants-chercheurs puissent être informés des universités d'été, du matériel pédagogique multilingue disponible pour des enseignements ou des formations à l'intérieur des laboratoires. L'inventaire de l'ensemble des activités pédagogiques et de formation qui existent en Europe est sans doute l'activité qui fonctionne le mieux, c'est peut-être aussi la plus facile.

Cela se veut aussi un réseau technique. Sur ce point-là, je pense que DARIAH est peut-être un peu plus fragile. Un certain nombre d'acteurs proposent une infrastructure d'authentification, la possibilité de créer des identifiants pérennes, la possibilité de labelliser des données pour les rendre conformes, sécurisées, interopérables, ce qui recoupe ce que disait Jean-Marie Pierrel sur le fait que lorsque vous versez vos données dans un entrepôt il faut vérifier qu'elles vont pouvoir être pérennisées d'un point de vue des formats. Les hollandais proposent un flux de travail qui est utilisable pour cela.

Le réseau se veut aussi pluridisciplinaire, ce qui à mon avis est un des points forts de DARIAH. Souvent, les chercheurs me disent qu'ils connaissent très bien les réseaux propres à leur discipline. Certes. Mais un archéologue ne connaît pas nécessairement bien les réseaux des linguistes ou des ethnologues, ou inversement. Les assemblées générales de DARIAH réunissent chaque année environ 200 personnes de l'ensemble de l'Europe, et sont la possibilité d'échanger sur des disciplines qui n'ont pas nécessairement l'habitude de confronter à la fois leurs savoirs, leurs outils, leurs méthodologies et leurs techniques. Voici quelques exemples de réseaux : *Advanced Research Infrastructure for Archaeological Dataset Networking in Europe* (ARIADNE) pour l'archéologie, *Collaborative EuropeaN Digital Archive Infrastructure* (CENDARI) pour l'histoire médiévale, Dispositif d'Information sur les travauX pour les Infrastructures Télécom (DIXIT), *Network for Digital Methods in the Arts and Humanities* (NeDIMAH), Histoire de l'Holocauste, etc.

Je cherche à vous montrer que c'est un lieu d'échange qui va permettre de confronter et d'apprendre un certain nombre de techniques. C'est également un lieu de partage de questions : quels sont les outils de préservation à long terme ? Quel genre de format je dois choisir ? Cela concerne évidemment les linguistes mais pas seulement. J'ai des exemples en tête de sources orales concernant les ethnologues (notamment au niveau des métadonnées) et pas seulement les linguistes. Ce type de question ne peut pas être vu uniquement sous une focale disciplinaire, mais nécessite de croiser des regards. DARIAH est un lieu qui permet d'échanger sur ces questions. Une question mentionnée plusieurs fois est celle des données personnelles et de l'anonymisation. Ces questions récurrentes ne concernent pas que les linguistes, mais aussi les ethnologues, les anthropologues, et ne relèvent pas des mêmes lois

en Hollande, en France, en Allemagne. Cela permet là encore de croiser les savoirs, et dans le cas où les données sont mises en exposition en Europe, de vérifier la conformité avec la loi française et la loi allemande en même temps, par exemple.

DARIAH propose un certain nombre de principes : le libre accès (avec préservation des problèmes juridiques et de données privées), la certification des entrepôts (sécurisation, pérennisation), le développement de l'archivage à long terme (les politiques dans les différents pays européens n'étant pas les mêmes, cela pose là aussi des problèmes de frontière : en France par exemple certaines données ne peuvent pas être stockées hors des frontières), et enfin, l'une des grandes idées de DARIAH, la décentralisation de l'ensemble (pas de volonté d'une direction centrale qui imposerait à la fois ses formats et ses principes).

Au niveau de l'organisation institutionnelle, DARIAH est un ERIC. C'est une production européenne exonérée de la TVA et dont l'objectif est de faciliter l'installation d'infrastructures sur le long terme : la feuille de route est de vingt ans, même si tous les partenaires ne se sont pas impliqués sur une telle durée, ce qui peut quelque fois poser quelques soucis. Le projet politique a été lancé en 2006 et l'ERIC a été officialisé l'année dernière : sa construction aura donc demandé huit ans.

En ce qui concerne la gouvernance, DARIAH cherche à se rapprocher du fonctionnement de la recherche. Le conseil d'administration est composé de trois directeurs : Laurent Romary, Tobias Blanke et Conny Kristel, auquel s'ajoute une assemblée générale et un coordinateur par pays. La coordinatrice pour la France est Sophie David, qui appartient au CNRS et qui est affectée à la TGIR Huma-Num. Il y a donc une imbrication entre DARIAH et la TGIR Huma-Num. L'idée de départ de DARIAH était celle d'une infrastructure découpée en quatre centres de compétences virtuelles, dans lesquels les partenaires échangeraient des informations, avec un découpage qui reprenait une idée très informatique avec une couche physique, l'infrastructure technique, une couche logicielle, l'infrastructure des contenus, une couche de formation. Ceci ne s'est pas avéré pertinent et n'a pas fonctionné. DARIAH a été amené à évoluer vers l'idée qu'il ne fallait pas avoir un découpage en couches, mais quelque chose qui soit à la fois plus décentralisé, et plus proche d'un réseau.

Je vais passer sur les quinze membres fondateurs, mais j'insiste quand même sur le fait que tous les pays européens ne sont pas représentés. En revanche, et je pense que c'est un élément intéressant concernant la recherche et à la diffusion des connaissances, de « petits pays » se sont impliqués. Par petit pays, j'entends au niveau PIB. En effet, la contribution dans DARIAH est proportionnelle au PIB du pays, il y a donc deux grands acteurs, la France et l'Allemagne, et de petits pays : Malte, Chypre, la Serbie, la Slovénie. Cela démontre qu'il y a un intérêt de ces pays pour DARIAH, ils sont très demandeurs de ce que la France et l'Allemagne peut leur apporter.

En ce qui concerne la place de la France dans DARIAH, elle est ancrée sur la TGIR Huma-Num, elle accueille le siège social de l'ERIC qui sera à Paris, et elle est impliquée dans ce qui s'appelle encore les centres de compétences, avec quatre personnes dont deux appartiennent encore à la TGIR Huma-Num : Nicolas Larrousse, plutôt sur la partie architecture technique, Sophie David, Aurélien Berra qui s'occupe de la partie formation et moi-même sur les contenus.

Pour insister et conclure sur ce qui, je pense, peut vous intéresser le plus, le fonctionnement de DARIAH évolue. Au départ et jusqu'à l'année dernière, les chercheurs apportaient des contributions : un contributeur pouvait décider d'apporter une ressource orale, un corpus qu'il souhaitait mettre à la disposition de l'ensemble de la communauté de DARIAH, avec un certain nombre de contraintes (documentation, accès, formations pour l'utilisation des données). Suivant les recommandations d'un certain nombre d'instances, nous allons évoluer vers des groupes de travail thématiques rassemblant des chercheurs de tous les pays à la place du découpage par couche. 18 propositions de groupes de travail se sont mis en place et rassemblent des chercheurs et enseignants-chercheurs de différents pays. Cela fonctionne un peu sur les techniques que vous connaissez sans doute de *working group* : deux ou trois chercheurs émettent une proposition qui est diffusée, des chercheurs se regroupent, ils décident de *délivrables*, d'une durée de vie identifiant l'apport au bout de deux ou trois ans dans DARIAH, et de ce qui sera mis à la disponibilité des chercheurs.

Certains groupes concernent plus particulièrement les technologies des langues, comme *Natural Language Processing (NLP) for DARIAH* et *Lexical Resources, Guidelines and Standards*.

Pour conclure, ce qui à mon avis est important dans DARIAH et ce que cela peut vous apporter, c'est qu'il s'agit d'un lieu où l'on donne de la visibilité aux recherches qui sont menées dans les laboratoires. Cela me semble assez clair au niveau européen. C'est également un lieu de rencontre pour commencer à construire un projet de recherche dans Horizon 2020, le portail français pour la recherche et l'innovation (H2020), parce que je pense que DARIAH peut servir de tremplin si on veut constituer un consortium de plusieurs pays.

La très grande infrastructure de recherche HumaNum

Stéphane Pouyllau, Très grande infrastructure de recherche des humanités numériques (TGIR HumaNum)

Plutôt que de brosser toute la surface qu'occupe l'infrastructure de recherche Huma-Num pour les humanités numériques, puisque Jean-Luc Minel juste avant moi a largement parlé de l'activité « européenne » de l'infrastructure au travers de la participation de la France dans DARIAH, je vais principalement faire un balayage rapide de notre infrastructure et des services que l'on met à disposition, puis j'essaierai de faire le lien avec ce que Jean-Marie Pierrel a dit tout à l'heure à propos de l'infrastructure ORTOLANG.

Tout d'abord, un point de contexte : Huma-Num est ce qu'on appelle une « très grande infrastructure de recherche ». La France compte une vingtaine de ces dispositifs financés par le ministère de l'Enseignement supérieur et de la Recherche.

Notre infrastructure n'est pas une infrastructure disciplinaire ou regroupant des disciplines proches, mais une infrastructure couvrant l'ensemble des disciplines des sciences humaines et sociales, à l'inverse de ce qui a été montré tout à l'heure par Jean-Marie Pierrel. Nous sommes composés principalement d'une structure portée par une unité mixte de services regroupant un peu plus d'une dizaine d'ingénieurs, localisée sur deux sites, à Paris et à Villeurbanne. La création est assez récente et issue de la fusion du Très Grand Équipement (TGE) Adonis, qui préexistait, et de l'infrastructure Corpus qui regroupait un ensemble de communautés scientifiques organisées autour des questions du numérique.

Huma-Num a une mission très simple : faciliter le tournant du numérique dans l'ensemble des pratiques au cœur de la recherche en sciences humaines et sociales. Nous sommes là pour accompagner les programmes de recherches, les laboratoires, autour des questions qui touchent leurs données numériques. Nous avons principalement deux objectifs : faciliter l'appropriation des questions du numérique dans les équipes de recherche, c'est-à-dire outiller

et donner des instruments au cœur des équipes de recherche, et développer des services pour mettre ces outils et instruments à disposition au juste niveau, et surtout au bon moment, dans le processus de recherche.

Quelques enjeux autour de ces objectifs sont importants à rappeler. Aujourd'hui, la recherche en sciences humaines et sociales est passée dans un monde massivement numérique. On manipule tous les jours des données numériques, qui ont une certaine fragilité dont il faut se préoccuper : le cycle de vie de ces données numériques n'est pas le même que celui des données analogiques ou physiques. Il nous faut construire des instruments, des infrastructures qui vont avoir le rôle de préservation, d'outillage, de traitement, etc. De la même façon que dans le monde analogique, physique, nous avions des bibliothèques, des services d'archives, des instruments pour préserver les données, nous avons besoin aujourd'hui de construire des instruments qui vont s'occuper des données numériques.

Nous développons aujourd'hui autour de ces enjeux trois activités principales :

• Une activité d'animation des communautés scientifiques en sciences humaines et sociales autour et sous la forme de consortiums disciplinaires dotés d'une mission très claire, celle d'animer leur communauté disciplinaire sur le plan de l'appropriation d'outils, des standards, des méthodologies, tournant autour de la donnée numérique et de son traitement.

