L'identification de langue, un outil au service du corse et de l'évaluation des ressources linguistiques

Laurent Kevers

Cet article a été publié dans la revue 'Traitement Automatique des Langues', « Diversité Linguistique », 62(3) : p.13-37, ATALA, mars 2022. La publication originale est disponible sur le site www.atala.org : https://www.atala.org/content/diversité-linguistique-linguistic-diversity-natural-language-processing.

L'identification de langue, un outil au service du corse et de l'évaluation des ressources linguistiques

Laurent Kevers*

* *UMR CNRS 6240 LISA, Università di Corsica - Pasquale Paoli*
Avenue Jean Nicoli, 20250 Corte, France

RÉSUMÉ. *La constitution de corpus est une des premières priorités que rencontrent les langues peu dotées. L'émergence de ressources issues d'Internet, de tailles de plus en plus imposantes et couvrant de nombreuses langues, peut laisser penser que ce point est désormais résolu, ce qui n'est pas le cas. À la suite de Caswell* et al. *(2021), qui ont évalué plusieurs ressources de grande envergure, dont une disposant de contenu corse, nous avons mené une analyse de deux corpus incluant cette langue : An Crúbadán et W2C. Parallèlement à une évaluation manuelle, nous avons estimé la possibilité d'utiliser un ou plusieurs modules d'identification de langue afin de filtrer le contenu de ces ressources, ce qui s'avère possible mais au prix d'un rappel peu élevé. Pour cette tâche, nous avons testé et réentraîné divers systèmes afin de les adapter au mieux au corse. Ce travail nous permet de mettre à disposition un modèle capable d'identifier le corse ainsi que 17 autres langues européennes.*

ABSTRACT. *The constitution of corpora is one of the first priorities faced by less-resourced languages. The emergence of Internet-based resources of increasing size and covering more and more languages may suggest that this issue has been resolved, but this is not the case. Following Caswell* et al. *(2021), who evaluated several large resources, including one with Corsican content, we conducted an analysis of two corpora including this language: An Crúbadán and W2C. In parallel to a manual evaluation, we considered the possibility of using one or more language identification modules to filter the content of these resources, which turns out to be possible but at the cost of low recall. For this task, we tested and re-trained various systems in order to adapt them to Corsican. This work makes it possible to provide a model allowing the identification of 17 European languages as well as Corsican.*

MOTS-CLÉS : *corpus, qualité, identification de langue, langues peu dotées, corse.*

KEYWORDS: *corpora, quality, language identification, less-resourced languages, Corsican.*

1. Introduction

Dans le domaine des technologies de la langue et du traitement automatique du langage (TAL), une très faible minorité des plus de 7 000 langues répertoriées est dotée de manière satisfaisante de ressources et d'outils. Les langues « numériquement délaissées » font partie, à des degrés divers, de la catégorie des langues « peu dotées ». Selon Joshi *et al.* (2020), qui définissent une nomenclature à six niveaux[1], seules sept langues intègrent le niveau le plus élevé, 65 sont reprises dans les trois niveaux intermédiaires, alors que la vaste majorité des autres langues se situe dans les deux catégories les plus basses !

D'une manière générale, les perspectives pour ces langues sont plutôt pessimistes. Certains estiment qu'à peine 500 langues vivantes pourraient subsister à l'horizon 2100 (Landragin, 2018), alors que d'autres vont même jusqu'à considérer que seules 250 langues pourraient atteindre le statut de « survivant numérique » (Kornai, 2013).

De nombreuses initiatives ont cependant été prises et ont donné lieu à des recommandations techniques et méthodologiques – telles que celles de Berment (2004), Soria *et al.* (2013) ou Ceberio Berger *et al.* (2018) – ainsi qu'à des développements concrets pour diverses langues, dont certaines langues régionales de France (Bernhard *et al.*, 2019 ; Millour, 2020). La situation reste néanmoins préoccupante pour de nombreuses langues.

Parmi les actions à entreprendre pour améliorer le statut de ces langues, la constitution de corpus textuels non bruités, disposant si possible de métadonnées ou d'annotations, et de préférence exploitables au niveau légal, est une priorité. Ces ressources sont importantes car elles permettent de documenter les langues et de mettre au point des outils, en particulier grâce aux techniques d'apprentissage artificiel.

Pour répondre à ce besoin de corpus, il a été de plus en plus courant, depuis environ deux décennies, d'essayer de constituer des ensembles de textes à partir d'Internet. En témoignent diverses manifestations scientifiques, telles que les ateliers Web as Corpus[2], ou plusieurs outils dédiés à cette activité (Baroni et Bernardini, 2004 ; Kilgarriff *et al.*, 2014). Grâce à l'augmentation des capacités de calcul, de transfert et de stockage, d'imposantes collections de documents, telles que ParaCrawl (Esplà-Gomis *et al.*, 2019) ou Common Crawl[3], ont vu le jour. Étant donné leur taille, des outils permettant l'extraction ciblée d'une fraction de leur contenu sont apparus (Roziewski et Stokowiec, 2016 ; Wenzek *et al.*, 2019). Des sous-corpus, qui revendiquent des textes « nettoyés » et organisés par langue – dont certaines sont considérées comme peu dotées – ont également été mis à disposition, entre autres C4Corpus (Habernal *et al.*, 2016), OSCAR (Suárez *et al.*, 2020) ou mC4 (Xue *et al.*, 2021).

La création de corpus pour les langues peu dotées serait-elle un problème résolu ?

1. Allant de « 0 » – très faiblement ou non dotées – à « 5 » – fortement dotées.
2. Voir la page du *Special Interest Group* ACL SIGWAC : `https://www.sigwac.org.uk/`.
3. `https://commoncrawl.org/`

Cela ne semble pas acquis car, à diverses reprises, des doutes ont été émis par rapport à la qualité de ces ressources. Cette impression a été objectivée, en particulier par Caswell *et al.* (2020), Caswell *et al.* (2021) ou Tahir et Mehmood (2021).

D'autre part, le problème des droits d'exploitation est souvent évacué, soit en l'ignorant totalement, soit en le reportant sur l'utilisateur final.

Dans cet article, nous nous plaçons dans le contexte d'une langue peu dotée issue du groupe italo-roman : le corse. Après un point sur sa situation et sur les ressources disponibles dans le contexte des technologies de la langue (section 2), nous nous intéressons à quelques corpus de grande taille qui revendiquent la présence plus ou moins importante de contenu en corse. L'un d'eux a déjà été inspecté par Caswell *et al.* (2021). Comme suggéré et encouragé par les auteurs de cette étude, nous avons examiné deux autres ressources afin d'évaluer si l'identification de langue y est fiable et d'expérimenter une méthode de filtrage visant à ne conserver que le contenu corse. Nous avons choisi de travailler à l'aide de logiciels d'identification de langue, ce qui nous permet, dans le même temps, de faire progresser l'outillage fondamental du corse. Nous nous penchons donc tout d'abord sur ces outils (section 3), avant d'aborder l'évaluation des ressources dans un second temps (section 4).

2. La situation du corse

Le corse fait partie des langues considérées comme étant « en danger » par l'Unesco (Moseley, 2010). L'utilisation du corse au quotidien au travers des applications numériques standard – correction orthographique, traduction automatique, vocalisation, moteurs de recherche, etc. – est de fait très limitée, voire inexistante.

Plusieurs facteurs viennent compliquer l'utilisation du corse dans le domaine numérique. Des variations dialectales sont observées, tant à l'oral qu'à l'écrit. On peut néanmoins identifier cinq aires principales (Dalbera-Stefanaggi, 2002 ; Dalbera-Stefanaggi, 2007), entre lesquelles l'intercompréhension des locuteurs est assurée. Malgré la mise en œuvre d'une approche polynomique (Marcellesi, 1984), l'écriture du corse ne bénéficie pas d'une normalisation. Enfin, le rapport de diglossie qu'entretient le corse avec le français peut se traduire par l'apparition du phénomène d'alternance codique ou d'une tendance à la francisation du lexique, notamment chez les plus jeunes. D'une manière générale, malgré une volonté politique de soutenir la langue, on observe un recul de la pratique du corse au profit du français.

En ce qui concerne les technologies de la langue et le TAL, le corse est versé dans l'avant-dernier groupe de la classification de Joshi *et al.* (2020). 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, en général de faible ampleur, et dont peu sont disponibles dans un format standard et permettant d'accéder aisément aux données brutes. Ce constat est confirmé par l'inventaire de Kevers *et al.* (2021). Au-delà de quelques corpus récemment créés spécifiquement pour le corse (Kevers et Retali-Medori, 2020), il existe cependant certaines ressources de grande dimension

qui revendiquent l'existence de contenu en corse : An Crúbadán[4] (Scannell, 2007), W2C[5] (Majliš et Zabokrtský, 2012), ou Common Crawl[6] (de 2008 à nos jours).

3. Identification de langue

L'identification de langue, c'est-à-dire l'attribution d'une étiquette représentant la langue d'un texte, est un composant fondamental pour le TAL. À ce titre, il est communément intégré aux chaînes de traitement destinées à gérer plusieurs langues. L'identification correcte de la langue d'un document permet d'appliquer les ressources et méthodes d'analyse les plus appropriées et performantes possible.

Dans le contexte particulier du traitement des langues peu dotées, ce composant revêt toute son importance, car il peut contribuer à la constitution des ressources de base, dont les corpus font partie.

