Outiller une langue peu dotée grâce au TALN : l'exemple du corse et de la BDLC

Laurent Kevers Florian Guéniot A. Ghjacumina Tognotti Stella Retali-Medori
UMR CNRS 6240 LISA, Università di Corsica - Pasquale Paoli
Avenue Jean Nicoli, 20250 Corte, France
kevers_l,gueniot_f,tognotti_a,medori_e@univ-corse.fr

RÉSUMÉ

Nos recherches sur la langue corse nous amènent naturellement à envisager l'utilisation d'outils pour le traitement automatique du langage. Après une brève introduction sur le corse et sur le projet qui constitue notre cadre de travail, nous proposons un état des lieux concernant l'application du TAL aux langues peu dotées, dont le corse. Nous définissons ensuite les actions qui peuvent être entreprises, ainsi que la manière dont elles peuvent s'intégrer dans le cadre de notre projet, afin de progresser vers la constitution de ressources et la construction d'outils pour le TAL corse.

ABSTRACT

Tooling up a less-resourced language with NLP : the example of Corsican and BDLC

Our research on the Corsican language naturally leads us to consider the use of NLP tools. After a brief introduction on Corsican and the project that constitutes our working environment, we propose an overview about the use of NLP for less-resourced languages, including Corsican. We then define the actions that can be undertaken, as well as how they can be integrated into our project, in order to progress towards the constitution of resources and the construction of tools for Corsican NLP.

MOTS-CLÉS : langues peu dotées, corse, ressources linguistiques, lemmatisation, POS.

KEYWORDS: less-resourced languages, Corsican, linguistic resources, lemmatisation, POS.

1 La langue corse et sa diffusion numérique

Le corse est une langue issue du latin et s'inscrit dans l'ensemble italo-roman ; il a connu divers contacts et influences linguistiques. Il présente en particulier des affinités avec le toscan médiéval liées spécialement à la domination pisane (IX^{ème}-XIII^{ème} s.). Le toscan constitue le superstrat du corse, et celui-ci connaît des emprunts à d'autres variétés italo-romanes voire romanes ainsi qu'aux langues germaniques et à l'arabe.

Sur le plan dialectal quatre, voire cinq aires sont identifiables (Dalbera-Stefanaggi, 2002, 2007), et l'aire méridionale extrême franchit même les frontières de la Corse puisqu'elle se prolonge en Gallura, dans le nord de la Sardaigne. Ces cinq aires dialectales constituent toutefois un *continuum* et ne font pas obstacle à intercompréhension entre les locuteurs des diverses variétés voire avec les variétés centrales et méridionales de l'Italie.

Les affinités génétiques et historiques que le corse a entretenues pendant de nombreux siècles avec le toscan en ont fait, avec l'italien qui lui a succédé, la langue d'écriture des insulaires depuis le Moyen

Âge jusqu'à l'émergence d'une écriture consciente en langue corse au XIX[e] s. L'orthographe du corse est, de ce fait et avec des adaptations, fondée sur le système graphique de l'italien [1]. Toutefois, et malgré la mise en œuvre d'une approche polynomique permettant d'englober l'ensemble des variantes dialectales (Marcellesi, 1984), l'écriture de la langue n'est pas normée ce qui implique une certaine difficulté pour son traitement automatique.

Aujourd'hui en rapport de diglossie avec le français, l'usage du corse tend à régresser et le développement d'outils est nécessaire pour sa préservation, sa valorisation, sa transmission et sa promotion [2]. Une politique au service de la langue corse est active sur le territoire insulaire notamment pour son développement par les nouvelles technologies. Si plusieurs outils et ouvrages existent pour la description linguistique des parlers corses ou encore pour leur apprentissage, l'inscription du corse dans les humanités numériques reste lacunaire [3].

En particulier, les sites et applications relatifs à la traduction, au vocabulaire et à la syntaxe contiennent peu de données en confrontation à la richesse et la complexité de la langue corse. Cette richesse se retrouve en revanche sur les bases de données telles que la *Banque de Données Langue Corse* (BDLC) et *Infcor* [4]. Cette dernière a été conçue dans un cadre associatif par l'ADECEC qui œuvre dans le domaine de la langue et la culture corses ; elle a été déclinée sous la forme d'une application smartphone mise à disposition du grand public et propose de nombreuses fiches lexicales avec des modes d'interrogation multicritères [5]. Concernant la BDLC, il s'agit d'un outil conçu dans un contexte scientifique associé à la réalisation d'un atlas linguistique, le NALC, comme nous allons l'évoquer.

