Réalisations, obstacles et perspectives pour l'outillage du corse

Laurent Kevers, Alice Millour

Réalisations, Obstacles et Perspectives pour l'Outillage du Corse

Laurent Kevers[1] Alice Millour[2]
(1) UMR CNRS 6240 LISA, Università di Corsica - Pasquale Paoli, Avenue Jean Nicoli, 20250 Corte, France
(2) Université Paris 8, 2 rue de la Liberté, 93526 Saint-Denis cedex, France
kevers_l@univ-corse.fr, am@up8.edu

RÉSUMÉ
Présentation des ressources et outils linguistiques développés depuis 2019 pour le corse.

ABSTRACT
Achievements, challenges and perspectives for the tooling up of Corsican.

Presentation of the language resources and tools developed since 2019 for Corsican.

MOTS-CLÉS : langues peu dotées, corse, corpus, ressources linguistiques, outils, TAL.

KEYWORDS: less-resourced languages, Corsican, corpora, lexical resources, tools, NLP.

1 Contexte et objectifs

Le corse, une des 24 langues de France présente sur le territoire métropolitain, est considéré comme « en danger » par l'UNESCO (Moseley, 2010) et est communément repris parmi les langues peu dotées (Leixa *et al.*, 2014; Joshi *et al.*, 2020).

Il s'agit d'une langue présentant une variation dialectale qui s'étend sur quatre, voire cinq aires géographiques (Dalbera-Stefanaggi, 2002, 2007), qui dépassent même les frontières de la Corse en allant jusqu'en Gallura (nord de la Sardaigne). Il existe une intercompréhension entre les locuteurs des diverses aires, celles-ci formant un *continuum* qui se prolonge jusqu'aux variétés centrales et méridionales de l'Italie. Malgré la mise en oeuvre d'une approche polynomique permettant d'englober l'ensemble des variantes dialectales (Marcellesi, 1984), l'écriture de la langue n'est pas standardisée.

Cet article fait le point sur des travaux entrepris depuis 2019 et qui ont pour objectif de constituer un socle de ressources et d'outils accessibles, autant que possible, selon les principes de la science ouverte, et ce afin d'améliorer progressivement le support du corse dans les outils numériques.

Notre ambition est de constituer des ressources et de développer des outils spécialisés pour le TAL [1], mais aussi à destination des apprenants et du grand public.

Cette action, qui est ancrée au sein de la Banque de Données Langue Corse [2] (BDLC), a également pour but l'ajout de fonctionalités au portail BDLC et le développement de modules d'aide au traitement des données linguistiques brutes.

1. https://bdlc.univ-corse.fr/tal/
2. https://bdlc.univ-corse.fr/

2 État de l'art

Depuis 1986, la BDLC recueille, stocke, analyse et restitue des données dialectales relatives aux savoir-faire et aux traditions culturelles corses, par le biais d'enquêtes de terrain en Corse et dans le nord de la Sardaigne. La BDLC contient plus de 2 000 ethnotextes totalisant 317 512 tokens, ainsi que près de 120 000 entrées lexicales contenant la « question » (forme en français), la « réponse » (forme en corse), le « lemme », et la « commune » (localisation).

En dehors du projet BDLC, plusieurs initiatives de différentes natures ont déjà porté sur cette langue peu dotée : le projet Interreg *INTERTESTU* (Chiorboli, 1995) ; le travail mené par l'association ADECEC [3] ; les traducteurs automatiques *Okchakko* [4] ou *Google Translate* [5] ; etc. Un inventaire plus complet des ressources et outils disponibles pour le corse est proposé par Kevers *et al.* (2021).

Les résultats pérennes, diffusés de manière ouverte et directement exploitables pour le TAL étant néanmoins de faible ampleur, l'impulsion décisive permettant la création progressive, cumulative et cohérente des ressources et outils n'a pas eu lieu.