3.1. *Un point sur l'état de l'art*

En raison de l'existence de nombreuses solutions incluant une grande variété de langues, l'identification de langue est une tâche qui est parfois considérée comme résolue. Nous n'allons pas en faire ici un panorama complet, d'autant que d'autres, tels que Jauhiainen *et al.* (2018), s'y sont attelés avant nous.

Les premières approches se sont d'abord intéressées à l'exploitation de listes de mots-clés représentatifs pour chaque langue (Ingle, 1976 ; Giguet, 1995 ; Rehurek et Kolkus, 2009). L'utilisation de n-grammes est ensuite apparue et constitue probablement, avec diverses déclinaisons, le moyen le plus utilisé pour traiter le problème (Dunning, 1994 ; Cavnar et Trenkle, 1994 ; Kerwin, 2006 ; Nakatani, 2010 ; Lui et Baldwin, 2012 ; Majliš, 2012 ; Nakatani, 2012 ; Takçı et Güngör, 2012 ; Brown, 2013). Ces dernières années, des outils basés sur des approches neuronales (Jaech *et al.*, 2016) ou impliquant des plongements lexicaux (Joulin *et al.*, 2017) ont également vu le jour. Quelques-uns de ces systèmes sont exposés plus en détail à la section 3.2.2.

Si ces différentes approches ont en général abouti à des résultats satisfaisants, de nombreux points pour lesquels des progrès doivent encore être apportés ont été mis en évidence. Hughes *et al.* (2006) et Jauhiainen *et al.* (2018), à plus de dix ans d'intervalle, identifient tous deux le support des langues peu dotées, la détection ouverte de langues[7], la prise en compte de documents multilingues, ainsi que les effets des prétraitements, comme n'étant pas encore totalement maîtrisés. D'autres points tels que le support d'un nombre élevé de langues, la distinction entre langues proches et

4. http://crubadan.org/
5. https://ufal.mff.cuni.cz/w2c
6. https://commoncrawl.org/
7. C'est-à-dire la possibilité pour un système de gérer des langues qui lui sont inconnues.

dialectes, ainsi que l'identification de langue pour les textes courts sont également relevés comme problématiques par Jauhiainen *et al.* (2018).

Dans le cadre du traitement automatique du corse, nous sommes confrontés à plusieurs de ces points. Il s'agit d'une langue peu dotée, pour laquelle des variations dialectales sont enregistrées, et qui peut également souffrir d'une certaine proximité linguistique avec l'italien, ainsi que de la situation de diglossie avec le français, celle-ci pouvant se matérialiser par l'alternance de ces deux langues dans certains textes.

Notre objectif prioritaire étant de disposer rapidement d'un module d'identification de langue performant pour le corse, nous avons effectué un inventaire et une évaluation de plusieurs systèmes existants. Certains d'entre eux supportent le corse en standard – nous avons donc pu les utiliser tels quels – d'autres ont dû être adaptés, voire complètement réentraînés.

3.2. *Évaluation et mise au point d'un outil adapté pour le corse*

Au-delà de l'obtention d'un module d'identification de langue le plus précis possible pour le corse, notre travail permet d'aborder plusieurs autres questions. La première concerne la possibilité de mettre au point un système performant de détection de langue au niveau européen en incluant au moins une langue régionale. La seconde porte sur la problématique des textes courts et la capacité des outils à les traiter correctement. Cette question est importante car un outil performant sur de très courtes séquences de caractères pourrait avoir un intérêt dans le traitement de l'alternance codique. Enfin, nous avons effectué la comparaison entre plusieurs solutions entraînées avec des quantités de données plus ou moins élevées, et en imposant, pour certaines évaluations, un équilibre entre les différentes langues.

3.2.1. *Choix des langues cibles*

Le choix des langues à prendre en compte a été conditionné par plusieurs paramètres. Tout d'abord, nous désirions nous placer dans un contexte européen et donc sélectionner en priorité des langues officielles de l'Union européenne. Le second critère est plus d'ordre pratique puisqu'il concerne la disponibilité de données exploitables au niveau technique et légal. Nous avons privilégié le choix d'une source unique pour toutes les langues – la collection collaborative de phrases Tatoeba[8] – afin de travailler sur des textes similaires et de limiter le plus possible tout biais dû à la nature des documents[9]. Nous avons conservé 17 langues sur les 24 officielles de l'UE. Il s'agit de celles qui disposent, dans les données Tatoeba, de plus de 100 000 tokens et de plus de 500 000 caractères. Seuls les textes corses n'ont pas pu être issus de cette source étant donné le très faible nombre de phrases disponibles. Nous avons donc utilisé trois

8. Les données sont publiées sous licence CC BY 2.0 FR sur la page `https://tatoeba.org/fr/downloads`. Le jeu de données a été téléchargé le 24/05/2021.

9. Notre préoccupation étant aussi de ne pas utiliser une ressource trop proche de celles que nous évaluons à la section 4.

corpus disponibles au format XML TEI : A Piazzetta, A Sacra Bìbbia et Wikipedia[10]. Les détails chiffrés relatifs à ces données sont repris au tableau 1.

Langue	**Codes**	**Phrases**	**Tokens**	**Caractères**
Anglais	eng - en	1 473 300	11 260 699	59 491 198
Italien	ita - it	787 115	4 616 453	27 386 085
Allemand	deu - de	549 024	4 332 461	26 982 232
Français	fra - fr	465 299	3 481 188	19 917 382
Portugais	por - pt	385 560	2 697 491	15 215 273
Espagnol	spa - sp	337 010	2 375 025	13 492 307
Hongrois	hun - hu	319 107	1 699 406	11 303 334
Néerlandais	nld - nl	142 681	914 262	5 133 999
Finnois	fin - fi	126 167	632 363	4 676 176
Polonais	pol - pl	109 845	582 059	3 790 406
Lituanien	lit - lt	59 254	282 044	1 862 723
Tchèque	ces - cs	56 177	285 837	1 669 959
Danois	dan - da	49 322	315005	1 723 968
Suédois	swe - sv	41 424	238 152	1 292 812
Grec	ell - el	33 981	180 062	1 069 296
Roumain	ron - ro	24 928	157 743	897 319
Bulgare	bul - bg	24 503	142 862	811 300
Corse	cos - co	-	2 314 619	12 619 463
		Corse - A Piazzetta	516 509	3 001 680
		Corse - A Sacra Bìbbia	867 627	4 144 117
		Corse - Wikipedia	930 493	5 473 666

Tableau 1. *Liste des langues sélectionnées*

3.2.2. *Outils*

Pour la sélection des logiciels, nous avons choisi des outils qui proposent en standard les langues ciblées, ou qui peuvent être réentraînés à partir de nos données. La disponibilité de ces systèmes sous une licence ouverte a également été un point d'attention. Afin de disposer d'une valeur de référence, nous avons codé un système effectuant une identification à partir d'un décompte de mots-clés[11] caractéristiques de

10. Disponibles sous licences CC BY-NC-SA 4.0 et CC BY-SA 3.0 sur `https://bdlc.univ-corse.fr/tal/index.php?page=res`

11. À l'exception de la liste corse, créée pour l'occasion, les mots-clés utilisés sont ceux exploités par Lucene (`https://github.com/apache/lucene`) et proviennent du projet Snowball (`https://github.com/snowballstem`) ou de Jacques Savoy (`http://members.unine.ch/jacques.savoy/clef/`). Le nombre de mots-clés par langue varie entre 78 et 393.

chaque langue. Notre sélection, incluant une estimation du niveau de *Technology readiness level* (TRL[12]), est reprise au tableau 2[13].

Nom	**Type**	**Nb. Lg.**	**Date**	**Licence**	**TRL**
Référence	mots-clés	18	-	Dév. personnel	3
Systèmes utilisables sans modification					
YALI	n-grammes	18 (122)	2019	BSD	4
WhatLang	n-grammes	1 475	2015	GPL v.3	4
CLD2	n-grammes	161	2015	Apache v.2	9
FastText	n-grammes	176	2019	MIT, CC BY-SA 3.0	8-9
CLD3	n-grammes	213	2020	Apache v.2	9
Systèmes nécessitant l'ajout du corse					
LibTextCat	n-grammes	18 (163)	2015	BSD	9
Lang. Detect.	n-grammes	18 (53)	2014	Apache v.2	4
Systèmes nécessitant un réentraînement complet					
Langid.py	n-grammes	18 (97)	2017	BSD	4
Ldig	*infinity-gram*	18 (17)	2013	MIT	4
FastText*	n-grammes	18 (176)	2019	MIT	8-9

Tableau 2. *Liste des logiciels sélectionnés*

Nous avons choisi de tester plusieurs logiciels utilisables sans aucune modification. Le premier, YALI[14] (Majliš, 2012), utilise un modèle de langue qui repose sur les listes des cent n-grammes d'octets les plus fréquents pour chaque langue. Chaque n-gramme est accompagné par une probabilité. Le système recherche, parmi les 122 langues supportées, corse inclus, celle obtenant la somme de probabilités la plus élevée en fonction des n-grammes rencontrés. Le logiciel est disponible sous licence BSD.