2 Le projet *Banque de Données Langue Corse* (BDLC)

Le *Nouvel Atlas Linguistique et ethnographique de la Corse* (NALC) a été conçu par le CNRS en 1975 et confié à Marie-José Dalbera-Stefanaggi en 1981 à l'ouverture de l'Université de Corse. En 1986, répondant à une demande de l'Assemblée Régionale de Corse, a été créée la *Banque de Données Langue Corse* (BDLC) qui s'est articulée tout naturellement avec l'atlas projeté [6].

Le NALC-BDLC accueille des données linguistiques en lien avec des savoir-faire et des traditions culturelles corses sur l'ensemble du territoire insulaire. Lors d'enquêtes sur le terrain auprès de locuteurs locaux, ces données sont collectées grâce à des questionnaires thématiques [7] constitués de listes de mots en français [8] : par exemple « la vigne », « le cep », « tailler la vigne », « le tonneau ». À partir d'une question individuelle telle que *« Cumu si dici* "tailler la vigne" *in corsu ? »* (fr. :

1. Cf. notamment Retali-Medori (2015) pour une synthèse sur la question.
2. Selon les recommandations de (UNESCO, 2003).
3. La représentation de cette langue sur le web peut se retrouver sous diverses formes : interfaces en langue corse (les moteurs de recherche *Qwant* et *Google*, ou des réseaux sociaux tels que *Facebook*), sites ou applications tels que *Google Translate* ou *Wiktionnaire*, applications pour smartphone (orientées tourisme, éducation ou traduction), ressources pédagogiques en ligne par le biais de sites d'apprentissage de la langue corse, blogs relatifs à la langue et à la culture corses.
4. *Banca di dati di a lingua corsa* : `http://infcor.adecec.net`
5. Même si l'ensemble des fiches lexicales nécessiterait une révision des informations relatives à la variation, aux signifiés, à la morphologie et à l'étymologie, le matériel contenu dans cette base est incontournable et pourrait aider aux développements de nouveaux outils.
6. Une synthèse de l'histoire du projet est présentée par Dalbera-Stefanaggi & Retali-Medori (2015). Le programme est dirigé depuis 2015 par S. Retali-Medori. Une collection de semi-vulgarisation intitulée *Detti è Usi di paesi, matériaux et analyses extraits de la Banque de Données Langue Corse* a en outre été créée en 2006 autour du NALC.
7. Les thèmes développés dans la BDLC pour les recueils lexicaux sont : l'élevage, l'agriculture, l'homme, la maison et la vie quotidienne, la nature, le village ou la ville et les croyances.
8. Les questionnaires ont été créés au début du programme par le biais d'enregistrements préalables réalisés dans l'île sur différents domaines techniques ou culturels. À partir de leurs transcriptions, a été établie la liste des mots, des signifiés, nommée en d'autres termes par M. J. Dalbera-Stefanaggi le « responsaire » (Dalbera-Stefanaggi, 1992, p. 397).

« Comment dit-on “tailler la vigne” en corse ? »), la collecte des traductions correspondantes en corse est permise et un entretien semi-dirigé entièrement en corse et relatif aux pratiques s'engage afin de recueillir aussi des ethnotextes (témoignages). Ces données sont ensuite traitées et analysées –sur le plan linguistique– et mises en ligne sur le site `http://bdlc.univ-corse.fr`.

Si cette base de données constitue un réel outil pour le développement du TAL appliqué au corse, une des difficultés majeures provient de la riche variation caractérisant la langue corse. Selon les signifiés, des variations lexicales importantes sont attestées : par exemple pour désigner l'acte d'épamprer la vigne, 25 lemmes ont été collectés. Ces lemmes vont à leur tour connaître des transcriptions variables notamment en conséquence de la non-normalisation de la langue corse et d'une production réalisée par différents transcripteurs au cours des 30 années d'existence du programme. Les différents choix d'écritures répondent à des objectifs tels que :

- valoriser la variation : par exemple pour nommer « la jarre », selon la prononciation dans les localités enquêtées, nous trouverons les formes *cerra* et *gerra* issues du même étymon ;
- indiquer graphiquement dans les textes l'aperture des voyelles des proparoxytons (*tróvula* : « écuelle » ; *còmpulu* : « abri » ; *pèrgula* : « treille » ; *tépidu* : « tiède ») ou l'accentuation des hiatus (*durmìa* : « il dormait ») ;
- représenter des enclitiques tels que *fanne* (« en faire ») correspondant à *fà* (« faire ») + *ne* (« en »).

