Diagnostic et feuille de route pour le développement numérique de la langue occitane : 2015-2019

Rapport final du projet

media.kom
Komunikabideak & Kontseilua

elhuyar
Hizkuntza eta Teknologia

Billère, 26-11-2014

ADEFFO

RÉSEAU PYRÉNÉES

l'Europe s'engage dans le Massif des Pyrénées

Liberté • Égalité • Fraternité
RÉPUBLIQUE FRANÇAISE

Datar Pyrénées

REGION AQUITAINE

RÉGION MIDI-PYRÉNÉES

la Région Languedoc Roussillon

Table des matières

1 Présentation

L'objectif de ce rapport est de présenter les résultats de l'étude sur la situation actuelle de l'occitan dans le domaine des technologies de la langue, ainsi que la feuille de route proposée par le groupe travaillant sur le projet pour le développement de ces technologies au cours de la période 2015-2019.

L'étude a été menée par Lo Congrès, et réalisée dans le cadre d'une d'intervention de l'ADEPFO pour la formation-développement. Le groupe de travail est composé de 21 stagiaires provenant de différentes organisations et institutions, et a une vaste expérience dans divers domaines de la promotion, de la recherche et du développement de l'occitan. De même, nous nous sommes appuyés sur les conseils des experts d'autres langues, notamment du basque, du breton, du catalan et du gallois. La coordination générale a été prise en charge par Media.kom, et la coordination technique et l'expertise dans le domaine des technologies de la langue ont été menées par la Fondation Elhuyar.

Voici les résultats de cette étude :

- un ensemble d'objectifs spécifiques et précis pour développer la technologie langagière de l'occitan, à obtenir pour fin 2019
- un calendrier annuel pour atteindre ces objectifs au cours de la période 2015-2019

2 Motivation

Il est incontestable qu'au jour d'aujourd'hui la société de l'information est une réalité qui a transformé de nombreux aspects de notre société. Les technologies de l'information et le monde numérique représentent un nouvel environnement pour la communication humaine.

Dans ce contexte, un nombre croissant d'acteurs et d'institutions a souligné et mis en valeur le besoin de développer les technologies langagières pour le développement moderne de la langue, et pour sa durabilité dans un environnement caractérisé par l'importance du numérique et de l'Internet. Selon l'étude menée par METANET pour les langues officielles de la Communauté européenne[1] :

> *« Sans les technologies de la langue, nous ne serons pas en mesure de donner aux utilisateurs des moyens de communiquer réellement interactifs, multimédias et multilingues dans un futur proche. »*

Voici les idées clés soulignées par METANET[2] :

- la réalité multilingue de l'Europe ;
- les technologies de la langue sont un facteur clé qui aidera à résoudre le problème des barrières linguistiques ;
- la technologie de la langue est une technologie clé pour la prochaine révolution de l'information ;
- grâce à un important travail de recherche, l'Europe peut préserver ses langues précieuses, jouir de la diversité linguistique, et jouer un rôle de premier plan dans la révolution informatique prochaine.

[1] http://www.meta-net.eu/whitepapers/volumes/french
[2] http://www.meta-net.eu/sra/key-messages?set_language=en

Cette nouvelle réalité est un défi pour toutes les langues, mais en est un particulièrement important pour les langues minoritaires qui manquent d'un statut et de ressources pour faire face à ce défi, ce qui les conduit à une situation de diglossie et de danger.

L'occitan est une des langues européennes qui se trouvent dans cette situation, et doit se doter d'un développement technologique approprié en maximisant l'utilisation des ressources disponibles, d'une manière planifiée et réaliste.

3 Objectif

L'objectif du projet est de réaliser :

- un diagnostic numérique des outils et des ressources linguistiques dans le domaine TIC (existantes et restant à réaliser) ;
- une feuille de route pour le développement des outils ou des ressources désormais nécessaires à la transmission et au développement de la langue occitane, et l'ordre de leur réalisation dans la période retenue (2015-2019).

L'obtention de ces objectifs permettra l'établissement d'une base technologique équilibrée et avancée pour l'occitan, et ouvrira la voie à de futurs développements, afin de réussir à relever les défis du monde numérique et de l'Internet, dans lesquels les technologies de l'information et de la communication présentent de plus en plus un profil qui incorpore le traitement automatique de la langue.

4 Technologies de la langue

Les technologies de la langue sont des technologies de l'information spécialisées dans le traitement du langage humain. Le but de ces technologies est de faciliter l'interaction homme-machine, et d'automatiser les processus dans lesquels la compréhension et la production du langage oral et écrit sont nécessaires, telles que la traduction automatique, les systèmes de dialogue ou l'extraction d'information. En outre, elles ont contribué à une révolution dans l'étude du langage, avec la capacité de traiter et d'analyser de grandes quantités de texte et d'obtenir des données sur l'utilisation du langage qu'il serait impensable de réaliser manuellement.

Par conséquent, un défi majeur est le traitement automatique de la langue à tous les niveaux d'analyse et de génération (lexical, morphologique, syntaxique, sémantique, pragmatique et discursif).

Les technologies langagières sont vastes et variées, et couvrent une large gamme d'outils, de ressources, d'applications, etc. Dans cette étude, nous nous sommes inspirés du cadre commun établi pour toutes les langues apparaissant dans la collection des livres blancs de META-NET[3].

Les deux catégories définies sont les ressources linguistiques et les outils linguistiques[4].