• Nous accompagnons ces consortiums et l'ensemble des chercheurs avec des services numériques qui ont une vocation générique. Nous sommes assez agnostiques sur le plan des disciplines scientifiques, les services que nous propulsons sont relativement génériques et peuvent être utilisés à la fois par des archéologues, des historiens, des géographes, etc.

• Et comme l'a dit Jean-Luc Minel, nous avons une mission de coordination de la participation de la France dans l'ERIC DARIAH, Dans cette mission de coordination, nous finançons DARIAH ce qui inclus, outre un apport financier, une contribution en nature, c'est-à-dire en actions de nos communautés de recherche, que nous sélectionnons et que nous faisons remonter dans l'infrastructure européenne.

Huma-Num a pour cœur de métier les données de la recherche et tout ce qui tourne autour, c'est-à-dire l'outillage des équipes qui manipulent et qui créent ces données. Nous faisons donc de l'animation et de la création de consortiums dans les communautés de recherche et les services qui entourent ces dispositifs.

Concernant les consortiums, nous en dénombrons aujourd'hui onze en fonctionnement. Ils n'ont pas tous été créés à la même époque. Un consortium disciplinaire est une labellisation d'une durée de quatre ans, sur un projet précis d'animation de la discipline. Un peu plus de 400 personnes dans les laboratoires de sciences humaines et sociales travaillent à leur animation. Leurs trois grandes missions sont de définir des politiques scientifiques dans la *curation* des corpus, des méthodes, des outils, des standards communs, et de former et échanger sur les pratiques numériques.

En ce qui concerne les services numériques, ils sont centrés comme pour les consortiums sur les données numériques, et se déclinent sous la forme de différentes actions qui vont du stockage à la diffusion, au signalement, à l'archivage, et à quelque chose que l'on essaie de développer depuis quelques mois maintenant : le fait de cultiver les données. En effet, il ne s'agit pas simplement de les créer, les documenter et les stocker dans un coin, il faut aussi les entretenir, notamment dans le temps. Cela chamboule le cycle de vie de la recherche, où dans le monde physique les documentalistes, les archivistes, les bibliothécaires s'occupaient des données quand on ne pouvait pas le faire. Avec le numérique et l'évolution des formats et des métadonnées c'est beaucoup moins évident. Nous n'avons pas forcément les bonnes structures pour le faire. Depuis quelques mois arrive donc cette idée de cultiver les données, c'est-à-dire de mettre en place des outils qui permettent aux chercheurs de le faire sans pour autant « plomber » leur temps de travail de recherche. Nos services numériques sont déclinés derrière, sous la forme de différents outils.

En ce qui concerne le stockage, nous outillons des communautés de recherche avec des plate-formes de partage de données, par exemple. Nous déployons cet outil dans beaucoup de communautés, pour permettre aux chercheurs de stocker des données hors les murs de leur laboratoire (sans pour autant les documenter d'ailleurs). Il s'agit d'une copie distante

du disque dur, comme beaucoup de nos collègues en utilisent déjà avec par exemple *Dropbox*. La différence avec ces services qui viennent du monde commercial, c'est que nous les déployons sur des serveurs qui sont à Huma-Num, dans notre infrastructure, ce qui apporte une certaine garantie que les données déposées ne sont pas recollectées par des moteurs de recherche commerciaux.

Nous avons d'autres outils de stockage, notamment un outil à destination des très grosses structures du type des maisons des sciences de l'homme, accueillant en leur sein plusieurs laboratoires, qui font déjà du stockage au niveau régional et qui ont besoin de faire une réplication à grande échelle. Cet outil ne s'adresse pas trop aux chercheurs individuels mais plutôt à des services informatiques qui veulent faire de la synchronisation de données ou de la réplication.

Une fois les données stockées, il est intéressant de pouvoir les traiter et les cultiver. Nous déployons donc des services et des outils comme des outils de sciences de l'information géographique (SIG), des conteneurs de données numériques comme la plate-forme ArkéoGIS, qui vient d'ailleurs d'une maison des sciences de l'homme et qui est en train d'être portée au niveau national via Huma-Num par l'équipe qui la fabrique. Nous prenons le relais derrière pour les autres communautés.

Pour cultiver ses données encore, on trouve la plate-forme *exec&share*, issue du monde de l'économie, qui permet de faire des publications scientifiques enrichies de calculs scientifiques et qui émane d'un laboratoire, ou encore la plate-forme intitulée Telemeta développée par le Centre de recherche en ethnomusicologie (CREM). C'est une contribution qui vient de l'un de nos consortiums et qui est en train d'être déclinée pour d'autres disciplines que celle dont elle est issue, à savoir l'ethnomusicologie.

J'ai rapidement balayé nos premiers services en vous montrant des choses qui arrivent du terrain, des laboratoires, des maisons des sciences de l'homme, de nos consortiums, mais nous sommes aussi attentifs aux collègues et c'est peut-être là qu'il y aura une différence entre Huma-Num et une infrastructure dédiée comme ORTOLANG. Nous devons aussi équiper des chercheurs et des programmes qui ne se sont pas forcément approprié des outils ou qui n'ont pas mis en place d'Équipex pour développer une plate-forme, et

qui aujourd'hui démarrent cette question de la donnée numérique. Nous cherchons à faire monter les outils qui viennent de ces communautés.

Cela prend la forme d'un tout nouveau service qui s'appelle *Nakala* (mot issu du swahili qui signifie « copie », « exemplaire »), ouvert en septembre dernier. Il s'agit d'un dispositif un peu différent du stockage de données que je vous ai montré tout à l'heure, dans lequel on demande de documenter les données suivant un certain nombre de méthodologies et de recommandations documentaires dont les modèles sont publiés et accessibles. L'idée de *Nakala* est à la fois d'offrir la possibilité de déposer des données sur une plate-forme et de les documenter, c'est-à-dire d'y adjoindre des métadonnées de façon normalisée pour pouvoir les réutiliser plus tard à l'aide d'identifiants pérennes, notamment dans les publications. C'est la possibilité également d'avoir des outils d'interopérabilité greffés aux réservoirs documentaires, comme le protocole pour la collecte de métadonnées de l'Initiative pour les Archives ouvertes (OAI-PMH), qui est un des vecteurs d'interopérabilité les plus utilisés dans le monde de l'enseignement supérieur et de la recherche, ou encore la possibilité de faire un pas en avant vers des méthodologies du web des données telles que la conversion automatique des métadonnées au format *Resource Description Framework* (RDF), la mise en place d'une base de données RDF de façon automatique (ce qu'on appelle un *triple store*). Il suffit de déposer et documenter les données pour que cette panoplie derrière se mette à disposition. Il est également possible de s'identifier dans la plate-forme Nakala pour pouvoir donner l'accès à des collègues qui sont dans le monde de l'enseignement supérieur et de la recherche, au travers de la fédération d'identités. Évidemment, la liaison est naturelle avec nos autres services comme le portail Isidore.

Pour résumer, le service *Nakala* est un simple conteneur d'informations qui permet de documenter les données et de les rendre interopérables avec un grand nombre de projets que vous connaissez sans doute, allant de Européana à Gallica en passant par les plate-formes d'édition électronique. Notre but est d'outiller des communautés qui n'ont pas forcément forgé un dispositif métier ou disciplinaire dans leur réseau de laboratoires.

Parmi les gens qui font déjà usage de *Nakala*, nous comptons le centre d'études franco-égyptiennes de Karnak qui a déposé l'intégralité de sa photothèque numérisée. Les données déposées dans Nakala sont ensuite

rééditorialisées avec leur propre portail de communication web. Une fois qu'on a un dispositif tel que je l'ai décrit rapidement, on peut presque réaliser un flux de travail d'archivage à long terme des données, ce qui est l'un de nos derniers services de stockage et d'archivage. Sur ce point, nous ne travaillons pas en direct avec les communautés, mais la plupart du temps avec des structures comme ORTOLANG, des equipex, des plate-formes d'édition électronique, ou des plate-formes de données qui existent déjà (ArkéoGrid et l'Institut de recherche et d'histoire des textes (IRHT) pour les données sur le texte) qui font office de *hubs* vers lesquels convergent des jeux de données. Au travers de notre dispositif, réalisé en partenariat avec le Centre Informatique National de l'Enseignement Supérieur (CINES) et les archives nationales, les données peuvent ensuite être archivées de façon intermédiaire pour une période allant de quinze à trente ans, puis à terme potentiellement migrer du CINES aux archives nationales (car toutes les données n'ont pas vocation à un archivage définitif aux archives nationales). On a donc un cycle complet, qui va du simple stockage jusqu'à l'archivage définitif avec les archives nationales dans le cadre de la loi sur les archives, avec un passage au travers de responsabilités documentaires, scientifiques au CINES et ensuite aux archives nationales.

Je terminerai la boucle par le signalement des données, car une fois qu'elles sont stockées, traitées et documentées, il faut pouvoir les diffuser. Cela passe chez nous par la plate-forme Isidore, le service le plus ancien que nous ayons à Huma-Num puisqu'il est en place depuis fin 2010. Ce dispositif s'appelle Isidore. Il prend la forme d'un site web qui permet par moissonnage et par enrichissement de signaler des métadonnées et des données classées et enrichies par croisement avec des référentiels scientifiques. Tout cela se fait dans le respect des normes et des recommandations du *World Wide Web Consortium* (W3C), comme la mise à disposition d'un triple store RDF interrogeable à l'aide du langage SPARQL permettant de démultiplier les usages dans les plate-formes qui se connectent à ces jeux de données. Établie depuis 2010-2011, nous venons de faire une mise à jour majeure avec le passage en multilingue. Dorénavant, lorsqu'une métadonnée ou une donnée entre dans la plate-forme Isidore, elle est enrichie sémantiquement sur trois langues : anglais, espagnol et français. Nous sommes en train de transformer notre moteur de recherche et de signalement de données en un outil de découverte des publications et corpus scientifiques en français pour des non-francophones. Le document reste en français et

les enrichissements sont en anglais et en espagnol, le tout aligné (puisque nous nous plaçons dans le cadre du web des données ou web sémantique).

Comme je l'ai dit tout à l'heure, on utilise tout le potentiel des outils du web sémantique. Les données et les informations publiées dans le web sémantique pourraient être réutilisées par exemple en déployant une version mobile de la plate-forme Isidore, qui agrégerait des données venant de la BNF, de DBpedia, de *Virtual International Authority File* (VIAF), c'est-à-dire un grand nombre de référentiels scientifiques qui sont eux aussi publiés dans le web sémantique, à la disposition des utilisateurs, mais qui sont souvent quelque peu méconnus, car il faut connaître un certain nombre de langages, en particulier SPARQL, pour pouvoir opérer des requêtes. Nous essayons en ce moment de les intégrer dans des interaction homme-machine (IHM), dans des interfaces pour que des chercheurs, des étudiants, des doctorants puissent avoir accès à ce signalement d'informations.

Je termine en disant que nous avons une petite partie recherche et développement (R&D) avec un projet que nous allons publier dans les semaines qui viennent, intitulé *Isidore Motor Constructor*. Il s'agit d'un environnement de développement autour de la plate-forme Isidore, qui permettra à des communautés scientifiques en archéologie ou en histoire par exemple, d'utiliser une partie des 3,5 millions de ressources que la plate-forme contient sur la thématique de recherche, et de construire un moteur de recherche relatif à une spécialité.