3 Réingénierie et modernisation des outils de la BDLC

En 1986, une première version de l'application BDLC a été développée en *standalone* en utilisant le système de gestion de bases de données 4D. Cette application proposait une fonction de recherche basée sur un ensemble fermé de mots en français, organisés par thèmes. Les réponses obtenues correspondaient à une ou plusieurs traductions en corse et comportaient la forme phonique, la forme graphique, le lemme et la localité d'origine de la forme, et pouvaient être illustrées par des sons ou des photos. Des données morphologiques (flexion et découpage morphématique) et étymologiques étaient également disponibles. La géolocalisation des réponses permettait la visualisation des variations dialectales à l'échelle de l'île avec des cartes de lemmes ou des variantes phoniques.

Une boîte à outils proposait plusieurs modules complémentaires : un module morphologique donnait accès à des tableaux verbaux et nominaux ainsi qu'à la segmentation des mots en morphèmes ; un module étymologique donnait la liste des étymons des termes corses, ou la liste des continuateurs corses d'un étymon, ainsi qu'un fichier morphématique étymologique (découpage des étymons en morphèmes). Il était aussi possible d'établir des requêtes de phonétique diachronique ou de morphologie diachronique (segmentation ou reconstruction en morphèmes des étymons).

L'obsolescence de l'application, et le fait qu'elle soit *standalone*, impliquait la coexistence de différentes versions selon les postes de travail. Une refonte complète a donc été réalisée en tirant parti des technologies web actuelles : PHP et Javascript pour l'applicatif, MySQL pour la base de données. Cette nouvelle version, reprenant en substance une grande partie des fonctionnalités de son aïeule, a connu différentes versions successives. En 2016, la dernière version a été développée en interne par Florian Guéniot (IGE, CNRS) et Aloïs Beck (Alternant, Université de Corse) dont l'objectif a été double : recoder le moteur de l'application afin de le rendre évolutif et modulaire, et moderniser l'interface graphique afin de le rendre adaptable aux supports de plus en plus variés. Cette refonte complète, y compris de la structure de la base de données, a été l'occasion de faciliter l'accès aux données pour les opérateurs par l'ajout d'un module de recherche multicritères. La base de données comporte actuellement 4.888 questions, 108.867 réponses et 1.288 ethnotextes.

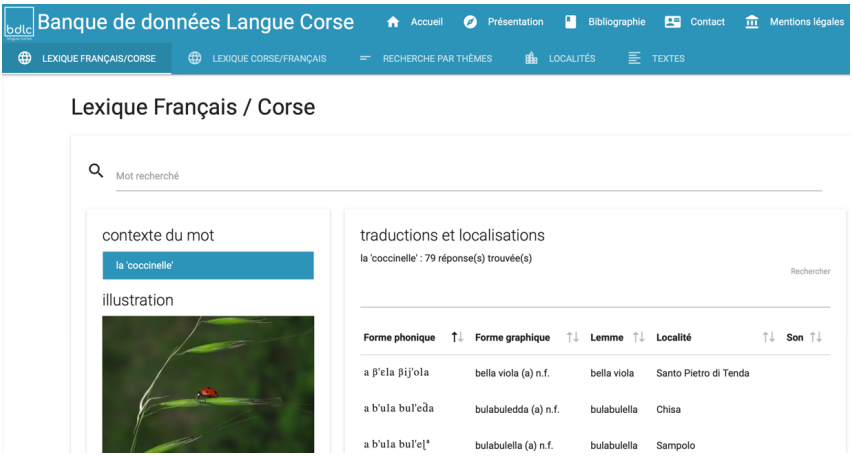

FIGURE 1 – Interface de consultation de la BDLC

4 Vers un TAL corse

Suite à ces évolutions techniques qui modernisent et pérennisent le projet BDLC, grâce à son corpus à la fois authentique et reflet de la complexité et de la richesse de la langue, et étant donné son enrichissement constant, la *Banque de Données Langue Corse* peut constituer un terreau intéressant afin d'œuvrer à la construction de ressources et d'outils pour le traitement automatique du corse.

4.1 État de la question

À notre connaissance, il n'existe que très peu de ressources et d'outils pour le TAL corse. Le rapport de l'ELDA de 2014 sur les ressources linguistiques consacrées aux langues de France (Leixa *et al.*, 2014) recense 93 ressources pour le corse. Plus d'un tiers de celles-ci sont des enregistrements issus du projet BDLC. Les deux tiers restants sont constitués de documents divers : blogs, articles scientifiques, sites institutionnels, sites de journaux, etc. On y retrouve également quelques lexiques, dont *Infcor* ou le *Wiktionnaire corse*, déjà mentionnés précédemment. En addition à cet inventaire, on peut touver quelques autres contributions, au rang desquelles figure le réseau sémantique *BabelNet* (Navigli & Ponzetto, 2012) qui propose un certain nombre d'éléments en corse, ou encore les ressources corses constituées au sein du *Crúbadán Project* (Scannell, 2007). À l'exception de ces dernières, la majorité des ressources disponibles ne sont pas directement exploitables en TAL.

