Feuille de route pour le développement numérique occitan

Benoît Dazéas

Lo Congrès permanent de la lenga occitana, Château d'Este, BP 326, 64141 Billère cedex

b.dazeas@locongres.org

Résumé

Le Livre blanc de META-NET, un réseau d'experts européens en technologies de la langue, alerte sur le risque « d'extinction numérique » de plusieurs langues européennes et de l'urgence pour elles de se doter rapidement de technologies de support. Cette étude propose également une grille de classification et d'évaluation des ressources et préconise des principes d'action tels que la création massive de données, la mutualisation ou encore le transfert technologique.
Dans ce cadre *Lo Congrès permanent de lenga occitana* a piloté la rédaction d'une feuille de route pour le développement numérique de l'occitan. Le rapport final fait état des ressources existantes et propose une planification de réalisation (2015-2019) des ressources de bases et des outils finaux.
La mise en place de cet ambitieux programme nécessitera la coordination des acteurs de transmission de l'occitan – politiques linguistiques, recherche scientifique et communauté du logiciel libre – ainsi que la mobilisation des différents crédits et fonds européens.

Abtract

Roadmap for the Occitan digital development
The META-NET White Paper Series, issued by the European network of experts in language technology, alerts to the risk of digital extiction for several European languages and the emergency to get equipped with supporting technologies. The study also presents a classification scale and assessment of the ressources and recommends taking actions, like massive data import, resource pooling, or technological tranfert.
According to this, *Lo Congrès permanent de la lenga occitana* steered the drawing up of a roadmap for the Occitan digital development. The final report lists the existing resources and schedules the release of each basic resource and final tools (2015-2019).
The implementation of this ambitious workprogram will require a teamwork of all partners involved in the transmission of Occitan – language policies, scientific research and open-source application community – as well as gathering various credits and European Funds.

Mots-clés : Langues régionales et minoritaires, occitan, feuille de route, ressources langagières, technologies de la langue, politiques linguistiques.
Keywords : Regional and minority languages, Occitan, language ressources, language technologies, language policies.

1 Introduction

1.1 L'occitan, une langue européenne

L'occitan est une langue romane parlée dans trois États de l'Union européenne (France, Espagne, Italie) sur un espace d'environ 150 000 km². Sur les 15 millions d'habitants concernés. Il est difficile d'en dénombrer les locuteurs ; à partir des différentes études[1] conduites ces dernières années, partielles et étalées dans le temps, on situe, selon les sources[2], le nombre de locuteurs entre plusieurs centaines de milliers à plusieurs millions de personnes.

En l'absence de standard imposé officiellement et du fait de la vitalité de certaines variétés dialectales, l'occitan peut être défini comme une langue polynomique, composée de six grandes variétés dialectales[3].

1 A noter l'enquête sociolinguistique conduite par la région Aquitaine en 2008 (http://www.aquitaine.fr/content/download/786/7753/file/Enquete_linguistique(1).pdf) ou encore celle de la région Midi-Pyrénées e ; 2010 (http://www.midipyrenees.fr/IMG/pdf/EnqueteOccitan.pdf).
2 http://www.univ-montp3.fr/uoh/occitan/une_langue/co/module_L_occitan_une%20langue_10.html
3 Auvergnat, gascon, languedocien, limousin, provençal, vivaro-alpin (BEC, 1995).

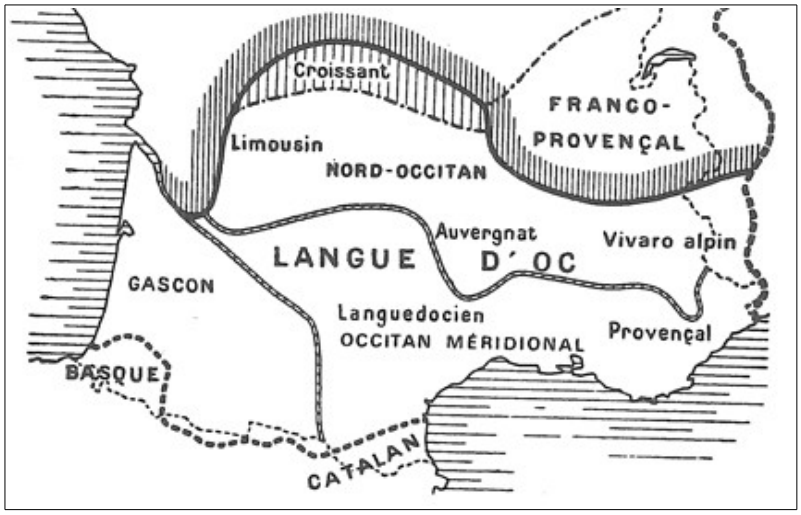

FIGURE 1 – Classification des dialectes occitans selon Pierre Bec

Co-officielle en Val d'Aran, la langue occitane bénéficie, à défaut d'une reconnaissance publique, du soutien des collectivités territoriales en France et fait partie des langues protégées par la loi sur les minorités linguistiques en Italie.