Pour les chiffres, nous couvrons un peu plus de vingt-deux disciplines avec 190 projets hébergés, 3,5 millions de ressources dans la plate-forme Isidore et un peu plus de 220 To de stockage à la fois dans Nakala et dans les autres outils que nous utilisons. Huma-Num est un réseau d'acteurs pour les données dans les communautés scientifiques et par les communautés scientifiques.

Que fait et peut faire la communauté scientifique ?

Table ronde animée par ***Jean-Marie Pierrel***
(Analyse et traitement informatique de la langue française (ATILF) – Université de Lorraine et CNRS), avec :

Gilles Adda, Institut des technologies multilingues et multimédia de l'information (IMMI) – Centre national de la recherche scientifique (CNRS)
Delphine Bernhard, Laboratoire linguistique, langues, parole (LiLPa) – Université de Strasbourg
Olivier Baude, délégation générale à la langue française et aux langues de France (DGLFLF) et laboratoire Ligérien de Linguistique (LLL) – Université d'Orléans
Eric de la Clergerie, Inria, équipe Alpage
Pascal Vaillant, Université Paris XIII

Jean-Marie Pierrel
Que fait et peut faire la communauté scientifique pour les langues de France ?

Au vu de cette thématique, il m'a semblé intéressant de rassembler autour de cette table ronde des personnes qui puissent intervenir sur différents plans : d'une part des scientifiques, qui travaillent sur des langues de France, comme Delphine ici présente, ou Pascal même si actuellement tu travailles plus sur les langues kanakes, ensuite des gens qui développent des outils de traitement automatique, je pense en particulier à Eric, Gilles pour les aspects oraux, et éventuellement moi-même, et enfin des gens qui travaillent plus sur l'organisation.

La présentation de ce matin a été faite un petit peu « à rebrousse-poil » : on aurait pu présenter les efforts européens avec DARIAH, puis ce qui se fait en France avec Huma-Num, et ce qui concerne spécifiquement la linguistique avec ORTOLANG, c'est-à-dire inverser les présentations. Cela aurait peut-être été plus évident.

La première question que je veux poser à cette table ronde est la suivante : quelles actions en cours faudrait-il conforter pour pouvoir développer des études et des recherches sur le traitement automatique des langues de France ? D'un côté, il y a ce que l'on pourrait faire dans la communauté scientifique, en particulier chez les linguistes, pour développer un peu plus les recherches sur les langues de France qui me paraissent sous-représentées. De grands organismes tels que le CNRS demandent de définir l'ère linguistique à laquelle on s'intéresse. Or, lorsque je dirigeais un laboratoire de linguistique française, on m'a dit que le français n'était pas une ère linguistique. Vous imaginez tout de suite que c'est encore pire pour les langues régionales.

J'aimerais poser la question à Delphine et Pascal : quel est, de votre point de vue, les actions qu'il faudrait mener pour développer des recherches sur les langues de France dans la recherche universitaire ?

Pascal Vaillant

Je crois qu'il est important pour la recherche d'avoir des équipes stables, sur une thématique donnée. J'ai retenu une réflexion à la fin de la présentation de Delphine et de Marianne hier. Marianne concluait en disant qu'on a souvent, sur certains projets, des titulaires qui n'ont pas tout leur temps à consacrer au travail sur leurs ressources, et à côté de ça, des gens qui sont payés à temps plein pour travailler sur les ressources mais qui sont sur des postes temporaires et ne vont pas rester. Je crois que l'une des plaies du développement de certains projets, c'est qu'ils se font par à coups. Pour qu'il y ait une continuité, il faudrait qu'il y ait une politique de création de compétences. En principe, la façon dont les universités sont structurées se prête bien à la stabilité. On a en revanche très peu de postes de personnel technique. Or, souvent, l'ossature est ce qui fait vivre ce genre de projets, quand on cherche à développer des ressources, accumuler des compétences et les mettre à disposition d'autres chercheurs et de la société autour, ce sont les ingénieurs d'étude et les ingénieurs de recherche, et c'est ce qui est devenu une denrée assez rare.

Dans les organismes de recherche, que je connais moins bien étant universitaire, la politique de recherche est nationale et les chercheurs peuvent évoluer au gré de leur carrière, changer d'endroit et de thématique. Nous ne sommes donc pas certains non plus, tant que l'équipe n'a pas atteint une certaine masse critique, qu'il peut y avoir un centre de compétences qui reste stable, fournissant des

ressources et permettant un développement sur le long terme dans telle ou telle langue. Dans le cas de l'université des Antilles et de la Guyane, au sein de laquelle j'ai travaillé il y a quelques années, le développement de ressources informatiques pour la langue créole reposait sur une personne. Ce n'est pas simple, car si la personne pour des raisons diverses se retrouve ailleurs, tout est à reprendre de zéro. Pour résumer, il faut des postes. Il y a une nécessité de stabilité avec un investissement en moyens humains sur le long terme passant par une équipe de masse critique minimale qui puisse avoir une mémoire.

Jean-Marie Pierrel

J'entends bien, mais je pense que nous ne pourrons pas développer les compétences sur la constitution de ressources dans les différentes langues régionales de France dans chaque équipe universitaire. Avant de passer la parole à Delphine pour qu'elle nous donne son point de vue, Gilles, tu pourrais peut-être nous en dire un mot, car tu travailles actuellement pour essayer de trouver comment faire émerger une structure prenant en charge la mutualisation de constitution de ressources, pour le français mais éventuellement pour les langues régionales de France. On s'aperçoit là qu'il y a quelques difficultés pour avoir des réponses à la hauteur des besoins.

Gilles Adda

Je crois que cette difficulté a été visible tout au long des exposés. D'abord, le mot « infrastructure » est ambigu : il veut souvent dire « réseau », sans impliquer la pérennité. Tout le monde est d'accord pour dire qu'il faut pérenniser et conserver les savoir-faire, et c'est surtout ça qui coûte de l'argent. Lorsqu'on forme des gens, ils acquièrent des compétences, des compétences juridiques par exemple, qui reviennent systématiquement en ce qui concerne les corpus afin de traiter les données en prenant en compte les problèmes d'éthique, de protection des données ou de *copyright*. C'est un problème politique que d'arriver à monter une vraie structure permettant de conserver ce savoir et de le pérenniser au-delà des quelques années de vie qu'ont normalement les infrastructures de recherche, les projets et les réseaux. Je pense que nous, scientifiques, nous pouvons essayer de pousser le plus possible, mais nous n'avons pas les moyens. C'est vraiment un message que nous avons entendu à de multiples reprises. La mutualisation est évidente, c'est un véritable besoin, tout comme la pérennisation, et j'espère que ce besoin de conservation est un des messages qu'on pourra mettre en exergue à la suite de ce colloque.

Jean-Marie Pierrel
Autrement dit, serait-on capables, dans un pays tel que la France, de mettre sur pieds un centre technique sur les ressources sur les langues régionales de France ? Il y a des centres techniques qui existent sur la métallurgie, dans beaucoup de domaines de l'industrie, mais à ma connaissance il n'y a pas de centre technique traitant des ressources linguistiques. Il y a des structures, des infrastructures, c'est ce qu'on fait dans ORTOLANG, qui sont à même de gérer, valoriser, diffuser des ressources, mais au niveau de leur création même il y a un vrai problème. Et lorsqu'on s'attaque à des langues dont la communauté est plus petite, je pense à l'alsacien par exemple, cela ne doit pas être évident de trouver des moyens pour constituer ces ressources.

Delphine Bernhard
Oui effectivement. J'aimerais aussi revenir sur quelque chose qui a été dit par Sébastien Quenot de la Collectivité territoriale de Corse, qui observait qu'il y avait d'un côté les informaticiens qui ont les compétences techniques, et de l'autre côté les personnes qui ont les compétences linguistiques en corse. Je pense qu'il faudrait aussi travailler sur les échanges entre les communautés, qui ne se connaissent pas forcément, et pas uniquement à l'intérieur de la communauté scientifique, en pensant aussi aux offices locaux qui ont fait des présentations hier. Il y en a un aussi à Strasbourg. Ils ont des besoins et essaient souvent de faire des choses mais ne savent pas très bien vers qui se tourner.

Alors, je ne sais pas très bien dans quel sens cela doit aller. Peut-être que cela doit venir de la communauté scientifique, des gens qui savent faire des ressources et des outils, qui doivent aller voir ces gens-là et leur proposer leur aide pour travailler ensemble. Je pense qu'il s'agit à la fois d'actions d'échange, mais aussi d'actions de formation, car on aura toujours besoin de gens qui ont des compétences dans les langues régionales concernées, qui ont peut-être plutôt des formations en linguistique, en sociolinguistique, et qui auront tout de même besoin d'un minimum de compétences « techniques », sans aller trop loin, pour pouvoir collaborer à ces projets. Échange et formation seraient donc pour moi deux choses importantes à mettre en place.

Jean-Marie Pierrel
Le côté interdisciplinaire est important. Si je reprends encore, excusez-moi, l'exemple d'ORTOLANG, je pense que c'est un projet qui n'aurait jamais vu le jour s'il n'y avait pas eu des informaticiens impliqués et des linguistes. Je rappelle qu'Olivier ici présent, qui est linguiste, fait partie d'ORTOLANG, et que pour ma part je suis professeur d'informatique depuis une éternité, mais c'est vrai qu'il y a un certain nombre d'équipes, de lieux où cette connexion entre les informaticiens et les linguistes s'est faite, d'autres lieux où cela ne s'est pas fait.

Eric de la Clergerie
Pour reprendre un des points sur la pérennité, le traitement d'une langue doit s'envisager sur un terme relativement long, à cause du développement des ressources, des corpus, des études. Cela ne veut pas dire pour autant que l'on n'obtient pas de résultats assez vite. Nous sommes capables de faire des outils qui fonctionnent relativement bien à des échéances assez brèves. En revanche, l'obtention de niveaux de qualité très bons et excellents s'envisage sur du long terme. J'étais un peu inquiet hier, car il y a eu une remarque qui disait qu'il fallait d'abord faire la première brique parfaitement avant de passer à la suivante. Or je crains que ce ne soit pas toujours la meilleure solution à suivre. La pérennité est absolument nécessaire, avec ce côté dynamique et pas seulement en ce qui concerne la conservation. Tout doit évoluer : les outils, les ressources, même éventuellement les corpus.

En ce qui concerne ce qu'a dit Delphine, je pense que c'est vraiment important qu'il y ait des points de contact qui existent entre différentes communautés. Personnellement, en tant que *taliste*, je ne suis pas linguiste à proprement parler, je suis donc intéressé pour intervenir sur de nouvelles langues, mais il faut que l'on me sollicite. C'était donc intéressant, lors de ce colloque, de discuter d'autres langues, de voir qu'il y a des choses qui peuvent démarrer.

À propos des points de contact, je pense que c'est aussi important parce qu'on peut avoir le risque pour certaines langues (mais pas en France), de considérer sa langue comme unique, spécifique, nécessitant des choses tout à fait hors norme. Il faut aller discuter avec les autres pour se rendre compte que ce n'est peut-être pas aussi vrai que ça, qu'il y a vraiment des échanges à faire, des idées à rependre. Ces points de discussion et d'échanges sont donc réellement nécessaires.

Olivier Baude
J'aimerais ajouter quelques remarques dans la continuité de ce qui est dit depuis le début.