3 Élaboration des ressources et outils pour le TAL corse

3.1 Approche générale

Notre travail vise à améliorer cette situation et s'oriente vers la constitution de corpus corses libres de droits, la création de modules pour le TAL, et enfin la mise à disposition de données et d'outils pour la recherche, ainsi que d'applications à visée pédagogique ou à destination du grand public.

L'approche adoptée s'appuie sur une interaction entre le projet BDLC proprement dit et les développements réalisés dans le cadre du TAL. La BDLC adopte une démarche de terrain qui lui permet de récolter des données linguistiques qui documentent la langue et ses variations, tant en diachronie qu'en synchronie. Les principaux résultats sont constitués par la base de données en tant que telle, par les outils d'accès et de visualisation des données, ainsi que par des publications scientifiques ou de vulgarisation. Les développements en TAL se déroulent à partir de données provenant de diverses sources qui doivent permettre le traitement le plus large et robuste possible de la langue, tout en essayant de respecter un cadre linguistique qui n'est rattaché à aucune norme strictement formalisée. Les résultats concrets sont constitués de ressources linguistiques – entre autres dictionnaires électroniques et corpus – et d'outils.

Notre volonté est que ces deux démarches aboutissent à une interaction et un enrichissent mutuel durant lesquels la BDLC apporte ses données lexicales et textuelles dialectales, ses connaissances linguistiques ainsi qu'un cadre structurant l'automatisation du traitement d'une langue non normalisée. De son côté le TAL propose des outils de traitement, de correction, d'exploration et d'exploitation des données, qu'elles soient brutes ou déjà dépouillées, ce qui ouvre de nouvelles perspectives pour l'accroissement de la base et pour l'enrichissement des connaissances linguistiques.

3. https://www.adecec.net/
4. http://www.okchakko.com/
5. https://translate.google.fr

3.2 Ressources lexicales et variation

Un premier lexique électronique a été extrait de la BDLC. Il comprend 21 108 formes simples ou composées se référant à 12 498 lemmes. Les formes verbales étant sous représentées dans la BDLC, un complément a été créé, en partie automatiquent, en partie manuellement (Kevers & Retali-Medori, 2020; Retali-Medori & Kevers, 2022). La couverture du lexique général a été estimée à 49% des occurrences du corpus d'ethnotextes de la BDLC, et à environ 16% pour les verbes [6]. Ces ressources, qui restent partielles et qui incluent parfois certaines erreurs ou incohérences, n'ont à ce stade pas encore été diffusées.

Du point de vue de la variation dialectale, les ressources lexicales produites par les linguistes de terrain de la BDLC présentent une richesse importante qui demande à être exploitée. Si les premières expérimentations de génération automatique de variantes dialectales sur la base d'entrées dialectales reliées au même lemme – inspirées de Millour & Fort (2019) – sont prometteuses, la méthode pâtit également du manque de cohérence recontré au sein de la BDLC.

L'exploitation du contenu lexical de la BDLC se heurte donc à divers obstacles. En particulier, la catégorie « lemme » pose des difficultés théoriques importantes dans le cadre d'une langue non standardisée. Ce point a fait l'objet de réflexions publiées dans Retali-Medori & Kevers (2022). La multitude d'acteurs ayant participé à renseigner la base au cours des années sans qu'un guide de saisie clair n'ait été imposé en est également une raison. La collaboration entre linguistes de terrain et spécialistes du TAL prend donc tout son sens dans ce projet, la ressource gagnant à être examinée et corrigée semi-automatiquement par des moyens informatiques afin de pouvoir être exploitée ultérieurement à des fins de production de ressources et d'outils pour le TAL.