Le corse rentre dans la catégorie des langues dites « peu ou mal dotées », ou encore « minoritaires ». Ces langues constituent un domaine de recherche actif. La conférence TALN a accueilli plusieurs événements dédiés à cette question, entre autres le workshop « Traitement automatique des langues minoritaires et des petites langues » (Streiter, 2003), ainsi que les workshops TALaRE, « Traitement Automatique des Langues Régionales de France et d'Europe » (Morin & Estève, 2013; Vergez-Couret *et al.*, 2015). De même, la conférence LREC a proposé de multiples workshops, dont les plus récents sont SaLTMiL (Alegria *et al.*, 2010; De Pauw *et al.*, 2012) et CCURL (Pretorius *et al.*, 2014; Soria *et al.*, 2016, 2018). Dernièrement, la revue TAL a également sorti un numéro thématique sur le sujet (Bernhard & Soria, 2018). La place de la langue corse dans ces publications est cependant faible.

4.2 Objectifs et moyens

Lorsque l'on désire initier ou améliorer le traitement informatique d'une langue peu ou mal dotée, il est logique de créer les ressources de base avant de s'attaquer aux outils. Ces ressources sont habituellement constituées de lexiques et/ou de corpus, annotés ou non, monolingues ou parallèles. Les outils sont souvent élaborés suivant une complexité croissante. On partira par exemple d'un détecteur de langues, déjà utile lors de la phase de constitution des corpus, pour développer ensuite des composants d'analyse morphosyntaxique et lexicale, pour enfin aller vers des applications de plus haut niveau telles que la correction orthographique ou la traduction automatique. Un aperçu d'actions à entreprendre pour améliorer la capacité digitale des langues est proposé par Ceberio Berger *et al.* (2018). Notre approche rejoint leurs recommandations.

Dans le cadre de notre projet, nous avons décidé d'avancer sur différents points en parallèle. En termes d'objectifs, nous désirons disposer en premier lieu :

- d'un dictionnaire électronique exploitable pour le TAL ;
- d'une interface de consultation de textes capable de générer des concordances et de répondre à des requêtes incluant des critères linguistiques (interrogation, éventuellement combinée, sur les formes, les lemmes, les catégories grammaticales et flexionnelles, etc.) ;
- d'un outil de détection de langue ;
- d'un outil d'annotation morphosyntaxique.

Du point de vue technique, ces objectifs impliquent :

- la constitution initiale et l'enrichissement progressif du dictionnaire électronique ;
- la définition d'une procédure de lemmatisation et son application à une base textuelle à intégrer dans l'interface de consultation (en l'occurrence, les ethnotextes issus de la BDLC) ;
- la mise en place de cette interface ;
- la création de corpus corses à des fins d'entraînement (e.a. pour la détection de langue et l'annotation morphosyntaxique) ;
- la mise au point des outils de détection de langue et d'annotation morphosyntaxique.

Nous avons donc en premier lieu construit une version initiale du dictionnaire à partir d'un export de la BDLC, ainsi que de quelques ajouts extérieurs en ce qui concerne les verbes [9]. Les données ont été organisées selon le format des dictionnaires du LADL [10] (Gross, 1989; Courtois, 1990; Silberztein, 1993). Cette première ressource permet l'initialisation du processus de lemmatisation. Elle sera, en retour, enrichie à l'issue de celui-ci. Actuellement, le dictionnaire compte 20.875 formes, dont 17.860 formes simples (se rapportant à 10.224 lemmes) et 3.015 formes composées (se rapportant à 2.244 lemmes). Lorsque ce dictionnaire est appliqué à notre corpus d'ethnotextes représentant environ 160.000 formes, dont un peu moins de 15.000 uniques, environ 49 % des occurrences sont reconnues. Pour ces éléments, plusieurs analyses concurrentes peuvent coexister (ambiguïté lexicale) et l'analyse correcte peut éventuellement être absente (incomplétude du dictionnaire). Notons encore qu'un traitement des formes non reconnues les plus fréquentes permet d'améliorer rapidement la couverture : les 20 premiers de ces éléments couvrent pas moins de 31 % du total des formes inconnues. À terme, ce dictionnaire constituera une ressource directement exploitable en TAL.

