Agence pour la Distribution des ressources Linguistiques et l'Evaluation

9, rue des Cordelières
75 013 Paris
Tél. : 01 43 13 33 33
Fax : 01 43 13 33 30
Web : www.elda.org
Email : info@elda.org

INVENTAIRE DES RESSOURCES LINGUISTIQUES DES LANGUES DE FRANCE

Auteur(s)	Jérémy LEIXA, Valérie MAPELLI et Khalid CHOUKRI
Organisme	ELDA
Adresse	9, rue des Cordelières 75013 Paris, France
E-mail	[leixa;mapelli;choukri]@elda.org
Date	22 septembre 2014
Version	V1.1

Historique du Document

Version	Date	Statut	Notes
0.1	15 Mars 2013	Projet	Structure du document, contenu et synopsis
0.2	13 Juin 2013	Projet	Commentaires de VM
0.3	19 Juin 2013	Projet	Corrections de JL suite aux commentaires
0.4	20 Juin 2013	Projet	Version révisée par VM
0.5	17 Juillet 2013	Projet	Version révisée par KC
1.0	22 septembre 2014	Projet	Correction par JL, révision VM
1.1	5 novembre 2014	Version finale	Révision et validation du document par la DGLFLF

Table des matières

1 RÉSUMÉ

En partenariat avec la Délégation générale à la langue française et aux langues de France (DGLFLF, ministère de la Culture et de la Communication), ELDA souhaite établir **un inventaire des ressources linguistiques existantes pour les langues régionales de France**, aussi bien en métropole que dans les DOM-TOM.

Dans un premier temps, nous avons procédé à la définition d'un périmètre de recherches, afin notamment d'utiliser des typologies adaptées en termes de technologies et de langues étudiées. C'est ainsi que pour les langues, nous nous sommes basés sur le classement établi par la DGLFLF, et qui comprend 76 langues différentes, en incluant la Langue des Signes Française. En nous basant sur le classement établi dans le Livre Blanc MetaNet[1], **six technologies de la langue** ont été prises en compte pour l'analyse : la traduction automatique, la synthèse et la reconnaissance vocale, la correction orthographique (ou assistance à la rédaction), l'analyse sémantique, l'analyse grammaticale, et la génération automatique de texte.

Rapidement, devant l'importante quantité de ressources identifiées à la fois sur Internet (au travers de sources de données telles que les sites institutionnels, les journaux, les sites TV…), et dans les grands catalogues comme celui d'ELRA, ou encore OLAC, METASHARE et LDC, il a fallu réfléchir à une solution efficace et utilisable facilement par la DGLFLF. Ainsi, un site Internet doté d'une interface ergonomique et associé à une base de données MySQL a été mise en place, et après plusieurs phases d'optimisation, elle est fonctionnelle et toutes les informations récoltées au cours de cet inventaire y ont été saisies.

Lors d'une réunion d'avancement avec la DGLFLF, plusieurs axes de réflexion ont été abordés pour la suite de l'étude, et des améliorations ont été proposées pour cette base de données. La principale recommandation concerne la volumétrie des données récoltées. En effet, pour le moment, une ressource est référencée par une adresse URL unique et peut aussi bien concerner un enregistrement audio de deux minutes, qu'un corpus aligné breton-catalan d'un million de mots. Ainsi, dans un souci d'optimisation, il est prévu que la base de données inclue une composante « volumétrie », où apparaîtra la taille des ressources (en nombre de mots, ou en temps d'enregistrement selon la nature de la ressource).

Enfin, parmi les langues et les technologies étudiées, il a été décidé de se focaliser sur quelques points en particulier. Ainsi, une analyse plus poussée a été menée sur les langues d'Outre-Mer, pour lesquelles peu de ressources existent dans un format utilisable immédiatement, et un accent particulier a été mis sur trois technologies de la langue : la traduction automatique, la synthèse / la reconnaissance vocale, et enfin la correction orthographique. Une étude de la faisabilité de ces technologies a été menée pour les langues ci-dessus, et il en ressort que la traduction automatique, ainsi que la correction orthographique **sont des technologies tout à fait envisageables pour l'occitan et le breton**, notamment grâce à l'existence de nombreuses ressources écrites. En ce qui concerne la synthèse et la reconnaissance vocale, la faisabilité est moindre car cela nécessite d'importants corpus de parole, qui ne sont pas disponibles actuellement. Enfin, **les langues d'Outre-Mer sont dans une situation de retrait** vis-à-vis de cette faisabilité, car trop peu de ressources existent, tant à l'écrit qu'à l'oral, pour pouvoir envisager de créer les technologies en question.

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

2 INTRODUCTION

2.1 CONTEXTE

En décembre 2011, se sont déroulés les États Généraux du Multilinguisme dans l'Outre-Mer. Ce rendez-vous a eu pour but de promouvoir les langues régionales dans les DOM-TOM, afin que celles-ci ne soient pas vouées à disparaître à court et moyen terme. Parmi les points abordés, la question de l'enseignement des langues d'Outre- Mer, mais aussi la position des langues d'Outre-Mer face au français, la langue de la République. Lors de ces États Généraux, deux projets Wikimédia avaient été présentés : Wikipédia, et le Wiktionnaire. À titre d'illustration, Wikipédia est le 5e site le plus consulté au monde, et le Wiktionnaire francophone contient la définition de milliers de mots dans 987 langues différentes. Ces deux projets pourraient donc constituer un formidable outil de promotion et d'expansion des langues d'Outre-Mer, dont seulement deux sont utilisées pour des versions de Wikipédia[2].

En 1992, le Conseil de l'Europe a publié, en complément de la convention européenne, la Charte Européenne des langues régionales ou minoritaires[3], l'objectif étant de protéger et favoriser les langues régionales et les langues des minorités en Europe. Cette charte repose sur plusieurs points, auxquels s'engagent les États l'ayant ratifiée : la reconnaissance des langues régionales et minoritaires en tant qu'expression de la richesse culturelle, le respect de l'aire géographique associée à une langue, faciliter et promouvoir l'usage **à l'oral comme à l'écrit** de ces langues, dans la vie publique et dans la vie privée, ou encore la lutte contre toute forme de discrimination ou d'exclusion à l'encontre de ces langues.

Au travers de ses activités, ELDA a participé à un grand nombre d'études et de projets dans le domaine des Technologies de la Langue, et grâce à sa participation à des projets européens majeurs, la société a acquis de grandes connaissances à propos de ce marché. Par exemple, nous pouvons mentionner les travaux menés par ELDA dans le cadre des projets NEMLAR/MEDAR, ainsi que le projet BLARK.

Par ailleurs, de 2004 à 2007, ELDA a participé au projet TC-STAR, financé par la Commission Européenne dans le cadre du Sixième Programme-cadre, qui met l'accent sur la recherche avancée dans le domaine de la traduction du discours (*Speech-to-speech Translation*, SST) dans un contexte multilingue. Cette technologie combine la reconnaissance automatique de la parole (*Automatic Speech Recognition*, ASR), la traduction de la parole (*Spoken Language Translation*, SLR), et la synthèse vocale (*Text-to-Speech*, TTS).

Au niveau national, ELDA s'est vue confier un certain nombre de projets ayant mené à une meilleure identification du marché des Technologies de la Langue. Nous pouvons notamment mentionner le Guide de Production réalisé par l'équipe d'ELDA en 2008 [F. Gandcher, O. Hamon, V. Mapelli, N.Moreau, N.Paulsson, D. Mostefa,*Réalisation d'un guide de production de ressources linguistiques pour la veille,* 2008].

Plus spécifiquement dans le domaine de la traduction automatique, ELDA a conduit une étude subventionnée par le Ministère français de la Recherche, ayant pour but d'établir un état de l'art des sociétés et des outils de traduction automatique sur le marché français. Pour ce travail, ELDA a établi un inventaire des acteurs français de la R&D, mais également en Europe et partout dans le monde. Nous avons identifié les principaux programmes de recherche (nationaux et internationaux), et avons analysé la participation française dans ce contexte.

2 http://blog.wikimedia.fr/etats-generaux-du-multilinguisme-4050
3 http://www.coe.int/t/dg4/education/minlang/default_fr.asp

2.2 OBJECTIFS

En partenariat avec la Délégation Générale à la Langue Française et aux Langues de France (DGLFLF), ce projet consiste à **réaliser un inventaire des ressources linguistiques existantes pour les langues de France**. Il s'agit dans un premier temps de mettre en place un comité de suivi pour définir et valider le périmètre de travail et les grands axes de recherche. Ensuite, il s'agit d'identifier les sources d'information et de les compiler avec un focus clair sur les ressources linguistiques pour les langues régionales de France en vue d'un usage en traitement automatique des langues et ingénierie linguistique.

La première étape du projet consiste à identifier des sources d'information pertinentes, et de faire une revue assez exhaustive des conférences, études, revues, sites web, etc. dont les informations portent à la fois sur le traitement de la langue et les langues régionales. Une attention toute particulière est portée à l'identification des laboratoires et centres de recherche dont l'activité peut toucher l'objet du projet.

Une fois les sources d'informations identifiées, l'équipe va collecter et compiler les informations utiles (inventaires existants, actes de conférences, listes/catalogues, rapports de projets, etc.) sous la forme d'une base de données avec une interface simple s'inspirant des catalogues et données existants, en particulier ceux développés par ELRA pour ses catalogues.

Ensuite, nous évaluerons l'adéquation des ressources identifiées pour le développement des technologies de traitement de la langue. Cela couvrira à la fois le traitement de la langue orale ou écrite (reconnaissance de la parole, traduction automatique, etc.). Il s'agira également d'évaluer la pertinence de leur utilisation par des acteurs travaillant essentiellement sur le français.

3 ÉTAT DE L'ART ET ETUDES ANALOGUES

3.1 POLITIQUE ACTUELLE ET CONSIDÉRATION VIS-À-VIS DES LANGUES RÉGIONALES

En dehors de la question de la Charte Européenne, de nombreux problèmes viennent entacher le processus d'identification et de reconnaissance de certaines langues régionales de l'Hexagone et d'Outre-Mer. Par exemple, malgré la loi du 25 mai 2013 sur la refondation de l'école, qui permet entre autres une plus grande considération des langues régionales dans l'enseignement, de grandes disparités existent d'une région à l'autre. De fait, il s'avère que la situation en Guyane et à Mayotte est problématique, car peu de moyens sont accordés pour la prise en compte et l'aide à la culture, alors qu'en Guadeloupe ou en Martinique, cela s'applique dans une moindre mesure.

De plus, en Nouvelle-Calédonie, se posent plusieurs problèmes : tout d'abord, la formation des enseignants est mal assurée, et la prise en compte de l'enseignement des langues kanakes s'en ressent. Ainsi, seules quatre langues kanakes sur les 28 existantes sont représentées au Baccalauréat. Ensuite, le statut des enseignants n'est pas reconnu au même niveau que ce qu'il peut être en métropole ou dans les autres départements et collectivités d'Outre-Mer, ce qui entraîne un problème de transmission de la langue aux jeunes générations.