Une ressource linguistique est un c, utilisée dans un processus par un outil de traitement linguistique. La ressource peut être l'entrée dans le processus, où

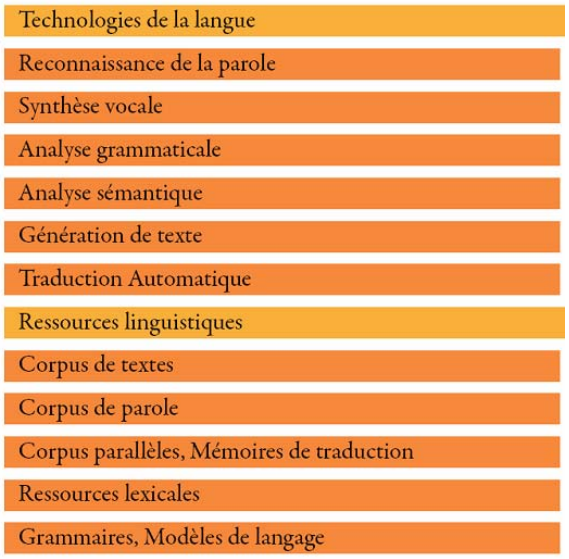
[3] Mariani, J., Paroubek, P., Francopoulo, G., Max, A., Yvon, F., & Zweigenbaum, P. (2012). *La langue française à l'Ère du numérique – The French Language in the Digital Age*. Springer.

[4] Littéralement, le rapport utilise les catégories *Technologies de la langue* et *Ressources linguistiques*. Cependant, le terme *technologies de la langue* est également utilisé de manière générique, et englobe les deux concepts ci-dessus. Voilà pourquoi nous avons choisi le terme *outil linguistique* par opposition à *ressource linguistique*.

elle peut être transformée, traitée ou enrichie ; elle peut également fournir les connaissances linguistiques dont l'outil a besoin pour traiter une entrée et générer une sortie.

Un outil linguistique est un programme informatique, un logiciel ou une application qui transforme une entrée en langage naturel dans une sortie traitée ou transformée, soit en langage naturel soit en un autre type de représentation.

Nous allons présenter brièvement ces ressources et outils.



4.1 Corpus

Un corpus linguistique est un ensemble, généralement très grand, d'exemples réels d'utilisation du langage. Pour être utilisé de manière efficace, le corpus doit contenir des documents numérisés et classés, avec un minimum de traitement linguistique (lemme et catégorie).

On peut classer les corpus selon différents critères :

- textuel (écrit) / oral / multimodal ;

- monolingue / bilingue / multilingue ;[5]
- de référence / spécialisé[6].

Le corpus est une ressource basique pour les technologies de la langue. Il et la source d'information pour construire des outils statistiques (TALN), des modèles de langage, et, en général, pour obtenir des données sur l'usage réel du langage, information essentielle pour l'élaboration de ressources lexicales comme les dictionnaires et les bases de données terminologiques. C'est pour cela que la lexicographie et la terminologie modernes sont basées sur l'utilisation de corpus.

4.2 Ressources lexicales

Les dictionnaires, les bases de données lexicales et terminologiques, les glossaires, les thésaurus, les bases de connaissance lexicales etc., sont des répertoires contenant des informations sur les mots ou les termes d'une ou plusieurs langues. Ils apportent les connaissances lexicales que tout outil langagier utilise pour traiter automatiquement le texte. Par exemple, les correcteurs orthographiques, les analyseurs morphologiques, syntaxiques et sémantiques, les systèmes de reconnaissance de caractères et de la parole. Tous ont besoin des connaissances lexicales pour identifier, étiqueter et traiter les mots.

---

[5] Les corpus bilingues et multilingues peuvent être parallèles ou comparables. Les corpus parallèles sont constitués de textes traduits qui peuvent être alignés au niveau du document, du paragraphe ou de la phrase. Les corpus comparables, cependant, contiennent des textes dans différentes langues qui partagent certaines caractéristiques telles que le domaine de la connaissance, le genre, le type de texte ou de discours, les dates de publication, etc.

[6] « Un corpus de référence est celui qui est conçu pour fournir des informations complètes sur une langue. Il vise à être suffisamment large pour représenter toutes les variétés concernées de la langue et le vocabulaire caractéristique, de sorte qu'il peut être utilisé comme une base pour les grammaires, les dictionnaires fiables, les thésaurus et autres documents de référence de la langue. » (Traduit en français à partir de la source : SINCLAIR, J. 1996. « Preliminary recommendations on corpus typology ». EAGLES. Consulté sur : http://www.ilc.cnr.it/EAGLES96/corpustyp/corpustyp.html). Par contre, un corpus spécialisé est un recueil de textes appartenant à une variante, une époque, une zone, un registre ou un domaine particulier de l'utilisation de la langue.

4.3 Grammaires

Les grammaires computationnelles sont nécessaires pour le traitement syntaxique du corpus, et la génération automatique de texte. Ces tâches sont essentielles, par exemple, dans les systèmes de traduction automatique à base de règles. En outre, un système qui vise à rendre le texte au niveau sémantique pour gérer ou extraire le contenu ou la signification doit considérer les informations de ce niveau qui sont encodées dans la structure syntaxique de la phrase (dépendances, rôle sémantique...).

4.4 Traitement de la parole

Les technologies de traitement de la langue parlée visent à élaborer les tâches suivantes :

- la reconnaissance de la parole, qui vise à transformer les énoncés des intervenants en un texte ou un autre format traitable ;
- la synthèse vocale, qui, à partir d'un un texte, génère un énoncé audible (généralement un fichier audio) ;
- la reconnaissance du locuteur ;
- l'identification de la langue parlée.

Il existe une grande variété de tâches où l'utilisation du traitement automatique de la parole peut contribuer d'une manière significative au développement d'applications avancées. Par exemple, la transcription de l'audio (nécessaire pour l'étiquetage des corpus oraux), le sous-titrage, la dictée, les systèmes d'interprétation, et en général, tous les systèmes interactifs homme-machine.

Parmi ces technologies, la reconnaissance vocale est celle qui comporte le plus grand défi technique, surtout si elle est destinée à être appliquée au discours commun (non formel).