Riche d'une littérature écrite millénaire, l'occitan est aujourd'hui présent dans la presse, sur Internet et à la télévision. Soutenue par une réseau associatif et institutionnel dense, elle est enseignée de la maternelle (enseignement immersif associatif ou bilingue public) jusqu'à l'Université.

1.2 Lo Congrès, une institution collégiale pour réguler l'occitan

Lo Congrès permanent de la lenga occitana est l'organisme interrégional de régulation de l'occitan. Il rassemble les institutions et fédérations historiques du territoire occitanophone et il est soutenu par la Délégation à la langue française et aux langues de France (ministère de la Culture et de la Communication) et les collectivités territoriales. Installé officiellement à l'hôtel de Région Aquitaine à Bordeaux en décembre 2011, il a pour mission de contribuer à la vitalité et au développement de l'occitan – appelé aussi langue d'oc – en travaillant à sa connaissance et à sa codification par la production des outils concernant les différents aspects de la langue (lexicographie, la lexicologie, la terminologie, la néologie, la phonologie, la graphie, la grammaire et la toponymie).

Lo Congrès possède deux organes assesseurs – le Conseil linguistique[4] (dont le président est Patrick Sauzet, linguiste et professeur à l'Université Toulouse 2) et le Conseil des usagers[5] – et agit selon des principes d'action tels que le respect de l'unité et de la diversité de l'occitan, la stabilité, la représentativité des régions linguistiques du territoire d'Oc, la collégialité des décisions et la diffusion de l'information.

Afin de répondre à la demande urgente des usagers, plus spécifiquement ceux du domaine de l'enseignement et de la formation pour adultes, le Congrès a développé une plate-forme numérique – locongres.org – rassemblant différents outils linguistiques de références : un multidictionnaire occitan (*dicod'Òc*), un conjugateur (*vèrb'Òc*), une base terminologique (*tèrm'Òc*), une base toponymique *(top'Òc*), un corpus textuel ainsi qu'un portail d'accès vers les différentes ressources occitanes en ligne.

Avec plus de 180 000 visites en 2014, le portail numérique *locongres.org* est pensé comme un service public en langue occitane : son accès est gratuit, multiplate-formes (*Windows*, *iOS*, *Androïd*, etc.), les formats libres et les licences

4 La Communauté scientifique est représentée au Congrès par le Conseil linguistique. Ce conseil assesseur est déjà constitué et ses membres sont à l'œuvre sur différents travaux. Toutes les régions occitanes y sont représentées. Le Conseil linguistique a un Président et un bureau élu, ainsi que des commissions qui travaillent pour les besoins du Congrès. Liste des membres : http://www.locongres.org/index.php/fr/lo-congres-fr/le-conseil-linguistique/membres

5 Le Conseil des usagers est un conseil assesseur du Congrès ayant pour fonction de représenter la demande sociale. Il rassemble des personnes qualifiées représentatives de la pratique sociale de la langue et qui sont réparties en trois secteurs : les transmetteurs (enseignement, cours pour adultes et formation professionnelle), les utilisateurs (écrivains, éditeurs, médias) et les institutionnels (opérateurs de politiques publiques).

contributives sont privilégiés.

1.3 Les enjeux du développement numérique

Lo Congrès fait partie d'une dynamique qui a permis au numérique occitan de se développer d'une façon générale ces dernières années : contenus encyclopédiques (*Wikipédia*[6]), patrimoine (*Occitanica*[7], *Sondaqui*[8], *trobadours d'Aquitaine*[9]), médias (*Octele*[10]), réseaux sociaux sont autant de secteurs désormais investis. Toutefois la langue occitane pâtit toujours d'un important retard numérique avec pour conséquence, une absence quasi totale dans des outils désormais courants (bureautique, téléphonie mobile, etc.). La prégnance croissante de ces technologies dans la vie quotidienne (travail, déplacements, consommation, éducation, vie sociale) font des technologies du langage un facteur supplémentaire de marginalisation pour une langue déjà minorisée.

