Facilitation de la mise en œuvre de la Charte européenne des langues régionales ou minoritaires par l'intelligence artificielle

Conseil de l'Europe,
Secrétariat de la Charte européenne
des langues régionales ou minoritaires

COUNCIL OF EUROPE
CONSEIL DE L'EUROPE

Publication du Conseil de l'Europe
Auteure : Miriam Gerken
Conception et rédaction : Secrétariat de la Charte européenne des langues régionales ou minoritaires
Image de couverture : Shutterstock
MIN-LANG(2022)4
Février 2022

Table des matières

Introduction **4**

1. Raisons générales d'utiliser l'IA pour faciliter la mise en œuvre de la Charte **5**

2. La traduction automatique **7**

2.1 Les méthodes de traduction automatique 7

2.2 Recours aux applications existantes 8

2.3 Développement de nouvelles applications 9

3. Autres applications du traitement du langage naturel (NLP) et utilité dans la mise en œuvre de la Charte **10**

3.1 Les langues régionales ou minoritaires dans la vie privée (article 7.1.d) 10

3.2 Les langues régionales ou minoritaires dans l'éducation (articles 7.1.g, 8.1.f ii, iii) 11

3.3 Les langues régionales ou minoritaires dans la justice (articles 9.1.a.i-iv, 9.1.b.i-iii, 9.1.c.i-iii, 9.1.d, 9.3) 13

3.4 Les langues régionales ou minoritaires dans l'administration et les services publics 13

3.4.1 Dialogueurs (Articles 10.1.a.i-iv, 10.2.a, 10.2.b, 10.3.a— c) 13

3.4.2 Recherche intelligente (article 10.2.g) 14

3.4.3 Synthèse vocale pour l'annonce des noms de rues (article 10.2.g) 15

3.4.4 Traduction automatique (articles 10.1.a.i-v, 10.1.b, 10.1.c, 10.2.a-f, 10.3.a-c, 10.4.a) 15

3.5 Les langues régionales ou minoritaires dans les médias 16

3.5.1 Génération automatique de sous-titres (articles 11.1.a.i-iii, 11.1.c.i-ii) 16

3.5.2 Extraction automatique d'informations (article 11.1.e.i-ii) 17

3.6 Les langues régionales ou minoritaires dans les activités et équipements culturels 17

3.6.1 Structuration des données (articles 12.1.g, 12.1.h) 17

3.6.2 Traduction automatique (articles 12.1.a, 12.1.b, 12.1.c) 18

3.6.3 Génération automatique de sous-titres (articles 12.1.b, 12.1.c) 18

3.7 Les langues régionales ou minoritaires dans la vie économique et sociale 18

3.7.1 Analyse des sentiments (article 13.1.c, 13.1.d, 13.2.b) 18

3.7.2 Traduction automatique (articles 13.1.a, 13.1.d, 13.2.a, 13.2.b, 13.2.d, 13.2.e) 19

3.8 Les langues régionales ou minoritaires dans les échanges transfrontaliers (articles 7.1.i, 14) 19

Perspectives **21**

Remerciements

La rédaction du présent rapport et les recherches préalables ont été effectuées par Miriam Gerken, lors de sa visite d'étude de février 2020 au Conseil de l'Europe, avec le concours du secrétariat de la Charte européenne des langues régionales ou minoritaires ; le document a été actualisé en mars 2022. Miriam Gerken a obtenu son diplôme de traductologie en 2018 et de linguistique informatique en 2021 aux universités de Hildesheim et de Bielefeld (Allemagne), avec spécialisation en traduction automatique. Elle est à présent experte en technologies linguistiques à la Deutsche Bahn AG. Le secrétariat de la Charte remercie Miriam Gerken d'avoir puisé dans sa connaissance approfondie de l'application de l'IA aux langues pour préparer un guide complet et pratique à l'intention des responsables politiques et des praticiens travaillant à la mise en œuvre de la Charte et à la promotion de l'emploi quotidien des langues régionales ou minoritaires dans la vie publique et privée.

Introduction

Des langues régionales ou minoritaires sont historiquement employées dans certaines parties de pays d'Europe par une minorité de la population. Leur situation démographique et juridique varie considérablement de l'une à l'autre. Beaucoup d'entre elles ont toutefois en partage une plus ou moins grande précarité.

La Charte européenne des langues régionales ou minoritaires du Conseil de l'Europe est pour le monde entier le cadre juridique de référence pour la promotion de ces langues. Elle soutient leur emploi dans plusieurs domaines de la vie publique : éducation, justice, administration et services publics, médias, activités et équipements culturels, vie économique et sociale et échanges transfrontaliers. Elle est ratifiée par 25 États[1], et signée (mais non ratifiée) par neuf autres[2].

Les travaux préparatoires de la Charte ont commencé en 1984. Le Comité des Ministres du Conseil de l'Europe a adopté le texte du traité en 1992. Depuis, de nouvelles technologies sont venues transformer les conditions de sa mise en œuvre au sein des États parties. L'internet, par exemple, a contribué à accroître l'offre de médias, ce qui s'est traduit par une certaine obsolescence des demandes de fréquences radio ou de créneaux horaires plus commodes de diffusion des émissions en langues régionales ou minoritaires.