3.3 Corpus

Dès le départ, nous avons pu disposer du corpus d'ethnotextes de la BDLC (317 512 mots). Celui-ci est représentatif d'un type de textes particulier : des retranscriptions d'entretiens oraux semi-dirigés. Ces textes ne peuvent à eux seuls constituer le substrat nécessaire à l'élaboration des outils de TAL. Dès lors, des corpus de portée plus générale ont été réunis et diffusés (Kevers & Retali-Medori, 2020) : l'encyclopédie collaborative *Wikipedia* en langue corse (919 382 mots), le *blog* journalistique *A Piazzetta* (504 225 mots), ainsi que la traduction corse de la *Bible* (770 560 mots). Ces corpus ont leurs propres caractéristiques. *Wikipedia* contient des documents issus d'un processus d'édition non centralisé qui pose la question de la qualité et la cohérence linguistique de ces textes. Au contraire, *A Piazzetta*, propose un matériau plus uniforme et contrôlé, mais qui intègre également de nombreux commentaires caractéristiques des *blogs*, ceux-ci présentant de fortes variations de la qualité du contenu linguistique, du registre employé, ainsi que des langues utilisées – le français étant fréquemment utilisé à côté du corse. Enfin, la *Bible* reste par nature un corpus très particulier. Par conséquent, la diversification des corpus reste donc un objectif du projet. Une convention conclue avec le réseau Canopé de Corse [7] nous permet de travailler actuellement à un nouveau corpus constitué d'oeuvres littéraires (adulte, jeunesse, enfants) ainsi que de documents relatifs au patrimoine et à l'histoire corses. Ce corpus, qui dépassera les 500 000 mots, sera diffusé très prochainement.

6. Ces chiffres sont sujets à variation en fonction du corpus d'application. Le fait qu'un mot soit reconnu par le dictionnaire ne signifie pas que l'ambiguïté lexicale ait été prise en compte : une forme reconnue par un dictionnaire peut disposer de plusieurs analyses concurrentes, sans que celles-ci soient forcément toujours adéquates pour toutes les occurrences du corpus.

7. https://www.reseau-canope.fr/canope-academie-corse/

Si les corpus sont encore en phase d'élaboration, nous avons d'ores et déjà mis en place une interface de consultation des ethnotextes de la BDLC au travers d'un concordancier [8]. Ce type d'outil permet une exploration potentiellement assez fine des corpus et la visualisation des résultats sous la forme KWIC (*Keywords in Context*). Le corpus peut être filtré selon les méta-données disponibles – thème et localisation – et interrogé au moyen de requêtes simples ou composées, faisant intervenir des formes brutes ainsi que diverses formes d'expressions régulières [9]. L'outil [10] est prévu pour tirer parti des futures annotations qui viendront enrichir nos corpus : identification de langue, parties du discours, lemmes...

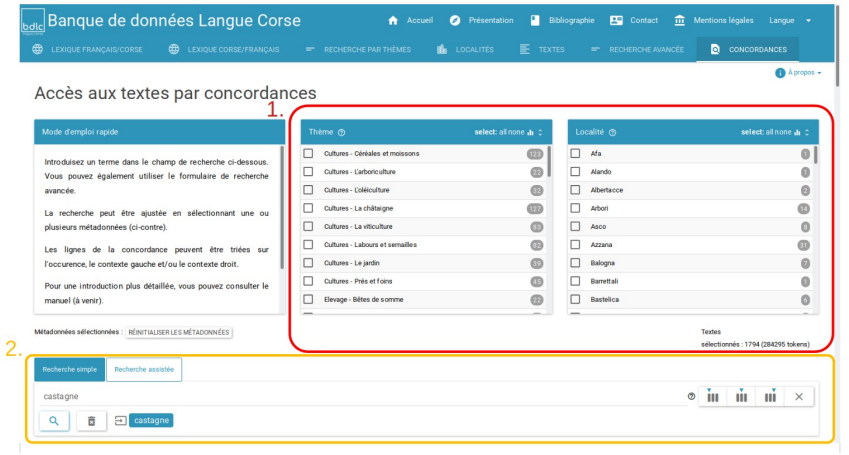

FIGURE 1 – Concordancier : filtrage par méta-données (1.) et zone d'introduction de la requête (2.)

Cette initiative, déjà annoncée dans Kevers & Retali-Medori (2020), constitue un premier enrichissement fonctionnel de la BDLC, à la fois à destination des chercheurs et des enseignants/apprenants. Il est prévu de l'étendre au futur corpus Canopé.