La première est liée à l'originalité de ce colloque. Les langues, que ce soit les langues de France ou régionales de France, peu importe l'appellation que je laisse à l'appréciation de la DGLFLF, ne sont pas un objet scientifique, en soi. Il n'y a pas de périmètre scientifique des langues régionales de France, voire des technologies pour les langues régionales de France. En revanche, il s'agit d'un véritable objet pour une politique linguistique et éventuellement pour une politique de recherche, comme on l'a vu très clairement hier. Cela me paraît important, car cela signifie qu'il faut qu'il y ait un lieu de contact non seulement entre des gens qui travaillent sur différentes langues mais aussi sur différents acteurs.

Les demandes fortes entendues hier en termes de technologies des langues de la part des collectivités constituent aussi un enjeu important. C'est pour ça que face à cela, une réponse de moyens, langue par langue, n'est pas suffisante et même pas acceptable. On l'a vu, il ne faut surtout pas réagir brique par brique, l'enjeu n'est pas de constituer des masses de corpus, de les annoter, puis de travailler dessus, de savoir ce qu'ils vont devenir et ce que les gens vont en faire. Il faut qu'il y ait un dialogue complet dès le début, dans une approche à la fois technologique, scientifique et politique.

Des éléments de réponse à ce sujet sont en train de se mettre en place, on l'a bien vu ce matin. On voit que sur certains aspects, la mutualisation de ressources est possible, et il n'y a pas de raison actuellement d'aller reconstituer x centres de ressources sur ces aspects de mutualisation. Par contre, il y a le besoin d'avoir des briques complémentaires, au niveau européen, au niveau national et au niveau régional, et là nous serons peut-être plus en contact avec certaines langues pour mutualiser des ressources et des outils, des technologies, mais aussi, de perspectives politiques et d'exploitation de ces ressources. Cela permettra à chacun d'avoir sa responsabilité dans le reste du travail qu'il y a à faire, qui est encore lourd.

Jean-Marie Pierrel
En allant au-delà de la simple mutualisation, ne croyez-vous pas qu'il faudrait aussi faire un effort pour mieux mutualiser nos modèles descriptifs de ressources ? Par exemple, dans le cas de ressources lexicales et syntaxiques, je suis désolé, mais je ne pense pas qu'il y ait une différence telle entre l'occitan et le français que deux modèles lexicaux différents soient nécessaires. L'instanciation est différente bien entendu, mais je ne suis pas sûr en ce qui concerne les modèles.

Il en va de même pour les modèles syntaxiques et pour les outils. Cela me fait rebondir sur la question que j'avais placée en troisième point, mais que l'on peut aborder dès maintenant : comment peut-on faire pour mieux voir ce qui a été réalisé sur le français ? L'étude du français et de son traitement automatique est une force de frappe disponible pour les modèles de traitement et de ressources qui peuvent s'appliquer à d'autres langues de France. D'après ce que j'ai entendu pendant ces deux jours, il m'a semblé que les occitans ont envie de développer des choses pour l'occitan, les Corses pour le corse, les Bretons pour le breton, et les francisants pour les francisants : je pense qu'ainsi on fait un peu fausse route. Peut-être que la DGLFLF devrait jouer un rôle institutionnel plus fort, pour provoquer des rencontres plus régulières de ces communautés.

Il est vrai que si vous travaillez sur le traitement automatique du français, il n'y a pas de raison que vous rencontriez des gens qui s'intéressent aux autres langues de France. Vous allez vous confronter à la concurrence internationale, vous allez être poussé à interagir avec les équipes qui travaillent sur l'anglais. Parallèlement à ça, il y a un gros trou en France dans une sorte de maillage existant entre nos langues sur le territoire national et les compétences existantes sur le traitement des langues.

Gilles Adda
Même travailler sur le français est le résultat d'une volonté politique. Quand on a commencé à travailler, on a « bien sûr » travaillé sur l'anglais. Au début des années 90, lorsqu'on évoquait l'importance de travailler sur la langue, sur la traduction, sur la reconnaissance, c'était sur l'anglais, ce qui nous a conduits naturellement à travailler dessus. Il y a eu une volonté politique et scientifique disant qu'une fois que l'on savait faire des choses sur l'anglais, il fallait aussi travailler sur le français.

Jean-Marie Pierrel
Oui. Mais excuse-moi, je ne suis pas complètement d'accord avec toi, car tu as une vision qui est peut-être celle d'un jeune, qui est arrivé dans le monde de la recherche au moment où se sont développées les méthodes statistiques, et où il fallait des gros corpus et les gros corpus étaient sur l'anglais, si vous le permettez à une vieille barbe comme moi ou Joseph qui pourrait sans doute en dire plus. Dans les années 70, quand on se réunissait entre acousticiens, linguistes et informaticiens, nous travaillions sur le français, avec des méthodes essentiellement symboliques. La communauté était forte à ce moment-là. Je l'ai vécu, et je l'ai regretté à certains moments. Les modèles statistiques ont montré qu'ils étaient plus performants pour obtenir des résultats, et comme les données en grand nombre existaient pour l'anglais, ta génération a travaillé essentiellement sur l'anglais avant de revenir sur le français. C'est peut-être là que l'on a fait une erreur stratégique.

Gilles Adda
Oui, je voulais aussi ajouter qu'on ne travaille pas que sur l'anglais et le français, mais sur des dizaines de langues. Aujourd'hui nous sommes beaucoup plus poussés pour travailler sur le pashto, sur le tamil, sur d'autres langues que le breton. C'est vrai que l'on a fait des systèmes de transcription en estonien, en France, dans notre laboratoire, alors que les besoins de systèmes de transcription des langues régionales, la gestion de la variation de l'occitan comme on peut le faire pour d'autres langues ne sont pas du tout traités.

Jean-Marie Pierrel
Mais tu dis que nous sommes poussés : par qui ?

Gilles Adda
D'abord par la disponibilité de corpus, puisque c'est effectivement très cher d'en constituer, ou de trouver des corpus développés par différentes agences. Par exemple et entre autres, l'agence *Defense Advanced Research Projects Agency* (DARPA) met à disposition des corpus dans certaines langues pour lesquelles les Américains jugent intéressant de développer des systèmes. La communauté travaille sur ces corpus. Cela encourage la recherche sur des sujets que les Américains pensent importants, ce sont eux qui montrent l'endroit où regarder, où travailler, parce que ce sont eux qui fournissent les corpus. C'est fondamental.

Gilbert Mercadier (depuis la salle)
Excusez mon enthousiasme, mais je pense que c'est bien que la salle puisse participer. Les interventions qu'on a vues depuis deux jours sont très riches, mais on a vu que les gens se connaissaient peu ou du moins pas assez : merci à la DGLFLF de nous avoir réunis. Il est évident qu'il faudrait sortir d'ici avec quelques idées sur la mutualisation, comme vous êtes en train de le dire. Nous n'avons pas, du moins au Congrès, mais je pense que je ne serai pas le seul à le dire, d'opposition à ce que les modèles de traitement utilisés pour le français soient appliqués aux langues de France, bien au contraire. Si la DGLFLF peut organiser et faciliter cette mutualisation, ce sera une excellente chose.

En ce qui concerne les corpus, vous nous expliquez que vous travaillez avec ce que vous proposent les Américains. Si je comprends bien, en plus du Coca-Cola, ils nous inondent de leurs idées et de leur corpus qui peuvent être traités. C'est bien. Mais je pense qu'on doit pouvoir vous donner pour l'occitan des corpus aussi importants que pour l'estonien, le tamoul ou le guarani. On ne peut donc que vous encourager à aborder enfin et avec zèle le grand champ des langues régionales, qui comme vous avez pu le comprendre est en friche, même s'il y a quelques défricheurs qui ont encore des moyens qui, s'ils ne sont du Moyen-Âge, sont au moins à améliorer.

Eric de la Clergerie
Je crois que l'on sous-estime le temps nécessaire pour fabriquer les corpus dont on parle, qui sont des corpus annotés. Ce n'est pas juste faire des collections de textes, de documents, cela prend beaucoup plus de temps. Mais c'est vrai que c'est un souci de se dire que quelque part, notre recherche française est pilotée par des instances américaines à travers leurs propres préoccupations. Plus que la disponibilité des données, c'est la question de la disponibilité des financements, sur un certain nombre de choses, qui ne sont peut-être pas sur les langues régionales, et c'est aussi la question des publications scientifiques. Cela peut être relativement difficile de publier sur les langues régionales de France par rapport à l'anglais, l'allemand, le chinois, l'arabe, etc. Ce sont des facteurs à prendre en compte pour arriver à inciter une partie de la recherche française à s'intéresser aux langues régionales de France parce que ce n'est pas forcément gagné de ce point de vue là.

Jean-Marie Pierrel
Mais il faudrait aller au-delà de ça. Je trouve qu'il y a dans nos agences de recherche, je pense à l'Agence nationale de recherche (ANR), un culte du résultat qui est important sans prise en compte de l'outillage qu'il faut mettre en place au départ pour obtenir ces résultats. J'ai été confronté à cela très concrètement sur des projets de constitution de ressources de base pour notre langue, le français. On m'a expliqué qu'il n'y avait aucune chance d'avoir des soutiens au niveau d'une agence comme l'ANR pour la constitution de telles ressources. J'en ai pris acte.

Deuxième point, je me suis dit, « dans ce cas-là, mutualisons ». Avant que sortent les projets, les appels d'offre Programme d'Investissement Avenir, j'avais pris des contacts pour voir dans quelle mesure un projet de constitution d'une infrastructure de mutualisation de ressources sur la langue aurait des chances d'être soutenu dans le cadre de l'ANR. On m'a dit que c'était impossible. Il y a peut-être une réflexion à avoir, par exemple de la part de la DGLFLF et vis-à-vis des organismes gouvernementaux pour que la transmission soit faite aux spécifications du côté de l'ANR, car si cela continue ainsi il y aura énormément de difficultés concernant les langues régionales.

Je sais qu'il y a eu quelques projets, RESTAURE en est un, qui sont passés récemment. Mais que d'échecs, pour si peu de réussite ! Essentiellement parce que dans des comités ANR (j'ai moi même eu à en subir les critères), nous sommes obligés de dire « oui, c'est un projet très intéressant, mais au bout du compte, en termes de résultats et de recherche scientifique fondamentale, qu'est-ce qu'il y a ? »

Martine Garnier, Agence nationale de la recherche (depuis la salle)
Je pense que c'est une réflexion de fond posée de façon récurrente à l'Agence nationale de recherche et dans les agences de financement globalement. Pour répondre, il y a quand même eu un côté Sciences Humaines et sociales avec un programme qui s'appelait CORPUS, et qui avait justement le mérite d'être complètement fléché et dédié corpus.

Jean-Marie Pierrel
Plus de la moitié des corpus produits ont été perdus.

Martine Garnier
Oui, mais c'est un autre sujet. Disons que la production de corpus a été soutenue.

Jean-Marie Pierrel
Il n'y avait pas d'infrastructure de mutualisation pour les mettre en place. J'ose espérer que maintenant ce ne serait plus le cas. D'autre part, il n'y avait pas de demande de l'agence d'assurer cette pérennisation, et ça c'est aussi un des points difficiles.

Martine Garnier
Honnêtement et sans vouloir être désagréable, je pense que c'est souvent ce qui est mis en avant pour expliquer que si l'on ne fait rien, c'est parce que rien n'est mis à disposition. Il faut un petit peu modérer.