En plein essor, l'intelligence artificielle offre aux États parties de nouvelles possibilités, mais leur confère aussi de nouvelles responsabilités dans l'application de la Charte. Hier du domaine de la science-fiction, elle a fait irruption dans la vie quotidienne de bien des gens aujourd'hui : elle leur conseille de partir plus tôt quand la circulation est dense sur leur trajet, leur suggère les films qu'ils pourraient aimer, ou leur permet de lire des messages dans des langues qu'ils ne connaissent pas. Mais qu'est-ce exactement que l'intelligence artificielle ? Et comment peut-elle apporter des solutions à un problème bien spécifique du monde réel : le soutien des langues régionales ou minoritaires ?

L'IA consiste d'une manière générale en machines intelligentes capables d'analyser leur environnement et de prendre des décisions sur cette base. Ces systèmes sont censés imiter des capacités humaines, comme l'apprentissage ou la résolution de problèmes par le traitement statistique de données dans des réseaux neuronaux. Leur application à des problèmes linguistiques, le « traitement du langage naturel » (ou *natural language processing*, NLP), est l'un des grands sous-domaines de l'IA et des sciences informatiques. Il porte sur les interactions homme-machine en langage naturel, et vise au développement de logiciels capables de lire, de traiter, d'analyser et en dernier ressort de comprendre le langage naturel dans toute sa complexité. Cela exige de gros volumes de données en langage naturel.

Le Conseil de l'Europe assiste les États parties dans la mise en œuvre de la Charte. Le présent rapport s'inscrit dans cet effort. Il montre comment des applications NLP (c'est-à-dire l'IA) peuvent faciliter l'usage quotidien et la promotion des langues régionales ou minoritaires, et aider ainsi les États parties à respecter les dispositions de la Charte qu'ils ont ratifiées. C'est pourquoi il suit dans sa structure les articles 7 à 14 de la Charte.

[1] Allemagne, Arménie, Autriche, Bosnie-Herzégovine, Croatie, Chypre, Danemark, Espagne, Finlande, Hongrie, Liechtenstein, Luxembourg, Monténégro, Norvège, Pays-Bas, Pologne, République tchèque, Roumanie, Serbie, Slovaquie, Slovénie, Suède, Suisse, Ukraine et Royaume-Uni.

[2] Azerbaïdjan, France, Islande, Italie, Malte, République de Moldova, Portugal, Fédération de Russie et Macédoine du Nord.

La Convention-cadre pour la protection des minorités nationales du Conseil de l'Europe définit les droits des personnes appartenant à des minorités nationales dans certains domaines, dont la langue et la culture. Elle a été signée par 39 États, dont plusieurs n'ont pas encore ratifié la Charte. Les mesures d'aide à la mise en œuvre de la Charte favorisent également la réalisation des droits linguistiques garantis dans la Convention-cadre, dont le présent rapport évoque les dispositions correspondantes.

Le Comité *ad hoc* sur l'intelligence artificielle du Conseil de l'Europe (CAHAI) a étudié la teneur potentielle d'un cadre juridique de l'intelligence artificielle fondé sur les normes du Conseil de l'Europe relatives aux droits de l'homme, à la démocratie et à l'État de droit. Au cours de ses travaux, de 2019 à 2021, il a constaté que l'IA peut aider les locuteurs de langues minoritaires à participer plus activement à la vie publique, aux débats, aux délibérations ou aux décisions, mais aussi mieux faire connaître ces langues ; il a notamment estimé que l'intégration des langues minoritaires dans les applications d'IA pourrait être un moyen de parer à la discrimination linguistique[3].

Le présent rapport pourrait ainsi contribuer à une réflexion plus large sur l'IA au sein du Conseil de l'Europe.

1. Raisons générales d'utiliser l'IA pour faciliter la mise en œuvre de la Charte

La présente section montre comment l'IA peut contribuer à la réalisation des buts généraux de la Charte, en préalable à l'examen du soutien qu'elle peut spécifiquement apporter à la mise en œuvre de telle ou telle de ses dispositions.

Les chercheurs parlent d**'« extinction numérique »** lorsqu'une langue ne parvient pas à s'implanter dans les nouveaux moyens de communication et les technologies nouvelles. Le phénomène peut affecter des langues peu répandues, mais aussi une langue qui, bien que majoritaire dans un pays, est minoritaire et menacée dans un autre. Ce risque impose de consacrer de vigoureux efforts à la percée des langues régionales ou minoritaires dans les technologies nouvelles. Bien mis à profit, les technologies et outils de communication modernes offrent un énorme potentiel de promotion de ces langues.

Les applications décrites ici sont toutes directement et aisément utilisables en situation réelle. Nombre d'entre elles aident à faire connaître les langues régionales ou minoritaires à un public plus large. La possibilité d'usage d'une langue régionale ou minoritaire dans des applications courantes d'IA, comme les assistants domotiques pour la maison intelligente, élargit considérablement son potentiel d'emploi dans la vie quotidienne, et encourage et motive ceux qui l'apprennent ; cela contribue à accroître le nombre de ses locuteurs. La présence d'une langue régionale ou minoritaire sur des plateformes internet augmente par ailleurs sa visibilité. C'est par exemple le cas du catalan, l'une des vingt langues les plus employées sur Wikipédia.