3.4 Outils

Ces différentes ressources ont permis de démarrer l'élaboration de certains modules de TAL, en particulier celui qui concerne l'identification de langue, qui s'avère utile lors de la constitution de corpus. La vérification et l'évaluation de la qualité des textes – comme déjà réalisé pour diverses ressources multilingues de grandes tailles (Kevers, 2022c) – ou l'annotation des documents multilingues [11] (Kevers, 2022a) permettent d'améliorer les corpus. À l'avenir, ce type d'analyse et d'annotation devraient être réalisés sur nos corpus bruts.

D'autre part, grâce à l'annotation manuelle d'un premier corpus de référence de 100 phrases, les premières expérimentations d'annotation morphosyntaxique ont récemment pu être menées. Si nous

8. https://bdlc.univ-corse.fr/concord/
9. Un manuel d'utilisation détaillé a été rédigé et mis à disposition : https://bdlc.univ-corse.fr/concord/docs/user_manual.pdf
10. Adapté et intégré à partir d'un développement réalisé par le Cental (https://uclouvain.be/fr/instituts-recherche/ilc/cental).
11. Spécification de la langue au niveau du mot.

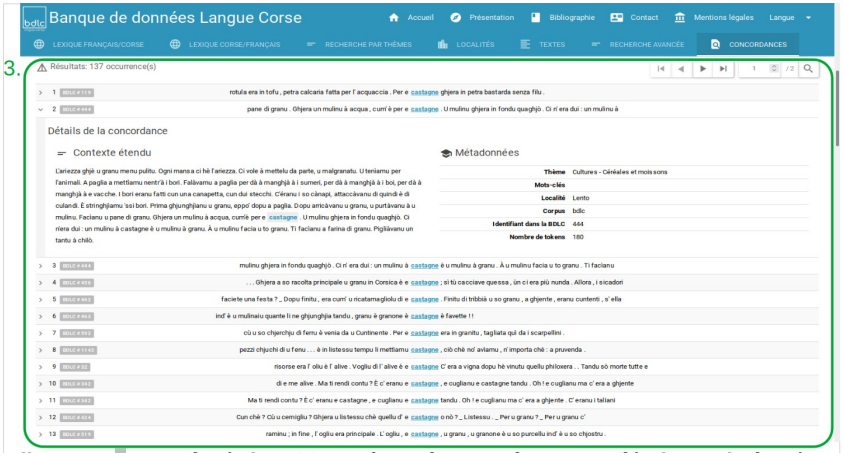

FIGURE 2 – Concordancier : exemple de résultat (3.)

pouvons envisager d'utiliser ce premier modèle d'annotation automatique comme un outil de pré-annotation, l'amélioration des performances reste nécessaire. L'annotation de nouvelles données devrait avoir lieu en 2023 et mener à une progression de la précision de l'outil. Ce module, une fois à maturité, pourra naturellement contribuer à l'annotation semi-automatique de nos corpus, et à leur interrogation au moyen des catégories POS dans le concordancier.

3.5 Questions transversales et méthodologiques

Ce projet est également l'occasion de soulever des questions transversales et communes à différentes langues peu dotées, entre autres la question des droits d'auteurs lors de la constitution de corpus (Kevers & Retali-Medori, 2019), qui a récemment connu un évolution positive avec la transposition française de la directive européenne 2019/790 (Maurel & Rennes, 2021).

L'adoption d'une démarche de science ouverte et le respect des principes FAIR [12] (Wilkinson *et al.*, 2016; Berez-Kroeker *et al.*, 2018) est un point important. Il convient en effet que les résultats des recherches soient reproductibles, que les ressources produites soient disponibles de manière ouverte et pérenne – au minimum pour la recherche – et que les différentes données puissent être correctement identifiées et citées, en particulier au travers des identifiants pérennes (Kevers, 2022b).

Citons aussi les approches par transfert depuis une langue mieux dotée, les méthodes non supervisées ainsi que les démarches participatives de *myriadisation* (Millour, 2020).