L'autre problème principal concernant l'accès aux ressources linguistiques, notamment pour les langues d'Outre-Mer, concerne le fait que bien souvent, ces langues sont de tradition or et font l'objet d'une transmission restreinte. Cette tendance va rapidement se confirmer dans cette étude, où la plupart des ressources identifiées pour les langues d'Outre-Mer sont des enregistrements de contes traditionnels, ou de récits de la vie de tous les jours. À l'inverse, nous allons retrouver très peu de ressources écrites. Quelques initiatives existent, comme le dictionnaire drehu-français mis en place par l'Académie des langues kanak, en Nouvelle-Calédonie, mais cela reste épisodique pour le moment. La plupart des contes traditionnels qui font la culture des Kanaks n'existent pas sous forme écrite, et leur transmission devient de plus en plus difficile, d'une part à cause du

vieillissement de ces populations, et d'autre part à cause du caractère ésotérique de certains de ces contes : bien souvent, les anciens ne transmettent cet aspect de leur culture que dans certaines conditions de discrétion ou d'intimité. Tout cela entraîne une rareté des corpus écrits disponibles pour les jeunes qui souhaitent apprendre la langue maternelle de leur région. Ce problème est autant qualitatif que quantitatif, car se pose aussi la question de l'accessibilité à ces corpus écrits. Les jeunes générations n'ont finalement pas d'accès à des corpus de qualité, ce qui entraîne naturellement la disparition de leur langue.

3.2 ÉTUDES SIMILAIRES

En ce qui concerne l'étude des langues minoritaires de France, ainsi que pour l'étude des besoins en ressources linguistiques, qui sont les 2 thèmes principaux de notre rapport, de nombreux travaux existent déjà, ainsi que plusieurs institutions et événements associés.

3.2.1. BLARK

Le concept BLARK (pour *Basic LAnguage Resource Kit* – Kit de base de ressources linguistiques) a d'abord été défini en Hollande. L'une des premières mentions du BLARK a été faite notamment dans un article rédigé par Steven Krauwer (Krauwer, 1998), faisant suite à une proposition de coopération commune entre ELSNET (European Network of Excellence in Language and Speech) et ELRA dans le cadre du 5e Programme-Cadre de la Commission européenne.

Ce concept a pu être mis en œuvre pour la première fois dans le cadre de l'initiative « Dutch Human Language Technologies Platform » (plate-forme pour les technologies de la langue hollandaises) pour la langue hollandaise, lancée en avril 1999 par le Dutch Language Union, un organisme inter-gouvernemental en charge de renforcer la situation de la langue hollandaise (Cucchiarini et al., 2001a) et (Cucchiarini et al. 2001b).

Plus récemment, la notion de BLARK a été adaptée à la langue arabe, dans le cadre du projet NEMLAR (Network for Euro-Mediterranean LAnguage Resources), suivi du projet MEDAR (cf section 3.2.2).

ELDA a développé un service interactif BLARK permettant d'identifier les différents besoins en termes de ressources linguistiques vis-à-vis d'applications spécifiques, et ce pour autant de langues possibles. Suite à sa propre expérience et aux différents rapports issus des autres initiatives telles que celle hollandaise, ELDA a implémenté et augmenté sa matrice d'origine qui consistait en un croisement entre les langues et les différents types de ressources qu'elle avait pu identifier.

Afin de comprendre les besoins plus clairement et de manière plus exhaustive, ELDA a étendu cette matrice à une liste d'applications et de modules potentiels pouvant être mis en relation avec les ressources linguistiques et langues nécessaires. Les matrices qui en résultent viennent en partie des travaux réalisés dans le cadre du projet NEMLAR. Deux matrices (Applications / Modules et Langues / Modules) ont été développées et sont disponibles et modifiables directement depuis les pages web.

En pratique, les matrices BLARK sont divisées en deux tableaux :

Le tableau « Applications/Modules » montre le niveau d'importance des modules nécessaires (ou non) à une application donnée, tant pour les technologies de l'écrit que celles de l'oral, et ce pour une langue donnée : important (+), très important (++), essentiel (+++) ou sans importance (ø).

close window	Customization to Different	Dialect/Language	Dictation	Embedded Speech	Emotion Identification	Emotion/Prosody Output	Generation Lips Movement
Acoustic Models	+++	+++	+++	+++	+++	+++	+++
Dialect/language Identification		+	+	+	+		
Emotion Identification		+	+	+		++	
Language Models		++	+++	++		++	
Lexicon Adaptation			+	+			
Lips Movement Reading		++					
Phoneme Alignment			+	+			
Pronunciation Lexicon			+++	+++			
Prosody Prediction						+++	
Prosody Recognition		+	+	+	+++		
Segmenter Speech/silence		++	++	++	++	+	
Sentence Boundary Detection		+	+	+	++	++	
Speaker Adaptation		+	++	++	+		

Figure 1 : exemple d'un tableau Applications/Modules pour les données orales de l'Arabe

display matrix in fullscreen	Annotated Written Corpus	Audio Data with Prosodic Markers and other	BNSC	Desktop/Microphone & High Quality	Non Vowelised Corpus	Onomastica (proper names)	Phonetic Lexicon	Telephony	Unannotated Written Corpora	VIsual Data (faces, lips, etc.)	Vowelised Corpus
Acoustic Models		+++	+++	+++				+++			
Dialect/language Identification		+	++	++		+	+	++			
Emotion Identification		+	+	+		+	+	+			
Language Models	++				++				+++		++
Lexicon Adaptation	+				+	+++	+++		+		++
Lips Movement Reading										+++	
Phoneme Alignment	++	++	++	++		+++	+++	++			+
Pronunciation Lexicon	+					+++	+++				++
Prosody Prediction	++	++				++	++				++
Prosody Recognition	++	+++		+		++	++	+			+
Segmenter Speech/silence		++	++	++				++			
Sentence Boundary Detection		++	++	++		+	+	++			
Speaker Adaptation		+	++	++				++			
Speaker Recognition/identification		+	+	+				+			
Speech Units Selection	++	+++		+		+	+	+			
Speech/non-speech Music Detection		++	++	+				+			
Word Boundary Identification		+	+	+		+	+	+			

Figure 2 : exemple d'un tableau Ressources/Modules pour les données orales de l'Arabe

Le tableau « Ressources/Modules » ci-dessus montre le niveau de nécessité en termes de ressources linguistiques à inclure ou non dans des modules spécifiques, tant pour les technologies de l'écrit que celles de l'oral, et ce pour une langue donnée : important (+), très important (++), essentiel (+++) ou sans importance (ø).

Ce système de matrices BLARK nous a servi de base d'étude pour la représentation des données identifiées lors de cet inventaire. En effet, ce type de tableau croisé nous permettra de visualiser rapidement l'importance de certaines ressources par rapport à d'autres. Cela nous permettra également de classifier de façon plus efficace les langues étudiées.

3.2.2. Autres études

D'autres études / projets d'importance capitale, ayant pour but de combler les manques en termes de ressources linguistiques, ont pu voir le jour ces dernières années. Qu'il s'agisse de réseaux d'excellence pour le traitement des langues naturelles, ou encore de consortiums chargés de promouvoir les langues, plusieurs d'entre eux peuvent être mis en avant :

- Le projet **NEMLAR**[4], fondé par la Commission Européenne, a été créé dans le but d'établir une collaboration forte entre les acteurs Européens et Méditerranéens, afin de faire avancer les technologies de la langue pour l'Arabe et d'autres langues régionales. Une suite lui avait été donnée avec le projet intitulé **MEDAR** (cf. medar.info/The_Nemlar_Project/index.html).
- Le projet **FLAReNet** (Fostering LAnguage Resources Network)[5] consiste également en la constitution d'un réseau efficace pour la production et la standardisation de ressources linguistiques. À la manière du projet MEDAR, FlaReNet a pour principales missions l'organisation de conférences et d'ateliers, et se base sur un réseau de communication important.
- Le réseau **METANET**[6] est lui aussi un réseau d'excellence dont la mission principale est la mise en place de fondations technologiques solides pour l'Europe multilingue. Grâce à METANET, les technologies de la langue pourront permettre la communication et la coopération entre les différentes langues, mais aussi assurer aux usagers d'une langue un accès égal à toute information ou source de connaissance, et enfin permettre l'avancée des technologies de l'information. À partir des travaux de ce réseau d'excellence, un livre blanc a été publié. Ce livre blanc MetaNet fait partie d'une collection d'ouvrages qui a pour objectif de faire connaître les technologies de la langue et leur potentiel. Il s'agit d'un ouvrage qui s'adresse aux journalistes, aux professionnels de l'industrie de la langue, aux communautés linguistiques, aux enseignants, et à tout le monde en général.
- **Les conférences LREC :** Tous les 2 ans, ELRA (l'Association Européenne pour les Ressources Linguistiques[7]) organise la conférence LREC (Language Resources Evaluation Conference[8]) avec l'aide de nombreuses institutions et entreprises œuvrant dans le domaine des technologies de la langue. Au cours de cette conférence, de nombreux thèmes sont abordés, dont celui des langues minoritaires.

4 http://www.medar.info/Mission/index.php
5 http://www.flarenet.eu/
6 http://www.meta-net.eu/
7 http://www.elra.info/
8 http://www.lrec-conf.org/

4 MÉTHODOLOGIE

4.1 CHOIX MÉTHODOLOGIQUES ET ORGANISATIONNELS : DÉFINITION DU PÉRIMÈTRE

Afin de dresser un inventaire des ressources linguistiques existantes pour les différentes langues régionales de France, il nous a fallu, dans un premier temps, définir le périmètre des informations que nous souhaitions mettre en avant dans cette étude. Ainsi, nous avons établi une liste de critères essentiels à la classification des différentes langues de France.

- La famille de langue : à quel groupe une langue (langues d'oïl, parlers corses, langues kanakes, etc.) appartient-elle ? Comment peut-on la classifier ?
- Le nombre de locuteurs : quelles informations sont disponibles quant au nombre (officiel ou estimé) de locuteurs pour une langue donnée ?
- Les modalités : la langue est-elle une langue uniquement orale (comme cela peut être le cas pour certaines langues d'Outre-Mer), est-elle également écrite, ou bien est-elle essentiellement visuelle, comme la langue des signes française (LSF) ?

Dans un second temps, il nous a fallu définir le périmètre des sources de données brutes vers lesquelles nous souhaitions nous diriger, et rapidement nous avons pu établir des catégories précises. Il a fallu poser clairement la définition de ce que nous entendions par « données brutes »: il s'agit de sites internet qui constituent des sources de données intéressantes, en opposition aux ressources linguistiques qui sont déjà collectées et formalisées. Ces sources contiennent des données textuelles, audio ou vidéo, tels que des sites TV et Radio. Les données qu'ils proposent ne sont pas, en l'état, utilisables en tant que ressources linguistiques, mais peuvent ensuite être transposées dans un formalisme adéquat pour la manipulation via un système automatisé (mise sous forme de corpus aligné, avec soit un alignement par mot/phrase/paragraphe, soit un alignement par segmentation après transcription). Nous avons concentré nos recherches sur le web, car la plupart de ces sources y sont facilement accessibles. Nous avons donc recensé les catégories suivantes :

- Les sites de médias (TV, radio, et journaux).
- Les sites institutionnels (sites académiques, souvent en version multilingue).
- Les catalogues de ressources (OLAC, METASHARE, ELRA, etc.)
- Les sites d'informations générales sur une langue donnée, qui même s'ils sont de moins grande ampleur, constituent souvent une source considérable de données, car ceux-ci sont régulièrement proposés en version multilingue.

C'est seulement suite à l'établissement de ces critères que nous avons pu nous pencher sur le critère suivant, essentiel pour l'inventaire : définir les langues à étudier.

4.2 RECHERCHE D'INFORMATIONS SUR LES LANGUES RÉGIONALES DE FRANCE ET D'OUTRE-MER

Afin de dresser une liste la plus fidèle possible à la réalité des langues régionales de France, nous nous sommes basés sur deux listes, celle proposée par le site Internet Ethnologue.com[9], et celle de la DGLFLF.