Avec WhatLang (Brown, 2013), les n-grammes de six, dix ou douze octets sont exploités dans une approche utilisant l'algorithme des k plus proches voisins et une mesure de similarité cosinus. Les n-grammes sont sélectionnés et filtrés selon divers facteurs : leur fréquence, leur taille, leur présence dans des n-grammes plus longs. Des indices négatifs (*stopgrams*) sont aussi utilisés pour certains n-grammes incon-

12. Échelle d'estimation de la maturité d'une technologie allant de 1, « principes de base observés », à 9, « système en production ». Une description complète utilisée par les projets européens Horizon 2020 peut être consultée sur : `https://ec.europa.eu/research/participants/data/ref/h2020/wp/2014_2015/annexes/h2020-wp1415-annex-g-trl_en.pdf`.

13. Le nombre renseigné correspond à la configuration exploitée pour nos tests. Une valeur entre parenthèses indique le nombre total de langues supportées par le système. La date correspond à la dernière modification de celui-ci.

14. `http://ufal.mff.cuni.cz/tools/yali`

nus d'un modèle de langue, mais présent dans d'autres. Enfin, un mécanisme de lissage peut être utilisé afin de favoriser la succession de chaînes de caractères dans une même langue. Initialement capable de reconnaître 1 100 langues, une version élargie à 1 475 langues, corse inclus, a ensuite été proposée[15] sous licence GPL v.3.

CLD2, Compact Language Detector 2[16], est un système bayésien naïf qui exploite des quadrigrammes de caractères. Il s'agit d'un composant de détection de langue développé par Google pour son navigateur Chromium. Il a été initialement présenté comme pouvant reconnaître 83 langues, mais une mise à jour lui permet d'en supporter 78 supplémentaires[17], dont le corse. Ce système est distribué sous licence Apache v.2.

Google a ensuite proposé CLD3[18], une solution neuronale exploitant toujours des n-grammes de caractères. Ce système dispose cette fois du support de 213 langues, incluant le corse, et est proposé sous licence Apache v.2.

Enfin, nous avons repris un dernier logiciel supportant le corse en standard. Il s'agit de FastText[19], une librairie issue du groupe de recherche de Facebook. Elle permet d'apprendre des représentations de données textuelles sous la forme de plongements lexicaux, et de créer des classifieurs de textes (Joulin *et al.*, 2017). Ceux-ci ont été utilisés pour produire un système d'identification de langue prenant en charge 176 langues, dont le corse. La librairie est diffusée sous une licence MIT, alors que les plongements lexicaux sont disponibles sous licence CC BY-SA 3.0.

L'approche TextCat, proposée par Cavnar et Trenkle (1994), est probablement l'une des plus connues. Des n-grammes de caractères de différentes longueurs – n allant de un à cinq – sont exploités pour construire des profils contenant une liste ordonnée des 300 éléments les plus fréquents pour chaque langue. Une mesure de distance, nommée *out-of-place*, consiste à déterminer la somme des distances relevées dans l'ordonnancement des n-grammes d'un document avec celui des différents profils de langue afin d'identifier la langue la plus probable. Diverses implémentations sont disponibles. Nous avons choisi celle intégrée à la suite bureautique Libre Office[20], distribuée sous licence BSD. Le système dispose de profils pour l'ensemble des langues ciblées, sauf pour le corse, pour lequel nous en avons généré un.

Nakatani (2010) propose un autre logiciel permettant la création de modèles de langue, et leur exploitation au travers d'une approche bayésienne naïve. Le logiciel, nommé Language Detection[21], dispose de modèles préentraînés pour 53 langues, mais sans y inclure le corse. L'outil permet cependant la génération de nouveaux modèles, ce que nous avons par conséquent effectué pour le corse. La licence d'utilisation accordée pour cette librairie est Apache v.2.

15. https://sourceforge.net/projects/la-strings/
16. https://github.com/CLD2Owners/cld2
17. https://github.com/CLD2Owners/cld2/wiki/CLD2-Full-Version
18. https://github.com/google/cld3
19. https://fasttext.cc/
20. https://github.com/LibreOffice/libexttextcat
21. https://github.com/shuyo/language-detection/

Enfin, nous avons encore retenu quelques systèmes qui ne supportent pas nativement le corse, mais qui disposent des outils permettant un réentraînement complet d'un modèle. Langid.py[22], proposé par Lui et Baldwin (2012), est le premier de ceux-ci. Ce logiciel utilise des n-grammes d'octets – n étant situé entre un et quatre – afin d'alimenter un classifieur de type bayésien naïf pour lequel une stratégie de sélection de caractéristiques basée sur le gain d'information est mise en place. À l'origine, 97 langues sont supportées, mais la version réentraînée par nos soins a été limitée à nos 18 langues cibles. Langid.py est distribué sous licence BSD.

Le dernier système pris en compte est Ldig[23] (Nakatani, 2012). Celui-ci est conçu pour analyser des documents très courts, tels que des *tweets*. Contrairement à la plupart des autres approches qui extraient des n-grammes de différentes longueurs – par exemple allant de deux à quatre – l'analyse du texte repose ici sur les concepts d'*infinity gram* et de *maximal substring* (Okanohara et Tsujii, 2009). Le principe consiste à énumérer l'ensemble des sous-chaînes de caractères pouvant constituer un document. Étant donné le nombre d'éléments potentiellement très élevé, ceux-ci sont rassemblés en classes d'équivalence, représentées par une sous-chaîne maximale. Ldig, qui est mis à disposition sous licence MIT, dispose du support pour 17 langues. Le modèle généré par nos soins en inclut 18, dont le corse fait bien entendu partie.

Nous avons également décidé de générer un nouveau modèle FastText* à partir de la librairie FastText, qui supporte pourtant officiellement toutes nos langues cibles. Le contraste entre les bons résultats généraux et les très mauvaises performances observées pour le corse nous a cependant incités à lui donner une seconde chance.

3.2.3. *Données*

Comme déjà mentionné, nous exploitons Tatoeba[24], pour les 17 langues européennes, ainsi que les trois corpus corses A Piazzetta, A Sacra Bìbbia et Wikipedia.

Nous avons défini quatre catégories de documents (TSML) : (T) *tiny*, jusqu'à 50 caractères, ce qui pourrait correspondre à une courte alternance de langue dans un texte multilingue ; (S) *small*, à partir de 51 et jusqu'à 300 caractères, soit environ la taille d'un *tweet* ; (M) *medium*, à partir de 301 et jusqu'à 3 000 caractères, taille que l'on pourrait comparer à une page de texte environ ; (L) *large* à partir de 3 001 caractères, catégorie qui représente les documents de plus d'une page. Cette partition permet d'évaluer l'efficacité pour différents contextes ou utilisations.

Le corpus Tatoeba reprend essentiellement des phrases isolées dont la longueur varie, allant jusqu'à plus de 1 500 caractères. A Piazzetta est composé des articles d'un blog journalistique, dont la majorité contient entre 300 et 3 000 caractères. A Sacra Bìbbia, la version corse de la Bible, est organisée selon plusieurs divisions hiérarchiques : parties, chapitres, versets. Des titres peuvent également apparaître à

22. `https://github.com/saffsd/langid.py`
23. `https://github.com/shuyo/ldig`
24. `https://tatoeba.org/fr/downloads`

différents endroits. L'unité de traitement retenue est le chapitre, mais les titres ont été traités séparément. La majorité des « documents » ainsi constitués fait moins de 50 caractères, mais un nombre non négligeable de ceux-ci disposent de 300 à 3 000 caractères, voire plus. Enfin, le corpus Wikipedia est composé en grande partie d'articles faisant entre 50 et 3 000 caractères. Tous ces corpus ont été utilisés pour l'apprentissage et pour les tests. La répartition des corpus selon les catégories TSML est détaillée au tableau 3.

Catégorie	**Tatoeba (17 lg.)**		**A Piazzetta**		**A Sacra Bìbbia**		**Wikipedia**	
T `[1,50]`	4 075 086	81,7 %	79	4,5 %	2 146	64,1 %	41	0,7 %
S `[51,300]`	905 294	18,2 %	223	12,8 %	14	0,4 %	2 760	48,2 %
M `[301,3000]`	4 317	0,1 %	1 179	67,6 %	542	16,2 %	2 532	44,2 %
L `[3001,[`	0	0 %	264	15,1 %	645	19,3 %	395	6,9 %
Total	4 984 697		1 745		3 347		5 728	

Tableau 3. *Répartition TSML des documents en fonction du nombre de caractères*

Le corpus Tatoeba est issu d'une base de données dans laquelle la langue est une métadonnée cruciale. Le corpus A Sacra Bìbbia provient lui d'un ouvrage édité en langue corse. Pour ces données, nous considérons l'attribution d'une langue comme fiable. Les deux autres corpus corses sont issus d'Internet et n'ont pas bénéficié d'une vérification. Ils ne peuvent donc pas prétendre au même niveau de fiabilité. Afin d'écarter les documents qui ne seraient pas majoritairement en corse, nous avons mis en place une étape de filtrage au moyen d'un détecteur de langue par mots-clés élaboré dans le cadre d'un précédent travail. Nous avons ainsi écarté 128 documents du corpus A Piazzetta (7 T, 26 S, 78 M et 17 L), et 141 de Wikipedia (1 T, 38 S, 96 M et 6 L).