L'IA aide les autorités à mettre rapidement une offre à la disposition des locuteurs de langues régionales ou minoritaires à un coût modique, et ainsi à mener une action résolue de promotion des langues régionales ou minoritaires afin de les sauvegarder — ce qui figure en bonne place parmi les principes et objectifs de la Charte (article 7.1.c).

[3] Voir Comité *ad hoc* sur l'intelligence artificielle (CAHAI) : Étude de faisabilité, CAHAI(2020)23, p. 19 ; Analyse de la consultation multipartite, CAHAI(2021)07, p. 22 ; Compilation des réponses à la consultation multipartite (F à M), CAHAI(2021)06, p. 80.

Toutes les applications évoquées dans ce rapport sont par nature transnationales et utilisables dans d'autres pays où se parle la même langue régionale ou minoritaire. Une fois une application développée pour une région et une langue, il est aisé de l'adapter à d'autres régions et à d'autres langues régionales ou minoritaires. Comme expliqué plus en détail ci-dessous, l'IA favorise ainsi les échanges et les contacts entre groupes vivant dans des pays où se parle la même langue régionale ou minoritaire, et entre leurs autorités.

Les domaines d'étude auxquels touchent l'IA et le NLP sont relativement nouveaux et porteurs. Aider les langues régionales ou minoritaires à y être présentes les aidera donc à survivre et à se développer comme langues vivantes. La promotion de la recherche sur ces langues est d'ailleurs prévue à l'article 7.1.h, qui s'applique à toute langue régionale ou minoritaire pratiquée dans un État partie. Le présent rapport indique plusieurs pistes de recherches futures. Le plus important, pour pouvoir développer des applications NLP utiles, est cependant de collecter des données en langage naturel et de constituer ainsi des **corpus** de textes (grandes bases textuelles dans une ou plusieurs langues, selon la tâche NLP à accomplir) en langues régionales ou minoritaires. La création d'applications NLP fonctionnelles est impossible sans un volume suffisant de données. La promotion d'études et de recherches sur les langues régionales ou minoritaires devrait donc commencer par la collecte de données en langage naturel.

Plusieurs projets et équipes de recherche ont déjà été créés. Le groupe de travail SALTMIL (*speech and language technology for minority languages*), par exemple, promeut la recherche et le développement en technologies de la parole et du langage pour les langues peu répandues, surtout d'Europe. *Ixa*, à l'Université du Pays basque (Espagne), développe et distribue des algorithmes, des outils et des ressources linguistiques permettant aux ordinateurs de traiter et de comprendre les langues humaines, en particulier le basque. Le groupe a déjà publié des corpus et des outils NLP pour le basque, comme un *wordnet* (base lexicale de relations sémantiques entre les mots), un correcteur orthographique et un *parser* (logiciel d'analyse syntaxique de phrases) ainsi que des outils de recherche et d'extraction d'informations, de traduction automatique et d'apprentissage des langues. L'université d'Helsinki/Helsingfors, quant à elle, étudie des technologies linguistiques comme le *treebanking* (corpus textuel d'informations sur les structures syntaxiques et sémantiques de la phrase) pour des langues minoritaires de Finlande.

Le plan d'action technologique pour le gallois de 2018 offre un bon exemple de soutien public à des technologies linguistiques. Il s'inscrit dans la stratégie générale du gouvernement gallois de promotion de l'usage quotidien du gallois et d'accroissement du nombre de ses locuteurs actifs. Il encourage et oriente le développement de technologies informatiques applicables au gallois. Le gouvernement gallois finance dans ce but de multiples projets, comme la programmation d'applications d'apprentissage du gallois pour enfants, une version en gallois d'*OpenStreetMap* (carte du monde collaborative gratuite en *open source*), une base de données vocales en gallois ou le développement de *Macsen*, un assistant numérique en gallois. Il favorise en outre l'utilisation de corpus parallèles existants et d'autres outils, telles les listes de mots vides (liste de mots fréquemment utilisés dans une langue, comme prépositions et articles) ou les analyseurs syntaxiques pour projets de recherche. Il organise par ailleurs des conférences annuelles pour chercheurs à l'université de Bangor (Pays de Galles), et crée des postes universitaires réservés à ce domaine de recherche. Ces actions ont un effet notable de sensibilisation aux technologies actuelles et futures au service du gallois et à cette langue en général. Elles donnent de bons résultats et pourraient servir d'exemples dans d'autres pays.

2. La traduction automatique

La traduction automatique peut se révéler d'une grande utilité dans la mise en œuvre de la Charte. Nombre d'engagements portent sur la traduction de documents en langues régionales ou minoritaires. Ces exigences pourraient être satisfaites en moins de temps et à moindres frais par traduction automatique. Les sections suivantes examinent donc la traduction automatique, ses méthodes, ses applications existantes, les langues régionales ou minoritaires qu'elle traite déjà et les façons de créer de nouvelles applications de traduction dans des domaines particuliers.