Ce phénomène est décrit et analysé dans une étude réalisée par META-NET, un réseau de recherche rassemblant différentes institutions, universités et centres de recherche et dont la mission principale est la mise en place de fondations technologiques solides pour une Europe multilingue. Son Livre blanc[11] fait un état actuel des ressources et technologies du langage pour trente langues européennes dans six domaines (la traduction automatique, la synthèse et la reconnaissance vocale, la correction orthographique, l'analyse sémantique, l'analyse grammaticale et la génération automatique de texte) et propose également une grille commune de classification et d'évaluation des ressources et outils numériques. Les résultats de l'étude sont particulièrement alarmants : les éditeurs soulignent l'écart croissant entre les « grandes » et les « petites » langues, il est indispensable d'équiper toutes les langues (y compris les plus petites et les moins dotées) des technologies de base nécessaires, sans quoi ces langues sont condamnées à « l'extinction numérique ». Pour ce faire, **l'étude préconise la création massive de données, la mutualisation au niveau européen, le transfert technologique entre les langues, l'interopérabilité des ressources, des outils et des services** (REHM, USZKOREIT, 2002 ; SORIA *et al.*, 2013).

Pour ce qui concerne la France, faisant suite à la proposition d'action n°26 du rapport Jacques ATTALI intitulé *La francophonie et la francophilie, moteurs de croissance durable*[12] et remis au Président de la République en août 2014, la Délégation générale à la langue française et aux langues de France (ministère de la Culture et de la Communication) a souhaité lancer un nouveau volet du programme *Technolangue*[13], dont la première édition a permis, entre 2003 et 2005, d'accompagner plusieurs projets d'outillage en traitement automatisé pour la langue française. Ce nouveau programme vise à compléter l'outillage pour la langue française et à développer de nouvelles technologies de traitement pour les langues de France. Dans ce cadre, la DGLFLF, en partenariat avec le CNRS et ELDA, a organisé les 19 et 20 février 2015 un colloque[14] sur le thème du développement des technologies en faveur des langues régionales de France afin de constituer un atelier de réflexion, réunissant une cinquantaine de participants, dont des experts scientifiques, des représentants des collectivités territoriales et des membres d'associations de soutien aux langues régionales.

La DGLFLF a également réalisé en partenariat avec ELDA un inventaire[15] des ressources linguistiques des langues régionales en reprenant le classement établi dans le Livre Blanc de META-NET (SORIA, MARIANI, 2013). Ces travaux comprennent également une étude de la faisabilité de l'application des technologies : l'occitan, qui serait proche du breton en termes de classification, disposerait ainsi des ressources suffisantes pour développer des outils de traduction automatique ou la correction orthographique.

2 Diagnostic et feuille de route pour le développement numérique de la langue occitane

Parallèlement à ces travaux, *Lo Congrès* a initié une recherche-action associant à la fois ses partenaires publics (DGLFLF et collectivités membres du Congrès[16]) et les opérateurs de recherche, de transmission et de diffusion de la langue (Université, opérateurs de missions publiques, formation professionnelle, médias, etc.) autour de la question de la stratégie de développement numérique pour l'occitan, plus spécifiquement pour le domaine du Traitement Automatique des Langues (TAL). L'objectif était, en huit mois de travaux (avril-novembre 2014), de réaliser un

6 http://oc.wikipedia.org
7 http://www.occitanica.eu
8 http://www.sondaqui.com
9 http://www.trobar-aquitaine.org
10 http://www.octele.com
11 http://www.meta-net.eu/whitepapers/press-release-fr
12 Rapport téléchargeable sur le site de la documentation française : http://www.ladocumentationfrancaise.fr/rapports-publics/144000511/
13 http://www.technolangue.net
14 Programme du colloque : http://tlrf2015.sciencesconf.org/
15 Rapport téléchargeable sur le site de la DGLFLF : http://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
16 Ministère de la Culture et de la Communication-DGLFLF, Régions Aquitaine, Midi-Pyrénées, Languedoc-Roussillon, Départements des Pyrénées-Atlantiques et des Hautes-Pyrénées, ville de Toulouse.

diagnostic (inventaire) des ressources et outils linguistiques existants et une feuille de route de développement 2015-2019. Des experts internationaux des TAL pour les langues basque, bretonne, catalane et gallois ont été également associés aux travaux.

L'inventaire des ressources linguistiques[17], réalisé en collaboration avec ELDA, a été publié en juin 2014 (il reste un chantier ouvert). Avec près de 250 références, il permet de faire le diagnostic suivant :

— Il existe très peu d'outils de technologie de la langue en occitan. Les outils existants se concentrent dans la catégorie des correcteurs orthographiques.