Les données présentées ci-dessus ont été réparties en deux ensembles, l'un pour l'apprentissage et l'autre pour le test. Nous n'avons pas de phase de paramétrage qui nécessiterait un ensemble de validation. Afin de minorer les éventuels effets dus à la répartition des données d'apprentissage et de test, nous avons mis en place une validation en cinq plis. Pour les expériences nécessitant un équilibrage des données en fonction de la langue et de la taille des documents, le respect de ces contraintes entraîne une utilisation partielle des données à notre disposition.

Pour chaque pli, un ensemble de test équivalent à environ 150 000 caractères par langue a été réservé, avec une répartition équilibrée entre les différentes catégories TSML. En cas de pénurie pour l'une des catégories, plusieurs documents ont été rassemblés afin de constituer un texte « artificiel » de longueur suffisante. La taille des documents pouvant varier à l'intérieur de chaque catégorie, le nombre de documents n'est pas forcément identique pour chaque langue et pour chaque pli, mais en pratique une certaine régularité a été observée[25]. Pour le corse, en raison de documents

25. La contribution au jeu de test d'une langue va de 136 à 172 documents (153 en moyenne). L'ensemble complet, hors corse, contient entre 2 536 et 2 712 documents (2 605 en moyenne).

en moyenne un peu plus longs, nous avons dû augmenter la taille des ensembles de test à 200 000 caractères afin que le nombre de documents ne chute pas trop[26]. Nous avons également veillé à exploiter les trois corpus de manière équivalente. Pour les ensembles d'apprentissage, deux versions ont été définies. La première (section 3.2.4) est limitée à environ 500 000 caractères par langue[27], répartis équitablement entre les différentes catégories TSML, ce qui nous permet d'obtenir des ensembles d'apprentissage équilibrés en termes de longueur de document et de représentation de chaque langue. Pour la seconde version (section 3.2.5), nous avons décidé d'utiliser la totalité des données à notre disposition, ce qui aboutit à des ensembles déséquilibrés.

En plus de ces données, qui permettent une évaluation sur des documents de même nature et de même source que ceux ayant servi à l'apprentissage, nous avons également mené un test sur des textes complètement différents (section 3.2.6). Nous avons choisi d'exploiter à cet effet le *Digital Corpus of the European Parliament*, DCEP (Hajlaoui *et al.*, 2014), dans sa version nettoyée des balises HTML/XML (*STRIP*)[28]. Il s'agit de documents de différentes natures[29] produits dans le cadre du travail du Parlement européen entre 2001 et 2012. Nous avons sélectionné un sous-corpus ne contenant que les documents disponibles pour l'ensemble de nos 17 langues cibles. À nouveau, nous avons réalisé une répartition des documents en plusieurs catégories en fonction de leur longueur. Nous avons cependant constaté que les documents de moins de 200 caractères s'avéraient très souvent vides de contenu linguistique, ce qui nous a obligés à revoir les catégories précédemment définies. Un autre filtrage a également été mis en place pour écarter les documents ne contenant que des références ou codes, sans inclure véritablement un texte en langue naturelle. Finalement, nous avons sélectionné 3 759 documents disponibles pour chacune des langues[30], répartis de la manière suivante : 198 entre 201 et 500 caractères, 558 entre 501 et 1 000 caractères, 1 882 entre 1 001 et 2 000 caractères, et enfin 1 121 entre 2 001 et 3 000 caractères. Notons enfin que ce corpus ne contient pas de document en corse.

L'ensemble des données textuelles a été soumis à un prétraitement identique. Celui-ci a permis de décapitaliser l'ensemble du texte, de décoller les signes de ponctuation et de normaliser les espacements.

Les résultats des évaluations sont repris aux sections suivantes. Le premier test a pour objectif d'évaluer les différents outils en utilisant des ensembles d'apprentissage équilibrés limités à 500 000 caractères (section 3.2.4). Le deuxième test permet, pour les systèmes identifiés comme les plus performants, d'investiguer les différences obtenues en passant à un ensemble d'entraînement maximal non équilibré (section 3.2.5).

26. Soit 115 documents en moyenne (entre 101 et 129 documents).

27. Cette limite s'est imposée, étant donné la quantité de données disponibles pour la langue la moins dotée dans le cadre de cette évaluation (voir tableau 1).

28. Téléchargeable à l'adresse `https://wt-public.emm4u.eu/Resources/DCEP-2013/DCEP-Download-Page.html`

29. Communiqués de presse, motions, procès-verbaux des sessions plénières, règlement intérieur, rapports, et questions écrites au Parlement.

30. Soit un total de 63 903 documents toutes langues confondues.

Enfin, le troisième test a la vocation de recouper et de vérifier les résultats sur des documents d'une autre nature que ceux ayant servi à l'apprentissage (section 3.2.6).

3.2.4. *Première évaluation : données d'apprentissage limitées mais équilibrées*

L'objectif de cette évaluation est de faire un premier tri parmi l'ensemble des outils que nous avons sélectionnés. Comme exposé précédemment, en plus du système de référence, nous disposons de cinq outils à utiliser sans aucune modification, de deux outils pour lesquels l'ajout du corse a dû être effectué, sans toucher aux autres langues, alors que trois autres outils ont nécessité un réentraînement complet. Pour chacun de ces cinq derniers outils, l'apprentissage de cinq modèles a été effectué à partir d'ensembles de documents d'environ 500 000 caractères chacun. Cinq ensembles de test totalement disjoints ont été utilisés pour l'évaluation. Nous rapportons, au tableau 4, la moyenne des résultats obtenus pour les cinq jeux de données.

Tous les systèmes sont globalement meilleurs que la référence, qui est particulièrement à la traîne pour les documents les plus courts (T et S), ce qui est logique pour une approche par mots-clés. Notons cependant que, pour les documents plus longs, ce système se place dans la même fourchette de valeurs que les autres.

Concernant les 17 langues autres que le corse, les outils pour lesquels nous avons réentraîné un modèle offrent les meilleurs résultats, atteignant une précision de plus de 99 %. Il faut cependant nuancer cette affirmation, puisque la version standard de FastText propose de très bonnes performances. À l'inverse, Langid.py et son modèle complètement réentraîné donnent des résultats assez faibles, s'écartant en moyenne très peu de la référence.

Notons que les systèmes réentraînés ont pu tirer un avantage de la proximité des ensembles d'apprentissage et de test, ce qui est d'ailleurs également le cas de FastText qui intègre Tatoeba dans son ensemble d'apprentissage. D'autre part, les logiciels utilisés de manière standard proposent en général le support d'un nombre plus important de langues, ce qui peut jouer en leur défaveur.

En ce qui concerne le corse, on observe des résultats assez moyens pour les systèmes standard (entre 83,25 % et 90,30 % de précision), voire vraiment catastrophiques pour FastText (précision de 1,24 %). Ce résultat, pouvant provenir d'un fort déséquilibre dans les données d'apprentissage (Siewert *et al.*, 2020), nous a d'ailleurs poussés à effectuer un réentraînement complet de ce système, que nous notons FastText*. Les deux logiciels pour lesquels nous avons effectué un apprentissage uniquement pour le corse s'en sortent mieux, avec une précision d'un peu plus de 92 %. En fin de compte, les meilleures performances sont obtenues avec les systèmes réentraînés complètement pour l'ensemble des langues, à l'exception de Langid.py (83,72 %). Ldig permet de grimper jusqu'à une précision de 94,98 %, alors que FastText* surclasse finalement tous les autres outils en atteignant 97,92 % de précision.

Les meilleurs résultats combinés pour les 18 langues sont logiquement à mettre à l'actif de Ldig (99,12 % de précision) et FastText* (99,42 % de précision).