4.2.1 Les critères du site web Ethnologue

Le site web Ethnologue.com propose une liste de 25 langues de France (cf. Figure 3), classées selon différents degrés de développement, de « langue courante » à « langue éteinte ». Parmi ces langues, nous avons choisi d'écarter celles décrites comme éteintes (shuadit et zarphatique), et de ne conserver que les plus importantes. C'est ainsi que la première liste de langues minoritaires retenues se compose de l'alsacien, le basque, le breton, le catalan, le corse, la langue des signes française (LSF), le ligurien, l'occitan, le picard, et le romani.

Cette liste, bien que très complète, ne nous a servi que de base pour établir notre classement de langues pour cette étude. C'est ainsi que nous avons choisi de compléter ces informations avec la liste proposée sur le site de la DGLFLF.

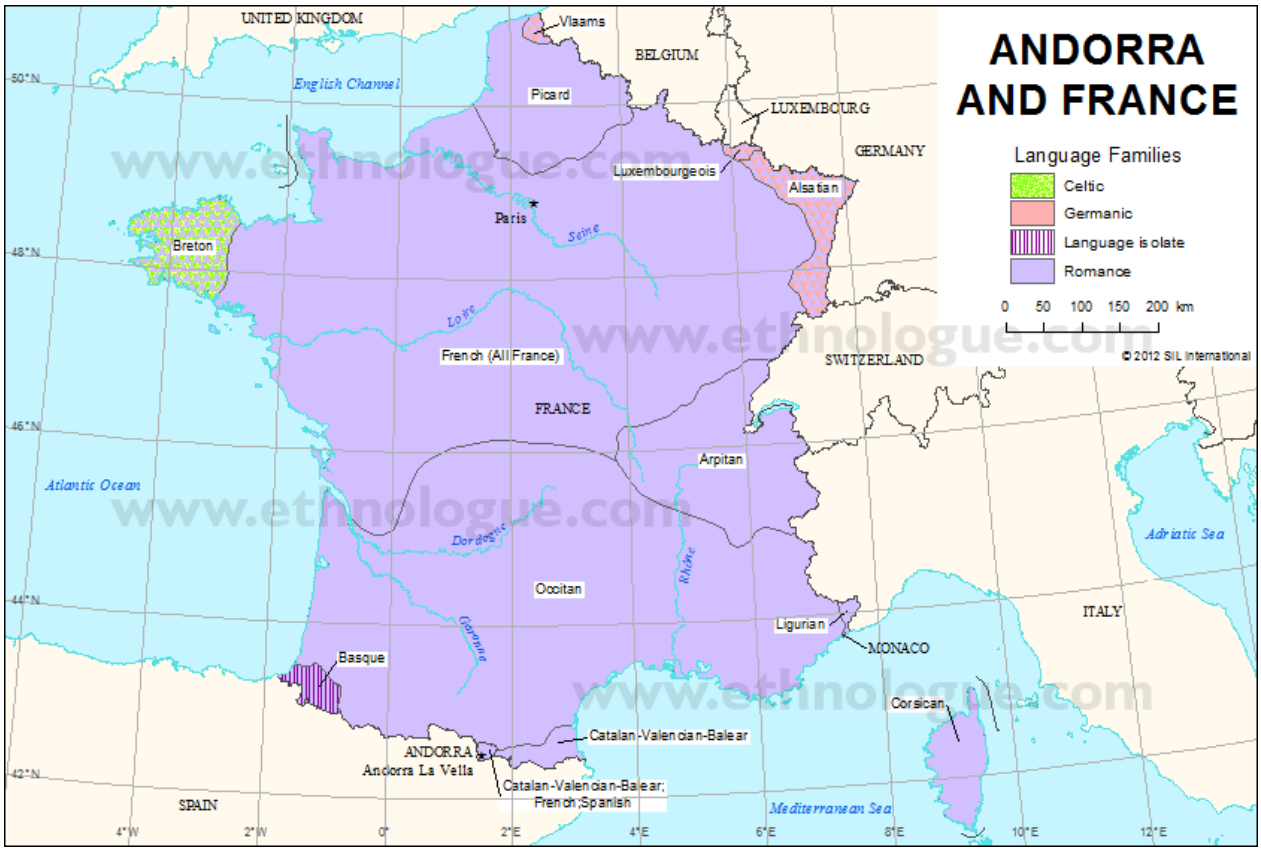

Figure 3 : Cartographie des langues régionales de France métropolitaine (source : ethnologue.com)

9 http://www.ethnologue.com/country/FR/languages

4.2.2 Les critères de la DGLFLF

Là où le site Ethnologue fournit un classement des langues de France selon divers degrés d'extinction, la DGLFLF dresse une liste exhaustive des langues.

4.2.3 Typologie des langues étudiées dans le cadre de ce projet

Après avoir confronté les informations identifiées sur les langues, nous avons dressé une liste de 84 langues minoritaires regroupées en trois grandes parties, qui constituent l'objet de cette étude.

4.2.3.1 Les langues de France Métropolitaine

- **Langues régionales (23)**: alsacien, basque, breton, catalan, corse, flamand occidental, francique mosellan, francoprovençal, langues d'oïl (franc-comtois, wallon, champenois, picard, normand, gallo, poitevin-saintongeais [dans ses deux variétés : poitevin et saintongeais], lorrain, bourguignon-morvandiau), langue d'oc ou occitan (gascon, languedocien, provençal, auvergnat, limousin, vivaro-alpin).
- **Langues non-territoriales (6) :** arabe dialectal, arménien occidental, berbère, judéo-espagnol, romani, yiddish.

4.2.3.2 Les langues d'Outre-Mer

- **Zone caraïbe (14) :**
  Créoles à base lexicale française : guadeloupéen, guyanais, martiniquais ;
  Créoles bushinenge de Guyane (à base lexicale anglo-portugaise) : saramaca, aluku, njuka, paramaca ;
  Langues amérindiennes de Guyane : galibi (ou kalina), wayana, palikur, arawak (ou lokono), wayampi, émerillon ; Hmong.
- **Réunion (1) :** créole réunionnais (à base lexicale française).
- **Nouvelle-Calédonie (28) (cf. Figure 4)** :
  Grande terre : nyelâyu, kumak, caac, yuaga, jawe, nemi, fwâi, pije, pwaamei, pwapwâ, langue de Voh-Koné, cèmuhi, paicî, ajië, arhâ, arhö, ôrôwe, neku, sîchë, tîrî, xârâcùù, xaragurè, drubéa, numèè ;
  Iles Loyauté : nengone, drehu, iaai, fagauvea.

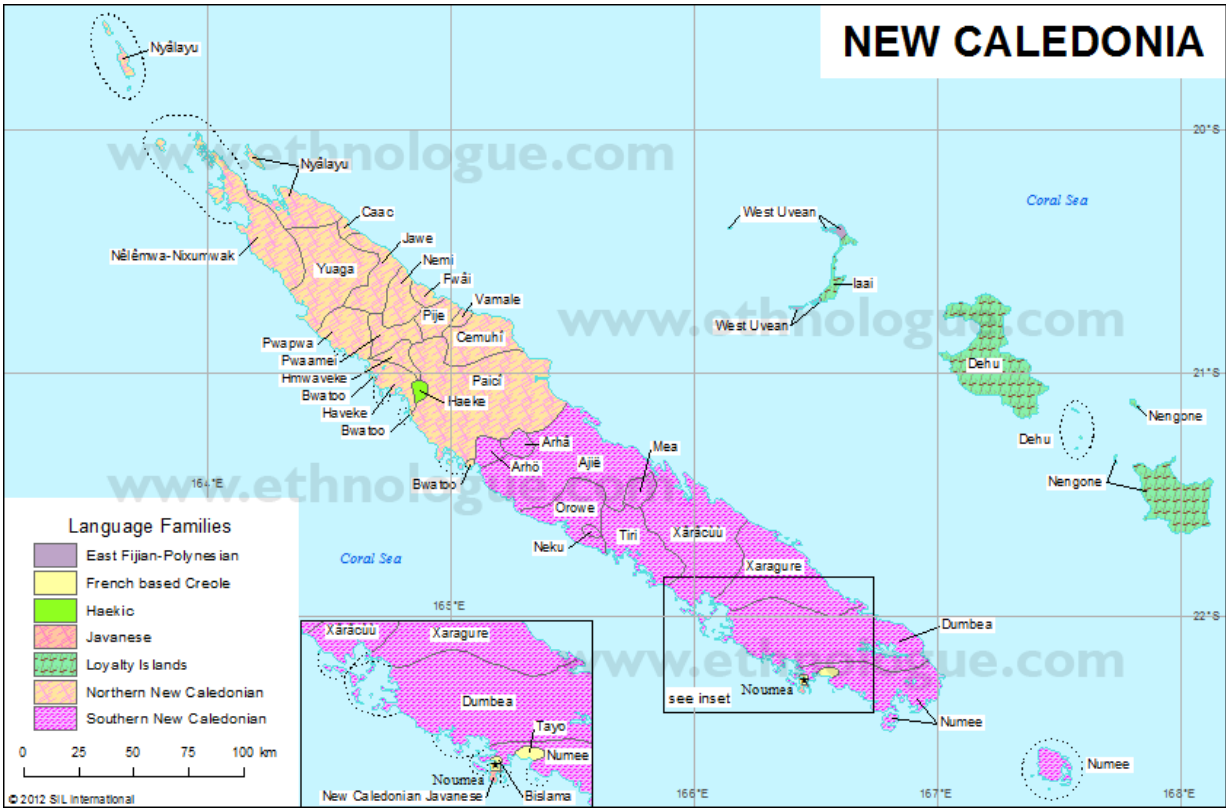

Figure 4 : Cartographie des langues régionales de Nouvelle-Calédonie (source : ethnologue.com)

- **Polynésie française (7) (cf. Figure 5)**: tahitien, marquisien, langue des Tuamotu, langue mangarévienne, langues des Îles Australes : langue de Ra'ivavae, langue de Rapa, langue de Ruturu.