Je pense que ce programme CORPUS a montré qu'il y avait ce besoin. Il est vrai que dans le cadre de l'appel à projets générique, la programmation a été complètement redéfinie, et qu'il y a une vraie difficulté à soutenir des projets qui sont soit sur de la production, de la constitution de corpus d'une part, et au sens plus large, d'outils et ressources pour la recherche, car je pense que les outils et ressources pour la recherche sont aussi une vraie question qui se pose.

C'est effectivement une question de moyens. Lorsque le budget est réduit de façon importante, il y a une sorte de crispation dans les comités d'évaluation comme tu l'as dit, de la part de la communauté (je rappelle que ce n'est pas l'ANR qui sélectionne les projets, mais la communauté). Ce sont donc des projets qui, la plupart du temps, ne sont pas mal notés, bien au contraire, qui ont un poids fort dans le cadre du critère d'impact, mais avec une sélection maintenant à 10% des projets, je dirais que l'on se trouve juste en dessous de ceux qui sont pris.

C'est vrai, c'est une réalité, et nous essayons de voir comment y remédier. Tu disais qu'il y a quand même des projets qui sont soutenus. C'est le cas lorsqu'ils intègrent dans leur annexe technique des verrous particuliers et pas uniquement de la production de corpus. Nous verrons ce qui ressortira de la stratégie nationale de la recherche. Beaucoup de choses se disent et se font autour des données, mais je ne suis pas sûre que les langues

soient particulièrement en tête de pont aujourd'hui dans le cadre de ces sujets, malheureusement. Quoiqu'il en soit, plus il y aura de dépôts de projets sur ces sujets-là, plus leur importance sera visible. Pour pouvoir soutenir une thématique, il faut qu'elle soit présente.

Jean-Marie Pierrel
Il faut peut-être aussi que la communauté s'intéressant aux outils et ressources pour la recherche soient plus présents dans certains comités de l'ANR. Pour CORPUS c'était le cas, mais en dehors de cela, sur des thématiques plus génériques comme actuellement, peu de gens en sont porteurs.

Martine Garnier
Il y a une volonté importante d'essayer de rapprocher les communautés Sciences Humaines et sociales (SHS) et Sciences des technologies de l'information et de la communication (STIC), notamment dans le cadre de l'appel à projets générique, d'avoir dans la programmation un effet miroir. Il y a notamment un axe « numérique au service du patrimoine » qui fait vraiment écho à un axe culture et patrimoine côté Défi 8, donc SHS, où, si le terme « langues régionales » n'apparaît pas, la langue en tant qu'élément du patrimoine est bien présente, et heureusement.

Édouard Geoffrois, Agence nationale de la recherche (ANR) (depuis la salle)
Merci de me permettre de rebondir aussi. Ingénieur de la direction générale de l'armement (DGA), ma carrière est depuis récemment affectée à l'ANR. Je voudrais préciser qu'il faut bien distinguer deux types de projets : les projets de recherche avec une composante production de ressources, et les projets de création de ressources et d'infrastructures. Je rappelle que l'ANR applique la stratégie du ministère de la Recherche, il est donc important aussi de véhiculer certains messages au niveau du ministère, et pas de l'ANR.

L'ANR est là pour financer la recherche et pas les infrastructures de recherche. On en pense ce que l'on en veut, mais le fait est là. Un projet de recherche qui inclut une composante de ressources nécessaires à la recherche a donc toutes ses chances, même si cela s'optimise. Votre échange montre qu'il faut peut-être plus de personnes sensibles à ces questions de ressources dans les comités pour améliorer encore les choses. C'est encore faisable.

Le soutien aux infrastructures pures ne relève pas des mécanismes de l'ANR tels qu'ils sont construits, et on ne peut pas le lui reprocher. C'est une question politique de provenance des financements.

Martine Garnier
J'ajoute à ce que dit Édouard qu'il s'agit du ministère de la Recherche et ses alliances, car les alliances sont maintenant particulièrement impliquées dans la programmation.

Edouard Geoffrois
Tout à fait. Il faut approcher le ministère de la Recherche et les alliances, pour soulever ces questions-là. Je souhaite préciser également que le financement de l'ANR passe par des mécanismes de financement qui sont fondamentalement ascendants, contrairement à ceux de la défense, puisque le cas a été cité pour les États-Unis. Cependant, il faut être bien conscient que la politique industrielle aux États-Unis se fait essentiellement par les crédits de défense, alors qu'en Europe et en France en particulier, même si la DGA peut jouer un certain rôle, l'essentiel des crédits passe par le civil, qui utilise des instruments différents.

La conclusion essentielle à mes yeux, après avoir essayé d'inciter au montage de projets du type que vous décrivez pendant des années, c'est qu'il manque un organisme qui se dote de la mission de produire, et qui accepte d'être emmené avec d'autres dans des projets de recherche. La création d'une telle entité passe par une décision si ce n'est politique, une décision d'organisme, et l'ANR ne peut pas en être porteuse, même si des gens comme Martine et moi sommes à titre individuel pleinement en phase avec cela. C'est un périmètre plus large qu'il faut envisager.

Jean-Marie Pierrel
Je suis tout à fait d'accord, mais vous savez aussi bien que moi que les organismes de recherche sont eux-mêmes frileux. J'espère donc que les gens de la DGLFLF écoutent bien ce qui se dit, car il me semble que c'est bien le vecteur de la DGLFLF qui pourrait porter ce message disant qu'il est indispensable de disposer d'un endroit, de moyens, d'une structure pour aider à la production de ressources sur le français et les langues de France, sans quoi on risque de ne pas pouvoir les outiller.

Edouard Geoffrois
Exactement. Pour abonder, c'est effectivement pour simplifier au niveau politique de pousser à la création de telles structures. Il y a la DGLFLF pour le français et les langues de France, et d'autres entités de niveau politique aussi, que ce soit le ministère de la Recherche pour les questions d'infrastructures recherche, le ministère de l'Industrie aussi éventuellement, en impliquant des industriels.

Tu citais Jean-Marie le modèle des centres techniques industriels : il faut voir à quoi cela correspond. Ce sont des organismes qui sont essentiellement financés par les industries du domaine, et nous ne sommes pas exactement dans ce modèle-là, pour l'instant. Il faudrait avoir une cotisation industrielle, mais dans les domaines qui nous intéressent, l'évolution est tellement rapide que c'est difficile d'avoir une base qui fonctionne comme cela. Le modèle général est bon, mais l'outil centre technique et industriel ne convient pas exactement. C'est un autre outil qu'il faut créer, en concertation avec tous les acteurs politiques concernés.

Khalid Choukri (depuis la salle)
J'ai l'impression, sans remonter dans les années 80, que c'est un débat que nous sommes en train de revivre une énième fois. On est en train d'imaginer que les chercheurs vont du jour au lendemain devoir agir en tant que prestataires de service, produire des ressources qui pour eux sont des commodités. Ce n'est pas leur objectif. C'est comme s'ils voulaient conduire une Ferrari mais qu'on leur donnait une Renault en leur disant qu'en plus il leur faudra raffiner le pétrole pour pouvoir rouler avec.

C'est un débat qui a déjà eu lieu, le rôle du chercheur est d'une certaine manière bien défini, le rôle des institutions, aussi bien de financement que de politiques linguistiques, devrait être défini. Les agences de service, y compris au sein du CNRS, devraient avoir aussi un rôle à jouer. Mais il y a d'autres alternatives. Je viens juste de regarder le catalogue d'ELRA, qui compte un certain nombre de ressources en basque, en catalan, et d'autres langues régionales, sauf qu'elles proviennent de Catalogne et du pays basque du sud et pas du nord. La raison pour laquelle ces ressources sont aujourd'hui encore disponibles et facilement trouvables est que les gouvernements régionaux de Catalogne et du pays basque, et fut un temps la commission européenne, avaient posé les mêmes

règles : le programme de recherche prévoyait une clause pour l'archivage des ressources au sein d'une agence qu'était ELRA à l'époque, pérenne et rendant les choses disponibles.

Je pense que si les relations entre le monde de la recherche, le monde de la politique et de la politique linguistique et les agences (y compris les agences privées), ne sont pas un peu redéfinies, nous aurons sans doute cette discussion dans dix ans, dans quinze ans, dans vingt ans. Je comprends que l'ANR en ait assez de remettre la main à la poche systématiquement, car les rares projets où les corpus ont été produits et mis à disposition sont ceux où des chercheurs considéraient que c'était un investissement important et qu'il fallait stocker les données à la fin du projet de recherche. Beaucoup de ressources ont été mises sur des sites internet en imaginant que ce serait suffisant. En quelques années les serveurs ont disparu et les données avec. Le monde de la recherche doit aussi faire, si ce n'est son *mea culpa*, au moins son auto-critique.

Jean-Marie Pierrel
Tout à fait. Ce faisant, les choses bougent un petit peu, avec la création de la TGIR Huma-Num ou la mise en place d'un equipex tel que celui que je dirige actuellement, fortement connecté à la TGIR.

Philippe Boula de Mareüil (depuis la salle)
Nous sommes dans un pays à l'histoire très centralisée, qui n'est pas très connu pour avoir soutenu les langues régionales, sinon nous serions peut-être ici dans un autre cadre. Je ne sais pas quel cadre institutionnel il faudrait donner. S'il y avait dans je ne sais quelle infrastructure réseau, commission de la DGLFLF, une volonté affirmée de développer les technologies et plus généralement les langues régionales, nous n'en serions pas là. C'est une volonté que nous pouvons appeler de nos vœux.

Cependant, tant que nous ne sommes pas dans ce cas de figure, c'est peut être trop se reposer sur un mouvement descendant. Des choses très importantes ont été dites. Dans l'état actuel, si informaticiens et linguistes n'arrivent pas à se mettre autour d'une table pour développer des systèmes, c'est qu'il y a peut-être non seulement un manque de communication, mais un manque de force vive, parce que la population corse et le nombre de chercheurs est limité. Or, on l'a vu, sur le continent aussi il y a des

volontés de développer des outils pour les langues régionales. De même, Delphine Bernhard parlait de l'alsacien, ou d'autres langues régionales : est-ce qu'avoir un système de synthèse de la parole suffit à l'Office de la langue bretonne ? Ou bien n'est-il pas possible de mutualiser les efforts, partager nos ressources ? C'est là le problème du fonctionnement du monde académique. Il faut faire remonter les bonnes volontés pour mutualiser nos efforts sans attendre qu'une décision politique nous soit favorable.

François Yvon, LIMSI-CNRS (depuis la salle)
Le débat s'est beaucoup focalisé sur les ressources, mais je ne voudrais pas qu'on ait l'impression que pour traiter l'occitan ou le corse il faille repartir à zéro, prendre un corpus, étiqueter les catégories grammaticales et développer de nouveau des banques d'arbres. Nous avons quand même quarante ans de recherche en traitement des langues, des ressources et des connaissances considérables sur un grand nombre de langues ont été acquises, dont certaines ne sont pas si éloignées de langues bien documentées. Nous avons des systèmes pour l'allemand, l'italien, le néerlandais. Ne peut-on pas mettre cela à profit pour outiller, équiper, transférer une partie de ces connaissances sur des langues qui sont très apparentées ? Ce que je veux dire, c'est que je pense qu'il y a aussi une recherche à faire sur les modèles, et sur le développement de modèles qui soient plus génériques, capables de se transporter d'une langue à une autre, d'abstraire les propriétés de langues qui sont proches.