		Utilisation standard					LEARNcos		LEARNall		
	Réf.	**YALI**	**WhatLang**	**CLD2**	**FastText**	**CLD3**	**TextCat**	**Lang. Detect.**	**Langid.py**	**Ldig**	**FastText***
cos	0,7810	0,9030	0,8678	0,8325	0,0124	0,8820	0,9216	0,9210	0,8372	0,9498	0,9792
eng	0,9469	0,9469	0,9635	0,9800	0,9977	0,9516	0,9447	0,9694	0,8916	0,9882	0,9953
ita	0,8851	0,8709	0,8659	0,8625	0,9940	0,9425	0,8709	0,9186	0,7706	0,9892	0,9904
deu	0,9702	0,9714	0,9845	0,9798	0,9976	0,9833	0,9738	0,9738	0,8452	0,9917	0,9976
fra	0,9267	0,9397	0,9493	0,9597	1,0000	0,9434	0,9268	0,9764	0,8625	0,9847	0,9703
por	0,9002	0,8886	0,8861	0,9436	0,9917	0,9589	0,9130	0,9554	0,8509	0,9800	0,9941
spa	0,9179	0,9369	0,8869	0,9357	0,9952	0,9381	0,9060	0,9583	0,8476	0,9869	0,9976
hun	0,8709	0,9282	0,9661	0,9552	0,9912	0,9409	0,9724	0,9840	0,9327	0,9937	0,9924
nld	0,9736	0,9575	0,8865	0,9300	0,9796	0,9641	0,9861	0,9913	0,9367	0,9988	0,9988
fin	0,8722	0,9894	0,8186	0,9869	0,9935	0,9882	0,9972	1,0000	0,9537	1,0000	1,0000
pol	0,7654	0,9672	0,9673	0,9622	0,9933	0,9755	0,9920	0,9960	0,9636	0,9986	0,9986
lit	0,8899	0,9655	0,9497	0,9773	0,9929	0,9632	0,9672	0,9944	0,9689	0,9986	0,9958
ces	0,8508	0,9472	0,9458	0,9672	0,9816	0,9376	0,9913	0,9942	0,9765	0,9943	0,9971
dan	0,9000	0,9556	0,7591	0,8928	0,9609	0,9083	0,9455	0,9830	0,8894	0,9957	0,9929
swe	0,9474	0,9644	0,9427	0,9373	0,9842	0,9728	0,9674	0,9777	0,9478	0,9944	0,9971
ell	0,8479	1,0000	0,9859	1,0000	1,0000	0,9986	1,0000	1,0000	1,0000	1,0000	1,0000
ron	0,8508	0,9484	0,8513	0,9382	0,9769	0,9554	0,9630	0,9868	0,9594	0,9971	0,9985
bul	0,9709	1,0000	0,9588	0,9182	0,9972	0,9713	1,0000	1,0000	1,0000	1,0000	1,0000
17 lg.	0,8992	0,9516	0,9158	0,9486	0,9899	0,9585	0,9598	0,9800	0,9175	0,9936	0,9951
18 lg.	0,8927	0,9489	0,9132	0,9422	0,9355	0,9542	0,9577	0,9767	0,9130	0,9912	0,9942
T	0,6759	0,8234	0,7236	0,8046	0,9267	0,8381	0,8494	0,9155	0,7293	0,9689	0,9781
S	0,9126	0,9707	0,9438	0,9724	0,9551	0,9841	0,9786	0,9906	0,9119	0,9971	0,9985
M	0,9930	0,9980	0,9888	0,9970	0,9565	0,9970	0,9971	0,9991	0,9977	0,9997	1,0000
L	0,9997	0,9997	0,9982	0,9997	0,9571	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000

Tableau 4. *Résultats des évaluations pour l'ensemble des outils de détection de langue (apprentissage à 500 k caractères)*

Enfin, l'analyse des résultats obtenus par catégorie de taille de document, nous confirme que la majorité des outils rencontre des difficultés avec les documents courts (T ou S). Les deux meilleurs systèmes mis en évidence ci-dessus – Ldig et FastText* – permettent cependant d'arriver à une très bonne précision pour les documents S (respectivement 99,71 % et 99,85 %), alors que les documents T restent tout de même légèrement en retrait (96,89 % et 97,81 % de précision).

Au vu de ces résultats, Ldig et FastText* nous semblent les outils les plus intéressants, tant au niveau de leurs résultats sur l'ensemble des 17 langues et sur le corse, que de l'efficacité atteinte pour les documents les plus courts.

3.2.5. *Deuxième évaluation : maximisation des données d'apprentissage*

Cette seconde évaluation nous permet d'observer l'effet d'une augmentation des données d'entraînement et de la perte de leur équilibrage. Les ensembles d'apprentissage limités à 500 000 caractères laissent cette fois la place à la totalité des données à notre disposition, telles que décrites au tableau 1. Nous conservons les cinq mêmes ensembles de test que lors de la première évaluation. Ces données n'apparaissent bien entendu pas dans les ensembles d'apprentissage étendus. Cette évaluation a été menée pour les deux meilleurs outils mis en évidence jusqu'ici, Ldig et FastText*. Les résultats détaillés peuvent être consultés au tableau 5, dans lequel sont également repris, à titre de comparaison, les résultats obtenus lors de la première évaluation.

Les différences observées sont plutôt faibles, ce qui est assez logique étant donné les bonnes performances déjà obtenues précédemment. Pour les 17 langues hors corse, Ldig enregistre une progression de 0,54 %, en proposant une précision de 99,90 %. Au contraire, FastText* ne semble pas profiter des données supplémentaires : la précision enregistrée à 99,37 % équivaut à une régression de 0,14 %.

La situation est cependant différente lorsque l'on s'intéresse spécifiquement aux résultats obtenus pour le corse. La précision proposée par Ldig progresse de 1,54 % pour s'établir à 96,52 %, alors que FastText* enregistre, dans ce cas, une amélioration de 1,18 % pour atteindre une précision de 99,11 %.

L'examen des résultats obtenus en fonction des catégories de taille de documents permet de mettre en évidence, pour les deux outils, une amélioration pour la catégorie qui concerne les plus petits documents (T). En revanche, là où Ldig s'améliore pour la catégorie S, et maintient des résultats identiques pour les catégories M et L, on observe, en ce qui concerne FastText*, une légère régression pour les catégories S et M, alors que la catégorie L reste à la précision maximale.

Pour cette évaluation, les évolutions de précision sont contrastées et relativement modestes. Ldig progresse en différents points, alors que FastText* montre à la fois des améliorations et des régressions. Au final, Ldig propose, à 99,71 %, la précision la plus élevée pour les 18 langues (99,36 % pour FastText*), alors que FastText* conserve la meilleure précision pour le corse (99,11 % contre 96,52 % pour Ldig).

	LEARNmin		LEARNmax	
	Ldig	**FastText***	**Ldig**	**FastText***
cos	0,9498	0,9792	0,9652	0,9911
eng	0,9882	0,9953	1,0000	0,9952
ita	0,9892	0,9904	0,9988	0,9856
deu	0,9917	0,9976	1,0000	1,0000
fra	0,9847	0,9703	1,0000	0,9753
por	0,9800	0,9941	0,9988	0,9965
spa	0,9869	0,9976	1,0000	0,9988
hun	0,9937	0,9924	0,9988	1,0000
nld	0,9988	0,9988	0,9988	0,9976
fin	1,0000	1,0000	0,9986	0,9988
pol	0,9986	0,9986	0,9987	0,9972
lit	0,9986	0,9958	0,9986	1,0000
ces	0,9943	0,9971	0,9972	0,9901
dan	0,9957	0,9929	0,9971	0,9873
swe	0,9944	0,9971	0,9985	0,9804
ell	1,0000	1,0000	1,0000	1,0000
ron	0,9971	0,9985	0,9985	0,9914
bul	1,0000	1,0000	1,0000	0,9986
17 lg.	0,9936	0,9951	0,9990	0,9937
18 lg.	0,9912	0,9942	0,9971	0,9936
T	0,9689	0,9781	0,9905	0,9811
S	0,9971	0,9985	1,0000	0,9948
M	0,9997	1,0000	0,9997	0,9983
L	1,0000	1,0000	1,0000	1,0000

Tableau 5. *Résultats des évaluations pour les outils Ldig et FastText*, apprentissage à 500 k caractères (LEARNmin) et avec l'ensemble des données (LEARNmax)*

3.2.6. *Troisième évaluation : corpus DCEP*

Avec cette troisième évaluation, nous voulons vérifier que les résultats obtenus lors des tests précédents ne sont pas (trop) influencés par la proximité des ensembles d'apprentissage et de test utilisés jusqu'ici. Nous avons donc choisi de mettre à l'épreuve les résultats des systèmes qui semblent les plus intéressants jusqu'à présent en les confrontant à un ensemble de test totalement différent et non exploité lors de l'apprentissage. Il s'agit du *Digital Corpus of the European Parliament*, DCEP.

Nous avons sélectionné Ldig et FastText*, entraînés avec les données d'apprentissage maximales. Étant donné l'évolution contrastée de FastText* entre les deux premiers tests, nous avons également choisi d'observer son comportement dans sa version entraînée avec l'ensemble d'apprentissage limité à 500 000 caractères. Bien

que l'ensemble de test soit ici unique – composé par la totalité des données issues de DCEP – les outils choisis disposent chacun de cinq modèles générés lors des étapes précédentes. Les chiffres détaillés présentés au tableau 6 constituent donc à nouveau une moyenne des résultats obtenus avec les cinq modèles disponibles.

	FastText*(LEARNmin)	**Ldig(LEARNmax)**	**FastText* (LEARNmax)**
eng	0,9723	0,9779	0,9305
ita	0,9636	0,9874	0,9043
deu	0,9944	0,9982	0,9941
fra	0,9313	0,9287	0,6764
por	0,9984	0,9984	0,9894
spa	0,9984	0,9985	0,9831
hun	0,9955	0,9979	0,9969
nld	0,9936	0,9968	0,9431
fin	0,9978	0,9987	0,9790
pol	0,9986	0,9950	0,9692
lit	0,9974	0,9961	0,9614
ces	0,9921	0,9980	0,8074
dan	0,9978	0,9974	0,9308
swe	0,9972	0,9979	0,6700
ell	0,9939	0,9983	0,9951
ron	0,9917	0,9936	0,8596
bul	0,9987	0,9892	0,6180
17 lg.	0,9890	0,9910	0,8946
T	0,9945	0,9936	0,9380
S	0,9889	0,9900	0,9233
M	0,9910	0,9921	0,8874
L	0,9847	0,9893	0,8805

Tableau 6. *Résultats des évaluations sur le corpus DCEP*

La précision s'établit pour Ldig à 99,10 %. La majorité des langues bénéficient d'une précision supérieure à 99 %, à l'exception du bulgare (98,92 %), de l'italien (98,74 %), de l'anglais (97,79 %) et du français (92,87 %). La version « minimale » de FastText* suit à peu de choses près la même tendance, avec une précision globale de 98,90 % un peu moins élevée, alors que les langues qui ne franchissent pas le seuil des 99 % sont cette fois limitées à l'anglais (97,23 %), l'italien (96,36 %) ainsi que le français (93,13 %). La version « maximale » de FastText* a donné lieu à de moins bons résultats. La précision globale n'atteint pas les 90 % (89,46 %) et plusieurs langues aboutissent à des résultats très décevants, en particulier pour le bulgare (61,80 %), le suédois (67,00 %) et le français (67,64 %). Nous n'avons pas d'explication objective pour éclairer ces chiffres, mais nous notons cependant que le comportement de cette configuration s'inscrit dans le prolongement de celui observé lors de l'évaluation précédente. Enfin, remarquons que les résultats en retrait pour le français sont observés de manière cohérente pour les trois outils.