Le deuxième chantier entamé est celui de la lemmatisation. Cette tâche répond à un triple objectif. D'une part, permettre une interrogation des textes sur des critères linguistiques, ainsi qu'une restitution des résultats sous la forme de concordances. D'autre part, nous visons la constitution d'un

9. Les principales formes de *esse* (« être »), *avè* (« avoir »), *andà* (« aller »), *dà* (« dire »), *fà* (« faire »), *stà* (« être », état).
10. Les entrées sont enregistrées dans des fichiers textes sous le format suivant : `forme,lemme.codes_grammaticaux_sémantiques:code_flexionnels/commentaire`.

corpus annoté permettant, dans le futur, de réaliser des apprentissages artificiels (e.a. un étiqueteur morphosyntaxique, cf. *infra*). Enfin, comme déjà exposé, la lemmatisation permettra l'élaboration progressive d'un lexique électronique exploitable en TAL. La définition de la procédure de lemmatisation s'appuie sur l'expérience du projet GREgORI [11]. Celui-ci a, depuis des années, mis au point des méthodologies et des outils pour l'aide à la lemmatisation de textes en grec ancien et dans les principales langues de l'orient chrétien (Kevers & Kindt, 2004; Kindt, 2018). Ces outils, qui peuvent être transposés pour le traitement du corse, permettent une automatisation partielle de la lemmatisation (Kindt, 2012). En pratique, nous avons défini le déroulement de la lemmatisation en deux grandes étapes (figure 2) : l'« étude lexicographique », durant laquelle les formes inconnues du dictionnaire sont ajoutées à celui-ci, et la « désambiguïsation », qui permet de ne conserver qu'une et une seule analyse pour chaque forme du texte. Lors de cette seconde étape, des ajouts au dictionnaire sont encore envisageables. C'est le cas des formes qui n'ont pas été prises en compte lors de l'étude lexicographique, car déjà présentes dans le dictionnaire, mais pour lesquelles l'analyse adéquate n'est pas encore proposée. Concrètement, ces traitements sont mis en œuvre au moyen du logiciel Unitex [12] (Paumier, 2016). Le processus de lemmatisation représente un effort conséquent, mais il permet de travailler au plus près des données et d'en améliorer la qualité, ce qui est important en vue de leur consultation à des fins scientifiques et de leur utilisation pour l'apprentissage artificiel.

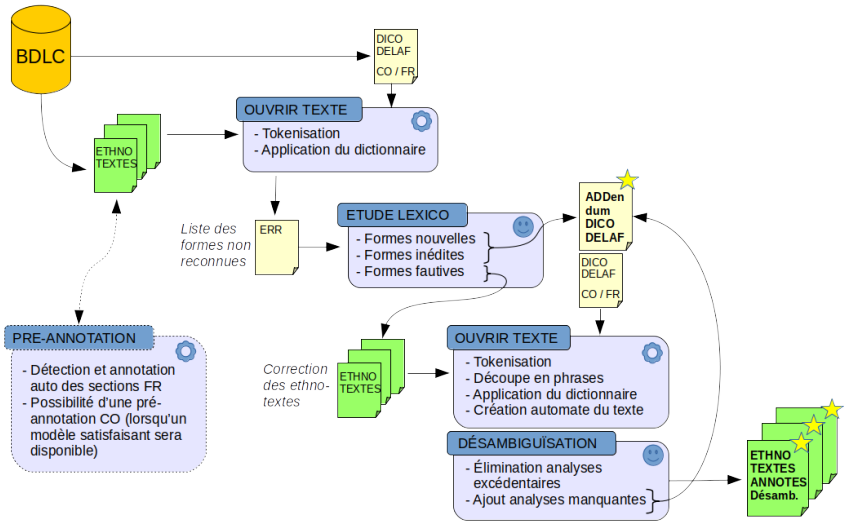

FIGURE 2 – Processus de lemmatisation

Cette procédure est en cours de précision sur différents points importants liés aux variations dialectales et à la non normalisation : choix du lemme entre les différentes versions attestées selon les régions, gestion des variations dûes à l'utilisation variable des accents. Différentes actions sont également entreprises afin de rassembler des données lexicales supplémentaires, extérieures à la BDLC, susceptibles de venir enrichir la version initiale du dictionnaire. En effet, suite à nos premières analyses, nous avons constaté qu'il serait intéressant, en termes d'efficacité, de démarrer la lemmatisation avec

11. Centre d'études orientales, Institut Orientaliste de l'université de Louvain-la-Neuve (Belgique) : `https://uclouvain.be/fr/instituts-recherche/incal/ciol/gregori-project.html` ; assisté par le Centre de traitement automatique du langage : `https://uclouvain.be/fr/instituts-recherche/ilc/cental`.