Il y a aussi un travail méthodologique à faire sur les méthodes de traitement. La question n'est pas seulement celle de la quantité de ressources, il s'agit aussi de développer des modèles pour mutualiser le savoir qu'on connaît déjà, parce qu'il y a à notre disposition un maillage qui se fait au niveau des langues. On peut y recourir, pour avoir très rapidement je pense des systèmes qui marcheront très bien pour beaucoup de langues régionales.

Jean-Marie Pierrel
Il faut peut-être aussi que nous apprenions à mieux nous connaître. J'ai un exemple très simple vécu récemment et qui m'a fortement interpellé. Le projet ARBRES[1], concernant le breton, demandait de l'aide à Huma-Num pour accueillir leur site et prendre en charge l'infrastructure autour de ce projet. ORTOLANG étant considéré comme spécialiste pour les langues

1 http://arbres.iker.cnrs.fr/index.php?title=Arbres:Le_site_de_grammaire_du_breton

au sein d'Huma-Num, la question est renvoyée vers moi. Alors que je réponds positivement à la demande, on me répond ensuite que nous sommes trop marqués langue française, et qu'il n'est pas possible pour ARBRES d'accepter de figurer sur un site où le français est présent. Il faut avoir conscience que ce type de réaction nous fait perdre des moyens, de l'argent et de l'énergie.

Pascal Vaillant
Je voulais dire quelque chose depuis tout à l'heure mais François Yvon en a dit une partie. Concernant le degré de mutualisation possible, nous avons évoqué la mutualisation en termes d'infrastructure, mais il y a aussi la mutualisation en termes de ressources linguistiques. Sans refaire le match méthodes statistiques versus méthodes symboliques, j'ai vu un exposé il y a quelque temps qui montrait qu'avec des méthodes statistiques, même sur des corpus non-annotés, il est possible d'attraper quelques fruits sur les branches basses.

En ce qui concerne les ressources, il y a effectivement des dialectes apparentés qui permettent de développer des ressources en partie communes. L'approche méta-grammaire développée par exemple par Marie Candito sur le français et l'italien marche assez bien sur des dialectes romans. Des gens comme Dominique Estival en Suisse l'ont appliquée ensuite à des familles de dialectes apparentés. On peut transposer des méthodes de l'allemand pour l'alsacien, je crois que Delphine a essayé de le faire à une époque. François l'a dit en partie.

En revanche il faut avoir conscience que quand on parle de langues de France, si l'on inclut les langues de l'Outre-Mer, on a des choses qui ne sont plus du tout transposables. Les créoles, pour reprendre l'exemple que je connais, sont apparentés au français historiquement, génétiquement et lexicalement, mais au niveau de la grammaire, y compris même à la subdivision traditionnelle en catégorie grammaticale, je doute que ce soit transposable. Quant aux langues kanaks et amérindiennes, elles ne fonctionnent plus sur les mêmes principes et on ne peut probablement pas transposer les choses.

Olivier Baude
En deux mots, pour faire un lien avec ce qui a été présenté ce matin et

l'ensemble des discussions que l'on a eues ici, il y a bien sûr un problème de moyens, un problème politique important. Je rappelle que les TGIR en sciences humaines et sociales représentent 1 % des TGIR en France et fonctionnent avec deux millions d'euros de budget par an. Or si je peux me permettre, la DGLFLF fonctionne avec un budget plus petit que ça. Ce n'est pas une prise en charge politique lourde, mais ce n'est pas une simple demande de moyens considérant que la DGLFLF ou les TGIR apporteront une solution en disant « Les technologies pour les langues de France c'est maintenant. ». C'est clairement plus compliqué.

Par contre, pour être moins pessimiste, nous avons vu ce matin qu'il y a des outils, que la situation a changé, comme le disait Khalid tout à l'heure. Maintenant, il n'y a plus aucun problème de corpus qui seraient perdus sur un site internet. Il y a à la fois une TGIR et un équipex sur les ressources linguistiques, c'est donc réglé, mais ce n'est pas la seule réponse qu'il y ait à apporter. Jean-Luc a montré l'intérêt qu'il pouvait y avoir à être au niveau européen sur un échange et une construction de réseau. C'est un outil à prendre en main.

Il y a également deux consortiums de linguistique dans le cadre de la TGIR, pouvant servir de lieu de discussion ou de groupe de travail sur les technologies sur les langues de France. Ce qu'il faut, c'est un lieu de discussion scientifique où l'on va un peu plus loin qu'un simple périmètre scientifique, où le ministère de la Culture et les collectivités territoriales sont présents.

Jean-Marie Pierrel

Nous allons devoir nous arrêter pour laisser du temps à l'autre table ronde. La communauté ici représentée, qui comprend à la fois des chercheurs mais aussi des acteurs associatifs qui peuvent apporter beaucoup sur les langues de France, n'est pas suffisamment créée, encore aujourd'hui, pour imaginer qu'elle va pouvoir se retrouver sans l'incitation ne serait-ce que de l'organisation de choses de ce type-là. Nous avons besoin de ces tables rondes. Je voudrais donc dire à la DGLFLF qu'il faut permettre des lieux de rencontre. Il n'y en a pas beaucoup aujourd'hui.

Traitement des langues régionales : que peuvent faire les acteurs publics ou privés en charge de l'accompagnement des langues régionales et les collectivités territoriales ?

Table ronde animée par ***Benaset Dazeàs***
(Lo Congrès permanent de la lenga occitana), avec :

Olier Ar Mogn, Office public de la langue bretonne / Ofis publik ar Brezhoneg
Nourdine Combo, Conseil général de Mayotte
Gaëtan Crespel, Dastum
David Grosclaude, Conseiller régional d'Aquitaine
Sébastien Quenot, Collectivité territoriale de Corse / Capiserviziu di u Cunsigliu linguisticu

Benaset Dazéas
Après la table ronde des scientifiques, nous avons maintenant une table ronde rassemblant plutôt des acteurs de terrain dans le sens de la transmission et du développement des langues régionales : des gens venant des collectivités, qui mènent des politiques publiques, et des acteurs de terrain qui œuvrent dans le domaine de la socialisation ou même de l'enseignement. J'ai entendu des choses extrêmement intéressantes ce matin, sur des questions que je me pose sur la mutualisation et le pilotage du développement de ces ressources. Je pense que nous aurons l'occasion d'y revenir.

Nous avons tous beaucoup entendu parler du livre blanc de Meta-Net[1], qui

1 http://www.meta-net.eu/whitepapers/press-release-fr

est assez alarmant quant à la situation des langues de France en termes de technologies du langage. Les éditeurs soulignent l'écart croissant entre les langues et le fait qu'il est désormais indispensable d'équiper les plus marginalisées de technologies de base. Nous savons qu'il est possible de créer des ressources de manière efficace, à condition d'avoir une coordination, une planification des travaux, la création massive de données, la mutualisation des efforts au niveau européen, le transfert technologique entre les langues, l'interopérabilité des ressources, ainsi que des outils et des services. Le développement des technologies de support est un besoin urgent et même vital pour les langues minorisées. C'est donc un enjeu majeur pour les politiques linguistiques.

À partir de là, comment les différents acteurs privés et publics de la transmission des langues de France peuvent-ils contribuer à élaborer une stratégie collective et à avoir des réalisations concrètes, avec quels outils, et bien sûr quels moyens ? Pour parler de cela, nous avons Gaëtan Crespel de Dastum, Olier Ar Mogn de l'Office Public de la Langue Bretonne, David Grosclaude, conseiller régional d'Aquitaine et administrateur de l'Office Public de la Langue Basque et instigateur du futur Office Public de la Langue Occitane, Sébastien Quenot de la collectivité territoriale corse, et Nourdine Combo du conseil général de Mayotte. Nourdine Combo, est-ce que vous pouvez nous dire quelle est la situation des technologies, des ressources numériques des langues de Mayotte ?

Nourdine Combo

Les deux langues régionales de Mayotte sont le shi-mahorais et le shibushi. L'une est une langue bantoue et l'autre est une langue austronésienne, variante du malgache. La numérisation, l'écriture et l'apprentissage de ces deux langues sont vraiment nouveaux, depuis les années 2000. Cela tombe bien, car étant à l'ère du numérique, le Conseil général et les associations locales cherchent à promouvoir ces langues à travers le traitement automatique. L'objectif actuellement est de recenser les besoins et de voir ce qui existe pour pouvoir satisfaire les demandes. Pour l'instant, nous n'avons pas de technologies particulières, mais nous espérons en avoir dans les années à venir, ou si nécessaire, collaborer avec des centres de recherche pour en créer.

Benaset Dazéas

Une question pour Olier Ar Mogn : hier, lorsque vous avez expliqué la

situation de la langue bretonne vous avez présenté l'Office Public de la Langue Bretonne, qui est une institution autour de laquelle gravitent des associations développant entre autres des technologies, des logiciels, mais pas seulement. Le paysage est donc très marqué par une institution, qui est l'office, et puis ces petites associations. Comment peut-on structurer, fédérer, pérenniser ces structures qui malgré tout restent fragiles, créer ce qu'il manque, et comment les collectivités territoriales peuvent-elles appuyer ce développement-là ?

Olier Ar Mogn
Effectivement, il y a un mot qui est revenu souvent ce matin, c'est celui de pérennité. C'est un souci que l'on a, car les initiatives personnelles ou associatives peuvent être très intéressantes mais sont souvent fragiles. En ce sens-là, l'Office public de la langue bretonne joue déjà un rôle pour épauler tout ce travail. Mais clairement, pour dire les choses de façon très simple, notre objectif est de disposer le plus rapidement possible d'un service « nouvelles technologies ». Ce service ne fera pas le travail de recherche qui a été évoqué lors de la table ronde précédente, mais il pourra servir d'interface entre les besoins du terrain et les personnes capables d'apporter des solutions techniques, d'assurer le suivi et la mise à jour de la localisation des logiciels, qui est une question importante. Il faut qu'il y ait des structures pérennes capables d'assurer cette pérennisation.

Par ailleurs, cela dévie un peu de la question, mais je pense qu'il faut évoquer la nécessité de l'implication des collectivités. Or, nous venons de vivre une réforme régionale, sur laquelle on peut vraiment se poser des questions. Il va y avoir des incidences directes sur les politiques linguistiques. Notre région reste amputée d'un département. Je ne veux pas parler pour d'autres régions, mais par exemple, quelle politique linguistique pour l'alsacien, qui est une grande région telle qu'elle a été dessinée ? Je pense que c'est une question qu'il fallait évoquer aujourd'hui.

Benaset Dazéas
David, un avis sur cette question de la réforme territoriale ?

David Grosclaude
Sur la réforme, il y a beaucoup à dire. Je suis administrateur de l'Office public de la langue basque, mais je ne vais pas parler de la langue basque pour

laquelle je m'inquiète moins que pour la langue occitane, dans la mesure où aujourd'hui l'Office public de la langue basque est assez exemplaire. De plus, il y a, de l'autre côté de la frontière, un gouvernement basque qui fait de gros efforts et qui finance cet Office public de la langue basque de façon assez conséquente, même si la crise connue par les communautés autonomes en pays basque et en Espagne en général a un peu fait baisser les financements.