À l'issue de ces trois évaluations, nous pouvons constater que Ldig présente des résultats particulièrement intéressants. La précision pour le corse est satisfaisante, et cet outil a pu enregistrer une (légère) progression suite à l'extension des données d'apprentissage. Les résultats se sont également maintenus à un niveau élevé sur des données de nature différente. De plus, une précision supérieure à 99 % a pu être observée pour toutes les catégories de documents, y compris celles représentant les plus courts.

4. Évaluation de ressources

4.1. *Ressources concernées*

Notre objectif est d'obtenir une estimation de la qualité de plusieurs ressources d'assez grande envergure, qui incluent du contenu en corse, et qui ont été constituées à partir d'Internet. La première est proposée par le Crúbadán Project[31] (Scannell, 2007), qui met à disposition des corpus moissonnés pour plus de 2 000 langues. Ces dernières ont été identifiées à l'aide d'une mesure cosinus, et ponctuellement d'un classifieur bayésien naïf, à partir d'un ensemble de trigrammes de référence collectés manuellement. Ce corpus est décliné sous différentes formes : des trigrammes de caractères, des mots simples, ainsi que des bigrammes de mots. Une information de fréquence accompagne chacun des éléments repris dans ces listes. Notre intérêt se porte surtout sur les bigrammes de mots, les autres formes ne se prêtant pas bien à une identification de langue. Cette ressource contient, pour le corse, 541 423 caractères répartis en 50 000 bigrammes, les plus courts n'étant cependant pas toujours pertinents[32].

Le corpus W2C[33] (Majliš et Zabokrtský, 2012) a été constitué à partir de Wikipédia et d'autres sources sur Internet. Il contient plus de 54 Go de données concernant 120 langues, dont l'identification a été réalisée par le système YALI entraîné sur un ensemble initial extrait de Wikipédia. La partie corse, d'une taille de 20 Mo, contient 90 405 lignes représentant chacune un « document ». La taille de ceux-ci varie entre 15 et 188 568 caractères. Le corpus totalise 2 765 040 mots et 16 848 279 caractères.

Enfin, nous nous intéressons aussi aux moissonnages effectués par le projet Common Crawl[34]. La proportion de pages en corse est estimée à 0,24 % de la récolte CC-MAIN-2021-21, ce qui représente 64 146 pages[35]. Les différents moissonnages étant en partie cumulatifs, le nombre total de pages identifiées comme étant exprimées en corse est difficile à déterminer. De plus, étant donné la taille de la ressource, il n'est pas aisé d'accéder au contenu relatif à une langue précise. Divers corpus dérivés de

31. `http://crubadan.org/`
32. À titre d'illustration, le bigramme le plus fréquent est « . \n ».
33. `https://ufal.mff.cuni.cz/w2c`
34. `https://commoncrawl.org/`
35. Ce nombre peut varier d'un moissonnage à l'autre. Les statistiques peuvent être consultées à l'adresse `https://commoncrawl.github.io/cc-crawl-statistics/plots/languages.csv`

Common Crawl ont cependant vu le jour et proposent des sous-ensembles organisés par langues. C'est en particulier le cas du corpus mC4[36] (Xue *et al.*, 2021), qui dispose d'un contenu pour 101 langues, et a été constitué à partir de 71 moissonnages de Common Crawl. Pour l'identification de langue, alors que Common Crawl exploite CLD2, le choix de mC4 s'est porté sur CLD3. La partie corse compte 494 913 extraits, dont on peut estimer qu'ils représentent environ 100 millions de caractères. Le corpus Common Crawl est donc abordé par l'intermédiaire de mC4.

4.2. *Méthodologie*

Les corpus à évaluer ont été soumis à trois outils de détection de langue testés précédemment (section 3.2). Nous avons choisi d'utiliser l'outil de référence, le seul basé sur l'utilisation de listes de mots-clés, le logiciel standard CLD3, ainsi que le système Ldig réentraîné avec le maximum de données pour les 17 langues européennes plus le corse. Ce choix est un compromis entre la diversité des approches ainsi que les performances observées lors de nos tests. CLD3 permet, en outre, de couvrir un nombre élevé de langues, ce qui n'est pas le cas des deux autres outils. Enfin, Ldig nous est apparu comme étant le système proposant le meilleur équilibre entre l'identification du corse et des autres langues, y compris sur des documents très courts.

Il n'est évidemment pas possible de faire une évaluation exhaustive des corpus. Nous avons donc suivi la méthodologie proposée par Caswell *et al.* (2021) et extrait 200 éléments pour chaque corpus. Cet échantillon a été équilibré en sélectionnant des documents en fonction du nombre d'outils les ayant reconnus comme exprimés en corse, soit 25 extraits pour chacune des huit combinaisons possibles. Les données non étiquetées ont été proposées à un expert pour annotation. Cette tâche consiste à attribuer au texte une des catégories suivantes[37] : C pour « (langue) correcte », M pour « (langues) mixées », AL pour « autres langues » et B pour « bruit », c'est-à-dire du contenu non linguistique tel que « ®ªsìálivvàáè nrp¬sõ ». À partir de ces informations et de l'identification automatique de langue, il est possible d'estimer si les différentes combinaisons d'outils peuvent constituer un moyen adéquat pour séparer le bon grain – corse – de l'ivraie – les autres langues et le bruit.

4.3. *Évaluation du corpus mC4*

Ce corpus ayant déjà été évalué par Caswell *et al.* (2021), nous nous faisons ici l'écho de leur travail et reprenons les résultats mis en évidence. Leur audit se base sur cent phrases – issues des 494 913 contenues dans le corpus pour le corse – jugées par un expert. La proportion de phrases effectivement exprimées en corse est de 33 %, dont 2 % sont cependant très courtes, et 2 % sont de faible qualité. Les deux tiers

36. Le jeu de données C4 multilingue est documenté à l'adresse : `https://www.tensorflow.org/datasets/catalog/c4#c4multilingual`

37. Celles-ci s'inspirent des catégories utilisées par Caswell *et al.* (2021).

restants se répartissent pour 48 % en phrases exprimées dans une autre langue, et pour 19 % en éléments non linguistiques. Étant donné ces résultats, et ceux observés pour d'autres langues peu dotées, les auteurs recommandent la plus grande prudence lorsqu'il est fait usage de cette ressource.

4.4. *Évaluation du corpus An Crúbadán*

Nous nous intéressons ici aux bigrammes de mots du corpus An Crúbadán. Un premier filtrage a été effectué afin d'écarter les éléments les plus courts, qui présentent une faible qualité linguistique, soit 25 374 éléments de moins de dix caractères. Les 24 626 bigrammes restants sont caractérisés par une longueur moyenne de 13,42 caractères. Les résultats obtenus suite à l'examen de l'échantillon par l'expert, ainsi que par les trois systèmes d'identification de langue sont présentés au tableau 7.

	C	**M**	**AL**	**B**	**Total**
Jugement expert	184 92,00 %	9 4,50 %	7 3,50 %	0 0,00 %	200
Identification automatique par les trois outils sélectionnés Les échantillons diffèrent par le nombre d'outils ayant reconnu le document comme corse (N). La proportion réelle de chaque combinaison dans le corpus complet est donnée entre parenthèses.					
N = 3 : Ldig + CLD3 + Réf. (8,63 % du corpus)	23 92,00 %	2 8,00 %	0 0,00 %	0 0,00 %	25
N = 2 : Ldig + CLD3 (28,86 % du corpus)	23 92,00 %	2 8,00 %	0 0,00 %	0 0,00 %	25
N = 2 : Ldig + Réf. (2,08 % du corpus)	24 96,00 %	1 4,00 %	0 0,00 %	0 0,00 %	25
N = 2 : CLD3 + Réf. (0,82 % du corpus)	24 96,00 %	1 4,00 %	0 0,00 %	0 0,00 %	25
N = 1 : Ldig (13,50 % du corpus)	19 76,00 %	1 4,00 %	5 20,00 %	0 0,00 %	25
N = 1 : CLD3 (16,0 % du corpus)	25 100,00 %	0 0,00 %	0 0,00 %	0 0,00 %	25
N = 1 : Réf. (0,89 % du corpus)	23 92,00 %	2 8,00 %	0 0,00 %	0 0,00 %	25
N = 0 (29,22 % du corpus)	23 92,00 %	0 0,00 %	2 8,00 %	0 0,00 %	25

Tableau 7. *Évaluation du corpus An Crúbadán*

Les jugements, fournis par notre expert, nous révèlent un niveau de qualité plutôt satisfaisant. En effet, 92 % des éléments examinés ont été jugés comme corrects. À cela, il faut ajouter 4,5 % pour lesquels il y a un mélange de corse et d'autres langues.