— Les ressources recensées sont plus nombreuses, mais peu peuvent être réutilisées :

> — La plupart des ressources lexicales sont constituées de dictionnaires numérisés, mais la plupart sont anciens ou non validés au niveau linguistique. Les autres demandent un gros traitement avant de pouvoir être utilisés en traitement automatique des langues (TAL).
>
> — Les corpus sont nombreux, mais il faut également les trier en fonction de leur qualité linguistique. Par ailleurs, ils ne sont pas, pour la plupart, directement utilisables pour créer des outils (les corpus oraux ne sont pas transcrits, les corpus textuels ne sont pas annotés).

— Les grammaires sont destinées à une utilisation papier davantage qu'à une utilisation informatique.

Il est donc indispensable de créer des ressources linguistiques de base avant de pouvoir développer des outils à partir de ces ressources.

Ressources linguistiques	**Recensées**	**Utilisables en informatique***
Corpus monolingues de textes	27	2
Corpus monolingues de parole	28	0
Corpus parallèles	1	1
Corpus multimédias et multimodaux	24	0
Lexiques	73	8
Bases terminologiques	21	2
Tesauri, Wordnets, ontologies	1	1
Toponymie	3	3
Grammaires, modèles de langage	30	0
Outils de technologie du langage	**Recensés**	
Reconnaissance de l'écriture	0	
Reconnaissance de la parole	0	
Synthèse vocale	0	
Analyse grammaticale	6	
Analyse sémantique	0	
Génération de texte	0	
Traduction automatique	2	
Recherche et extraction d'information	0	
Autres outils	**Recensés**	
Logiciels disponibles en occitan	6	
Outils numériques pour apprendre l'occitan	4	

FIGURE 2 – Diagnostic du développement numérique de la langue occitane (synthèse)

17 http://inventari.locongres.org

La feuille de route pour le développement numérique occitan 2015-2019[18], basée sur le Livre blanc META-NET, a été présentée officiellement le 28 novembre au Congrès à Billère ; cette dernière propose une stratégie harmonisée sur cinq ans (2015-2019) pour le développement des technologies linguistiques de l'occitan, et fournit un cadre cohérent duquel peuvent dériver des actions concrètes pour sa mise en œuvre. Heureusement, l'occitan n'est pas à un point de développement zéro et les initiatives prises à ce jour devraient constituer un point de départ intéressant.

Il est donc essentiel de réutiliser les ressources et les outils existants, ce qui permettra de cibler les efforts à venir. Cette stratégie, basée sur l'optimisation de la coopération entre les différents acteurs, exigera un effort collectif pour veiller à atteindre les objectifs fixés en 2019.

	2015	**2016**	**2017**	**2018**	**2019**
Ressources linguistiques					
Corpus textuels					
Monolingues					
Corpus spécialisés (10-25)			x*		x**
Corpus web (5 millions)			x		
Parallèle (2-5)				x	
Ressources lexicales					
Base lexicale monolingue	x***		x		
Base lexicale bilingue			x		
Grammaires					
Base grammaticale/syntaxique		x			
Outils linguistiques					
Traitement de la parole					
Ressources pour la reconnaissance de la parole					x
Synthèse vocale					x
Détection automatique de la langue					
Détecteur de l'occitan	x				
Détecteur des variantes de l'occitan		x			
Analyse grammaticale					
Correcteurs orthographiques					
Correcteur orthographique polyvalent (toutes les variantes)		x			
Clavier prédictif et autocorrection			x		
Analyseurs					
Lemmatiseur-analyseur morphologique		x			
Analyseur syntaxique					x
Analyse sémantique					
Base de connaissance lexicale				x	
Traduction automatique					
Traducteurs automatiques					
oc --> fr (toutes les variantes)			x		
fr--> oc					x
Transcripteur automatique entre variantes			x		

18 Rapport complet : http://locongres.org/images/docs/feuille_route_numerique_occitan_fr.pdf

Logiciels					
OS + Applications principales			x		

** Corpus monolingue : première version (10 millions de mots), ** Corpus monolingue : deuxième version (25 millions de mots), *** Base lexicale monolingue : première version basique nécessaire pour le développement du correcteur et du lemmatiseur*

FIGURE 3 – Feuille de route 2015-2019 pour le développement numérique de la langue occitane

3 Données de cadrage

3.1 L'occitan, langue romane et européenne

L'occitan est une langue romane proche du catalan, langue déjà bien outillée. Il serait intéressant d'étudier les possibilités de transfert technologique entre les deux langues. Il existe un traducteur automatique catalan-occitan languedocien (*Opentrad*[19]) – pouvant encore être amélioré, permettant par exemple d'envisager – avec une phase postérieure de correction manuelle – un traitement massif de corpus textuels. De même, il est envisageable et même souhaitable de s'appuyer sur les développements réalisés pour les autres langues romanes – dont les syntaxes par exemple restent finalement assez proches – et en premier lieu le français.