4.5 Analyse grammaticale/syntaxique

La mission de l'analyse grammaticale est d'enrichir le texte avec des informations autant sur les mots (lemme, catégorie, flexion, cas) que sur la structure de la phrase (syntagmes, dépendances entre les composants, structure superficielle ou profonde).

Le niveau de connaissance que nous avons sur les composants et la structure des phrases dans un texte détermine en grande partie les tâches dans lesquelles on peut s'engager. Par exemple, dans un système de traduction sur la base de règles, l'analyse syntaxique du texte source est un facteur essentiel dans la qualité du résultat. Par ailleurs, les systèmes d'extraction de terminologie ou d'unités multi-mots (telles que les collocations), présentent une nette amélioration s'ils peuvent faire usage de texte étiqueté au niveau syntaxique, ou du moins, morphosyntaxique.

4.6 Analyse sémantique

L'analyse et le traitement de texte au niveau sémantique est aujourd'hui l'un des défis les plus importants de la technologie de la langue. Tout système qui veut interagir en langue naturelle avec l'utilisateur ainsi que les systèmes d'information, d'extraction et de gestion de contenu doit extraire et traiter les représentations sémantiques des énoncés. Toutefois, cela est une tâche extrêmement complexe, et l'état actuel de la technologie est encore loin de cet objectif, bien que ces dernières années il y ait eu des progrès remarquables.

Parmi les nombreux domaines de recherche et les activités liées aux tâches de traitement sémantique, il faut souligner la désambiguïsation du sens des mots et des autres unités linguistiques du texte, qui est habituellement appelée l'étiquetage sémantique. Pour ce faire, une stratégie consiste à relier chaque unité textuelle avec une entité conceptuelle représentée dans une base de connaissances ou d'ontologie.

Au-delà de la sémantique formelle, nous trouvons des disciplines telles que la pragmatique et l'analyse du discours. L'approche computationnelle a commencé à être appliquée aussi dans ces domaines poussés.

4.7 Génération de texte

Les systèmes de génération de texte visent à transformer des contenus représentés formellement (par exemple, les registres d'une base de données) dans des textes en langage naturel qui sont syntaxiquement et sémantiquement corrects.

Ces systèmes ont de nombreuses applications existantes ou potentielles : production de rapports automatisés, systèmes d'information hospitaliers, rédaction de bulletins météorologiques, etc. Normalement, il s'agit d'applications intégrées dans des environnements de logiciels plus larges.

4.8 Traduction automatique

La traduction automatique est l'une des premières et plus importantes aspirations de l'ingénierie linguistique. C'est un domaine dans lequel on a beaucoup œuvré et qui atteint un degré élevé de développement, malgré les grandes difficultés de la tâche.

Aujourd'hui, les systèmes de traduction automatique sont principalement de deux types : systèmes statistiques et systèmes fondés sur des règles. Dans le premier cas, il est nécessaire pour obtenir des résultats acceptables d'avoir une énorme quantité de textes traduits et alignés. Cette condition n'est généralement pas remplie dans le cas des langues minoritaires, et donc le choix habituel se porte sur les systèmes fondés sur des règles, qui ont besoin d'analyseurs syntaxiques précis et des ressources lexicales bilingues à large couverture.

Parmi les nombreuses applications, on établit habituellement une distinction générale entre celles conçues pour l'assimilation et celles conçues pour la dissémination. Dans le premier cas, on est face à un document dans une langue qu'on ne connait pas, ou peu, et le seul but est la compréhension du texte. Par contre, la dissémination a lieu quand une personne ou une organisation veut diffuser un document traduit automatiquement à l'attention d'un certain public. Il va sans dire que la qualité de la traduction ne doit pas être obligatoirement parfaite dans le premier cas (à condition qu'elle ne fausse pas le contenu), tandis que dans le second les conditions sont plus exigeantes, et un document traduit automatiquement est rarement délivré sans révision et correction.

5 Méthodologie

5.1 Bases méthodologiques

La méthodologie de l'étude présentée ici est basée sur la mise en place de différentes expertises, connaissances et expériences dans le domaine de l'occitan et des technologies langagières, de sorte que tout le monde travaille ensemble à l'élaboration d'un plan réaliste et aussi complet que possible.

Le groupe de travail a été conçu et sélectionné pour atteindre cet objectif, et le processus de planification de l'étude a également répondu à cette exigence. Nous avons établi une méthodologie participative, avec les moyens nécessaires pour garantir la participation directe et la contribution de l'ensemble des stagiaires.

5.2 Structure du groupe de travail

Fonction	Responsables/Participants	Structure
Coordination du projet	Agus Hernan	Media.kom
Coordination Lo Congrès	Bénaset Dázeas Aure Séguier	Lo Congrès
Coordination technique sur les Technologies de la Langue	Antton Gurrutxaga Igor Leturia	Fondation Elhuyar – Langue & Technologie
Dynamisation de séances de travail	Jon Abril	Elhuyar Services-conseils
Groupe des stagiaires	Natalia Beillé	CFPÒ
	Brigitte Bouncer	Vistedit
	Joan Breç Brana	InÒc Aquitània
	Sébastien Bujeaud	InÒc Aquitània (stagiaire)
	Jean-Jacques Casteret	InÒc Aquitània
	Perrine Charteau	Lo Cirdòc
	Clarisse Harbouzit	InÒc Aquitània
	Sèrgi Javaloyes	Edicions Reclams
	Sandra Juan	Lo Cirdòc
	Mélanie Larché	InÒc Aquitània
	Domenja Lekuona	Ràdio País
	Estel Llansana	CFPÒ – Mieidia-Pirenèus
	Florence Marcouyre	Lo Congrès permanent de la lenga occitana