12. `https://unitexgramlab.org`

un lexique électronique plus couvrant que celui dont nous disposons actuellement.

Au delà du traitement des ethnotextes de la BDLC, nous travaillons également à la constitution de plus gros volumes de textes dans le but de créer des corpus, mono- ou multilingues (3ème chantier). La récolte de documents sur le web et la recherche de corpus déjà constitués font partie de nos préoccupations actuelles. Le recours à des techniques de *crowdsourcing* appliquées à la construction de ressources linguistiques (Millour & Fort, 2018) pourra également être envisagé, non seulement pour produire des corpus annotés en parties du discours, mais aussi des lexiques et corpus multilingues. Ces ressources pourront nous être utiles pour la mise au point d'outils (annotation morphosyntaxique, détection de langue, cf. *infra*), mais aussi à plus long terme, nous permettre de travailler à l'élaboration d'applications de haut niveau telles que la traduction automatique [13], principalement vers l'italien et le français.

L'entraînement et l'utilisation d'un étiqueteur morphosyntaxique, tel que le Tree Tagger (Schmid, 1994), le Stanford Part-Of-Speech Tagger (Toutanova *et al.*, 2003) ou encore d'un outil tel que Wapiti (Lavergne *et al.*, 2010), sont provisoirement laissés de côté. Au fur et à mesure de la production de textes lemmatisés, des modèles de plus en plus complets pourront cependant être entraînés. Dès qu'un niveau de qualité suffisant sera atteint, l'étiqueteur pourra venir effectuer une pré-annotation en support du processus de lemmatisation (figure 2). Parallèlement, nous testons également la possibilité d'exploiter un étiqueteur déjà entraîné pour une langue proche et mieux dotée, en l'occurrence l'italien à l'aide du Tree Tagger. Ce type de démarche a entre autres déjà été expérimentée avec des résultats intéressants pour l'alsacien et l'occitan (Bernhard & Ligozat, 2013; Vergez-Couret, 2013; Bernhard *et al.*, 2018).

Enfin, la quatrième action, en cours actuellement, concerne la mise au point d'un détecteur de langue capable de reconnaître le corse. En plus de l'intérêt intrinsèque de cet outil, il nous sera à nouveau utile pour les tâches déjà initiées, e.a. la lemmatisation (figure 2) et la constitution de corpus. Nous nous intéressons dans un premier temps à la détection de la langue principale du texte, mais visons également la détection de segments de différentes langues à l'intérieur du document (les textes de la BDLC mixent parfois le corse et le français). Pour ce travail, nous nous appuyons sur l'étude de Jauhiainen *et al.* (2018) et avons commencé à tester et apprendre des modèles pour différents outils, sur la base des premiers corpus récoltés. Au vu des premiers résultats, nous devrions être en mesure de nous rapprocher assez rapidement des performances au niveau de l'état de l'art.

5 Conclusion

Nos recherches sur la langue corse nous amènent naturellement à envisager l'utilisation d'outils pour le traitement automatique du langage. L'état des lieux de ce domaine, et plus particulièrement de son application au corse, nous a révélé le manque de ressources et d'outils en la matière. Nous pensons cependant que notre projet peut constituer un terreau intéressant afin d'œuvrer à leur construction. Nous avons donc esquissé les grandes lignes des actions à entreprendre pour nous faire progresser vers la mise en œuvre du TAL corse : constitution de corpus annotés par la lemmatisation semi-automatique, création concomitante d'un lexique exploitable pour le TAL, expérimentation d'outils de base, dont un détecteur de langues et un analyseur morphosyntaxique. Les premières étapes ont été entamées, les plus ambitieuses suivront progressivement. À plus long terme, une application telle que la traduction automatique est envisagée. Les travaux réalisés sur d'autres langues régionales ou

13. Un premier essai, non satisfaisant, utilisant le *deep learning* (Tensor Flow) pour la traduction automatique, nous a montré la nécessité de disposer de données conséquentes pour cette tâche.

peu dotées, dont nous avons dressé un bref aperçu, nous guideront dans ce cheminement.