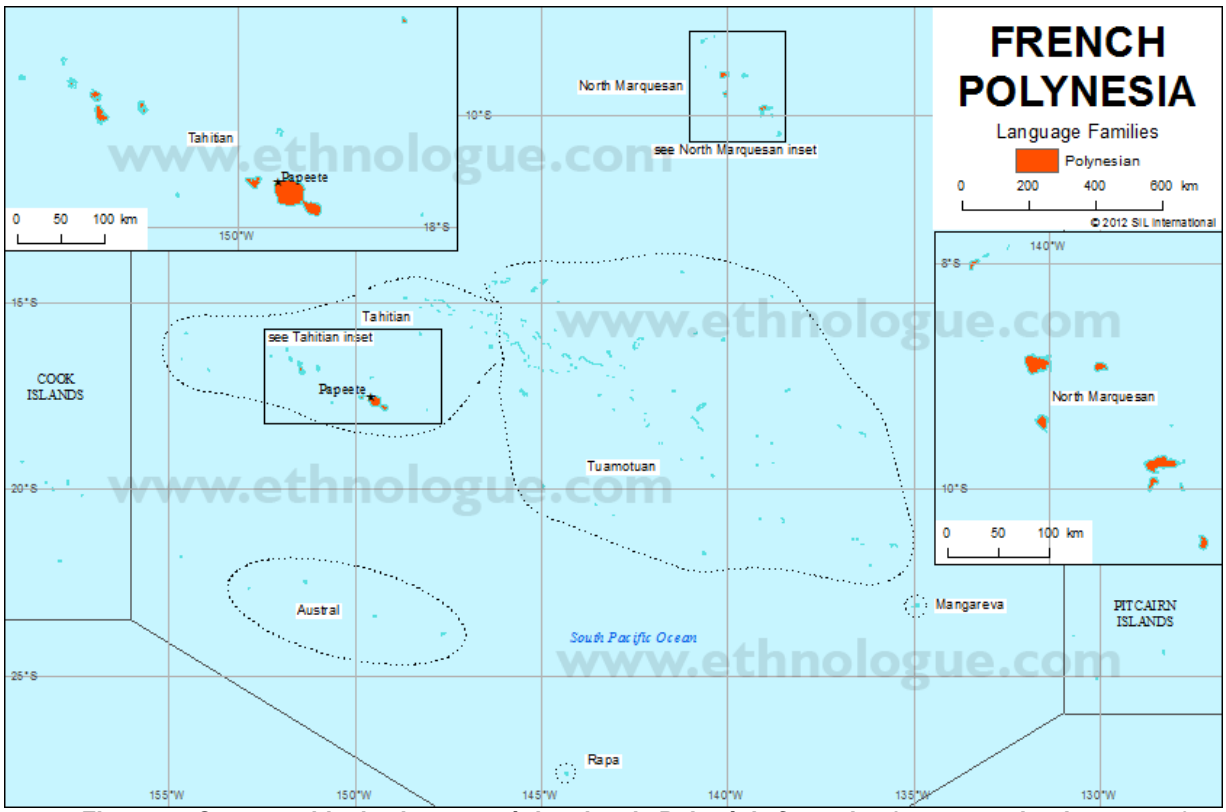

Figure 5 : Cartographie des langues régionales de Polynésie française (source : ethnologue.com)

- **Iles Wallis et Futuna (2)** : wallisien, futunien.
- **Mayotte (2)** : mahorais (shimaoré), malgache de Mayotte (shibushi)

4.2.3.3 La langue des signes française (LSF)

Selon le site de la DGLFLF, la langue des signes française est considérée également comme une langue de France, puisqu'elle est pratiquée par un certain nombre de citoyens français.

Cette langue connaît des variations régionales, bien qu'utilisée à une échelle nationale, et restant dans l'ombre de l'institution parisienne de la rue Saint-Jacques (l'Institut National des Jeunes Sourds de Paris). Ainsi, il existe des « dialectes » de la LSF à Nancy, Rouen, Chambéry, Lyon ou encore Montpellier.

4.3 TYPOLOGIES

Nous avons choisi d'établir des typologies pour chacun des domaines concernés par cet inventaire, à savoir : les sources de données, les ressources linguistiques envisagées et les technologies prises en compte pour l'analyse. Ces typologies sont nécessaires dans la mesure où elle vont permettre de définir un cadre pour cette étude, ce qui nous aidera par la suite lors des phases d'analyse et de classification. Ces typologies nous permettent aussi de mieux entrevoir les critères dont nous avons besoin pour mettre en place la base de données qui sera décrite par la suite.

4.3.1 Typologie des sources de données

Comme il a été évoqué précédemment, les sources de données que nous avons pu identifier sur internet constituent un moyen fiable de récolter beaucoup de ressources linguistiques potentielles, même si cela implique des phases de négociations et de « nettoyage » de ces données brutes. Nous avons donc distingué plusieurs types de sources :

- **Les sites de médias** : il s'agit de sites de différents types, à savoir : des sites TV, radio ou encore de journaux. Par exemple, le site internet de la chaîne TV et Radio www.la1ere.fr constitue une formidable source de données brutes potentielles, puisque les émissions radio diffusées dans les régions d'Outre-Mer peuvent être écoutés directement sur le site. On peut donc imaginer récupérer les droits d'exploiter ces fichiers audio, afin d'en tirer des corpus de conversations en langues d'Outre-Mer. Pour les sites de journaux, nous nous sommes penchés sur la presse locale, qui bien souvent, est traduite en plusieurs langues.
- **Les sites institutionnels** : il s'agit pour la majeure partie de sites académiques ou publics, qui sont bien souvent en version multilingue. Ceci constitue donc un élément intéressant quant à la constitution de corpus parallèles pour des technologies de traduction automatique. Le site de l'Académie des langues kanak (http://www.alk.gouv.nc/portal/page/portal/alk/), institution de Nouvelle-Calédonie dont le but est d'aider à la promotion et au développement des langues kanakes, contient de nombreuses ressources dans ces langues. Il doit donc être tout à fait possible de constituer des données textes à partir de données brutes récupérées sur ce site.
- **Les sites généralistes** : du petit site associatif au grand site culturel, il existe des sources de données brutes (exemple : http://www.montraykreyol.org/) où toutes sortes de sujets sont abordés. Présentation de livres, chants traditionnels sous la forme de fichiers audio, ces sites sont bien souvent une source intéressante de données brutes qu'il est intéressant d'exploiter.
- **Les catalogues de ressources**: une grande partie des ressources identifiées dans la base de données provient des catalogues de ressources linguistiques que proposent ELRA (http://catalog.elra.info/index.php?language=fr), OLAC (http://www.language-archives.org/), LDC (http://www.ldc.upenn.edu/Catalog/) ou encore METASHARE (http://metashare.elda.org/repository/search/). Les données disponibles dans ces catalogues, à l'inverse des autres types de sources, sont ici disponibles dans des formats qui en font des ressources linguistiques utilisables immédiatement dans des systèmes automatisés. Ce qui n'est pas le cas des données brutes que l'on peut identifier sur les autres sites.

4.3.2 Typologie des ressources linguistiques envisagées

Afin d'apporter le plus de clarté possible quant à la cible des ressources linguistiques que nous souhaitons collecter pour les besoins de ce projet, il a également fallu définir précisément quels types de ressources linguistiques nous souhaitions prendre en considération. Nous nous sommes inspirés du Livre Blanc de MetaNet pour cette typologie :

- **Corpus de textes** : un corpus de textes est constitué d'un ou plusieurs fichiers au format texte. Ce ou ces fichiers peuvent être au format brut, ou bien être annotés et étiquetés pour les besoins d'un système.
- **Corpus de parole** : un corpus de parole sera, lui, constitué d'enregistrements audio de conversations, ou d'émissions TV ou radio, par exemple.
- **Corpus parallèle** : un corpus parallèle, ou aligné, sera souvent constitué de corpus de textes traduits dans différentes langues, et ces corpus seront alignés de façon à pouvoir entraîner par exemple des systèmes de traduction automatique.
- **Corpus multimédia** : un corpus multimédia, à l'inverse d'un corpus de textes ou d'un corpus de parole, contient des fichiers avec plusieurs médias, comme par exemple des enregistrements vidéo, où l'on a à la fois l'image et le son, et éventuellement les sous-titres extraits dans un fichier texte.

- **Lexique** : un lexique (ou dictionnaire) est une liste de mots accompagnés de leur définition qui pourra par exemple servir de base à un système de correction orthographique, ou de synthèse vocale s'il est accompagné de données phonétiques.
- **Grammaire** : une grammaire va définir les règles de fonctionnement d'une langue tant à l'écrit que pour le discours. Une grammaire sera particulièrement utile pour un système de synthèse et de reconnaissance vocale.
- **Thésaurus** : Un thésaurus est un type particulier de langage documentaire. Il est constitué d'un ensemble structuré de concepts représentés par des termes, pouvant être utilisés pour l'indexation de documents dans une banque de données bibliographiques ou dans un catalogue de centre de documentation, à des fins de recherche documentaire.

4.3.3 Typologie des technologies de la langue dans le cadre de cette étude

Là encore, cette typologie est basée sur le livre blanc MetaNet, et traitera de technologies telles que le traitement de la parole, la génération automatique, ou la traduction automatique.

- **Traduction automatique :** un système de traduction automatique (tel que Moses[10], ou Apertium[11]) permet de traduire automatiquement un texte d'une langue vers une autre. Ce type de technologie nécessite de grandes quantités de données afin que les systèmes de traduction automatique puissent s'entraîner.
- **Synthèse et Reconnaissance vocale :** à partir de corpus audio et de grammaires de langages, un système de synthèse vocale peut créer de la parole, tandis qu'un système de reconnaissance vocale peut reconnaître des instructions laissées par un locuteur humain. Dans le cas de la synthèse vocale, nous partons d'un texte pour aller vers de la parole (technologie *Text-to-Speech*), alors que pour la reconnaissance vocale, il faut partir de la parole pour aller vers un texte (technologie *Speech-to-Text*).
- **Correction Orthographique / Assistance à la rédaction** : il existe plusieurs types d'outils de correction, selon l'aspect du langage qu'ils vont traiter. Nous avons donc d'un côté les correcteurs orthographiques qui vont comparer les mots d'un texte à ceux d'un dictionnaire. Si ce mot y est présent, alors il est validé. Tandis qu'un correcteur grammatical vérifiera que les mots, bien qu'ils soient présents dans un dictionnaire, sont conformes aux règles de grammaire et de sémantique de la langue en question. De gros corpus de textes sont essentiels à l'apprentissage de tels systèmes. Parmi les correcteurs les plus connus, Aspell (du projet GNU) est un correcteur orthographique très répandu. Nous avons également ces mêmes correcteurs dans les moteurs de recherche tels que Google ou Yahoo. Cordial et Antidote figurent parmi les correcteurs grammaticaux les plus utilisés.
- **Analyse grammaticale :** à l'aide de lexiques ou de grammaires, la phase d'analyse grammaticale d'un texte va permettre de mettre en avant les éventuelles erreurs au niveau d'un mot, d'une proposition ou d'une phrase.
- **Analyse sémantique :** L'analyse sémantique d'un message est la phase de son analyse en se basant sur le sens des éléments (mots) du texte, par opposition aux analyses lexicales ou grammaticales qui décomposent le message à l'aide d'un lexique ou d'une grammaire.
- **Génération automatique :** la Génération Automatique de Texte (GAT) est particulièrement utile dans les systèmes de questions-réponses, où le processus va se faire en trois temps : analyse et compréhension de la question, puis recherche d'informations pour constituer des éléments de réponse, et enfin génération d'une réponse à la question. De manière générale, la GAT va partir du sens pour former du texte.

10 http://www.statmt.org/moses/
11 http://www.apertium.org/

5 CONSTITUTION D'UNE BASE DE DONNÉES

Lors de cette étude, l'une des questions fondamentales a été de savoir comment structurer et représenter les résultats de l'inventaire des ressources linguistiques pour les langues régionales de l'Hexagone et d'Outre-Mer. Il a été rapidement décidé d'organiser ces ressources sous la forme d'une base de données. Cela nous a permis de structurer le mieux possible ces ressources, afin que chaque utilisateur puisse s'y retrouver, et puisse consulter à sa guise les résultats de cet inventaire.

5.1 IMPLÉMENTATION TECHNIQUE ET DÉFINITION D'UN METADATA

La première étape de la constitution de cette base de données a été de définir la façon dont nous voulions structurer les données inventoriées. Nous avons au départ répertorié les ressources sous la forme d'un simple fichier au format tableur, puis, devant l'ampleur des ressources et des informations disponibles, il a rapidement fallu définir un nouveau formalisme pour ces données. Ainsi, nous avons décidé de compiler toutes ces informations sous la forme d'une base de données, telle qu'elle est disponible aujourd'hui. Pour sa mise en place, une structure MySQL couplée à une interface simple et ergonomique nous semblait la meilleure solution pour une utilisation optimale par un utilisateur final. Ainsi, la base est désormais accessible via une adresse internet, tout en étant hébergée chez ELDA. Ce choix de mise à disposition via Internet permettrait à n'importe quel utilisateur ayant un profil enregistré sur la base de pouvoir y saisir des informations concernant de nouvelles ressources linguistiques qui n'auraient pas encore été inventoriées.