Ce qui m'a frappé dans ce que j'ai entendu tout à l'heure et dans ce que j'entends maintenant, c'est qu'on définit des choses prioritaires, mais ça fait trente ans que je m'occupe de l'occitan et ça fait trente ans que je n'entends parler que de choses prioritaires. Chaque fois qu'il y a une nouvelle technologie, elle est prioritaire. On se retrouve donc dans une situation assez inquiétante à toujours courir derrière quelque chose, en ayant le sentiment que ce sera déjà obsolète quand on l'aura atteint, parce qu'on sera toujours en retard sur les autres langues.

Aujourd'hui j'ai suivi avec intérêt le rapport qui nous a été remis sur l'occitan, au Congrès Permanent, où l'on voit où nous en sommes, mais j'ai peur que cela nous montre surtout où nous n'en serons pas quand il faudra.

Ce que j'exprimais le jour de la remise de ce rapport, c'est que nous avons intérêt à viser une étape plus loin, et qu'il va falloir aller plus vite et franchir les étapes plus rapidement que les autres, ceci à la condition qu'on ait les moyens financiers et politiques. Les moyens financiers, vous savez ce qu'il en est, les moyens politiques, vous savez aussi ce qu'il en est. Nous sommes aujourd'hui dans un pays où la question de la diversité linguistique n'est pas vraiment une question qui passionne la haute administration et le gouvernement central et pas toujours non plus les collectivités.

J'ai participé à la rédaction de rapport Caron, souhaité par Aurélie Filipetti : on attend encore la première application de quelques mesures mentionnées dedans. La DGLFLF a fait quelques efforts, elle a publié le petit recueil sur le code des langues régionales, mais je n'ai pas vu beaucoup d'enthousiasme de la part des sphères politiques. Quant aux collectivités dont on parlait à l'instant, je me retrouve dans une région Aquitaine qui va devenir une grande région Aquitaine, nous rassemblant avec des Limousins et quelques Charentais également de langue occitane avec lesquels nous

allons pouvoir travailler. Je ne sais pas si cela va dynamiser la politique linguistique, mais cela va lui donner une autre dimension, d'autant plus avec un projet d'office public que j'espère voir se créer officiellement dans les jours qui viennent par décret ministériel, ce qui constituera peut-être politiquement un outil supplémentaire. Je l'espère.

Financièrement, je ne sais pas comment on pourra être meilleurs qu'aujourd'hui, parce que ce travail que vous avez évoqué aujourd'hui, en effet, coûte de l'argent, demande des chercheurs et la mobilisation de compétences. Comme le Congrès permanent l'a fait avec nos voisins basques, nous arrivons à faire un certain nombre de choses, ce qui nous a permis de faire du transfert de compétences technologiques à notre profit. Ce qui m'inquiète et qui m'intéresse dans les discussions d'aujourd'hui, c'est de savoir quel regard cela va nous amener à porter sur un sujet qui nous préoccupe, quand on fait de la politique linguistique : celui de la transmission d'une langue.

Notre vision de la transmission est complètement changée alors que politiquement l'on nous renvoie à des schémas anciens. On nous dit que nos langues ne se parlent plus, ou mal, et qu'elles ne se transmettent plus en famille. Mais tout ce dont on vient de discuter depuis deux jours, et tout ce que l'on vient d'entendre, prouve que la question de la transmission ne se pose plus aujourd'hui comme elle se posait il y a cinquante ans. Ce ne sont plus ni l'école ni la famille, finalement, qui sont les éléments majeurs de transmission de la langue. Nous avons besoin aujourd'hui de technologies pour les transmettre, pour les faire apprendre et les faire connaître. Nous sommes nous-mêmes encore sur des schémas anciens, parce que l'on parle beaucoup de transmission par l'école, on nous renvoie à la transmission familiale en nous disant que c'est de là que nous tirerions notre légitimité. Je demande aujourd'hui combien de langues se transmettent uniquement par l'école et par la transmission familiale : certainement encore beaucoup, mais les choses sont en train de changer.

La question que je me pose est celle des moyens que l'on va nous donner pour être une langue moderne, c'est-à-dire pouvoir se transmettre autrement que par la famille et l'école, et des moyens politiques qu'on va nous donner pour qu'elle puisse aussi continuer à se transmettre par la famille et par l'école.

Les collectivités comme les nôtres n'ont pas la réponse, parce qu'elles n'ont pas les moyens politiques. Et va se discuter aujourd'hui ou demain l'idée selon laquelle les régions ont la compétence partagée avec les autres collectivités sur les langues régionales. Je ne sais pas si vous voyez où nous en sommes : on parle de technologies modernes et on ne sait même pas qui fait quoi. Je ne suis donc pas très optimiste, mais d'un autre côté je suis très volontariste, alors j'espère que nous arriverons à faire quelque chose. Il est nécessaire de faire évoluer les mentalités politiques de nos collègues élus sur ces questions de langues, notamment de faire la jonction entre nouvelles technologies et langues, ce qui va demander beaucoup de travail. Une partie de nos élus ne savent même pas que nos langues s'écrivent, il est difficile de leur expliquer qu'il faut que l'on entre dans ces nouvelles technologies.

J'ai essayé de donner un sentiment général de ce que je peux ressentir aujourd'hui, dans les difficultés qui sont les nôtres face au sujet très pointu que vous avez traité et dont certains éléments m'échappent, mais je le reconnais avec facilité.

Benaset Dazéas

Justement, concrètement, comment les offices publics en tant que maîtrise d'ouvrage public, peuvent-ils être pertinents pour offrir des cadres pérennes pour développer les technologies ? Et puisqu'on parle de moyens, de crédits, comment ces offices peuvent faire office de levier, je pense notamment aux fonds européens ?

David Grosclaude

Pour ce qui est de l'occitan, nous avons à la fois l'inconvénient et la chance d'être sur plusieurs régions et d'être un gros morceau sur les plans démographique et territorial. L'Office public est fait pour rassembler des régions et des moyens, et permettre d'avoir des opérateurs capables d'aller chercher éventuellement un certain nombre de financements européens. Cette question-là aussi a été évoquée. Nous avons des machines qui fonctionnent bien, comme le Centre interrégional de développement de l'occitan (CIRDOC), nous avons nos régions, nous aurons je l'espère l'Office public et nous avons un certain nombre de compétences que l'on peut rassembler. Avec une réelle volonté politique, nous aurons un budget nous aussi pour essayer de dynamiser cela.

Nous les occitans avons constaté que nous avions fait un certain nombre de choses mais que nous avions du retard. C'est pour cela que je parlais du basque tout à l'heure : l'Office public de la langue basque accomplit des choses en lui-même, mais il est en plus adossé à une langue qui a une officialité et un gouvernement avec une politique extrêmement dynamique en pays basque.

Pour les occitans, à part le Val d'Aran (7 000 habitants), nous n'avons pas grand-chose et l'on peut peut-être espérer travailler un peu plus étroitement avec nos amis catalans, puisque nous sommes très proches, même s'il faut reconnaître que ce n'est pas toujours facile.

Je mets donc de l'espoir dans la création de l'Office public, cela permettra peut-être de réunir d'ici peu tout l'espace occitan. Nous avons voulu la création du Congrès permanent de la lengua occitana, qui est un lieu où les discussions peuvent avoir lieu sur ce sujet, notamment sur ce qu'est l'occitan. J'entendais des discussions sur ses variétés. C'est un sujet qu'il faut manipuler avec précaution parce qu'il ne recoupe pas nos réalités politiques et qu'il a de temps en temps tendance à servir d'épouvantail à tous ceux qui nous disent que l'occitan est compliqué. Nous allons donc essayer de faire comprendre qu'il y a un objet qu'est la langue occitane, qu'il faut travailler dessus, que les problèmes de variantes sont notre affaire et qu'effectivement, ensemble, nous pouvons aller chercher des financements ailleurs.

Vous savez mieux que moi au Congrès permanent qu'aller chercher de l'argent pour travailler sur l'occitan sur des sujets qui ne sont pas toujours visibles du grand public et pas très valorisants pour les politiques, ce n'est pas passionnant, ce n'est pas une réalisation qui permet de l'électoralisme, mais on sait que c'est indispensable. Il faut travailler avec les pédagogues, qui ont besoin de matériel et de passer au numérique. Les auteurs sont de moins en moins nombreux, les éditeurs ont un mal fou à se dire qu'il faut qu'ils passent aussi à un certain nombre de technologies.

Enfin, il reste l'utilisation de la langue. C'est très bien d'avoir des traducteurs, de la reconnaissance vocale, d'avoir un certain nombre de choses, mais il faut que la langue soit utile, et utilisée. C'est un vrai problème politique.

Benaset Dazeas
Olier, sur cette question justement d'effets structurants et de montages financiers, comment cela se passe de votre côté à l'Office de la langue bretonne ?

Olier Ar Mogn
J'abonde tout à fait dans ce que vient de dire David. Je voudrais juste ajouter l'idée que toutes nos langues restent hors statut réel de protection et de promotion. Cela implique un effet en cascade. Si nos langues bénéficiaient d'un réel statut, nous pourrions très bien imaginer comme je l'évoquais hier dans mon intervention que l'État fasse pression pour la localisation de logiciels, pour que les opérateurs téléphoniques proposent automatiquement comme le disait aussi Jeremy l'effet « un clic », c'est-à-dire la possibilité, à l'achat d'un portable, d'avoir l'interface en français, en corse, en breton, en occitan sans faire un seul effort.

Il est vrai que si nous en tant que bretons on s'occupe des relations avec tous les opérateurs tout seuls (Microsoft, etc), on y arrivera difficilement. Je pense qu'il y a vraiment quelque chose à imaginer au niveau de l'État français pour cela.

Benaset Dazeas
Gaëtan Crespel, je sais que pour des raisons pratiques l'on a tendance à séparer langue et culture, alors que l'on sait très bien que ce sont des choses interdépendantes. Je voulais donc savoir comment, à Dastum, les technologies que vous pouvez développer croisent les données linguistiques et les données patrimoniales.

Cela pose aussi une autre question, qui est celle de comment assurer la meilleure disponibilité de ces corpus, notamment en termes de format, licence, etc. Je pense que vous avez des corpus importants, audio et vidéo, pour pouvoir développer ultérieurement des technologies ? Et enfin, pouvez-vous nous dire un mot sur Dastum ?

Gaëtan Crespel
Tout à d'abord, j'en profite pour présenter Marie-Barbara Le Gonidec, ethnomusicologue, membre du bureau de Dastum qui m'accompagne. J'adresse un grand merci à la délégation générale à la langue française et

aux langues de France de nous donner la parole et de nous avoir invités.

En quelques mots, Dastum est une association loi 1901 créée en 1972. Dastum signifie « recueillir » en breton. Lorsque Alan Stivell passe à l'Olympia en 1972, il déclenche tout un phénomène vis-à-vis de la musique traditionnelle bretonne. Une pléthore de groupes se constitue et reproduit quasiment à l'identique un très petit nombre de chansons voire de musiques, ce qui pousse quelques Bretons à se réunir en se disant qu'il faut développer les répertoires. Ils décident donc d'aller sur le terrain enregistrer chansons et musiques chez les porteurs de tradition, sous entendu les anciens. De là se constituent des collections d'enregistrements sonores, et face à l'affluence, la nécessité d'organiser les archives se pose très rapidement, d'où la formation de l'association Dastum.