2.1 Les méthodes de traduction automatique

Un logiciel de traduction automatique traduit par lui-même un texte d'une langue en une autre. C'est l'un des volets de la **linguistique informatique** auquel a été consacré le plus de recherches. Elle doit résoudre des problèmes complexes pour décoder le sens dans la langue de départ ou langue source, puis le réencoder dans la langue d'arrivée ou langue cible. Or le décodage du sens d'une phrase impose la compréhension sémantique d'une langue, ce qui n'est pas (encore) réalisable : on sait traiter un langage naturel (NLP), mais pas le comprendre. L'incapacité du logiciel à comprendre le fait buter sur des obstacles comme l'ambiguïté (lexicale et syntaxique), la traduction des noms propres (par exemple d'institutions) et les difficultés structurelles, sans même parler des différences culturelles entre langues source et cible. Ces difficultés proviennent toutes du fait que la traduction ne consiste évidemment pas à transposer des mots dans une autre langue, mais du sens, qui dépend lui-même du contexte de chaque langue.

La traduction automatique n'en donne pas moins déjà de bons résultats dans certains domaines. Ce sont surtout les sujets dans lesquels la formulation est standardisée et répétitive, avec des structures syntaxiques similaires, sans ambiguïté : bulletins météorologiques ou formulaires, par exemple.

Il existe trois grandes approches méthodologiques de la traduction automatique.

La **traduction automatique basée sur des règles** s'appuie sur un lexique prédéfini et des règles explicites de formation grammaticale des phrases dans la langue source et la langue cible. La complexité de ces descriptions linguistiques et le peu de fiabilité des traductions obtenues font que cette méthode est maintenant en général considérée comme dépassée.

Dans la **traduction automatique statistique**, le logiciel utilise des données textuelles traduites en langue de départ et en langue d'arrivée (corpus parallèles) et en déduit des probabilités : le mot « le film » est par exemple souvent suivi de « était ». Lorsqu'il reçoit un nouveau segment, l'algorithme utilise ces probabilités pour sélectionner la traduction statistiquement la meilleure, c'est-à-dire celle qui présente la plus grande probabilité. C'est un système plus facile à mettre en œuvre que le précédent, car il ne nécessite pas de règles définies manuellement, uniquement des corpus parallèles. La traduction automatique statistique ne donne cependant pas de très bons résultats pour toutes les paires de langues, en particulier celles dont les syntaxes sont très différentes.

La **traduction automatique neuronale** est utilisée dans la plupart des applications de traduction automatique d'aujourd'hui. Elle recourt à l'IA et aux réseaux neuronaux pour « apprendre » à apparier des phrases, comme des neurones biologiques apprennent des connexions. Un réseau commence par coder la phrase d'entrée en **vecteurs de mots** : des représentations vectorielles qui permettent au système de transférer les relations sémantiques entre les mots dans un espace mathématique. Un écart de sens entre deux mots apparaît ainsi sous forme de différence de valeur entre leurs

représentations vectorielles. Par exemple, en soustrayant les valeurs du vecteur du mot « homme » au vecteur du mot « roi » et en ajoutant les valeurs du vecteur du mot « femme », on obtient un vecteur égal à celui du mot « reine ». C'est un progrès considérable en traduction automatique et l'une des raisons pour lesquelles la traduction automatique neuronale donne de bons résultats aujourd'hui. Après cet encodage, le réseau décode les vecteurs de mots dans la langue d'arrivée. Cela est rendu possible par les neurones mentionnés ci-dessus, qui ont « appris » les connexions entre la langue de départ et la langue d'arrivée dans des corpus parallèles, comme pour la traduction automatique statistique. L'apprentissage des réseaux neuronaux (neurones connectés) à partir de données s'appelle l'**entraînement**. La traduction automatique neuronale donne de meilleurs résultats que les systèmes antérieurs, tout en nécessitant moins de mémoire que la traduction automatique statistique. C'est pourquoi c'est elle qu'utilisent la plupart des systèmes de traduction automatique actuels.

2.2 Recours aux applications existantes

Nombre de systèmes de traduction automatique utilisés aujourd'hui prennent déjà en charge des langues régionales ou minoritaires couvertes par la Charte. Le tableau suivant passe en revue les principales, avec les langues traduites (par ordre alphabétique).

Système TA	Langues régionales ou minoritaires prises en charge	Nombre de langues
DeepL	Bulgare, tchèque, danois, finnois, français, allemand, grec, hongrois, italien, lituanien, polonais, roumain, russe, slovaque, slovène, suédois.	16
Google Translate	Albanais, arménien, basque, biélorusse, bosniaque, bulgare, catalan, croate, tchèque, danois, finnois, français, frison, galicien, allemand, grec, hongrois, irlandais, italien, kurde, lituanien, macédonien, polonais, roumain, russe, gaélique écossais, serbe, slovaque, slovène, suédois, tatar, turc, ukrainien, gallois, yiddish.	35
Microsoft Translator	Bosniaque, bulgare, catalan, croate, tchèque, danois, finnois, français, allemand, grec, hongrois, irlandais, italien, kurde, lituanien, macédonien, polonais, roumain, russe, serbe, slovaque, slovène, suédois, tatar, turc, ukrainien, gallois.	27
PROMT	Finnois, français, allemand, grec, italien, russe, tatar, turc, ukrainien	9
Watson Language Translator (IBM)	Basque, bosniaque, bulgare, catalan, croate, tchèque, danois, finnois, français, allemand, grec, hongrois, irlandais, italien, lituanien, polonais, roumain, russe, serbe, slovaque, slovène, suédois, turc, ukrainien, gallois.	25
Yandex	Albanais, arménien, basque, biélorusse, bosniaque, bulgare, catalan, croate, tchèque, danois, finnois, français, galicien, allemand, grec, hongrois, irlandais, italien, lituanien, macédonien, polonais, roumain, russe, gaélique écossais, serbe, slovaque, slovène, suédois, tatar, turc, ukrainien, gallois, yiddish.	33

Cette liste n'est aucunement exhaustive. D'autres sociétés offrent des services de traduction automatique ; le marché évoluant rapidement, de nouvelles pourraient bientôt voir le jour. DeepL, par exemple, n'est apparu qu'en 2017 et s'est rapidement fait une bonne réputation.