	Juliette Minvielle	InÒc Aquitània
	Cristina Nougué-Debat	CAPÒC – CDDP
	Patrick Pouzet	Collectif ça-i
	Vincenç Rivière	Universitat Tolosa – Joan Jaurès
	Laurent Soleil	CFPÒ – Mieidia-Pirenèus
	Franck Taurinya	Lo Cirdòc
	Lauriane Tresserre	FIMÒC
	Stéphane Valentin	Octele
	Cedric Valmary	CFPÒC
	Marianna Vergez-Couret	Universitat Tolosa – Joan Jaurès
Experts d'autres langues	Basque: Xabier Artola	Groupe de recherche Ixa (Ixa taldea) – Faculté d'Informatique – Université du Pays Basque
	Breton: Olier Ar Mogn	Directeur scientifique – Office de la Langue Bretonne
	Catalan: Asunción Moreno	Centre de Recerca TALP – Universitat Politècnica de Catalunya
	Gallois: Francis M. Tyers	The Arctic University of Norway – Department of Language and Linguistics

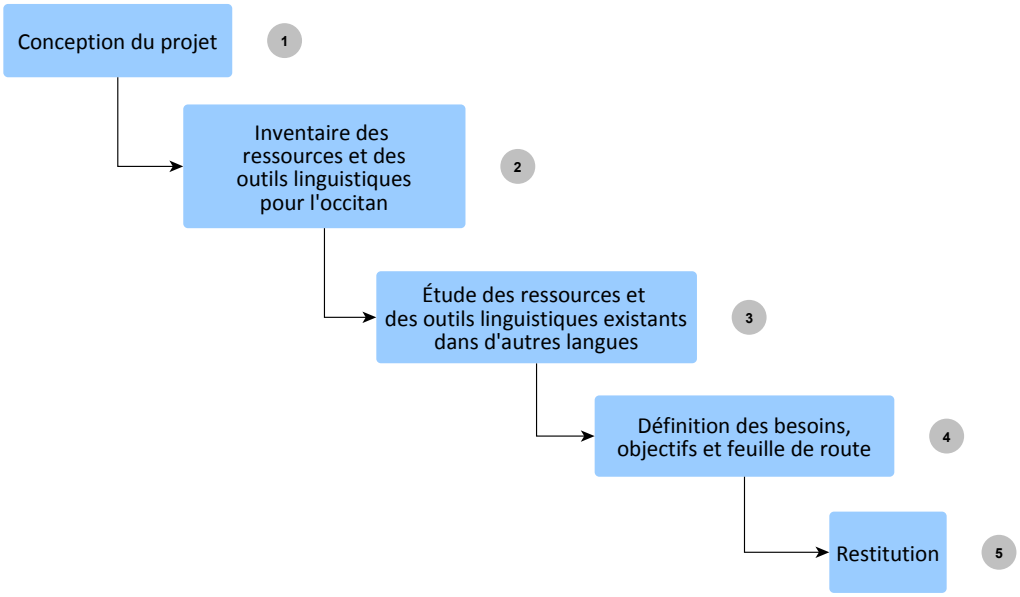
5.3 Comité de pilotage

	Fonction	Organisme
David Grosclaude	Président du Comité de Pilotage	Conseil Régional d'Aquitaine
Joël Maitia	Animateur technique	ADEPFO
Agus Hernan	Responsable Pédagogique	Media.kom
Benoît Dazeas	Représentant des stagiaires	Lo Congrès permanent de la lenga occitana
Thibaut Grouas	Chef de mission langues et numérique	DGLFLF
Jean-François Sibers	Chef du service des collections, de l'information et de la communication	DRAC Aquitaine
Jérémie Obispo	Chargé de mission langues régionales	Région Aquitaine
Elise Dudézert	Chef de projet	Région Aquitaine
Estève Cros	Chargé de mission langue occitane	Région Midi-Pyrénées
Philippe Vialard	Chargé de mission langues régionales	Région Languedoc-Roussillon
Colette Laplechère	Chargé de mission langue occitane	Conseil Général des Pyrénées-Atlantiques
Robert Tamburello	Chargé de mission langue occitane	Conseil Général des Hautes-Pyrénées
Magali Blenet	Chargée de mission langue occitane	Ville de Toulouse

Les missions du comité de pilotage:

- impliquer tous ceux qui jouent ou peuvent jouer un rôle pour la réussite du projet ;
- faciliter une coordination efficace des moyens mis en œuvre ;
- veiller à l'adéquation de la formation aux exigences du projet et aux besoins des stagiaires ;
- s'assurer de la cohérence de la démarche.

5.4 Processus d'analyse et décision - Phasage



Le processus d'analyse et de prise de décisions a été organisé en 5 étapes :

À l'étape de la conception, nous nous sommes mis d'accord sur les objectifs du projet, la méthodologie d'intervention et le planning de réalisation, dans le cadre d'un processus participatif avec les stagiaires représentants des partenaires de Lo Congrès,

Pour parvenir à un diagnostic commun, qui précède la définition de besoins et l'établissement d'objectifs, nous avons traversé deux phases. Tout d'abord, la mise au point d'un inventaire de ressources et d'outils disponibles aujourd'hui pour l'occitan. Notre objectif était d'obtenir une image aussi fidèle que possible du point de départ du projet. Deuxièmement, nous avons jugé nécessaire d'avoir une connaissance de la technologie d'autres langues qui sont dans une situation sociolinguistique plus ou moins semblable à celle de l'occitan; de même, grâce à la contribution d'experts dans ces langues, le groupe de travail a obtenu des informations directes sur le processus grâce auquel elles sont

arrivées à leur stade actuel de développement. Les langues étudiées sont le basque, le breton, le catalan et le gallois.

Une fois que le groupe de travail a eu étudié les informations recueillies, il a été possible de se mettre d'accord sur un diagnostic.

Le but de l'étape suivante était d'identifier les besoins prioritaires sur la base du diagnostic et de préciser les objectifs détaillés auxquels nous devions parvenir à la fin de la période prévue. La dernière démarche a été la planification annuelle pour parvenir à ces objectifs, et cette planification a finalement abouti à la feuille de route.