Références

I. ALEGRIA, N. BEL, L. BORIN, H. LOFTSSON, F. SANCHEZ-MARTINEZ, K. SCANNELL, T. TROSTERUD, P. LANGGA, P. MEURER, S. MOSHAGEN, E. NAVAS & D. TOMAS, Eds. (2010). *7th SaLTMiL Workshop on Creation and use of basic lexical resources for less-resourced languages (LREC2010 Workshop)*.

BERNHARD D. & LIGOZAT A.-L. (2013). Es esch fàscht wie Ditsch, oder net ? Étiquetage morphosyntaxique de l'alsacien en passant par l'allemand. In *TALaRE, Traitement Automatique des Langues Régionales de France et d'Europe. Atelier TALN 2013*, p. 209–220, Les Sables d'Olonne, France.

BERNHARD D., LIGOZAT A.-L., MARTIN F., BRAS M., MAGISTRY P., VERGEZ-COURET M., STEIBLÉ L., ERHART P., HATHOUT N., HUCK D., REY C., REYNES P., ROSSET S., SIBILLE J. & LAVERGNE T. (2018). Corpora with Part-of-Speech Annotations for Three Regional Languages of France : Alsatian, Occitan and Picard. In *Proceedings of the 11th edition of the Language Resources and Evaluation Conference (LREC'18)*, Miyazaki, Japon.

BERNHARD D. & SORIA C. (2018). Traitement automatique des langues peu dotées. *Traitement Automatique des Langues*, **59**(3).

CEBERIO BERGER K., GURRUTXAGA HERNAIZ A., BARONI P., HICKS D., KRUSE E., QUOCHI V., RUSSO I., SALONEN T., SARHIMAA A. & SORIA C. (2018). *Digital Language Survival Kit. The DLDP Recommendations to Improve Digital Vitality*. The Digital Language Diversity Project. Accessible à l'adresse <http ://www.dldp.eu/sites/default/files/documents/DLDP_Digital-Language-Survival-Kit.pdf>.

COURTOIS B. (1990). Un système de dictionnaires électroniques pour les mots simples du français. *Langue Française*, **87**, 11–22.

DALBERA-STEFANAGGI M.-J. (1992). Le Nouvel Atlas Linguistique de la Corse et son articulation sur une base de données. In *Atlanti Linguistici italiani e romanzi : esperienze a confronto. Atti del Congresso Internazionale (Palermo 3-7 Ottobre 1990)*, p. 395–402, Palermo : Centro di Studi Filologici e Linguistici Siciliani.

DALBERA-STEFANAGGI M.-J. (2002). *La langue corse*. Number 3641 in Que sais-je ? Paris : PUF.

DALBERA-STEFANAGGI M.-J. (2007). *Nouvel atlas linguistique et ethnographique de la Corse : Volume 1, Aréologie phonétique, édition revue et corrigée*. Ajaccio : Paris : Comité des travaux historiques et scientifiques - CTHS, Alain Piazzola edition.

DALBERA-STEFANAGGI M.-J. & RETALI-MEDORI S. (2015). Trente ans de dialectologie corse : autour du programme Nouvel Atlas Linguistique et Ethnographique de la Corse et Banque de Données Langue Corse. In S. RETALI-MEDORI, Ed., *Actes du colloque Tribune des chercheurs, études en linguistique*, volume 6 of *Corse d'hier et de demain - Nouvelle série*, p. 17–25, Bastia, France : Société des Sciences Historiques et Naturelles de la Corse.

G. DE PAUW, G.-M. DE SCHRYVER, M. L. FORCADA, K. SARASOLA, F. M. TYERS & P. W. WAGACHA, Eds. (2012). *8th SaLTMIL & AfLaT 2012 Workshop, Language Technology for Normalisation of Less-Resourced Languages (LREC 2012 Workshop)*.

GROSS M. (1989). La construction de dictionnaires électroniques. *Annales de Télécommunications*, **44**, 4–19.

JAUHIAINEN T., LUI M., ZAMPIERI M., BALDWIN T. & LINDÉN K. (2018). Automatic Language Identification in Texts : A Survey. *arXiv :1804.08186 [cs]*.

KEVERS L. & KINDT B. (2004). Vers un concordanceur-lemmatiseur en ligne du grec ancien. *L'Antiquité Classique*, **73**, 203–213.

KINDT B. (2012). *Traitement automatique de l'ambiguïté en grec ancien. Outils informatiques et ressources linguistiques.* Thèse de doctorat., Université catholique de Louvain.

KINDT B. (2018). Processing Tools for Greek and Other Languages of the Christian Middle East. *Journal of Data Mining and Digital Humanities*, **jdmdh :4184**. Special Issue on Computer-Aided Processing of Intertextuality in Ancient Languages,.