Le but premier de ce site Internet étant de manipuler le plus simplement possible les données récoltées, il a fallu définir un formalisme simple, mais efficace, afin de pouvoir décrire de façon claire et systématique les ressources qui ont été identifiées.

L'étape suivante de cette réflexion consistait à définir l'ensemble des métadata nécessaires à la bonne représentation des informations saisies. À chaque ressource ajoutée à la base, nous associons tout un ensemble de caractéristiques, afin de pouvoir structurer et classer efficacement les ressources linguistiques inventoriées, et ainsi permettre la génération de statistiques de toutes sortes. Ces statistiques sont consultables, et permettent d'avoir un aperçu des tendances relatives aux ressources collectées.

Les deux tables principales de cette base de données sont les tables « Sources » et « Ressources », et sont alimentées par la plupart des autres tables. Ainsi, les informations relatives à la langue, à la famille de langue, au nombre de locuteurs et aux modalités font toutes l'objet d'une table à part, dont les informations n'ont a priori plus besoin d'être modifiées. De plus, lors de la saisie d'une nouvelle ressource entrée dans la base de données, il suffit de sélectionner la langue pour que toutes les informations associées soient automatiquement intégrées dans la fiche de la ressource. Ainsi, les 2 tables regroupent les informations suivantes :

- **Ressources linguistiques**
  - Langues (famille de langues, répartition linguistiques, nombre de locuteurs, modalités)
  - Ressources (nom, description, URL, source, type)
  - Applications existantes et potentielles
  - Disponibilité (fournisseur, copyright)
  - Volumétrie (estimation de la taille de la ressource)

- **Sources de données brutes**
  - Source (nom, description, URL, source, type)
  - Langues
  - Applications potentielles
  - Contact

5.2 VISUALISATION DES INFORMATIONS

5.2.1 Profils / Accès / Tableau de bord

Pour les besoins de gestion de la base de données, 2 profils de connexion ont été créés. Le premier est le profil « Administrateur », qui donne accès à l'intégralité des tables et des options de la base. Le second profil est un profil de type « Lecteur », qui est uniquement à but consultatif, et pour lequel les seules tables accessibles sont « Sources » et « Ressources ».

Lors de la saisie de l'adresse de la base de données, l'utilisateur arrive à l'écran de connexion illustré par la figure n°6. Là, il est invité à saisir son identifiant ainsi que son mot de passe, afin de pouvoir accéder à la base.

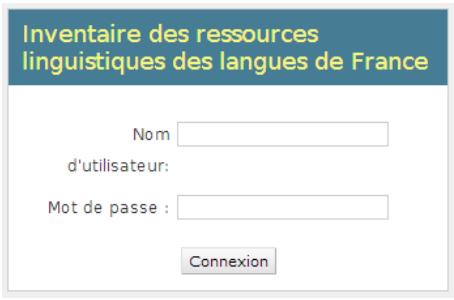

Figure 6 : écran de connexion à la base de données

La figure n°7 nous montre que sous le profil « Administrateur », la totalité des tables sont accessibles, et modifiables (exemple : il est possible de modifier des informations sur le nombre de locuteurs d'une langue quelconque en accédant à la table Langues, puis à la langue en question).

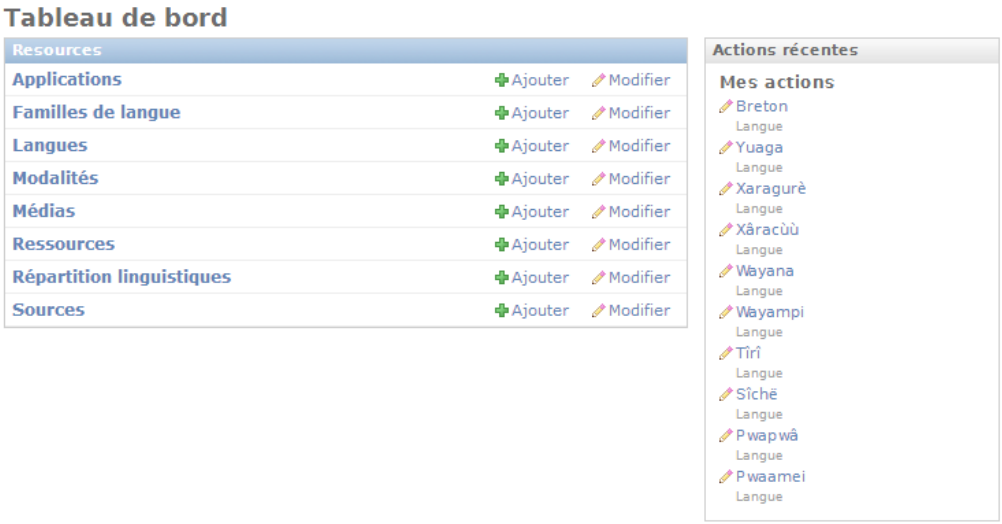

Figure 7 : Tableau de bord du profil « Lecteur »

La figure n°8, en revanche, nous montre l'aspect du tableau de bord du profil « Lecteur ». Comme indiqué, nous pouvons voir que seules sont accessibles les tables « Sources » et « Ressources ». Aucune autre action n'est disponible.

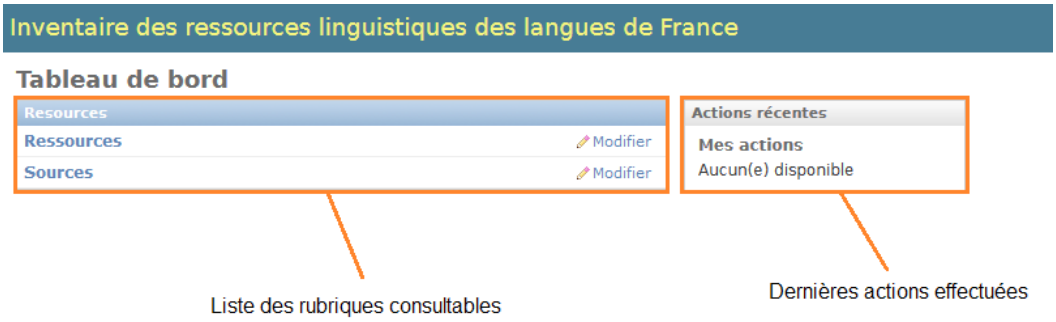

Figure 8 : Tableau de bord du profil « Administrateur »

Sur la figure suivante (figure n°9), nous avons un aperçu du tableau de bord d'une personne connectée sous le profil « Administrateur ». Dans la rubrique « Ressources », nous voyons les diverses informations affichées et les options disponibles. Au centre, nous avons la liste des ressources saisies dans la base, ainsi que les différentes informations qui leur sont liées. Nous avons ainsi, pour chaque entrée, des indications concernant la langue concernée, la famille de langue à laquelle la ressource appartient, des informations sur le type de ressource dont il s'agit, et des informations d'ordre « légal » à propos du fournisseur, et des droits d'usage de la ressource.

Ensuite, en haut à droite, sont affichées les options accessibles : aller sur la page des Statistiques, ajouter une ressource, ou bien importer une ressource au format OLAC (une grande partie des ressources provenant de ce catalogue, une fonction d'import automatique a été ajoutée à la base, afin de pouvoir ajouter un grand nombre de ressources en une seule fois).

Enfin, sur la partie droite de l'écran, on peut voir les différents filtres que l'on peut appliquer à la liste des entrées de la base de données. Ces filtres sont déroulants, de façon à afficher les options disponibles sans gêner la bonne lisibilité des informations à l'écran. À côté de chaque filtre se trouve le nombre d'entrées recensées pour ce type de filtre.

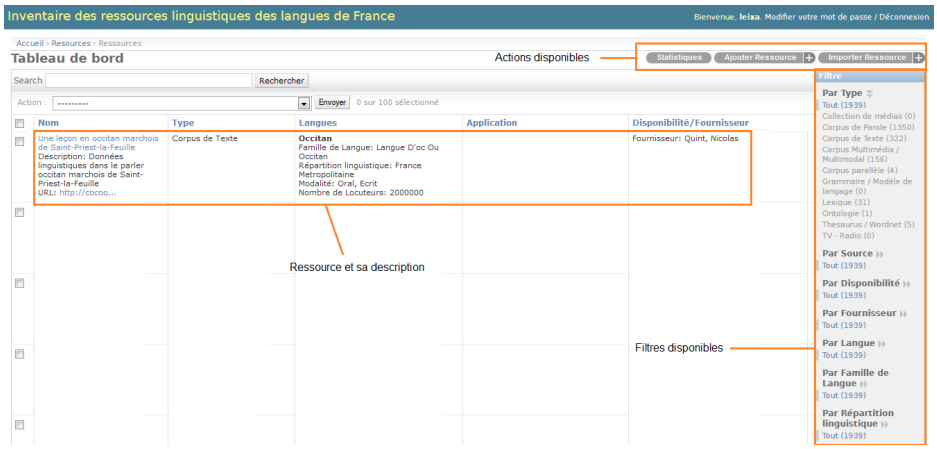

Figure 9 : Tableau de bord du profil "Administrateur", avec une ressource et sa description

5.2.2 Saisie et extraction de l'information

L'intérêt principal de cette base de données concerne bien sûr la manipulation de l'information, c'est pourquoi les deux options majeures de cette interface sont la saisie et l'extraction des données. En effet, n'importe quel utilisateur disposant de droits suffisants peut saisir de nouvelles entrées dans la base de données. Le formulaire de saisie se présente sous la forme de champs à remplir, de cases à cocher et d'éléments à sélectionner dans des listes déroulantes. Ainsi, lorsque l'utilisateur à fini de saisir les métadata de sa ressource, il peut l'enregistrer afin que celle-ci soit intégrée à la base, et apparaisse ainsi dans la liste des ressources disponibles à la consultation. Toutes les informations que l'on retrouve dans le tableau de la figure n°9 sont à renseigner lors de la création d'une fiche de ressource.

À l'inverse, il est également possible d'extraire des informations de la base de données sous forme tabulaire. En effet, dans la liste déroulante se trouvant en tête de tableau, l'utilisateur a la possibilité, après avoir sélectionné les entrées qui l'intéressent, de les exporter au format tableur. Bien sûr, cette fonction d'export est compatible avec les différents, afin de n'exporter que les informations souhaitées. Il suffit ensuite d'enregistrer le fichier ainsi généré où l'utilisateur le souhaite.

5.2.3 Exploitation de l'information : Statistiques, graphiques, filtres

En termes d'options, nous avons voulu que les informations recensées dans cette base de données soient suffisamment claires et exhaustives pour que quiconque puisse y accéder aisément. Ainsi, des graphiques sont disponibles à la consultation dans une rubrique dédiée, et les utilisateurs ont le choix du mode de représentation pour les statistiques désirées (camembert, graphique type « barres », etc.): nombre d'entrées par langues, nombre d'entrées par famille de langue, ou encore le type d'entrées le plus présent dans la base de données.