Enfin, les résultats de l'étude sont présentés dans ce rapport.

Nous allons exposer dans la section 6 Diagnostic les travaux réalisés dans les phases 2 et 3; ceux de l'étape 4 seront exposés dans la section 7 Feuille de route pour 2015-2019..

6 Diagnostic

6.1 Situation de la langue occitane

Les caractéristiques et l'état de l'occitan sont bien connus par la communauté à laquelle ce rapport est adressé, mais il est intéressant de rappeler, même brièvement, les aspects particulièrement pertinents pour cette étude.

- L'espace linguistique occitan : 14 millions de personnes répartis sur trois États de l'UE (France, Espagne, Italie).
- 10 % en seraient des locuteurs, et près de 50 % auraient des compétences linguistiques plus ou moins avancées (locuteurs passifs).

- Aujourd'hui environ 80 000 élèves sont scolarisés en occitan du primaire à l'université.
- L'occitan est aussi une langue de recherche, enseignée et étudiée dans plusieurs universités à travers le monde.
- Néanmoins l'occitan a pris beaucoup de retard dans le numérique et les technologies du langage.

Actuellement, la plupart des classifications reconnaissent plusieurs grandes variantes de l'occitan, sans que l'une d'entre elles soit considérée comme une norme standard. Lors de la formulation des objectifs, il nous faut donc prendre en compte le fait que les ressources et outils développés doivent sur le principe, et après prise en compte des besoins des usagers et des contraintes techniques et budgétaires, répondre aux besoins pour différentes variantes

6.2 Inventaire des ressources et des outils linguistiques pour l'occitan

Pour le répertoire des ressources et des outils, la classification suivante a été choisie :

Ressources linguistiques

- Corpus
  - Corpus de textes
    - Monolingues (anciens et modernes)
    - Corpus parallèles, mémoires de traduction
  - Corpus de parole
  - Corpus multimédia et multimodaux
- Ressources lexicales
  - Lexiques

- Bases terminologiques
- Thesauri, wordnets, ontologies
- Bases de toponymes
- Patronymie

- Grammaires, modèles de langage

Outils linguistiques

- Reconnaissance de l'écriture (OCR)
- Reconnaissance de la parole
- Synthèse vocale
- Analyse grammaticale
- Analyse sémantique
- Génération de texte
- Traduction automatique
- Recherche et extraction d'information

Autres outils

- Logiciels en occitan
- Outils numériques pour apprendre l'occitan

6.2.1 Sommaire des résultats

6.2.1.1 Ressources linguistiques

Ressource linguistique			Recensées	Pensées pour une utilisation informatique
Corpus	Corpus de textes		27	2
	Corpus de parole		28	1
	Corpus multimédia et multimodaux		24	0
	Corpus parallèles, mémoires de traduction	Mémoires de traduction	1	1
Ressources lexicales	Lexiques/	Dictionnaires monolingues	4	4
		Dictionnaires bilingues en ligne	13	4
		Dictionnaires bilingues en PDF, scannés ou autres	27	0
		Dictionnaires étymologiques	4	0
		Dictionnaires PDF numérisés non océrisés	21	0
		Dictionnaires manuscrits	3	0
		Dictionnaires étymologiques non océrisés	1	0

	Bases terminologiques	Dictionnaires et lexiques spécialisés	18	2
		Dictionnaires et lexiques spécialisés non océrisés	3	0
	Toponymie	Bases de données toponymiques	3	3
	Thesauri, wordnets, ontologies		1	1
Grammaires, modèles de langage	Grammaires en PDF, scannées ou autres		30	0
	Conjugaison		6	3

Note : La plupart des dictionnaires et une partie des grammaires recensés sont anciens, de qualité linguistique inconnue, et difficilement utilisables informatiquement.
La qualité des corpus (de référence ou autre) n'a pas encore été définie.

6.2.1.2 Outils linguistiques

Outil linguistique		Recensés
Reconnaissance de l'écriture (OCR)		2
Reconnaissance de la parole		0
Synthèse vocale		0
Analyse grammaticale	Correcteurs orthographiques	4
	Divers	2
Analyse sémantique		0
Génération de texte		0
Traduction automatique		2
Recherche et extraction d'information		0

6.2.1.3 Autres outils

Outil	Recensés
Logiciels disponibles en occitan	6
Outils numériques pour apprendre l'occitan	4

6.2.2 Évaluation

Dans le tableau suivant, on montre l'évaluation faite par le groupe de travail sur l'état de la technologie de la langue en occitan (valeurs utilisées pour l'évaluation : 0/ Inexistant ; 1/ Quelques-uns ; 2/Acceptable ; 3/ Beaucoup).

Ressource linguistique	Évaluation
Corpus de textes	1
Corpus de parole	1
Corpus multimédia et multimodal	1
Corpus parallèles, mémoires de traduction	1
Ressources lexicales	1
Lexiques	1
Bases terminologiques	1
Toponymie	1
Thesauri, wordnets, ontologies	1
Grammaires, modèles de langage	1
Outils linguistiques	
Reconnaissance de l'écriture (OCR)	1
Reconnaissance de la parole	0
Synthèse vocale	0
Analyse grammaticale	1
Analyse sémantique	0
Génération de texte	0
Traduction automatique	1
Recherche et extraction d'information	0

Autres outils	
Logiciel en occitan	1
Outils numériques pour apprendre l'occitan	1

6.2.3 Conclusions

Nous avons tiré les conclusions suivantes de l'analyse de l'inventaire :

- 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 il y en a peu qui peuvent être réutilisées :
  - La plupart des ressources lexicales consistent en des 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 informatique.
  - 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 faudrait donc créer des ressources linguistiques de base avant de pouvoir développer des outils à partir de ces ressources.