L'occitan est également une langue partagée par trois États (France, Espagne et Italie), la question de son développement est donc « d'intérêt » européen, avec de plus des opérateurs de missions publiques basés sur des territoires éligibles aux fonds de coopération transfrontalières. La langue occitane est également voisine au sens géographique de deux autres langues dites « minorisées » – encore que co-officielles et disposant déjà d'institutions linguistiques et de technologies de support proche de certaines langue d'État : le catalan et le basque. Sur les questions linguistiques, on a vu ces dernières années se développer des deux côtés de la frontière des actions d'échange, de mutualisation voire même de transfert de technologie : on peut mentionner le partenariat entre le *Congrès*, *Elhuyar*[20] et la société *Media.kom* – avec le soutien de l'Eurorégion Aquitaine-Euskadi – autour de la création d'un dictionnaire référentiel et orthographique occitan (le *Basic*[21]), un corpus textuel en ligne[22] ainsi que la première version de traducteur automatique occitan-français (en partenariat avec la société de presse *Vistedit* et le soutien du fonds SPEL, en cours de développement) mais également le diagnostic et la feuille de route de développement numérique de l'occitan 2015-2020 (avec la participation de représentants de l'Université du Pays Basque et de l'Université Polytechnique de Barcelone). *Le Cirdòc – médiathèque occitane* collabore depuis plusieurs années avec les centres de ressources des occitanophones italiens (vallées du Piémont) et espagnols (Val d'Aran) dans le cadre du développement de la plate-forme numérique *Occitanica*[23]. Il est également, dans le cadre de l'Eurorégion Pyrénées-Méditerranée, chef de file du forum interrégional « patrimoine et création » (réseau d'opérateurs autour du patrimoine et de la création). Enfin, on peut citer également le partenariat entre *l'InÒc Aquitaine* et *l'Institut d'Études occitanes* (Toulouse) avec le *Termcat* (Barcelone) autour de la question du développement de la terminologie en langue occitane[24].

3.2 La question des standards

Les technologies du langage utilisent des standards (graphiques, lexicaux, grammaticaux). Or, l'occitan a la particularité de ne pas avoir de variante dite « standard » mais est au contraire d'être composé de plusieurs « grandes variantes » (auvergnat, gascon, languedocien, limousin, provençal et vivaro-alpin). De même, pour ce qui concerne la graphie, Lo Congrès utilise, codifie et diffuse la graphie dite « classique[25] » par ses productions (dont un dictionnaire d'orientation pan-occitan – *lo Basic* – et un conjugateur). De par sa mission d'organisme de régulation de la langue occitane confiée par ses partenaires publics, *Lo Congrès* s'emploie à répondre aux besoins de la transmission de la langue en stabilisant les formes en graphie classique autour des grands espaces linguistiques. Cela n'écarte en rien les autres systèmes, les technologies permettant d'envisager le développement de transcripteurs graphiques et dialectaux. Il conviendra d'en étudier la faisabilité technique et budgétaire, ainsi que les besoins réels.

3.3 La recherche scientifique

L'occitan dispose d'un réseau international de chercheurs (rassemblés au sein de l'Association internationale d'études

19 *Opentrad* est une plateforme de traduction automatique basée sur le moteur au code source ouvert *Apertium* : http://www.opentrad.com
20 https://www.elhuyar.eus
21 http://locongres.org/index.php/fr/lo-congres-fr/les-chantiers/la-basic/introduction
22 http://corpus.locongres.org
23 http://www.occitanica.eu
24 Dans le cadre de la céation d'un lexique du transport touristique pour plusieurs langues latines minorisées (occitan, corse, sarde, etc.) : http://www.termcat.cat/ca/Diccionaris_En_Linia/36/Fitxes/
25 Autres graphies : mistralienne, fébusienne, escolo dóu Po, etc.

occitanes et pour certains présents au Conseil linguistique du Congrès), plusieurs départements à l'Université ainsi que deux laboratoires actifs dans les domaines des TAL :

— CLLE est un laboratoire de recherche en psychologie et en linguistique. Sa composante, CLLE-ERSS (Cognition, Langues, Langage, Ergonomie – Equipe de Recherche en Syntaxe et Sémantique, basée à l'Université Toulouse 2) a pour visée la description scientifique et la modélisation des langues naturelles (phonologie, morphologie, syntaxe, sémantique, lexique, discours). Depuis sa création, CLLE-ERSS s'engage dans la constitution et l'exploration de grands corpus langagiers, écrits ou oraux.