Tous les services ci-dessus sont accessibles en ligne avec un navigateur web standard. DeepL et PROMT proposent également une version de bureau de leur produit, avec traduction hors ligne. La version de bureau de PROMT n'est toutefois pas gratuite.

L'utilisation de Microsoft Translator et de Watson Language Translator nécessite l'ouverture d'un compte. Le service est gratuit (avec un plafond de mots traduits pour Watson Language Translator).

Ces logiciels présentent des degrés variables de précision en fonction de la langue source et de la langue cible choisies. Dans certaines langues, en particulier celles pour lesquelles il existe peu de données, ils peuvent ne servir qu'à se faire une idée générale du contenu d'un texte ou d'un document pour décider si la traduction par un traducteur professionnel se justifie (**raffinage, écrémage ou *gisting***). On voit donc que chaque système se prête à certaines utilisations, chacun a ses avantages et ses inconvénients : aucun ne saurait être vu comme une panacée.

2.3 Développement de nouvelles applications

Les services de traduction automatique existants couvrent de nombreuses langues régionales ou minoritaires, mais il n'en existe pas pour plusieurs d'entre elles. Or la traduction automatique serait là aussi possible.

Comme indiqué dans la première section, c'est dans un domaine bien circonscrit, avec un vocabulaire limité et des formules toutes faites, que la traduction automatique fonctionne le mieux. C'est justement le cas pour de nombreuses traductions que demande la Charte (documents juridiques ou financiers, formulaires administratifs, etc.), et il serait donc possible de développer des services de traduction automatique qui fourniraient probablement de bons résultats dans ces domaines.

L'essentiel à cette fin est de disposer de données parallèles (textes identiques alignés en langue source et en langue cible, de préférence traduits par un traducteur professionnel). Il devient alors possible d'entraîner un logiciel déjà fonctionnel avec ces données (plusieurs services permettent d'intégrer ses propres données d'entraînement dans un logiciel existant), ou alors de développer son propre service de traduction en recourant à l'un des nombreux didacticiels disponibles en ligne. Ce pourrait aussi être l'occasion de promouvoir la recherche et l'étude des langues régionales ou minoritaires, car dès lors que l'on dispose de données, le développement du système est par exemple à la portée d'une équipe pluridisciplinaire d'étudiants de programmation et d'informatique.

3. Autres applications du traitement du langage naturel (NLP) et utilité dans la mise en œuvre de la Charte

3.1 Les langues régionales ou minoritaires dans la vie privée (article 7.1.d)

La vie sociale, notamment celle des jeunes, se déroule en grande partie aujourd'hui sur les **plateformes de réseaux sociaux** en ligne, qui intègrent de nombreuses applications de l'IA dans leurs services. Ces médias ont joué un rôle particulièrement important de maintien de la communication et des rapports sociaux pendant la pandémie de covid-19. Il serait tout à fait naturel de s'intéresser à eux et d'y encourager l'emploi des langues régionales ou minoritaires à protéger et promouvoir. De nombreux réseaux sociaux sont d'ailleurs déjà disponibles dans des langues régionales ou minoritaires. Le tableau ci-dessous montre les langues prises en charge par les trois plus grandes plateformes de réseaux sociaux.

Réseau	Langues régionales ou minoritaires prises en charge	Nombre de langues
Facebook	Albanais, arménien, basque, biélorusse, bosniaque, bulgare, catalan, croate, tchèque, danois, finnois, français, frison, galicien, allemand, grec, hongrois, irlandais, italien, kurde, lituanien, macédonien, polonais, roumain, russe, serbe, slovaque, slovène, suédois, tatar, turc, ukrainien, gallois.	33
Instagram	Bulgare, croate, tchèque, danois, finnois, français, allemand, grec, italien, polonais, roumain, russe, serbe, slovaque, suédois, turc, ukrainien.	17
Twitter	Tchèque, danois, finnois, français, allemand, grec, hongrois, italien, polonais, roumain, russe, suédois, turc, ukrainien.	14