LAVERGNE T., CAPPÉ O. & YVON F. (2010). Practical Very Large Scale CRFs. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, p. 504–513, Uppsala, Sweden : Association for Computational Linguistics.

LEIXA J., MAPELLI V. & CHOUKRI K. (2014). *Inventaire des ressources linguistiques des langues de France*. ELDA. Accessible à l'adresse <http ://www.elda.org/media/filer_public/2014/12/17/rapport_dglflf_05112014-1.pdf >.

MARCELLESI J.-B. (1984). La définition des langues en domaine roman : les enseignements à tirer de la situation corse. In *Actes du Congrès de Linguistique et de Philologie Romanes 5*, p. 307–314, Aix-en-Provence.

MILLOUR A. & FORT K. (2018). À l'écoute des locuteurs : production participative de ressources langagières pour des langues non standardisées. *Traitement Automatique des Langues, numéro spécial sur le traitement automatique des langues peu dotées*, **59**(3).

E. MORIN & Y. ESTÈVE, Eds. (2013). *TALaRE 2013 - Traitement Automatique des Langues Régionales de France et d'Europe. Atelier TALN 2013*, Les Sables d'Olonne, France.

NAVIGLI R. & PONZETTO S. P. (2012). BabelNet : The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, **193**, 217–250.

PAUMIER S. (2016). *Unitex 3.1 User Manual*. Université Paris-Est Marne-la-Vallée. Accessible à l'adresse http ://releases.unitexgramlab.org/3.1/man/Unitex-GramLab-3.1-usermanual-en.pdf.

L. PRETORIUS, C. SORIA & P. BARONI, Eds. (2014). *CCURL 2014 : Collaboration and Computing for Under - Resourced Languages in the Linked Open Data Era, LREC 2014 Workshop*.

RETALI-MEDORI S. (2015). La documentation corse. In E. R. MARIA ILIESCU, Ed., *Anthologies, textes, corpus et sources des langues romanes*, number 7 in Manuals of Romance Linguistics, p. 558–564. Tübingen : De Gruyter.

SCANNELL K. P. (2007). The Crúbadán Project : Corpus building for under-resourced languages. In C. FAIRON, H. NAETS, A. KILGARRIFF & G.-M. DE SCHRYVER, Eds., *Proceedings of the 3rd Web as Corpus Workshop*, volume 4 of *Cahiers du Cental*, Louvain-la-Neuve, Belgium.

SCHMID H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of International Conference on New Methods in Language Processing*, Manchester, UK.

SILBERZTEIN M. D. (1993). *Dictionnaires électroniques et analyse automatique de textes. Le système INTEX*. Paris : Masson.

C. SORIA, L. BESACIER & L. PRETORIUS, Eds. (2018). *CCURL 2018 : 3rd Workshop on Collaboration and Computing for Under-Resourced Languages, Sustaining knowledge diversity in the digital age (LREC 2018 Workshop)*.

C. SORIA, L. PRETORIUS, T. DECLERCK, J. MARIANI, K. SCANNELL & E. WANDL-VOGT, Eds. (2016). *CCURL 2016 : Collaboration and Computing for Under-Resourced Languages : Towards an Alliance for Digital Language Diversity (LREC 2016 Workshop)*.

O. STREITER, Ed. (2003). *Traitement automatique des langues minoritaires et des petites langues, Actes du Workshop TALN 2003*, Batz-sur-Mer. ATALA.

TOUTANOVA K., KLEIN D., MANNING C. D. & SINGER Y. (2003). Feature-rich Part-of-speech Tagging with a Cyclic Dependency Network. In *Proceedings of HLT-NAACL 2003*, p. 173–180 : Association for Computational Linguistics.

UNESCO, GROUPE D'EXPERTS SPÉCIAL DE L'UNESCO SUR LES LANGUES EN DANGER (2003). *Vitalité et disparition des langues - UNESCO Bibliothèque Numérique*. Paris : Organisation des Nations Unies pour l'Education, la Science et la Culture. Version française accessible à l'adresse <https ://ich.unesco.org/doc/src/00120-FR.pdf>.

VERGEZ-COURET M. (2013). Tagging Occitan using French and Castillan Tree Tagger. In *Less Resourced Languages, new technologies, new challenges and opportunities*, p.6̃, Poznan, Poland.

M. VERGEZ-COURET, D. BERNHARD, A.-L. LIGOZAT, J.-M. ELOY & C. REY, Eds. (2015). *TALaRE 2015 - Traitement Automatique des Langues Régionales de France et d'Europe. Atelier de TALN 2015*, Caen, France.