— *BaTelÒc* (base textuelle occitane) est un projet transversal CLLE-ERSS, sous la responsabilité de Myriam Bras, visant la construction d'une base de textes annotée en langue occitane, en partenariat avec le CNRTL, le CROM, CIEL d'ÒC, IEO/IDECO, lo Congrès Permanent de la Lenga Occitana, lo CIRDOC avec le soutien financier de la Région Midi-Pyrénées.

— CLLE_ERSS participe également avec LILPA – Université de Strasbourg, LESCALP – Université d'Amiens, LIMSI – Université Paris-Orsay au projet *Restaure*[26] (Ressources Informatisées et Traitement Automatique pour les langues régionales) ; il s'agit d'un programme ANR (2015-2018, appel à projet 2014) visant à développer des ressources et des outils de traitement automatique des langues (TAL) pour trois langues de France : le picard, l'alsacien et l'occitan.

— Le laboratoire UMR 730 *base, corpus, langage* associé au CNRS de l'Université de Nice travaille sur le corpus et exploite les données à des fins théoriques autour de quatre thématiques : dialectologie-phonologie-diachronie, logométrie et corpus politiques, médiatiques et littéraires, linguistique de l'énonciation, langage et cognition. Il est éditeur de l'importante base *Thésoc*[27] (Thesaurus occitan) : il s'agit d'un important programme comprenant une base lexicale annotée de 1,2 millions de mots avec un volet cartographique permettant de visualiser les variantes dialectales et un module morphosyntaxique (MMS).

3.4 Les opérateurs institutionnels

Il est à noter que l'occitan a trois opérateurs de missions publiques possédant une expertise technique pour la collecte, la numérisation, le traitement et l'édition numérique ainsi que de nombreuses ressources et outils :

— Le Congrès est spécialisé dans le traitement numérique des ressources linguistiques occitanes. Il possède des ressources linguistiques de référence (dictionnaires, modèles de conjugaison), son équipe professionnelle (lexicographie, développement et webmastering) et ses différents prestataires lui permettant d'assurer la chaîne complète de traitement de données : numérisation, « parsage », formatage, développement applicatif et édition numérique.

— *Le Cirdòc – médiathèque occitane* est un établissement public travaillant à la sauvegarde, à la valorisation et à la diffusion du patrimoine occitan. Pôle associé à la BNF pour la langue et la culture occitanes, *le Cirdòc* développe des actions inter-régionales autour du patrimoine et de la création occitans : la médiathèque numérique www.occitanica.eu, le développement de la numérisation des documents en partenariat avec les institutions de toutes les régions et leur diffusion dans le cadre d'Occitanica, la création d'outils de connaissance et de développement de l'occitan (Répertoire des fonds occitans, Bibliographie occitane), la conception et prêts d'expositions ou de ressources documentaires (service Question/Réponse, service aux chercheurs). Avec plus de 80 000 titres du XVIe siècle à nos jours (manuscrits, archives, livres, revues, partitions, enregistrements sonores et audiovisuels, estampes, affiches, photographies, objets, etc.), il est le grand conservatoire de la langue et culture occitane.

— *l'InÒc Aquitaine* a pour mission régionale la valorisation des ressources numériques de l'occitan, qu'il s'agisse des pratiques vivantes et des savoir-faire (Patrimoine culturel immatériel) en Aquitaine ou de fonds patrimoniaux anciens. Ethnopôle[28], il conçoit et réalise pour ce faire des projets éditoriaux en ligne (sites internet) dans le cadre d'*Aquitaine Cultures Connectées* (anciennement BnsA) : *Sondaqui*[29] (patrimoine oral et festif aquitain), *Troubadours d'Aquitaine*[30] et collabore avec la *Banque Numérique des ressources Pyrénéennes* (BNRP). Dans sa mission régionale de préservation et de valorisation du patrimoine sonore et audiovisuel occitan et de la mémoire collective en Aquitaine. Il accompagne,

26 http://lilpa.unistra.fr/fdt/projets/projets-en-cours/restaure/
27 http://thesaurus.unice.fr/
28 L'appellation ethnopôle est un label du ministère de la Culture et de la Communication attaché à une institution qui, en matière de recherche, d'information et d'action culturelle, œuvre à la fois au plan local et au niveau national. À travers cette appellation, la mission du patrimoine ethnologique entend, dans le cadre propre à chaque structure, promouvoir une réflexion de haut niveau s'inscrivant tout à la fois dans les grands axes de développement de la discipline ethnologique et dans une politique de constitution des bases d'une action culturelle concertée. Site internet : http://www.culturecommunication.gouv.fr/Politiques-ministerielles/Patrimoine-ethnologique/Ethnologie-en-region/Ethnopoles
29 http://www.sondaqui.com
30 http://www.trobar-aquitaine.org

pour ce faire, les collectivités dans les différentes étapes du processus de sauvegarde des archives (inventaire, numérisation, description, collecte). Son pôle *Langue et Société* travaille sur la terminologie (lexiques spécialisées), la toponymie et mè, e un important travail de traduction.