Enfin, les aspects méthodologiques liés à la prise en compte des variations dialectales constituent également un point d'intérêt pour de nombreuses langues.

12. Faciles à trouver ; Accessibles ; Interopérables ; Réutilisables

3.6 Obstacles

Au-delà de ces questions, nous pouvons mettre en avant plusieurs difficultés qui ont été rencontrées durant ces dernières années. Tout d'abord, en dépit de la numérisation croissante de la société en général, il reste souvent difficile d'accéder à des contenus de qualité, libres de droits et dans un format structuré natif tel qu'XML. La constitution de corpus reste donc en partie dépendante de problématiques de conversion de formats – par exemple à partir de fichiers PDF – voire de la numérisation de documents imprimés.

D'autre part, nous avons aussi été confrontés à la disponibilité des ressources humaines adéquates. En effet, le projet BDLC s'appuie sur des linguistes spécialistes du corses, mais peu familiers avec les enjeux, méthodes et outils du TAL. La mobilisation de profils spécialisés en TAL n'est de plus pas toujours aisée – en raison de la disponibilité des personnes et des financements – et ceux-ci ne sont pas nécessairement compétents en corse. Enfin, le recrutement de stagiaires est rendu difficile car le cursus universitaire corse ne produit pas d'étudiant disposant de la double compétence linguistique et informatique, et l'intérêt des étudiants de chaque filière pour la discipline complémentaire est limité.

D'une manière générale, la création et la pérennisation de postes dédiés à la problématique des langues peu dotées – et du corse en particulier – sont compliquées, ce qui rend la poursuite des projets à moyen et long termes incertaine.

4 Conclusion

Nous avons résumé les progrès enregistré depuis 2019 et présenté les dernières avancées, en particulier le travail engagé pour la constitution d'un nouveau corpus d'environ 500 000 mots, les premières expérimentations relatives au traitement de la variation ainsi qu'à l'annotation morphosyntaxique.

Les interactions entre linguistique de terrain et TAL ont été mises en évidence, tout comme les différents obstacles auxquels nous avons été confrontés.

Malgré ces difficultés, nous avons progressé sur la production pérenne et ouverte de ressources et outils utiles au traitement automatique du corse. Nous désirons poursuivre ce travail, tout en contribuant le plus possible à une approche générique de l'outillage des langues peu dotées, en particulier durant le projet ANR DiViTal [13] (2022-2025) dont l'Université de Corse est un des partenaires.

Remerciements

Ce travail a été mené grâce au financement CPER : « Un outil linguistique au service de la Corse et des Corses : la Banque de Données Langue Corse (BDLC) » ainsi qu'une bourse post-doctorale de la Collectivité de Corse (CDC).

13. https://divital.gitpages.huma-num.fr/fr/

Références

BEREZ-KROEKER A. L., ANDREASSEN H. N., GAWNE L., HOLTON G., KUNG S. S., PULSIFER P., COLLISTER L. B., THE DATA CITATION AND ATTRIBUTION IN LINGUISTICS GROUP & THE LINGUISTICS DATA INTEREST GROUP (2018). *The Austin Principles of Data Citation in Linguistics*. Rapport interne Version 1.0. https ://site.uit.no/linguisticsdatacitation/austinprinciples.

CHIORBOLI J. (1995). *La gestion du territoire linguistique : INTERTESTU, une base textuelle littéraire et linguistique corse*. Centru di ricerche Corse Gruppulingua, Universuta di Corsica. OCLC : 490793344.

DALBERA-STEFANAGGI M.-J. (2002). *La langue corse*. Volume 3641 de 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 édition.

JOSHI P., SANTY S., BUDHIRAJA A., BALI K. & CHOUDHURY M. (2020). The State and Fate of Linguistic Diversity and Inclusion in the NLP World. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 6282–6293, Online : ACL. DOI : 10.18653/v1/2020.acl-main.560.

KEVERS L. (2022a). CoSwID, a Code Switching Identification Method Suitable for Under-Resourced Languages. In *1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages (SIGUL2022 @LREC2022)*, Proceedings of the the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages (SIGUL2022 @LREC2022), p. 112–121, Marseille, France : European Language Resources Association. Backup Publisher : European Language Resources Association.