La figure n°10 nous montre l'un des graphiques disponibles sur la page « Statistiques ». Celui-ci est divisé en deux parties. En haut, différentes options pour masquer ou visualiser le graphique, et pour en changer le type. On peut ainsi passer d'un graphique type « camembert » à un graphique type « barres ». En bas, le graphique est affiché, accompagné des statistiques en question.

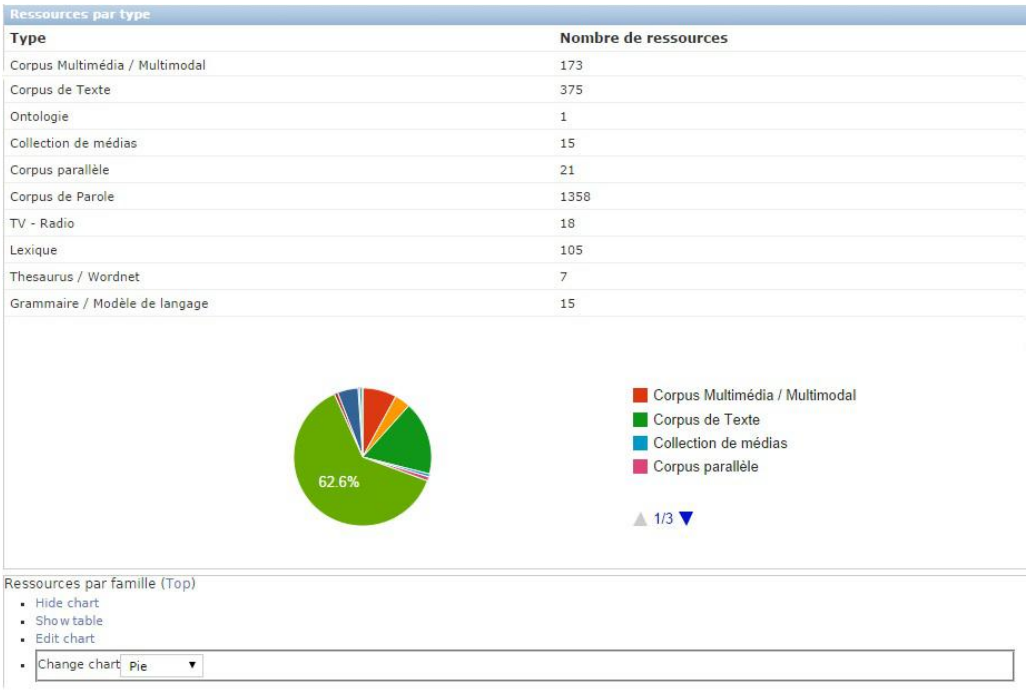

Figure 10 : Exemple d'un graphique pour le nombre de ressources par type

Les différentes statistiques disponibles sont : le nombre de ressources par langue, le nombre de ressources par type, le nombre de ressources par famille de langue, le nombre de ressources par modalité, le nombre de ressources par application potentielle / existante, et le nombre de ressources par fournisseur. Ces statistiques sont également consultables dans la partie Sources.

L'autre aspect de cette base concernant l'exploitation de l'information réside dans les différents filtres disponibles dans les tables des ressources et des sources de données. En effet, l'utilisateur peut trier les entrées affichées à l'écran selon tous les critères possibles. Plusieurs filtres peuvent être activés en même temps, si bien qu'il est tout à fait possible d'afficher à l'écran uniquement les corpus de texte pour le judéo-espagnol, mis à disposition par le site Internet OLAC, ou encore les lexiques en occitan. Toutes les combinaisons sont possibles.

6 RÉSULTATS DE L'INVENTAIRE

6.1 INVENTAIRE

Au terme de cet inventaire, nous avons pu établir plusieurs statistiques à propos des entrées saisies dans la base de données.

Actuellement, 1 532 ressources uniques, ainsi que 80 sources de données brutes sont recensées dans cette base de données. Il existe bien sûr certaines langues (notamment parmi les langues kanakes) pour lesquelles nous n'avons pas trouvé de ressources. Celles-ci ne figurent donc pas dans le tableau ci-dessous. Nous avons donc classé toutes ces ressources uniques par langue, et par type. Une ressource unique peut être recensée dans plusieurs types (une ressource peut contenir à la fois un corpus texte et un corpus de parole, par exemple). C'est ainsi que le tableau ci-dessous comptabilise 2 233 ressources, car celles-ci peuvent se répéter d'une catégorie à une autre en fonction de leurs caractéristiques.

LANGUES	CORPUS PAROLE	CORPUS TEXTE	CORPUS MULTIMÉDIA	CORPUS PARALLÈLE	LEXIQUE	ONTOLOGIE	THÉSAURUS	TOTAL
Langue d'oc ou occitan	365	92	29	33	102		1	622
Breton	403	17						420
Marquisien	75		116					191
Judéo-espagnol	129	35						164
Alsacien	127							127
Corse	36	42	11		2			91
Kumak	36	34						70
Langue des signes française		6	41					47
Catalan	7	9	6	1	19	1	3	46
Futunien	23	21						44
Wallisien	20	19						39
Réunionnais	31	4						35

LANGUES	CORPUS PAROLE	CORPUS TEXTE	CORPUS MULTIMÉDIA	CORPUS PARALLÈLE	LEXIQUE	ONTOLOGIE	THÉSAURUS	TOTAL
Nemi	15	16						31
Arabe dialectal	14	6		3	6			29
Fagauvea	13	13						26
Xaragurè	11	9						20
Numèè	9	9						18
Xâracùù	9	9						18
Picard	16	1						17
Basque	6	1	2		5		3	17
Guadeloupéen	7	7						14
Drehu	13	1						14
Iaai	8	5						13
Picard	16	1						17
Pije	1	1						2
Réunionnais	31	4						35
Romani	3	3						6
Wallisien	20	19						39
Wallon	1							1
Wayampi					1			1
Wayana	3	3						6
Xâracùù	9	9						18
Xaragurè	11	9						20
Yuaga	2	2						4

Table 1 : nombre de ressources par langue et par type dans la base de données

Parmi toutes ces ressources identifiées, le corpus de parole est le plus représenté. Il faut cependant rappeler que la plupart des corpus de parole collectés concernent des enregistrements assez courts. Il faut bien évidemment mettre en relation le nombre de ces ressources avec le volume de données qu'elles représentent. Et sur ce point, il est fort probable que certains lexiques ou corpus représentent plus de données que les corpus de parole.

Un autre type de ressource a été implémenté dans la base de données : les grammaires. Bien qu'il

ne s'agisse pas à proprement parler de ressources linguistiques au même titre qu'un corpus de parole ou qu'un lexique (la grammaire va définir les règles d'une langue), il est envisagé d'effectuer des recherches en amont des ressources pour identifier ces grammaires.

Dans les tables 1 et 2, présentées ci-dessous, nous pouvons voir la répartition de ces ressources en fonction des langues, mais aussi en fonction des familles de langues. Là encore, certaines de ces ressources peuvent apparaître dans plusieurs catégories à la fois, notamment dans le cas des corpus multilingues.

Breton	**420**	**Catalan**	**47**	**Fagauvea**	**26**
Gascon	**286**	**Langue des signes française**	**47**	**Vivaro-alpin**	**22**
Languedocien	**202**	**Futunien**	**44**	**Xaragurè**	**20**
Marquisien	**191**	**Wallisien**	**39**	**Numèè**	**18**
Judéo-espagnol	**164**	**Réunionnais**	**35**	**Xâracùù**	**18**
Alsacien	**127**	**Nemi**	**31**	**Basque**	**17**
Corse	**93**	**Provençal**	**30**	**Picard**	**17**
Kumak	**70**	**Arabe dialectal**	**29**	**Auvergnat**	**16**
Occitan	**56**	**Limousin**	**27**	**Drehu**	**14**
Iaai	**13**	**Italien**	**4**	**Wayampi**	**1**
Palikur	**11**	**Guyanais**	**4**	**Suédois**	**1**
Anglais	**11**	**Ajië**	**3**	**Néerlandais**	**1**
Français	**10**	**Francoprovençal**	**2**	**Wallon**	**1**
Espagnol	**9**	**Pije**	**2**	**Portugais**	**1**
Cèmuhi	**9**	**Langue des signes grecque**	**2**	**Galicien**	**1**
Aluku	**7**	**Grec moderne**	**2**	**Arménien occidental**	**1**
Romani	**6**	**Langue des signes allemande**	**2**	**Latin**	**1**
Wayana	**6**	**Nengone**	**2**	**Niçois**	**1**
Yuaga	**4**	**Langue des signes britannique**	**2**	**Roumain**	**1**

Mahorais (shimaore)	**4**	**Flamand Occidental**	**2**
Allemand	**4**	**Paicî**	**2**

Table 2 : Nombre de ressources par langue dans la base de données

Une nouvelle phase de remplissage de la base de données a été effectuée en août 2014, suite à la réalisation de deux inventaires, l'un pour la langue occitane (*Lo congrès permanent de la lenga occitana*), l'autre pour la langue corse (DGLFLF). Ces deux inventaires ont permis de largement agrémenter la base de données pour les langues nommées ci-dessus, de l'ordre de 252 ressources pour l'occitan, et 53 ressources pour la langue corse.

L'ensemble de l'inventaire, présenté sous forme de tableau, est disponible sur le CD qui est joint au rapport.

Langue d'oc ou occitan	**669**	**Catalan**	**47**
Langues celtiques	**420**	**Autre**	**37**
Langues non-territoriales	**200**	**Langues d'oïl**	**18**
Polynésie française	**191**	**Langues amérindiennes**	**18**
Grande-Terre	**177**	**None**	**17**
Dialectes allemands d'Alsace et de Moselle	**127**	**Basque**	**17**
Corse	**93**	**Espagne**	**9**
Wallis-et-Futuna	**83**	**Créoles bushinenge (base anglo-portugaise)**	**7**
Îles Loyauté	**55**	**Mayotte**	**4**
Créole à base lexicale française	**53**	**Francoprovençal**	**2**
LSF	**53**	**Flamand occidental**	**2**

Table 3 : Nombre de ressources par famille de langues dans la base de données

6.2 FOCUS ET ANALYSES

Afin de pouvoir établir, dans le cadre de cette étude, la faisabilité des technologies évoquées précédemment pour les langues régionales, nous avons décidé de nous focaliser sur quelques langues en particulier, et de ne sélectionner que certaines technologies afin de mener une analyse plus poussée sur une partie de l'inventaire. Ainsi, nous avons choisi de prendre en compte les domaines suivants : la traduction automatique, la synthèse et reconnaissance vocale, et la correction orthographique.

Concernant les langues, nous avons voulu traiter des langues les plus variées possibles, tant au niveau de la structure de la langue, qu'au niveau de leur statut et de leur évolution. Ainsi, nous avons choisi de traiter le breton, l'occitan et plusieurs langues d'Outre-Mer, dont le créole réunionnais.

6.2.1 Quelques langues d'Outre-Mer

Pour cette catégorie de langues, nous avons choisi d'inclure le créole réunionnais, mais aussi d'autres langues des départements et collectivités d'Outre-Mer comme le guadeloupéen, le martiniquais, le guyanais ou encore le wallisien et le futunien. Enfin, les langues kanakes, ainsi que les langues de Mayotte (mahorais et malgache) viennent compléter ce groupe.

Au cours de l'inventaire, nous avons rapidement constaté de grandes disparités dans l'existence, mais aussi dans la disponibilité des ressources linguistiques existantes pour ces langues d'Outre-Mer. Cela s'illustre parfaitement dans le tableau de statistiques donné précédemment. On y voit clairement que si des ressources sont disponibles pour le malaisien et le futunien, c'est loin d'être le cas pour les langues de Mayotte : 83 ressources recensées pour Wallis et Futuna, contre quatre pour Mayotte.