— Les 23 et 27 juin 2014, les Régions Aquitaine et Midi-Pyrénées ont respectivement approuvé en assemblée plénière la création de l'Office public de la langue occitane, un nouvel outil dédié à la promotion de la langue. L'objectif de ce G.I.P (Groupement d'Intérêt Public) sera d'assurer la sauvegarde et le développement de l'occitan, en travaillant à l'augmentation quantitative et qualitative du nombre de locuteurs. Il participera à la mise en œuvre d'une politique linguistique publique interrégionale. Cet outil commun, qui a pour vocation d'accueillir rapidement d'autres partenaires, comme l'État ou d'autres Régions, est en cours d'installation ; il remplacera à terme les régions membres dans leur soutien au Congrès, dont il deviendra le partenaire public privilégié.

3.5 La « communauté »

Bien que modeste, il existe une communauté numérique occitane , c'est-à-dire des particuliers participant à des projets collaboratifs et en code-source ouvert en langue occitane. Nous parlions plus haut de la communauté occitanophone de Wikipédia qui a produit près de 100 000 articles ainsi qu'une version occitane du Wikictionnaire (programme de dictionnaire collaboratif), Wikiccionari[31], possédant 30 000 entrées (ce qui au regard de la situation sociolinguistique n'est pas négligeable).

De même il existe une communauté travaillant à la localisation (traduction) de logiciels, malheureusement par faute de moyens pas toujours concertés et connus du grand public. L'association *Tot en Òc* a traduit *Firefox* et *Libre Office* en occitan, le système d'exploitation libre *Ubuntu* est traduit à 80 %. Nous pouvons citer également *Dicollecte*[32], un projet collaboratif visant à améliorer les dictionnaires orthographiques pour les logiciels libres (la version occitane comprend actuellement un peu plus de 65 000 entrées).

4 Vers un programme opérationnel

Le plan de développement numérique de la langue occitane est un programme complexe rassemblant différents acteurs et dispositifs administratifs et budgétaires. Il est indispensable d'être vigilant sur la qualité de la planification (ordre des réalisations) et de la coordination des différents acteurs[33]. Aussi, pour garantir le bon déroulement du programme, il conviendrait de s'assurer :

— au niveau de la maîtrise d'ouvrage du projet, de reconduire le Comité de pilotage politique de la feuille de route, qui pourrait s'élargir à d'autres participants (universités ? experts extérieurs ?). Espace de concertation entre des décideurs et des prescripteurs, il valide les objectifs, le calendrier et les moyens du programme.

— au niveau de la maîtrise d'œuvre, que le pilotage soit assuré par le Congrès permanent de la langue occitane. Il rassemble en son sein les différents acteurs concernés par le développement de cette feuille de route et a une vision transversale de la problématique du TAL. En tant qu'organisme de régulation de la langue occitane, il dispose également de ressources de référence en TAL indispensables le plaçant en position centrale pour la réalisation des différents objectifs. Pour des raisons de capacité et de moyens, *le Congrès* ne sera pas en mesure de porter seul le développement de la feuille de route. Cette maîtrise d'œuvre comportera donc sur les aspects techniques et opérationnels une assistance ainsi que des coproductions avec différents opérateurs : *le Cirdòc* (qui possède la surface financière nécessaire au portage et à la gestion des dossiers européens), Université de Toulouse Jean-Jaurès, *InÒc Aquitaine*, *Elhuyar* (pour les développements applicatifs) pour ne citer qu'eux.

Pour ce qui concerne la partie opérationnelle, le rapport propose un développement des ressources et outils que l'on pourrait diviser en trois catégories :

- Les ressources de base, dans lesquelles on distingue les corpus et les bases lexicales (monolingues et bilingues) ;
- Les outils intermédiaires (analyseur morphosyntaxique, analyseur syntaxique, reconnaissance d'entités nommées) ;
- Les outils finaux (correcteur orthographique, traducteur automatique).

La feuille de route et de son calendrier de réalisation d'objectifs permettent, en considérant les interdépendances entre ses différences ressources, de fixer des objectifs de réalisation.

31 http://oc.wiktionary.org
32 http://www.dicollecte.org
33 Les experts internationaux ont particulièrement insisté sur ces aspects lors de l'étude.