6.3 Expériences d'autres langues

6.3.1 Basque

6.3.1.1 Situation

Selon le livre blanc de METANET sur le basque :

- Dans le domaine des technologies de la langue, le basque possède un certain nombre de produits, de technologies et de ressources.
- L'industrie des technologies de la langue basque est bien établie et un nombre important de PME est actif dans ce secteur, bien que la plupart des technologies soient appliquées à la langue écrite. Leurs produits ont été – et sont encore – des outils efficaces qui soutiennent le processus de normalisation et de promotion de l'utilisation de la langue basque. Le basque n'a pas été inclus dans le catalogue de grandes entreprises, à l'exception de quelques actions spécifiques, et il est généralement pris en charge par le Gouvernement basque.
- Il existe plusieurs groupes de recherche travaillant sur le discours et le traitement du langage depuis 1988. Si le basque est aujourd'hui une exception dans la corrélation entre la taille de la langue et la TL cela est dû aux efforts coordonnés de ces groupes de recherche.
- La recherche et le développement pour les langues qui ont moins de ressources devraient être confrontés à des critères élevés de normalisation, à la promotion de l'*open-source* et à la réutilisation de ressources et d'outils déjà développés.

- Le basque est l'une des langues de l'UE nécessitant des recherches plus poussées afin que les solutions technologiques soient opérationnelles pour une utilisation quotidienne.
- Il est urgent de développer des technologies de haute qualité pour la langue basque, et cela est même capital pour sa préservation.

	Quantity	Availability	Quality	Coverage	Maturity	Sustainability	Adaptability
Language Technology: Tools, Technologies and Applications							
Speech Recognition	2	1	1	1	4	3	2
Speech Synthesis	2	3	4	4	4	3	3
Grammatical analysis	4	2.5	4	4	4	2.5	2.5
Semantic analysis	1	1.5	2	1	1	1	1
Text generation	1	0	0	0	0	0	0
Machine translation	3	5	2	3	3	2	2
Language Resources (Resources, Data and Knowledge Bases)							
Text corpora	2	4	3	2	3	4	2.5
Speech corpora	3	2	3	2	3	3	2
Parallel corpora	2	4	2	2	2	2	1
Lexical resources	4	4	4	5	5	4	3
Grammars	2	2	2	2	2	2	2

7: State of language technology support for Basque

6.3.1.2 Préconisations

Préconisations pour le développement des technologies de la langue :

- La conception et le développement des bases de la technologie langagière, des outils et des applications doit se faire d'une manière progressive et planifiée afin d'en tirer le meilleur bénéfice possible.
- Les ressources doivent être normalisées afin de pouvoir être utilisées dans des recherches variées, dans le développement de divers outils, dans des applications et des produits différents. On

peut par exemple adopter des TEI et des standards comme le XML en tant que base pour l'étiquetage linguistique des différents niveaux de traitement.

- Pour traiter une langue écrite, on ne peut se passer des trois ressources suivantes : les bases de données lexicales, les lemmatiseurs/*taggers*, les corpus.
- Importance de l'*open source*.

6.3.2 Breton

6.3.2.1 Situation

Il n'y a pas de livre blanc de METANET sur le breton (il y en a seulement pour des langues officielles ou co-officielles).

Ressources et outils existants :

- Divers dictionnaires lexicographiques et terminologiques en format numérique et en ligne.
- Correcteurs orthographiques et grammaticaux.
- Traduction automatique breton-français.
- Corpus textuel monolingue et parallèle breton-français.
- Corpus oraux.
- Grammaires computationnelles.
- Logiciels en breton.
- Contenu en breton en ligne.

Développements vitaux :

- Synthèse vocale.
- Reconnaissance automatique de la parole.

6.3.2.2 Préconisations

Voici les points positifs et négatifs dans le développement d'outils des technologies de la langue :

- Obstacles à surmonter
  - Tenue à jour des traductions.
  - Pérennité des traductions de commande.
  - Correction de la langue des traductions.
  - Stabilisation des choix terminologiques.
  - Financement des développements nécessaires.
  - Absence de travaux d'analyse de la grammaire et de la syntaxe.
  - Absence d'interface en breton pour les produits les plus utilisés : Microsoft, Twitter, Facebook…
- Points positifs
  - Nombreuses initiatives spontanées.
  - Nouvelle génération de brittophones demandeuse de nouvelles technologies en breton.
  - Vers une stabilisation de la terminologie de NT grâce, notamment, à l'action de l'Office public de la langue bretonne.
  - Collaborations internationales possibles autour de ces problématiques.

6.3.3 Catalan

6.3.3.1 Situation

Selon le livre blanc de METANET sur le catalan :

- Dans le cas de la langue catalane, nous sommes modérément optimistes quant à l'état actuel de soutien des technologies langagières.
- Il y a une communauté viable de recherche de TL en Catalogne, qui a été soutenue par les programmes de recherche espagnols et catalans.
- Un certain nombre de ressources et de technologies à la pointe ont été produites et distribuées pour le catalan.
- Toutefois, la portée des ressources et la gamme d'outils sont encore très limitées par rapport à celles existant pour la langue espagnole (et évidemment pour la langue anglaise) et elles ne sont tout simplement pas suffisantes en qualité et en quantité pour développer le type de technologies nécessaires pour soutenir une société véritablement multilingue.
- L'industrie des technologies de la langue catalane dédiée à la transformation des résultats de la recherche en produits est actuellement très faible. La plupart des grandes entreprises ont soit arrêté soit fortement réduit l'élaboration de TL, reléguant les langues parlées par seulement un petit nombre de personnes à un objectif secondaire.
- Nos résultats montrent que la seule alternative est de produire un effort considérable pour créer des ressources de TL pour le catalan, et de les utiliser pour faire avancer la recherche, l'innovation et le développement. Vu la nécessité d'avoir de grandes quantités de données et l'extrême complexité des systèmes de technologie de langue, il est essentiel de développer une nouvelle infrastructure et

une organisation de recherche plus cohérente afin de stimuler un plus grand partage et une coopération.