KEVERS L. (2022b). L'identifiant pérenne, une clé pour les bases de données linguistiques dans une perspective de science ouverte. In *XXXe Congreso Internacional de Lingüística y Filología Románicas*, La Laguna, Tenerife, Islas Canarias, Spain. HAL : hal-03722878.

KEVERS L. (2022c). L'identification de langue, un outil au service du corse et de l'évaluation des ressources linguistiques. *Traitement Automatique des Langues*, **62**(3). Numéro spécial " Diversité linguistique".

KEVERS L. & RETALI-MEDORI S. (2019). Copyright in the context of tooling up Corsican and other less-resourced languages. In *Proceedings of the 1st International Conference on Language Technologies for All*, p. 198–201, Paris, France : European Language Resources Association (ELRA).

KEVERS L. & RETALI-MEDORI S. (2020). Towards a Corsican Basic Language Resource Kit. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC2020)*, p. 2726–2735, Marseille, France : European Language Resources Association (ELRA).

KEVERS L., RETALI MEDORI S. & TOGNOTTI A. G. (2021). *A Survey of Language Technologies Resources and Tools for Corsican*. Rapport interne, UMR CNRS 6240 LISA, Université de Corse. https://hal.archives-ouvertes.fr/hal-03228733.

LEIXA J., MAPELLI V. & CHOUKRI K. (2014). *Inventaire des ressources linguistiques des langues de France*. Rapport interne, ELDA.

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.

MAUREL L. & RENNES S. (2021). La fouille de textes et de données à des fins de recherche : une pratique confirmée et désormais opérationnelle en droit français. https://www.ouvrirlascience.fr/la-fouille-de-textes-et-de-donnees-a-des-fins-de-recherche-une-pratique-confirmee-et-desormais-operationnelle-en-droit-francais.

MILLOUR A. (2020). *Myriadisation de ressources linguistiques pour le traitement automatique de langues non standardisées*. Theses, Sorbonne Universite.

MILLOUR A. & FORT K. (2019). Unsupervised Data Augmentation for Less-Resourced Languages with no Standardized Spelling. In *RANLP*, p. 776 – 784, Varna, Bulgaria. HAL : hal-02280002.

MOSELEY C., Éd. (2010). *Atlas of the World's Languages in Danger*. Paris : UNESCO Publishing. 3rd edn. Online version : http://www.unesco.org/culture/en/endangeredlanguages/atlas.

RETALI-MEDORI S. & KEVERS L. (2022). La morphologie dans la Banque de Données Langue Corse : bilan et perspectives. *Corpus*, **23**. Numéro thématique « Corpus et données en morphologie ».

WILKINSON M. D., DUMONTIER M., AALBERSBERG I. J. J., APPLETON G., AXTON M., BAAK A., BLOMBERG N., BOITEN J.-W., DA SILVA SANTOS L. B., BOURNE P. E., BOUWMAN J., BROOKES A. J., CLARK T., CROSAS M., DILLO I., DUMON O., EDMUNDS S., EVELO C. T., FINKERS R., GONZALEZ-BELTRAN A., GRAY A. J. G., GROTH P., GOBLE C., GRETHE J. S., HERINGA J., 'T HOEN P. A. C., HOOFT R., KUHN T., KOK R., KOK J., LUSHER S. J., MARTONE M. E., MONS A., PACKER A. L., PERSSON B., ROCCA-SERRA P., ROOS M., VAN SCHAIK R., SANSONE S.-A., SCHULTES E., SENGSTAG T., SLATER T., STRAWN G., SWERTZ M. A., THOMPSON M., VAN DER LEI J., VAN MULLIGEN E., VELTEROP J., WAAGMEESTER A., WITTENBURG P., WOLSTENCROFT K., ZHAO J. & MONS B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, **3**. Article number : 160018 (2016), DOI : 10.1038/sdata.2016.18.