La place importante occupée par le groupe « Grande-Terre » dans le tableau provient du fait que

certaines langues kanakes, comme le drehu, le kumak ou le nemi, ont fait l'objet d'une étude plus poussée, et par conséquent un nombre important de ressources audio a été référencé sur le catalogue OLAC. Il s'agit pour l'essentiel d'enregistrements audio de contes traditionnels, et de récits de la vie quotidienne. Comme on peut le voir dans les statistiques, il existe des disparités au sein même des langues kanakes. Cela peut s'expliquer en partie par le fait que certaines langues ne sont parlées que par très peu de personnes encore vivantes, et que bien souvent, ces mêmes langues ont un caractère sacré, traditionnel. Alors, les derniers représentants de ces langues ne sont pas forcément enclins à les transmettre aux chercheurs.

La situation du guyanais et des langues de Mayotte est un peu différente, dans la mesure où le manque de ressources est en grande partie dû au contexte sociopolitique. En effet, le guyanais, tout comme le mahorais et le malgache, souffrent d'un grave manque de reconnaissance et de support. La situation est très mauvaise concernant la prise en compte et l'aide à la culture dans ces pays, et par conséquent peu de moyens et de structures sont accordés pour la sauvegarde de ces langues. Cela implique également que peu de ressources linguistiques ont pu être collectées et mises à disposition des chercheurs.

De grands corpus existent néanmoins, notamment pour le réunionnais : le corpus VALIRUN, constitué par Gudrun Ledegen, contient environ 200 heures d'enregistrements audio de créole réunionnais et de français, ainsi que les transcriptions orthographiques alignées de ces deux langues. Dans cette base de données orales numérique de la langue française et créole réunionnaise débutée il y a six ans, Gudrun Ledegen procède à la sauvegarde d'enregistrements anciens (années 70) et actuels, à leur transcription et à leur analyse, procédant ainsi à l'étude de l'évolution des pratiques linguistiques orales françaises et créoles, ainsi que des différents contacts entre ces deux langues. Comme son nom l'indique, elle est parrainée par M. Francard et son équipe Valibel.

Ainsi, des ressources linguistiques existent, mais elles sont encore trop rares, et leur disponibilité pourrait être améliorée.

6.2.2 Le breton

De la famille des langues celtiques, le breton est bien représenté parmi les langues régionales de France. Comme on peut le voir dans les tableaux de statistiques, le breton figure dans les places les plus hautes du classement, avec notamment 420 ressources uniques inventoriées. À l'inverse de la plupart des langues d'Outre-Mer, on trouve parmi ces ressources de petits enregistrements audio de quelques minutes, mais également d'importants corpus alignés pouvant servir de base à des technologies de la langue. Parmi les ressources audio, nous avons par exemple les enregistrements effectués par M. Jean Le Dû lors d'une enquête dialectologique réalisée en Bretagne, en vue de constituer le Nouvel Atlas Linguistique de la Basse-Bretagne.

L'importance de la communauté bretonne, le nombre élevé de locuteurs bretons (172 000 en 2009), ainsi que l'effort entrepris pour faire perdurer la langue à travers l'enseignement, la culture et les médias, font que beaucoup de ressources linguistiques existent, et sont disponibles pour la communauté scientifique.

6.2.3 L'occitan

L'occitan se compose de plusieurs variétés dialectales : le gascon, le languedocien, le provençal, l'auvergnat, le limousin, et le vivaro-alpin. L'une des grandes difficultés de ce groupe concerne ses locuteurs. Il est très difficile d'obtenir des chiffres fiables concernant l'importance des communautés composant les différentes variétés du groupe. Ainsi, il est estimé qu'environ deux millions de personnes font partie des locuteurs de l'occitan, mais sans plus de précisions à propos de chaque variété dialectale.

Cela soulève un autre problème que l'on rencontre pour la plupart des langues régionales, que cela soit en France métropolitaine ou en Outre-Mer : il est difficile d'identifier clairement un groupe de locuteurs, tout comme il est compliqué d'identifier et de normer une langue maternelle régionale par rapport au Français. C'est ce qui se passe pour l'occitan. On peut remarquer que beaucoup de ressources ont été identifiées pour le gascon, faisant d'elle la variété la plus représentée de cet inventaire (286 entrées dans la base de données, allant de l'enregistrement audio au lexique de taille conséquente). De même, le languedocien et le limousin occupent des places importantes dans ce classement, avec respectivement 202 et 27 entrées. Enfin, les trois autres dialectes de l'occitan, à savoir l'auvergnat, le provençal et le vivaro-alpin, comptabilisent respectivement 16, 30 et 22 entrées (sachant que certaines de ces entrées sont des ressources multilingues). Ces chiffres illustrent donc ce qui se passe pour le limousin, l'auvergnat, le provençal et le vivaro-alpin : les communautés linguistiques sont difficilement identifiables, et les dialectes le sont tout autant, dans la mesure où ils ne sont pas clairement distingués au sein de l'ensemble occitan. De ce fait, il est compliqué de pouvoir collecter des ressources linguistiques pour ces sous-ensembles.

6.2.4 Répartition des langues du focus et statistiques

Dans la figure ci-dessous, inspirée des tableaux que l'on peut trouver dans le livre blanc de Meta -Net, nous avons voulu représenter la répartition des langues en termes de volume de données. On remarque donc que le français, à gauche, est plutôt bien représenté, et les langues kanakes, en comparaison, n'ont que peu de ressources linguistiques à leur disposition pour leur assurer une présence forte au sein des langues régionales de France métropolitaine et d'Outre-Mer. Ce tableau tente donc de rendre compte de la place des langues les unes par rapport aux autres, au vu des ressources identifiées lors de cet inventaire.

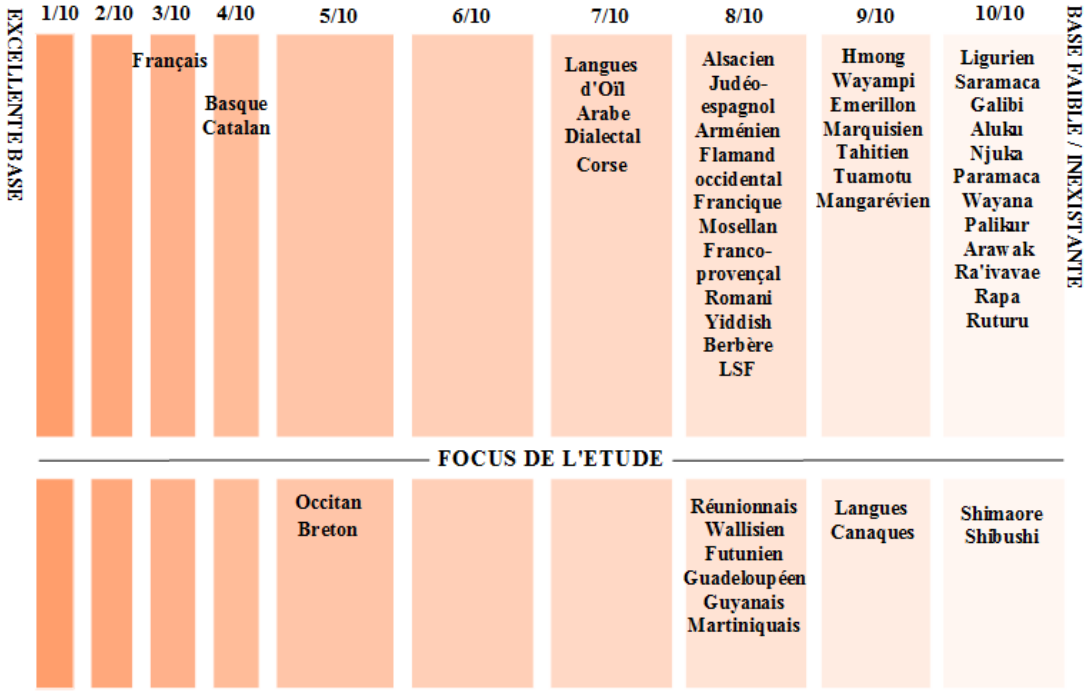

Figure 11 : Classement des langues régionales de France et d'Outre-Mer en termes de volume de ressources

Quant au nombre d'entrées pour les langues de ce focus, bien qu'elles aient déjà été présentées dans le tableau général, voici un récapitulatif des différentes entrées dans la base de données.

LANGUES	CORPUS PAROLE	CORPUS TEXTE	CORPUS MULTIMÉDIA	CORPUS PARALLÈLE	LEXIQUE	ONTOLOGIE	THÉSAURUS	TOTAL
Langue d'oc ou occitan	365	92	29	33	102		1	622
Breton	403	17						420
Langues kanakes	125	96	9		1			231
Futunien	23	21						44
Wallisien	20	19						39
Réunionnais	31	4						35
Guadeloupéen	7	7						14
Guyanais	4							4
Mahorais (shimaoré)	2	2						4

Table 4 : Répartition des types de ressources par famille de langue

6.3 APPLICATION DES TECHNOLOGIES À CES LANGUES : ÉTUDE DE LA FAISABILITÉ

Au vu des ressources disponibles pour les langues sur lesquelles nous nous sommes focalisés, il convient maintenant de mettre cela en perspective avec les différentes technologies de la langue que nous avons choisi pour ce focus, à savoir : la traduction automatique, la synthèse / la reconnaissance vocale et la correction orthographique.

Pour que des technologies existent, il y a une composante essentielle : les ressources linguistiques. Tous ces systèmes nécessitent des corpus, des grammaires ou des lexiques pour pouvoir fonctionner efficacement. Ainsi, cette étude de la faisabilité des technologies pour les langues régionales va se baser sur les statistiques évoquées précédemment.

6.3.1 Traduction automatique

La traduction automatique possède à l'heure actuelle plusieurs grands systèmes. Parmi les plus connus, nous retrouvons Google Traduction, Systran, Reverso ou encore PROMT. Il existe également des solutions gratuites et open source, telles que Moses, ou Apertium. Afin de permettre l'entraînement de ces systèmes, des corpus parallèles ont été créés dans diverses paires de langues. Les langues les plus représentées en France sont le catalan, le basque et bien sûr, le français. Ainsi, au vu des ressources existantes pour le breton et l'occitan, nous pouvons tout à fait envisager que soient créés des corpus alignés dans ces langues, afin d'alimenter un système de traduction automatique. Les corpus existent, et les aligner entre eux est une tâche qui peut être envisagée à court ou moyen terme.

6.3.2 Synthèse et reconnaissance vocale

Concernant la synthèse et la reconnaissance vocale, la situation est un peu différente, dans la mesure où les besoins de cette technologie ne sont pas les mêmes que pour la traduction automatique. Ici, les systèmes requièrent d'importants corpus de parole pour pouvoir fonctionner efficacement, car de nombreux phénomènes de l'oral ont besoin d'être appris par les systèmes de synthèse et de reconnaissance vocale afin d'être efficaces. En plus de ces corpus, les systèmes de synthèse / reconnaissance vocale nécessitent des lexiques ainsi que d'importants corpus de texte. Aujourd'hui, pour les langues sur lesquelles nous avons choisi d'appuyer notre analyse, cette technologie est difficilement envisageable à court ou à moyen terme, car toutes les ressources nécessaires ne sont pas disponibles, comme c'est le cas pour les corpus de parole. De ce fait, nous aurions tout d'abord besoin de constituer d'importants corpus oraux, avant de pouvoir les utiliser dans de tels systèmes, et cela représente une tâche importante et fastidieuse en soi.