Les autorités et les particuliers peuvent agir de quatre grandes façons pour que les langues régionales ou minoritaires soient présentes sur les réseaux sociaux. Tout d'abord utiliser les plateformes dans ces langues et encourager les jeunes à le faire, par exemple en classe de langue régionale ou minoritaire. La langue s'installe ainsi dans la vie sociale quotidienne (en ligne), ce qui rehausse sa modernité, son extension et sa visibilité. Les locuteurs de langues régionales ou minoritaires déjà traduites sur les plateformes de réseaux sociaux peuvent contribuer à l'amélioration de ces traductions, par exemple sur des forums de traduction en ligne. Il existe déjà une **communauté de traducteurs** qui aide à mieux intégrer de langues régionales ou minoritaires dans la structure de plateformes, surtout Facebook. Les usagers peuvent soumettre des traductions originales ou en améliorer d'existantes. Troisièmement, les locuteurs de langues régionales ou minoritaires encore non traduites peuvent demander qu'elles le soient et encourager d'autres personnes à faire de même, afin de montrer aux réseaux sociaux qu'il existe un besoin de traduction dans la langue concernée. C'est une bonne occasion pour les défenseurs des langues régionales ou minoritaires de soutenir et d'encourager l'usage de la langue sur les plateformes de réseaux sociaux. Enfin, ces plateformes sont un nouvel outil permettant de tisser des liens entre groupes de locuteurs ou d'apprenants de langues régionales ou minoritaires, en particulier de différents pays. Ces groupes peuvent organiser des réunions ou des actions, communiquer dans la langue régionale ou minoritaire ou simplement créer des réseaux de contacts. Cela facilite, encourage et promeut les échanges transfrontaliers.

3.2 Les langues régionales ou minoritaires dans l'éducation (articles 7.1.g, 8.1.f ii, iii)

L'enseignement et l'apprentissage d'une langue sont des façons très importantes, sinon les plus importantes, de faire en sorte qu'elle soit comprise et employée dans la vie sociale et d'assurer sa survie. Des **plateformes d'apprentissage en ligne** fondées sur l'IA peuvent faciliter l'**apprentissage des langues régionales ou minoritaires**. La pandémie de covid-19 a bien montré qu'elles jouent un rôle essentiel et devraient couvrir ces langues. Elles peuvent également contribuer à la mise en œuvre de l'article 14.2 de la Convention-cadre.

Les plateformes d'apprentissage en ligne sont des sites web ou des applications mobiles destinés aux personnes qui souhaitent acquérir ou améliorer la connaissance d'une langue. Elles utilisent divers supports (textes, enregistrements sonores, etc.) et souvent aussi des procédés de **ludification** (ajout d'éléments ludiques, comme récompenses ou concours, à des activités non ludiques) pour rendre l'apprentissage divertissant et efficace. La plupart exploitent les données de l'apprenant pour améliorer le produit. Nombre d'entre elles prennent déjà en charge des langues régionales ou minoritaires, comme le montre le tableau ci-dessous.

Nom	Langues régionales ou minoritaires prises en charge	Nombre de langues
Beelinguapp	Français, allemand, italien, russe, suédois, turc	6
Busuu	Français, allemand, italien, polonais, russe, turc	6
Clozemaster	Albanais, arménien, basque, biélorusse, bulgare, catalan, cornique, croate, tchèque, danois, finnois, français, galicien, allemand, grec, hongrois, irlandais, italien, lituanien, macédonien, polonais, roumain, russe, gaélique écossais, serbe, slovaque, slovène, suédois, turc, ukrainien, gallois, yiddish.	32
Gouttes	Bosniaque, croate, danois, finnois, français, allemand, grec, hongrois, italien, polonais, russe, serbe, suédois et turc.	14
Duolingo	Allemand, danois, finnois, français, grec, hongrois (stade expérimental), irlandais, italien, polonais, roumain, russe, gaélique écossais, suédois, tchèque, turc, ukrainien, gallois, yiddish (stade expérimental).	18
Memrise	Danois, français, allemand, italien, polonais, russe, slovène, suédois, turc.	9
Mondly	Bulgare, catalan, croate, tchèque, danois, finnois, français, allemand, grec, hongrois, italien, lithuanien, polonais, roumain, russe, slovaque, suédois, turc, ukrainien	19

Ces plateformes sont toutes accessibles sous forme d'application mobile. À l'exception de Beelinguapp et de Drops, toutes proposent également leurs services sur un site web. L'utilisateur ouvre un compte afin d'enregistrer sa progression dans l'apprentissage, qu'il peut reprendre à la dernière sauvegarde. Toutes les plateformes ci-dessus sont gratuites, même si l'usager peut par exemple éliminer les publicités en payant un abonnement. Il existe aussi des offres payantes, comme Babbel ou Rosetta

Stone, qui ne figurent pas dans la liste ci-dessus, l'obligation de paiement pouvant exclure certains apprenants.

D'autres particularités sont à noter en ce qui concerne les plateformes énumérées ci-dessus.