- Il y a aussi un manque de continuité dans le financement de la recherche et développement. Des programmes coordonnés à court terme ont tendance à alterner avec des périodes de financements rares ou inexistants. En outre, il y a un manque général de coordination avec les programmes d'autres pays de l'UE ainsi qu'au niveau de la Commission européenne. Nous pouvons donc conclure qu'il y a un très grand besoin d'une large initiative coordonnée destinée à surmonter les différences de maturité dans la technologie langagière pour les langues européennes dans leur ensemble.

	Quantity	Availability	Quality	Coverage	Maturity	Sustainability	Adaptability
Language Technology: Tools, Technologies and Applications							
Speech Recognition	3	3	3	3	3	3	2
Speech Synthesis	4	2	4	4	5	4	2
Grammatical analysis	3	2.5	4	4	4	2.5	2.5
Semantic analysis	1	1	2	1	1	1	1
Text generation	1	2	3	1	3	3	1
Machine translation	3	3	2	3	4	1	2
Language Resources: Resources, Data and Knowledge Bases							
Text corpora	3	2.5	3.5	3	3	2.5	2.5
Speech corpora	3	5	3	2	3	3	2
Parallel corpora	2	1	2	2	2	1	1
Lexical resources	2.5	2	3	2.5	3	3	2.5
Grammars	2	3	2	2	2	2	2

7: State of language technology support for Catalan

6.3.3.2 Préconisations

- Importance de disposer d'un BLARK (*Basic Language Resource Kit*).
- Nécessité d'une feuille de route mise à jour.

- Trouver des synergies avec d'autres centres de recherche.
- Promouvoir la coopération entre les chercheurs travaillant sur des langues similaires à l'occitan.
- Recherche de financement pour un projet commun : l'occitan, le catalan, le basque .... dans les différentes régions : Le Sud de la France, Le Pays Basque, La Catalogne, etc.

6.3.4 Gallois

6.3.4.1 Situation

Selon le livre blanc de METANET sur le gallois :

	Quantity	Availability	Quality	Coverage	Maturity	Sustainability	Adaptability
Language Technology: Tools, Technologies and Applications							
Speech Recognition	1	1	1	1	1	1	3
Speech Synthesis	1	2	2	2	2	2	3
Grammatical analysis	2	1	2	2	3	2	1
Semantic analysis	2	2	2	2	2	2	2
Text generation	2	2	2	2	2	2	2
Machine translation	3	3	3	2	1	1	2
Language Resources: Resources, Data and Knowledge Bases							
Text corpora	1	1	2	1	2	2	1
Speech corpora	4	3	4	4	4	4	3
Parallel corpora	3	3	2	3	3	4	3
Lexical resources	3	2	3	2	2	4	4
Grammars	4	3	3	3	3	5	4

6: State of language technology support for Welsh

6.3.4.2 Préconisations

Groupes de travail :

- Soutenir les groupes existants.
- Essayer de créer un groupe académique pour travailler sur les technologies du langage en occitan.

Outils :

- Unifier les correcteurs orthographiques pour l'occitan, en prenant en compte les variantes, si désiré.
- Le correcteur doit être cohérent et offrir les mêmes fonctionnalités pour toutes les variantes et toutes les plateformes.
- Lancer le développement d'un correcteur grammatical et syntaxique.

Mémoire de traduction :

- Suivre l'exemple d'OpenData Euskadi.
- Les mémoires de traduction sont très utilisées.
- Investissement minimum, gros impact.

Traduction automatique :

- Mettre à jour les systèmes existants vers une orthographe normative, élargir aux autres variantes.
- Lancer le développement de systèmes français-occitan et italien-occitan.
- Lancer le développement de « convertisseurs » entre variantes.

Localisation :

- Financer les efforts en cours.
- Promouvoir les efforts existants.
- Travailler sur une standardisation de la terminologie et des guides de style.

Publier les données sous licences libres est vital pour leur réutilisation.

6.3.5 Conclusions

La situation sociolinguistique des langues analysées est différente d'une langue à l'autre, ainsi que le statut juridique, et par conséquent, la capacité des communautés linguistiques à relever les défis du développement technologique également. Quoiqu'il en soit, un certain nombre de conclusions peuvent être tirées des expériences présentées.

- Il semble important de faire des démarches dans un ordre cohérent et de développer autant les ressources que les outils.
- L'intérêt d'une collaboration a été plusieurs fois souligné.
- L'importance des licences libres a également été soulignée.

7 Feuille de route pour 2015-2019

7.1 Définition des besoins prioritaires

Avant de préciser l'étape des objectifs, nous avons déterminé les besoins prioritaires pour la langue occitane, qui découlent de l'analyse de l'inventaire et des expériences d'autres langues.

Type de ressource linguistique	Ressource à créer
Corpus	Corpus textuels
Ressources lexicales	Base lexicale monolingue
	Base lexicale bilingue
Grammaires	Base grammaticale/syntaxique
Type de d'outil linguistique	**Outil à créer**
Traitement de la parole	Reconnaissance de la parole
	Synthèse vocale
Détection automatique de la langue	Détecteur de l'occitan
	Détecteur des variantes de l'occitan
Analyse grammaticale	Correcteurs orthographiques
	Analyseurs
Analyse sémantique	Base de connaissance lexicale
Traduction automatique	Traducteurs automatiques
	Transcripteur automatique entre variantes
Autres outils : logiciels	**Outil à créer**
Logiciels disponibles en occitan	OS + Applications principales

En prenant cet ensemble de priorités comme point de départ, le groupe de travail a dû préciser les objectifs stratégiques du développement des technologies langagières pour l'occitan. Nous détaillons ces objectifs dans la section suivante.