6.3.3 Correction orthographique

Enfin, nous avons choisi d'évoquer la correction orthographique (ou assistance à la rédaction). Cette technologie nécessite d'importants lexiques, ou dictionnaires, pour pouvoir fonctionner efficacement. Des correcteurs orthographiques existent déjà, notamment dans les logiciels de traitements de texte tirés des suites logicielles *LibreOffice* ou *Microsoft Office*. Dans le cas de *LibreOffice*, les ressources linguistiques existent également, du fait de l'aspect libre de ce logiciel : la communauté d'utilisateurs est largement invitée à participer aux traductions des dictionnaires requis par le système pour fonctionner, et de ce fait, les langues comme le breton ou l'occitan sont traitées et des lexiques ont été créés pour le correcteur orthographique de *LibreOffice*. De fait, cette technologie est tout à fait envisageable à court terme, car les ressources existent et les systèmes aussi.

6.3.4 Répartition des langues en fonction des technologies ciblées

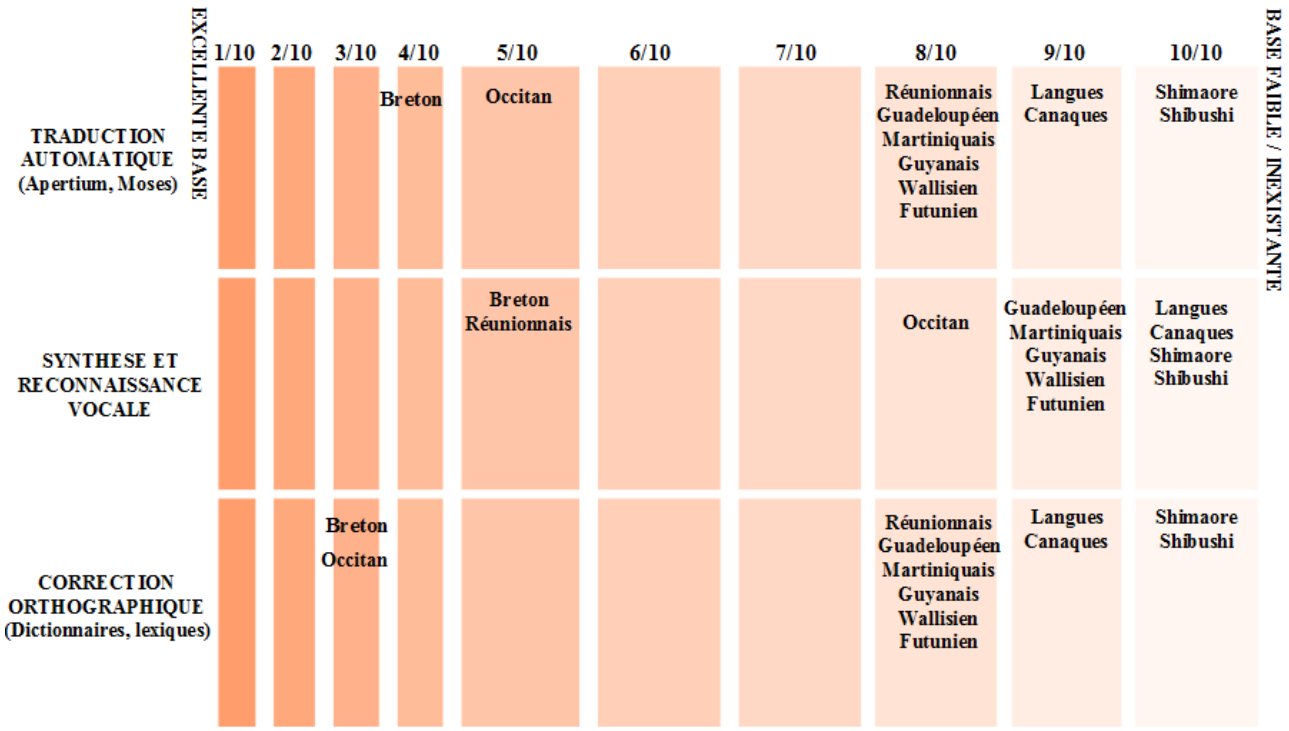

Figure 12 : Répartition de la faisabilité des technologies pour les langues inventoriées

Dans la figure ci-dessus, également inspirée des tableaux de Meta-Net, nous avons tenté de représenter la faisabilité des technologies étudiées pour les langues sur lesquelles nous avons appuyé notre analyse. On peut donc voir que la traduction automatique est tout à fait envisageable pour des langues comme le breton ou l'occitan, mais qu'elle l'est beaucoup moins pour les langues de Mayotte. Dans une autre mesure, la synthèse et la reconnaissance vocale sont plus difficilement envisageables à court terme, car les besoins ne sont pas les mêmes, et les ressources n'existent pas forcément pour toutes les langues régionales. Une fois encore, on peut voir que les langues d'Outre-Mer sont les moins bien représentées, et de telles technologies sont pour le moment peu envisageables. Enfin, la correction orthographique fait partie des technologies pour lesquelles on peut considérer qu'à court terme des langues soient traitées. En effet, de nombreux lexiques existent et sont disponibles, et les systèmes sont déjà au point. Ainsi, le breton et l'occitan sont une fois encore les plus à même de bénéficier de cette technologie, tandis que la plupart des langues d'Outre-Mer ne disposent que de peu de ressources linguistiques leur permettant d'envisager des solutions de correction orthographique.

7 BILAN ET RECOMMANDATIONS

Au vu des résultats de cet inventaire, nous pouvons affirmer que la situation est très différente d'une langue à une autre. Certaines langues régionales disposent d'importantes ressources linguistiques, comme le breton, le catalan ou encore le basque, tandis qu'une grande partie des langues d'Outre-Mer ne dispose que de peu de ressources. De plus, les chiffres ne rendent pour le moment pas compte du volume des données comprises dans la base. Une entrée peut désigner à la fois un fichier audio de deux minutes, comme un corpus parallèle de deux millions de mots, dans quatre langues différentes. La composante « volumétrie » est un élément qu'il faudra prendre en compte par la suite, afin d'étoffer cette étude, et d'obtenir des statistiques plus poussées quant au volume des ressources disponibles pour une langue donnée.

Il semble évident qu'en fonction des régions de France, il existe des disparités à identifier au niveau social, économique et linguistique, et il conviendrait de trouver des applications technologiques concrètes qui permettraient de venir combler ces manques. Comme nous l'avons évoqué, certaines politiques régionales font que les langues maternelles ne sont pas reconnues, comme en Guyane ou à Mayotte où peu de moyens sont accordés à l'expansion de ces langues. Dans d'autres cas, comme pour l'auvergnat, la principale difficulté concerne aussi bien l'identification de la communauté de locuteurs que la normalisation de la langue par rapport à la langue officielle de la République, le français.

Malgré cela, de nombreuses initiatives ont été mises en place, à l'image de la base de données Watreng, dont s'occupe le département scientifique et technique de l'Académie des langues kanak[12]. Cette base de données vise à regrouper un grand nombre de données numériques sur l'ensemble des langues kanakes. Et même si cette base n'en est encore qu'au stade du développement, nous pourrions probablement l'intégrer à la base de données de ce projet d'inventaire, afin de participer à la promotion de ce qu'entreprend l'Académie des langues kanak.

Dans le cas de langues créoles, il a été démontré que, via un enseignement bilingue créole-français, l'oral permet aux jeunes élèves de mieux appréhender l'écrit, tant en français qu'en créole. Il est donc envisagé, dans certaines régions d'Outre-Mer, de mettre en avant l'enseignement bilingue plutôt que monolingue (en français) afin de permettre une meilleure approche du créole, tant à l'oral qu'à l'écrit. À cela, nous pourrions ajouter le fait que très peu de corpus de qualité sont accessibles aux jeunes élèves d'Outre-Mer, ce qui freine d'autant plus la sauvegarde de ces langues qui sont déjà plus ou moins en voie d'extinction.

À cela viennent s'ajouter plusieurs questions : tout d'abord, faut-il normer une langue pour les

12 http://www.alk.gouv.nc/portal/page/portal/alk

besoins d'une technologie ? Quel pourrait être l'impact d'une telle normalisation sur la pratique d'une langue ?

Les Conseils de la Culture, de l'Éducation et de l'Environnement (CCEE) fondés en 1984 pour les DOM-TOM, puis en 2004 pour Mayotte, ont participé à l'établissement d'une écriture du mahorais (l'une des deux langues de Mayotte, de tradition essentiellement orale). Puis, actuellement, ils tentent de renouveler l'expérience pour une écriture consensuelle du réunionnais, à partir des quatre graphies existantes. Ce type d'initiative ne peut qu'être bénéfique au développement d'une langue, dans la mesure où l'écrit constitue un très bon moyen de transmission d'une langue à long terme. Dans le même ordre d'idée, un premier dictionnaire papier drehu-français a été récemment constitué, grâce à l'initiative de l'Académie des langues kanak basée en Nouvelle-Calédonie.

Toutes ces initiatives font partie des recommandations que nous pourrions faire afin d'assurer la sauvegarde des langues régionales, dans la mesure où elles permettront probablement la création de ressources linguistiques solides pour des systèmes de traduction automatique, de synthèse vocale, ou encore de génération de langue.

Autre point important : les technologies évoquées dans ce rapport sont en constante évolution, et cela permet aux acteurs du domaine linguistique informatique de traiter plus facilement et plus rapidement les données de langage. Nous pouvons d'ailleurs à ce sujet suggérer d'effectuer des productions « test » sur une langue sélectionnée parmi les langues d'Outre-Mer, afin de proposer un premier échantillon de ce que peut donner une ressource linguistique. Dans l'optique de cette production, il est tout à fait envisageable de cibler des applications liées aux besoins locaux, ainsi qu'aux capacités techniques locales (téléphones portables, mise en place d'un service d'accueil téléphonique avec reconnaissance vocale, système de recherche d'informations en bibliothèques, etc.).

Cette production peut évidemment se faire sur la base d'une collaboration entre organismes locaux et spécialistes du traitement de la langue. De nombreux organismes de recherche pourraient apporter leur expertise technique dans ce projet.

Enfin, nous suggérons d'approfondir cette étude lors d'une nouvelle phase, axée sur l'aspect technologique de l'inventaire, en proposant également un inventaire des systèmes traitant des langues d'Outre-mer.

8 ANNEXE

L'ensemble de l'inventaire, présenté sous forme de tableau, est disponible sur le CD qui est joint au présent rapport. Il faut observer que le nombre de ressources calculé est basé sur le nombre d'entrées dans la base de données. Une ressource peut contenir plusieurs types, langues, modalités, applications. De ce fait, chaque combinaison de caractéristique comptera comme une entrée dans le fichier tabulé.

9 BIBLIOGRAPHIE

Mariani et al., 2012	J. Mariani, P. Paroubek, G. Francopoulo, A. Max, F. Yvon, P. Zweigenbaum, *La langue française à l'ère du numérique,* 2012
Gandcher et al., 2008	F. Gandcher, O. Hamon, V. Mapelli, N.Moreau, N.Paulsson, D. Mostefa, *Réalisation d'un guide de production de ressources linguistiques pour la veille,* 2008
Ramchetty, 1998	R. Ramchetty, *Rapport sur l'état de la coopération régionale,* 1998