— **Beelinguapp** ne propose pas de « leçons » à proprement parler, mais la lecture et l'écoute parallèles de la langue de l'apprenant et de la langue qu'il apprend.
— **Clozemaster** propose de très nombreuses langues, mais il est surtout à conseiller aux apprenants avancés, car il ne propose pas de leçons à proprement parler, mais apporte du vocabulaire nouveau en contexte.
— **Drops** a prévu une application spéciale pour les enfants *(Droplets)*, qui recourt à des jeux spéciaux de dessin, sans publicité. Les parents peuvent en outre protéger les achats in-app par un mot de passe secondaire.
— **Duolingo** est probablement l'application d'apprentissage des langues la plus intéressante dans notre contexte ; il offre des cours de qualité dans de nombreuses langues régionales ou minoritaires, avec de nouvelles méthodes d'enseignement des langues (comme des histoires ou des podcasts). Il en existe une version spéciale pour les écoles : l'enseignant peut former des classes avec les comptes de ses élèves, suivre leur apprentissage et leur donner des devoirs.
— **Memrise** propose des cours « communautaires », créés par les utilisateurs. Plusieurs langues régionales ou minoritaires sont déjà couvertes, comme l'albanais, l'arménien, le basque, le catalan, le finnois, le grec, le hongrois, l'irlandais, le lituanien, le roumain, le gaélique écossais, le sâme du nord ou le gallois. L'ajout d'autres langues serait simple puisque chaque membre enregistré peut créer son propre cours, ensuite utilisable par n'importe quel autre usager de la plateforme.
— **Mondly** a également une application spéciale d'apprentissage des langues pour enfants appelée *Mondly Kids*. Les leçons sont simples, et la ludification est adaptée aux enfants.

Au-delà de cet apprentissage « classique » des langues, des applications *comme HelloTalk* mettent en relation des apprenants avec des locuteurs de langue maternelle pour leur permettre de parler ou de tchatter ensemble. Ces **tandems linguistiques en ligne** peuvent également intéresser des apprenants de langues régionales ou minoritaires.

Les plateformes d'apprentissage ont pour grand inconvénient de ne pas offrir toutes les combinaisons possibles de langues d'enseignement et de langues à apprendre. La plupart ne proposent que quelques langues d'enseignement, surtout l'anglais. Elles peuvent donc pour l'instant compléter, mais non pas remplacer, l'enseignement ordinaire. Elles ont toutefois de quoi attirer des apprenants plus âgés de pays étrangers qui souhaitent apprendre une langue régionale ou minoritaire et possèdent déjà une connaissance pratique de l'anglais, par exemple.

En résumé, de nombreux systèmes d'apprentissage des langues proposent et prennent en charge des langues régionales ou minoritaires. Ils présentent de nombreux avantages, en permettant par exemple à des personnes qui ne vivent pas dans la région, voire dans le pays, d'apprendre et de maîtriser une langue régionale ou minoritaire, et en rehaussant la visibilité d'une langue. Les États parties peuvent mettre ces offres à profit en faisant connaître l'existence de cours de langues en ligne sur ces applications, ou alors en créant de nouveaux cours de langues ou des fiches de vocabulaire pour une langue régionale ou minoritaire qui n'est pas encore couverte par une application. Cela mettrait les outils pédagogiques à la disposition d'un grand nombre d'apprenants potentiels, et donnerait ainsi une plus grande visibilité à la langue.

3.3 Les langues régionales ou minoritaires dans la justice (articles 9.1.a.i-iv, 9.1.b.i-iii, 9.1.c.i-iii, 9.1.d, 9.3)

L'article 9 traite de l'emploi des langues régionales ou minoritaires dans l'un des grands domaines d'action de l'État : la justice. Les applications NLP, notamment la traduction automatique, peuvent servir au traitement de documents en langues régionales ou minoritaires. La traduction automatique revêt un intérêt particulier au regard de l'article 9.1.d en évitant le coût, souvent élevé, de la traduction professionnelle.

Le fonctionnement de la traduction automatique et les applications existantes ont déjà été abordés à la section précédente. On l'a vu, la traduction automatique des documents juridiques serait très prometteuse, en raison de leur structure et de leurs formulations fixes et répétitives. Un investissement dans ce domaine pourrait donc donner de bons résultats.

Des applications de traduction automatique existantes peuvent également servir au **raffinage** *(gisting)***.** Cela peut se révéler utile pour des langues régionales ou minoritaires dans lesquelles la traduction automatique ne donne pas une précision insuffisante.

La traduction automatique peut par ailleurs contribuer à la mise en œuvre de l'article 10.3 de la Convention-cadre.

3.4 Les langues régionales ou minoritaires dans l'administration et les services publics

Les applications NLP peuvent faciliter l'emploi des langues régionales ou minoritaires dans l'administration et les services publics, ainsi que pour la communication avec eux. Les applications envisageables sont les dialogueurs (*chatbots)*, les systèmes de recherche intelligente, la synthèse vocale (pour les annonces de noms de rues) et là encore la traduction automatique. Elles peuvent de surcroît contribuer à la mise en œuvre des articles 10.2 (emploi des langues minoritaires avec les autorités) et 11.3 (toponymie) de la Convention-cadre.

3.4.1 Dialogueurs (Articles 10.1.a.i-iv, 10.2.a, 10.2.b, 10.3.a— c)

Les dialogueurs sont des logiciels capables de converser avec les clients pour répondre directement à leurs demandes et questions. Ils répondent à des questions standards, ou alors aiguillent la personne vers un agent humain. Ces **agents de dialogue** sont censés imiter le langage humain écrit ou parlé, comme dans une conversation ou une interaction avec une personne réelle. Ils commencent par traiter le texte tapé par le client, réagissent sur la base d'un algorithme d'interprétation, puis sélectionnent la réponse à donner. Certains dialogueurs repèrent des mots et des expressions clés et transmettent une réponse préparée ou programmée. Cette méthode se prête particulièrement bien aux informations simples, classables en catégories prévisibles. D'autres dialogueurs apprennent par IA de nouvelles réponses à partir de leurs interactions avec les clients. Un dialogueur est utilisable sur une messagerie très fréquentée, par SMS, en logiciel autonome ou sur un site web. De nombreuses entreprises proposent des services de programmation de dialogueurs. Si l'on dispose déjà de dialogues écrits, certains sites web sont dotés d'une interface graphique intuitive de création de dialogueurs simples par glisser-déposer.