7.2 Objectifs

Type de ressource	Ressource à créer	Objectifs
Corpus	Corpus textuels	Trois types de corpus spécialisés : a) corpus de textes validés, à partir de documents originellement imprimés ; objectif : 10 à 25 millions de mots ; b) corpus web : corpus construit automatiquement en utilisant des documents publiés sur Internet ; objectif : 5 millions de mots ; c) corpus parallèle français-occitan ; objectif : 2 à 5 millions. Caractérisation de textes inclus : variantes, genre, domaine de connaissances, périodes, qualité linguistique (à définir). Niveau d'analyse linguistique : catégorie, puis lemme ; analyse syntaxique (dépendances).
Ressources lexicales - Lexiques/Bases terminologiques	Base lexicale monolingue	Base de données lexicale pour l'occitan : a) entre 500 000 et un million d'entrées à terme (y compris les formes infléchies) ; b) information sur chaque entrée : entrée, lemme, catégorie grammaticale, informations morphosyntaxiques, variante, usage (hyperonymie, domaine, registre, géographique…) lien entre variantes. Base terminologique : le travail terminologique se fait en fonction des demandes liées aux politiques régionales ; les objectifs dans ce champ doivent également tenir compte des demandes du conseil des usagers de Lo Congrès.
	Base lexicale bilingue	Base de données lexicale bilingue : a) priorité français-occitan ; b) catalan-occitan ; c) mêmes champs que le base lexicale monolingue ; d) orientée vers la traduction automatique.
Grammaires	Base grammaticale/syntaxique	Grammaire descriptive de l'occitan et de ses différentes variantes, conçue comme un ensemble de règles pour son utilisation ultérieure en informatique (analyseurs morphologiques et syntaxiques automatiques).

Type de d'outil linguistique	Outil à créer	Objectifs
Traitement de la parole	Reconnaissance de la parole	Il est impossible de créer un véritable outil de reconnaissance de la parole d'ici 2019. Mais il faut commencer à identifier et créer les ressources nécessaires à son développement technologique. L'intérêt particulier de cet outil serait son utilisation pour le sous-titrage automatique.
	Synthèse vocale	Outil de synthèse vocale (TTS) : a) pour toutes les variantes principales ; b) deux locuteurs minimum, avec parité.
Détection automatique de la langue	Détecteur de l'occitan	Développement d'un détecteur de textes écrits en occitan, à partir des travaux faits jusqu'à présent.
	Détecteur des variantes de l'occitan	Développement d'un détecteur des variantes écrites de l'occitan (pour toutes les variantes).
Analyse grammaticale	Correcteurs orthographiques	Correcteur orthographique polyvalent : a) qui permette à l'utilisateur de choisir une ou plusieurs variantes comme référence pour la correction ; b) pour toute plateforme, avec priorité aux logiciels libres, mais également aux logiciels les plus utilisés : 1) Open Office, Libre Office, Microsoft Office ; 2) modules pour navigateurs ; c) suite Adobe, PAO Clavier prédictif et autocorrection pour toutes les variantes et pour les téléphones portables les plus utilisés (*Android, iPhone, Windows, Blackberry*... ou ceux à venir). Basé sur les fréquences des mots dans des corpus.
	Analyseurs	Développement d'un analyseur morphologique (*tagger*) robuste à partir des travaux faits jusqu'à présent.
Analyse sémantique	Base de connaissance lexicale	Développement d'un outil d'analyse syntaxique superficielle/profonde. Construction de la première version de base de connaissance lexicale pour l'occitan (15 000 concepts), en utilisant les entrées de la base de données lexicale comme source d'information pour les variantes de l'occitan.

Traduction automatique	Traducteurs automatiques	Développement, en réutilisant les travaux existants, des outils pour la traduction automatique : a) en priorité, sens occitan ⇢ français (toutes les variantes) ; b) français ⇢ occitan (le système permettra à l'utilisateur de choisir une ou plusieurs variantes comme référence(s) pour la traduction ; utiliser le transcripteur de variantes au besoin) ; c) besoin des outils basiques pour la paire occitan-anglais.
	Transcripteur automatique entre variantes	Développement d'un convertisseur automatique entre variantes, qui travaille non seulement au niveau du lexique mais aussi sur la syntaxe. Il permettra entre autres de réaliser certains outils pour une seule variante, et de pouvoir ensuite les utiliser avec toutes (ex : [traducteur français ⇢ occitan languedocien] + [transcripteur languedocien ⇢ gascon] = [traducteur français ⇢ occitan gascon]).
Autres outils : logiciels	**Outil à créer**	**Objectifs**
Logiciels disponibles en occitan	OS + Applications principales	Traduction des logiciels les plus utilisés et des logiciels libres (bureautique, navigateurs, mails, réseaux sociaux…).

7.3 Feuille de route

Lors de la configuration d'une feuille de route et d'un calendrier pour la réalisation des objectifs ci-dessus, nous devons considérer qu'il existe des dépendances entre eux. Ci-dessous, nous décrivons les dépendances plus significatives.

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

Compte tenu de ces dépendances, le groupe de travail a élaboré la proposition suivante comme une feuille de route pour le développement de technologies de la langue occitane au cours de la période 2015-2019. L'année proposée pour atteindre l'objectif (jalon) a été marquée par le symbole ✓.

Feuille de route 2015-209

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

* 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

La feuille de route proposée décrit une stratégie harmonisée 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 être le point de départ de la feuille de route présentée. Il est donc essentiel de réutiliser les ressources et les outils existants. Si cela est fait, il sera possible de cibler les efforts à venir pour les améliorer et pour développer de nouvelles ressources et outils.

Cette stratégie est basée sur la maximisation de la coopération entre les acteurs et les personnes intéressées par ce développement, et exigera un effort collectif pour veiller à ce que cette coopération soit fructueuse et propice à la situation que nous visons pour fin 2019.