Vous l'avez compris, nous ne sommes pas les deux pieds dans la langue. Bien que nous ayons une jambe plutôt du côté de la musique traditionnelle, voir de l'ethnomusicologie, Dastum travaille « dans les deux sens ». Nous répondons à la fois à l'individu, qui souhaite apprendre à jouer d'un instrument de musique traditionnelle et se constituer un répertoire (nous mettons à disposition sur internet d'un certain nombre d'enregistrements permettant de constituer son propre répertoire et d'avoir conscience de ce qui existait vis-à-vis de tel instrument ou tel style de musique), et dans l'autre sens, nous sommes très intéressés pour travailler avec le monde de la recherche. L'un des objets de Dastum, au-delà de la mise à disposition directe des sources, est bien entendu la valorisation. Dastum publie notamment des ouvrages et des CD de référence.

Par ailleurs, dans une autre dimension qui me semble importante ici, Dastum est, en Bretagne, identifié comme le professionnel de la conservation du son. Contrairement à ce que je découvre ici dans une logique « linguistique », nous avons dans ce domaine déjà un grand siècle d'invention de la gravure du son, ce qui fait que nous nous préoccupons depuis au moins un siècle de la conservation de ces enregistrements. Le plus vieil enregistrement que nous ayons à Dastum est un rouleau de cire datant de 1900.

Les mêmes problématiques qui ont été évoquées ici se posent : vulnérabilité des sources, nécessité d'avoir des technologies pour enregistrer le son et le lire. C'est toute l'histoire des archives audio-visuelles, photo, cinéma et

son. On en revient donc à la même logique qui se pose sur le patrimoine linguistique des technologies.

Cependant, comme le disait le cousin du Pays de Galles, les technologies pour les langues régionales sont-elles une fin ou un moyen ? Si personne ne les utilise, à quoi servons-nous aujourd'hui dans ce colloque ? Que ce soit au niveau de la DGLFLF, des collectivités territoriales ou des autres, les premières questions à prendre en compte, si le public s'intéresse désormais aux langues régionales, sont « qui parle ? », et « est-ce que nos technologies servent à quelque chose ? ». Il me semble que c'est le point le plus important. Je pense que la délégation a fait son travail en nous réunissant, pour prendre conscience que tous autant que nous sommes, dans toutes nos différences, il est important de se rencontrer, de se connaître. Nous n'avons pas forcément les réponses à toutes les questions, mais de notre rencontre peuvent naître éventuellement des solutions.

Vis-à-vis de ce que représente Dastum plus généralement en France, peut-être sommes nous mieux dotés en Bretagne et tant mieux si l'on peut servir de modèle, mais il faut quand même, par exemple pour l'occitan, savoir qu'il existe des ressources. À partir du moment où nous avons été identifiés comme spécialistes de la conservation du son, d'autres types de documents que des fonds sonores nous sont arrivés, relatifs à la musique, à la danse et à la chanson traditionnelle.

Si nous considérons aujourd'hui notre collection uniquement sur la question de la langue bretonne (mais l'on pourrait aussi s'étendre éventuellement à la question du gallo et du français), Dastum conserve désormais des enquêtes linguistiques. Certains enregistrements relèvent plus de l'ordre du récit de vie et du témoignage en langue bretonne.

Nous comptons dans nos archives tout ce qui est de l'histoire de ce que sont aujourd'hui les radios libres : c'était à l'origine des journaux parlés sur cassette, devenues des radios privées dans les années 80, dont nous conservons les archives, ce qui représente des corpus, comme vous pouvez l'imaginer, totalisant des dizaines voir des centaines d'heures.

Nous continuons d'enregistrer tout ce qui est de l'ordre des événements culturels, des veillées, des concours et bien entendu nous rassemblons les

éditions sonores diffusées, en quelque sorte tout ce qui relève du patrimoine édité, alors que le premier objet de Dastum est le son inédit.

De là, d'après ce que j'ai entendu depuis le début du colloque, nous avons beaucoup de problématiques identiques nous aussi. Nous faisons partie d'une fédération nationale, la Fédération des Associations de Musiques et de Danses Traditionnelles (FAMDT). Les régions qui ont besoin de faire valoir des langues régionales ne doivent pas oublier qu'il existe des équivalents de Dastum ailleurs. Hélas, nous n'en sommes pas au traitement et aux technologies de la langue, mais si l'on peut un jour partager et mettre à disposition ces fonds, nous le ferons volontiers. Aujourd'hui nous avons nous aussi des technologies qui font que nous sommes en ligne sur internet : c'est le site www.dastum.bzh, après le 3 615 Dastum qui existait dès 1990. Nous sommes moissonnables en quelques clics. Les fonds radio dont je parlais à l'instant ont déjà été diffusés et publiés.

Je pense que nous avons intérêt à faire se rejoindre le macro et le micro : il y a des problématiques à régler à l'échelle de la France vis-à-vis des autres pays et de sa place dans le monde, mais ce n'est pas pour autant que le citoyen doit être totalement exclu et que l'on en arrive uniquement à des logiques de recherche. Il faut que cela aille dans les deux sens, et je pense que l'on perd beaucoup de temps si l'on se restreint à un débat de Jacobins-Girondins : nous avons besoin des deux. Les régions ont des logiques que ne régleront pas des grands projets à l'échelle de la France. Le monde des chercheurs et celui des amateurs doivent se rejoindre pour renforcer nos arguments et être plus forts ensemble.

Benaset Dazeas

Une dernière question, pour Sébastien Quenot. Nous avons parlé à deux reprises, lors de la précédente table ronde, de l'informaticien corse qui ne rencontre pas le linguiste corse. Comment et dans quel cadre pourrait-on réaliser les perspectives de mutualisation et de réflexion commune au niveau de toutes les langues de France ? Est-ce que ce serait une mise en réseau ? Quelque chose de plus formalisé ? Comment est-ce que l'on pourrait faire représenter au sein de cette instance à la fois ceux qui conduisent les politiques publiques, que ce soit les collectivités territoriales, la communauté scientifique, mais aussi les usagers et les diffuseurs de langue ? Qu'est-ce que cela vous inspire pour arriver, à terme, à développer ces technologies pour le corse ?

Sébastien Quenot
À la différence des autres régions, nous sommes une île. J'ai pu mesurer ce matin le bonheur d'être insulaire, rapport à vos difficultés de vous constituer, d'identifier votre territoire et les locuteurs.

Pour nous, c'est un peu plus simple.

Ensuite, c'est la collectivité territoriale de Corse qui s'occupe des actions de préservation comme celles que Dastum réalise.

Troisième point, en ce qui concerne les questions d'équipement et de nouvelles technologies, nous avons l'exemple très pédagogique du pays basque que nous avions vulgarisé à l'époque des débats concernant la co-officialité. Si l'on compare simplement le nord et la communauté autonome, la communauté autonome dispose de droits linguistiques, d'outils, et ils ont davantage de locuteurs. Le nord dispose des mêmes outils que le sud, mais n'a pas de droits linguistiques ni d'usagers pour ces nouvelles technologies. Tout l'intérêt est donc justement de faire coïncider à la fois les acteurs publics, qui peuvent être les collectivités locales ou les offices, pour obtenir de nouveaux droits linguistiques auprès de l'État, puisque les mesures de co-officialité sont une question constitutionnelle importante en termes de droits linguistiques.

Concernant les outils, il faut que les collectivités locales essaient de piloter, qu'elles influent sur les acteurs. Nous avons eu durant l'automne dernier une grande polémique en Corse entre l'État et l'université. L'État ne voulait pas signer la convention tripartite (un autre statut particulier entre la collectivité territoriale, l'université, et l'État), car l'université voulait absolument faire figurer la mention de la co-officialité, donc le vote de l'assemblée de Corse, dans cette convention. Finalement, la convention a été signée, mais le terme de co-officialité n'est plus présent que dans une annexe de la convention entre l'État et l'université. Ce soir, le président de l'université ainsi que le président de notre conseil exécutif vont se rencontrer de façon à discuter des modalités de mise en œuvre de cet avenant à cette convention, dont l'objet sera la mise en place d'un groupement d'intérêt public entre la collectivité territoriale et l'université pour pouvoir créer des outils.
En 2012, on a créé le *Consiglio di Lingua Corsa*, qui est un échec. D'après notre expérience, je déconseille l'intégration des usagers, car cela devient

vite un défouloir. Notre idée partait de bons sentiments, nous avions créé différentes commissions dirigées par des politiques afin de construire un consensus linguistique, avec des universitaires, des experts de la langue, des représentants de la société civile, des membres extérieurs. Cependant, la diversité des profils rassemblés n'a pas fonctionné.

Ensuite se pose une autre question, celle de la valorisation du travail des universitaires dans ce groupement d'intérêt public. Un universitaire qui travaille sur un dictionnaire ne gagne pas beaucoup de reconnaissance, car ce ne sont pas des publications de rang A. Pourtant, ce sont les outils dont nous avons besoin sur nos territoires. Là aussi, je pense que le monde universitaire devrait s'adapter un peu aux besoins des acteurs sociaux de la population de façon à ce que l'on reconnaisse nos besoins, et ce d'autant plus à l'université de Corse, où l'on a peu d'enseignants-chercheurs spécialistes de ce domaine. Ils sont accaparés par des missions administratives et d'enseignement, comme cela a déjà été évoqué. Chez nous, étant une petite université de 4 000 étudiants, c'est encore plus accru.

J'écoutais ce matin Jacques Attali avec attention parlant avec science des droits linguistiques des russophones en Ukraine. Il nous disait qu'il fallait que les russophones puissent continuer à parler russe. Nous demandons la même chose mais chez nous. Ce qui semble naturel à l'extérieur apparaît problématique ici.

Voilà pour ces questions et ce que nous entendons construire, ce groupement d'intérêt public (GIP) entre la collectivité territoriale Corse et l'université, auquel il faudra aussi associer des partenaires privés je pense. C'est peut-être le premier étage de la fusée. Le deuxième étage de la fusée pourrait être de construire de même un autre GIP, mais entre nos GIP existants, de façon à pouvoir mener un lobbying un peu plus efficace lorsque l'on appelle Microsoft. Je pense qu'en ce moment c'est un peu plus intéressant en termes de restrictions budgétaires.

Dans le plan qui devrait être voté par l'assemblée de Corse le mois prochain, nous avons évalué la partie « nouvelles technologies » à 3 millions d'euros par an, ce qui est en fait la totalité de notre budget annuel. Ce n'est donc pas possible, nous n'allons pas pouvoir tout faire tous seuls. Il nous faut

ce type de mutualisation, et je pense que l'occasion d'y travailler de façon à être un peu plus efficaces est là, à la suite de ce type de réunion.

Je terminerai par une parenthèse, étant donné que les projets ANR sont très difficiles à décrocher, nous avons l'impression de ne vraiment pas intéresser. Après dix refus, nous abandonnons donc de ce côté-là pour chercher du côté de la région, de l'Europe.

Je pense que l'État aurait tout intérêt à donner une partie des fonds à cette cause d'utilité publique : en effet, 90 % de la population corse veut une société bilingue, ce qui est un chiffre important. Et lorsque l'État défend la francophonie dans le monde entier, c'est un argument de poids de dire que la diversité est aussi défendue sur notre propre territoire.

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

Achevé d'imprimer en mars 2016
sur les presses de l'imprimerie Corlet
à Condé-sur-Noireau (Calvados).
dépôt légal : mars 2016
ISBN 978-2-11-139348-6
Rencontres/2016/01/FR

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