OBJECTIF	RESSOURCE/OUTIL NÉCESSAIRE
Corpus monolingue	Numérisation, OCR et conversion de texte à un format standard traitable par un analyseur
Corpus web	Détecteur de l'occitan Détecteur des variantes de l'occitan
Corpus parallèle	Collection de documents bilingues Mémoires de traduction (TMX)
Base lexicale monolingue	Dictionnaires monolingues au format électronique (MRD)
Base lexicale bilingue	Dictionnaires bilingues au format électronique (MRD)
Correcteurs orthographiques	Base lexicale monolingue
Analyseur morphologique (*tagger*, lemmatiseur)	Base lexicale monolingue Base grammaticale
Analyseur syntaxique	Analyseur morphologique Base grammaticale/syntaxique
Base de connaissance lexicale	Base lexicale monolingue
Traducteurs automatiques oc → fr (toutes les variantes)	Base lexicale bilingue Base grammaticale/syntaxique
Transcripteur automatique entre variantes	Base lexicale monolingue Base grammaticale/syntaxique

FIGURE 4 – Outils et ressources : interdépendances

Dans l'attente d'un programme opérationnel détaillé de développement de la feuille de route, deux travaux sont d'ores et déjà lancés : un lexique ouvert des formes fléchies d'occitan (partenariat CLLE-ERSS/Lo Congrès permanent de la lenga occitana) concernant, pour commencer, les variétés gasconne et languedocienne et un traducteur automatique occitan (gascon et languedocien)-français sur la plate-forme Opentrad en partenariat avec la fondation Elhuyar.

Conclusion

Si la langue occitane pâtit d'un retard important en TAL, elle est également en capacité de se doter dans des délais raisonnables des premières technologies de support. La feuille de route de développement numérique 2015-2019 en détaille la faisabilité dans un calendrier opérationnel. On constate également que l'occitan dispose de réels atouts, tels que le dynamisme de la recherche universitaire et des acteurs institutionnels, sa proximité avec les autres langues romanes ou encore des perspectives en termes de transferts de technologies avec les langues géographiquement voisines.

Cependant, il est indispensable, comme le souligne le Livre blanc de META-NET, de mettre en place une planification rigoureuse et de se donner les moyens de mutualiser efficacement les ressources. Cette feuille de route, largement

diffusée depuis, a permis de sensibiliser et, il faut l'espérer, de mobiliser les différents agents de transmission de la langue à des enjeux vitaux pour la langue occitane.

Il s'agit d'un programme pluriannuel rassemblant des acteurs (universités, institutions, associations) ayant leurs propres contraintes administratives, techniques et financières et répartis sur un grand espace géographique (huit régions administratives françaises, une italienne et une autonomie espagnole) qui donc nécessitera un niveau élevé de coordination.

Enfin, même si elle n'est pas chiffrée, la feuille de route représentera, on le sait, un budget conséquent. Les montants toutefois pressentis induisent en amont un important travail d'ingénierie financière, avec la participation active des collectivités, en particulier les Régions, des universités et des opérateurs institutionnels. Il s'agit d'un pré-requis indispensable pour émarger sur la nouvelle programmation 2015-2020 des crédits FEDER et les crédits de coopération interrégionale et transfrontalière.

Remerciements

Nous remercions l'ADEPFO qui a financé l'étude ayant permis la rédaction de la feuille de route pour le développement numérique occitan, la fondation Elhuyar qui en a assuré le support technique ainsi que tous les experts qui y ont porté leurs concours. Nous remercions également Marianne Vergez-Couret ainsi que les relecteurs anonymes pour leurs conseils avisés.

Références

BEC P. (1995), *La langue occitane*, Number 105, Que sais-je ? Paris.

GURRUTXAGA A. (2014). *Diagnostic et feuille de route pour le développement numérique de la langue occitane* : 2015-2019. Rapport final de la fondation Elhuyar dans le cadre de l'étude-action ADEPFO. Billère.

LEIXA, J., MAPELLI, V. et CHOUKRI, K. (2014). *Inventaire des ressources linguistiques des langues de France.* Rapport réalisé par ELDA en partenariat avec la DGLFLF-ministère de la Culture et de la Communication. Paris.

REHM, G. & USZKOREIT, H., editors. (2012). *META-NET White Paper Series: Europe's Languages in the Digital Age*. Springer, Heidelberg etc. 32 volumes on 31 European languages

SORIA C., MARIANI J., ZOLI C. 2013. *Dwarfs sitting on the giants' shoulders – how LTs for regional and minority languages can benefit from piggybacking major langages.* In M.J. Norris, E. Anonby, M-O. Junker, N. Ostler and D. Patrick (Eds.). Proceedings of the XVII FEL Conference. Carleton University, Ottawa, Canada, 1-4 October 2013, pp. 73-79