Au final, seuls 3,5 % ne constituent pas du contenu corse et ne devraient idéalement pas apparaître dans la ressource.

En ce qui concerne la détection effectuée par les trois outils d'identification de langue, seuls 8,63 % des éléments sont détectés à l'unanimité comme du contenu corse. Au contraire, le cas le plus fréquent est celui où aucun logiciel n'attribue l'étiquette corse (29,22 %), suivi d'assez près par l'identification du corse par le duo Ldig + CLD3 (28,86 %). Vient ensuite la reconnaissance du corse par un seul outil, que ce soit CLD3 (16 %) ou Ldig (13,50 %). Les autres combinaisons, impliquant le système de référence par mots-clés, sont plus marginales.

La confrontation de l'analyse automatique à celle fournie par l'expert ne permet pas de dégager des critères qui offriraient la possibilité d'écarter la totalité du contenu exprimé dans une autre langue sans supprimer une part importante du contenu valide. En effet, ne conserver que les documents ayant été identifiés comme corse par deux outils au minimum éliminerait l'ensemble des AL et un tiers des M, mais ne nous offrirait environ que 40 % du contenu estimé en corse. Une approche plus souple, qui garderait les éléments ayant été identifiés par un outil au minimum, permettrait de faire grimper le rappel à environ 68 %, mais inclurait également quelques éléments exprimés dans d'autres langues. Un filtrage, impliquant une perte de contenu, est donc envisageable sur cette ressource qui, rappelons-le, a été évaluée plutôt positivement par l'expert, mais en ayant tout de même connu une amputation préalable de plus de la moitié de ses données.

4.5. *Évaluation du corpus W2C*

Contrairement au corpus An Crúbadán, W2C propose des textes plus conséquents. Les 200 éléments qui constituent notre échantillon ont une longueur moyenne de 312,31 caractères. Les évaluations réalisées par l'expert et par l'intermédiaire des trois outils d'identification de langue sont reprises au tableau 8.

Le premier élément à mettre en avant concerne la fiabilité de l'identification du corse dans le corpus W2C, qui est nettement en retrait par rapport à ce qui est observé pour An Crúbadán. En effet, à peine 19,50 % des éléments examinés ont été jugés conformes par notre expert. Même si l'on ajoute les 3 % pour lesquels une ou plusieurs autres langues ont été observées, le résultat n'est pas flatteur. La majorité des extraits (67 %) est en réalité exprimée dans une autre langue – l'italien, mais aussi d'autres langues italo-romanes, ainsi que le roumain – alors qu'une part non négligeable de l'échantillon est constituée par du bruit (10,50 %). Même si cette analyse ne porte que sur des données très partielles, les chiffres observés ne laissent que peu de place à l'incertitude quant à la qualité de l'identification du corse dans ce corpus.

La majorité des textes soumis aux trois outils d'identification de langue ont été marqués comme non corses par ceux-ci (62,64 %), ce en quoi ils ont raison, puisque ces documents ont également été écartés par l'expert. Il existe trois configurations pour lesquelles une partie des éléments de l'échantillon correspond effectivement à

	C	M	AL	B	Total
Jugement expert	39 19,50 %	6 3,00 %	134 67,00 %	21 10,50 %	200
Identification automatique par les trois outils sélectionnés Les échantillons diffèrent par le nombre d'outils ayant reconnu le document comme corse (N). La proportion réelle de chaque combinaison dans le corpus complet est donnée entre parenthèses.					
N = 3 : Ldig + CLD3 + Réf. (5,12 % du corpus)	20 80,00 %	3 12,00 %	2 8,00 %	0 0,00 %	25
N = 2 : Ldig + CLD3 (1,98 % du corpus)	11 44,00 %	1 4,00 %	13 52,00 %	0 0,00 %	25
N = 2 : Ldig + Réf. (1,75 % du corpus)	0 0,00 %	0 0,00 %	10 40,00 %	15 60,00 %	25
N = 2 : CLD3 + Réf. (0,03 % du corpus)	8 32,00 %	1 4,00 %	16 64,00 %	0 0,00 %	25
N = 1 : Ldig (23,57 % du corpus)	0 0,00 %	0 0,00 %	25 100,00 %	0 0,00 %	25
N = 1 : CLD3 (4,33 % du corpus)	0 0,00 %	1 4,00 %	24 96,00 %	0 0,00 %	25
N = 1 : Réf. (0,57 % du corpus)	0 0,00 %	0 0,00 %	19 76,00 %	6 24,00 %	25
N = 0 (62,64 % du corpus)	0 0,00 %	0 0,00 %	25 100,00 %	0 0,00 %	25

Tableau 8. *Évaluation du corpus W2C*

du contenu corse. Lorsque CLD3 et le système par mots-clés (Réf.) identifient un contenu corse sans que Ldig ne le fasse, ce qui ne concerne que 0,03 % du corpus, le taux de précision estimé par rapport à l'échantillon est de 32 % (64 % sont dans une autre langue). La précision de l'identification augmente jusqu'à 44 % lorsque ce sont Ldig et CLD3 qui s'accordent sur la détection du corse, alors que la méthode par mots-clés donne une autre langue. Ce cas de figure, qui ne concerne que 1,98 % du corpus, laisse tout de même 52 % de textes dans une autre langue. Finalement, seule la combinaison des trois outils permet d'atteindre une identification plus fiable, avec 80 % des documents réellement en corse dans l'échantillon, 12 % de documents présentant une ou plusieurs langues en plus du corse, et 8 % de documents dans une autre langue.

L'utilisation conjointe des trois outils pourrait donc constituer une méthode de filtrage imparfaite, mais exploitable, pour ce corpus. Cette configuration, pour laquelle on peut espérer une présence au moins partielle du corse dans 92 % des cas, ne s'est cependant présentée que pour 5,12 % des documents du corpus, soit 4 629 éléments. Le taux de rappel à l'échelle du corpus serait donc assez faible, de l'ordre de 24,16 %,

et ne permettrait de conserver qu'un ensemble d'environ 4 259 documents en corse sur les 17 629 potentiellement disponibles.

5. Conclusion

L'identification de langue est un problème parfois considéré comme résolu. De nombreux points, qui s'appliquent en partie au corse, ne sont cependant pas encore maîtrisés. Il existe néanmoins des outils capables de traiter cette langue peu dotée, mais les performances ne sont pas au niveau de celles observées pour les « grandes » langues. Les résultats enregistrés sur les documents courts restent également en retrait. L'entraînement spécifique d'outils, pour le corse et une série de 17 langues européennes, a montré des performances intéressantes. Les données et procédures de génération de modèles utilisées pour cet article ont été mises à disposition[38].L'ajout de variantes dialectales, de langues proches du corse, voire d'autres langues régionales européennes, pourrait constituer une évolution importante de cet outil.

En ce qui concerne les corpus de grande envergure issus d'Internet et revendiquant du contenu corse, nous constatons une fiabilité peu élevée. Dans la lignée de Caswell *et al.* (2021), nous recommandons d'utiliser ces ressources avec prudence et de s'assurer de leur contenu par des sondages et des évaluations manuelles. Face à une ressource fortement bruitée, un filtrage automatisé ou semi-automatisé pourra, dans une certaine mesure, être mis en place au moyen de systèmes d'identification de langue. L'obtention d'une précision satisfaisante nécessite l'utilisation combinée de plusieurs d'entre eux, au prix d'un faible rappel et d'une diminution importante du volume du corpus, ce qui peut être acceptable si le corpus de départ est très volumineux. L'entraînement d'outils de filtrage spécifiques pourrait également s'avérer judicieux.

D'autre part, nous avons mentionné que la majorité de ces ressources ne prennent pas, ou peu, en compte les aspects liés aux droits d'exploitation et aux droits d'auteur.

Par conséquent, la constitution de corpus pour les langues peu dotées est loin d'être une problématique réglée, à plus forte raison si la prise en compte de la dimension dialectale est souhaitée. La création, manuelle ou automatique, de ressources textuelles fiables et sécurisées au niveau juridique reste, à notre avis, une priorité.

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) ». Nous remercions Stella Retali-Medori d'avoir consacré le temps nécessaire à l'important travail de validation des données corses. Enfin, tous nos remerciements vont aux relecteurs pour leurs commentaires avisés et constructifs.

38. Voir `https://github.com/lkevers/ldig-models-TAL62-3`, ainsi que `https://bdlc.univ-corse.fr/tal/index.php?page=lgid` pour une démonstration en ligne.

6. Bibliographie

Baroni M., Bernardini S., « BootCaT : Bootstrapping Corpora and Terms from the Web », *Proceedings of the fourth international conference on Language Resources and Evaluation (LREC 2004)*, ELRA, 2004.

Berment V., Méthodes pour informatiser les langues et les groupes de langues « peu dotées », Thèse de doctorat, Université Joseph Fourier (Grenoble), May, 2004.

Bernhard D., Bras M., Erhart P., Ligozat A.-L., Vergez-Couret M., « Language Technologies for Regional Languages of France : The RESTAURE Project », *International Conference Language Technologies for All (LT4All) : Enabling Linguistic Diversity and Multilingualism Worldwide*, Collection of Research Papers of the 1st International Conference on Language Technologies for All, ELRA, Paris, France, p. 272-275, December, 2019.

Brown R. D., « Selecting and Weighting N-Grams to Identify 1100 Languages », *Text, Speech, and Dialogue. TSD 2013*, vol. 8082, p. 475-483, 2013.

Caswell I., Breiner T., van Esch D., Bapna A., « Language ID in the Wild : Unexpected Challenges on the Path to a Thousand-Language Web Text Corpus », *Proceedings of the 28th International Conference on Computational Linguistics*, International Committee on Computational Linguistics, Barcelona, Spain (Online), p. 6588-6608, December, 2020.

Caswell I., Kreutzer J., Wang L., Wahab A., van Esch D., Ulzii-Orshikh N., Tapo A., Subramani N., Sokolov A., Sikasote C., Setyawan M., Sarin S., Samb S., Sagot B., Rivera C., Rios A., Papadimitriou I., Osei S., Suárez P. J. O., Orife I., Ogueji K., Niyongabo R. A., Nguyen T. Q., Müller M., Müller A., Muhammad S. H., Muhammad N., Mnyakeni A., Mirzakhalov J., Matangira T., Leong C., Lawson N., Kudugunta S., Jernite Y., Jenny M., Firat O., Dossou B. F. P., Dlamini S., de Silva N., Balli S. C., Biderman S., Battisti A., Baruwa A., Bapna A., Baljekar P., Azime I. A., Awokoya A., Ataman D., Ahia O., Ahia O., Agrawal S., Adeyemi M., « Quality at a Glance : An Audit of Web-Crawled Multilingual Datasets », *arXiv :2103.12028 [cs]*, April, 2021. arXiv : 2103.12028.

Cavnar W. B., Trenkle J. M., « N-Gram-Based Text Categorization », *Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval*, p. 161-175, 1994.

Ceberio Berger K., Gurrutxaga Hernaiz A., Baroni P., Hicks D., Kruse E., Quochi V., Russo I., Salonen T., Sarhimaa A., Soria C., *Digital Language Survival Kit. The DLDP Recommendations to Improve Digital Vitality*, 2018.

Dalbera-Stefanaggi M.-J., *La langue corse*, nº 3641 in *Que sais-je ?*, PUF, Paris, June, 2002.

Dalbera-Stefanaggi M.-J., *Nouvel atlas linguistique et ethnographique de la Corse : Volume 1, Aréologie phonétique, édition revue et corrigée*, alain piazzola edn, Comité des travaux historiques et scientifiques - CTHS, Ajaccio : Paris, December, 2007.

Dunning T., Statistical Identification of Language, Technical report, 1994.

Esplà-Gomis M., Forcada M. L., Ramirez-Sanchez G., Hoang H., « ParaCrawl : Web-scale parallel corpora for the languages of the EU », p. 118-119, August, 2019.

Giguet E., « Multilingual Sentence Categorization according to Language », *Proceedings of the European Chapter of the Association for Computational Linguistics SIGDAT Workshop "From text to tags : Issues in Multilingual Language Analysis"*, p. 73-76, 1995.

Habernal I., Zayed O., Gurevych I., « C4Corpus : Multilingual Web-size Corpus with Free License », *in* N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis (eds), *Proceedings of the Tenth*

International Conference on Language Resources and Evaluation (LREC 2016), ELRA, Portorož, Slovenia, May, 2016.

Hajlaoui N., Kolovratnik D., Väyrynen J., Steinberger R., Varga D., « DCEP -Digital Corpus of the European Parliament », *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, ELRA, Reykjavik, Iceland, May, 2014.

Hughes B., Baldwin T., Bird S., Nicholson J., Mackinlay A., « Reconsidering language identification for written language resources », *Proceedings of the fifth international conference on Language Resources and Evaluation (LREC2006)*, ELRA, p. 485-488, 2006.

Ingle N. C., « A language identification table », *The Incorporated Linguist*, 1976.

Jaech A., Mulcaire G., Ostendorf M., Smith N. A., « A Neural Model for Language Identification in Code-Switched Tweets », *Proceedings of the Second Workshop on Computational Approaches to Code Switching*, ACL, Austin, Texas, p. 60-64, November, 2016.

Jauhiainen T., Lui M., Zampieri M., Baldwin T., Lindén K., « Automatic Language Identification in Texts : A Survey », *arXiv :1804.08186 [cs]*, April, 2018. arXiv : 1804.08186.

Joshi P., Santy S., Budhiraja A., Bali K., Choudhury M., « The State and Fate of Linguistic Diversity and Inclusion in the NLP World », *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, ACL, Online, p. 6282-6293, July, 2020.

Joulin A., Grave E., Bojanowski P., Mikolov T., « Bag of Tricks for Efficient Text Classification », p. 427-431, April, 2017.

Kerwin T., Classification of natural language based on character frequency, Technical report, Ohio State University, June, 2006.

Kevers L., Retali-Medori S., « Towards a Corsican Basic Language Resource Kit », *Proceedings of the 12th Language Resources and Evaluation Conference*, European Language Resources Association, Marseille, France, p. 2726-2735, May, 2020.

Kevers L., Retali Medori S., Tognotti A. G., A Survey of Language Technologies Resources and Tools for Corsican, Research Report, UMR CNRS 6240 LISA, Université de Corse, 2021.

Kilgarriff A., Baisa V., Bušta J., Jakubíček M., Kovář V., Michelfeit J., Rychlý P., Suchomel V., « The Sketch Engine : ten years on », *Lexicography*, vol. 1, n° 1, p. 7-36, July, 2014.

Kornai A., « Digital Language Death », *PLoS ONE*, October, 2013.

Landragin F., *Comment parler à un alien ? : Langage et linguistique dans la science-fiction*, BELIAL, October, 2018.

Leixa J., Mapelli V., Choukri K., *Inventaire des ressources linguistiques des langues de France*, ELDA, September, 2014. Accessible à l'adresse `http://www.elda.org/media/filer_public/2014/12/17/rapport_dglflf_05112014-1.pdf`.

Lui M., Baldwin T., « Langid.Py : An Off-the-shelf Language Identification Tool », *Proceedings of the ACL 2012 System Demonstrations*, ACL, Stroudsburg, PA, USA, p. 25-30, 2012. event-place : Jeju Island, Korea.

Majliš M., « Yet Another Language Identifier », *Proceedings of the Student Research Workshop at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, ACL, Avignon, France, p. 46-54, April, 2012.

Majliš M., Zabokrtský Z., « Language Richness of the Web », *Proceedings of the Eighth International Conference on Language Resources and Evaluation*, Istanbul, Turkey, May, 2012.

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

Millour A., Myriadisation de ressources linguistiques pour le traitement automatique de langues non standardisées, Thèse de doctorat, Sorbonne Universite, December, 2020.

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

Nakatani S., « Language Detection Library for Java », March, 2010. `https://www.slideshare.net/shuyo/language-detection-library-for-java`.

Nakatani S., « Short Text Language Detection with Infinity-Gram », May, 2012. `https://www.slideshare.net/shuyo/short-text-language-detection-with-infinitygram-1294944`.

Okanohara D., Tsujii J., « Text Categorization with All Substring Features », *Proceedings of SDM 2009*, p. 838-846, April, 2009.

Rehurek R., Kolkus M., « Language Identification on the Web : Extending the Dictionary Method », *n : CICLing '09 : Proceedings of the 10th International Conference on Computational Linguistics and Intelligent Text Processing*, p. 357-368, March, 2009.

Roziewski S., Stokowiec W., « LanguageCrawl : A Generic Tool for Building Language Models Upon Common-Crawl », *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, ELRA, Portorož, Slovenia, p. 2789-2793, May, 2016.

Scannell K. P., « 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*, vol. 4 of *Cahiers du Cental*, Louvain-la-Neuve, Belgium, 2007.

Siewert J., Scherrer Y., Wieling M., Tiedemann J., « LSDC - A comprehensive dataset for Low Saxon Dialect Classification », *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, International Committee on Computational Linguistics (ICCL), Barcelona, Spain (Online), p. 25-35, December, 2020.

Soria C., Mariani J., Zoli C., « Dwarfs sitting on the giants' shoulders – how LTs for regional and minority languages can benefit from piggybacking major languages », *XVII FEL Conference*, Ottawa, October, 2013.

Suárez P. J. O., Romary L., Sagot B., « A Monolingual Approach to Contextualized Word Embeddings for Mid-Resource Languages », *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 1703-1714, July, 2020.

Tahir B., Mehmood M. A., « Corpulyzer : A Novel Framework for Building Low Resource Language Corpora », *IEEE Access*, vol. 9, p. 8546-8563, 2021.

Takçı H., Güngör T., « A high performance centroid-based classification approach for language identification », *Pattern Recognition Letters*, vol. 33, nº 16, p. 2077-2084, December, 2012.

Wenzek G., Lachaux M.-A., Conneau A., Chaudhary V., Guzmán F., Joulin A., Grave E., « CCNet : Extracting High Quality Monolingual Datasets from Web Crawl Data », *arXiv :1911.00359 [cs, stat]*, November, 2019. arXiv : 1911.00359.

Xue L., Constant N., Roberts A., Kale M., Al-Rfou R., Siddhant A., Barua A., Raffel C., « mT5 : A Massively Multilingual Pre-trained Text-to-Text Transformer », *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, ACL, Online, p. 483-498, June, 2021.