Les dialogueurs sont répandus et très utilisés. Ils permettent d'automatiser des tâches liées à des demandes simples par le truchement d'une interface conversationnelle ; l'automatisation paraît ainsi moins mécanique, plus humaine. Ils sont utilisables dans de nombreux domaines impliquant des **interactions avec la clientèle**. Ils peuvent être utiles dans l'administration en permettant à l'administré

de prendre un rendez-vous, de télécharger un formulaire, de trouver la réponse à une question type, de programmer un rappel pour un rendez-vous ou d'obtenir une information importante, comme les heures d'ouverture. De nombreuses administrations les utilisent déjà pour étoffer leurs services et simplifier leur travail : les villes de Berlin, Bonn et Würzburg en Allemagne par exemple, des services administratifs finlandais (immigration, administration fiscale, bureau des brevets et de l'enregistrement) ou encore le bureau des transports de Londres au Royaume-Uni. La Commission européenne a publié un rapport sur le recours aux dialogueurs dans les services publics (*Architecture for public service chatbots*) pour aider les administrations désireuses de recourir à des dialogueurs dans la fourniture de leurs services[4].

La communication des dialogueurs étant entièrement disponible sous forme écrite, il est possible de la **combiner avec la traduction automatique**. Il serait ainsi aisé d'adapter un dialogueur existant dans une langue régionale ou minoritaire. Le recours aux dialogueurs permettrait aux administrations de faire en sorte que les locuteurs de langues régionales ou minoritaires puissent facilement s'adresser à elles, soumettre une demande écrite et recevoir une réponse écrite dans leur langue.

3.4.2 Recherche intelligente (article 10.2.g)

L'internet se prête à de nombreuses activités liées à la **recherche de toponymes** (réservation de voyages, recherche d'un itinéraire ou prévisions météorologiques, par exemple). Certains sites web permettent de le faire en utilisant des toponymes en langue régionale ou minoritaire. Ils utilisent pour cela des techniques de recherche intelligente. Par exemple, si on tape un toponyme dans une langue régionale ou minoritaire dans le champ de recherche de booking.com, le site procède automatiquement à une recherche croisée sur d'autres moteurs de recherche, comme Google, pour trouver davantage de résultats. Les diverses versions d'un nom de lieu peuvent ainsi être mises en correspondance, grâce aux résultats de recherche d'autres sites web, comme Wikipédia.

La recherche intelligente repose donc sur des méthodes de combinaison de plusieurs moteurs de recherche et de détermination des correspondances des résultats obtenus avec le monde réel : objets, personnes ou lieux, etc. L'IA a fait considérablement progresser ce mode de recherche, qui accepte maintenant les variantes, par exemple orthographiques, et les fautes de frappe. La recherche peut s'appuyer sur :

- **l'extraction de mots-clés** (détermination des phrases clés dans un texte par analyse de la fréquence d'apparition de mots et de leur cooccurrence avec d'autres mots) ;
- **la reconnaissance d'entités** (classification des entités nommées dans un texte en catégories prédéfinies, comme personne, toponymes, expression du temps, etc.) ;
- **les relations entre entités** (lien d'une entité nommée avec la base de connaissances sur les entités du monde réel) ;
- **un robot d'indexation** (collecteur automatique et systématique explorant d'autres sites web, généralement dans le but de les indexer pour connaître le contenu des pages afin d'y récupérer des informations exploitables en cas de besoin et de mettre à jour le contenu ou l'index du contenu du site) ;

[4] Commission européenne, Direction générale de l'informatique et programme ISA2. (2019). *Architecture for public service chatbots*. Consultable en ligne à https://joinup.ec.europa.eu/sites/default/files/news/2019-09/ISA2_Architecture%20for%20public%20service%20chatbots.pdf

- **la similarité sémantique/le plongement lexical** *(word embeddings,* transposition des rapports sémantiques entre mots dans un espace mathématique, comme pour la traduction automatique).

Ces méthodes montrent que les sites web ne fonctionnent pas en vase clos, qu'ils s'appuient les uns sur les autres pour chercher des résultats plus pertinents pour leurs utilisateurs. Cela explique également pourquoi les toponymes en langues régionales ou minoritaires sont reconnus sur certains moteurs de recherche. Il est assez difficile pour les personnes extérieures à l'entreprise concernée, d'influer sur la recherche. Il serait bon d'encourager le recours à la recherche intelligente dans les moteurs de recherche qui ne l'utilisent pas encore.

3.4.3 Synthèse vocale pour l'annonce des noms de rues (article 10.2.g)

La synthèse vocale est la **création de parole artificielle** par concaténation de segments enregistrés dans une base de données. Plus le domaine est spécifique, plus les segments enregistrés peuvent être longs. Ce peuvent être de simples combinaisons de voyelles et de consonnes, ou alors des mots complets, comme pour les annonces horaires. L'apprentissage automatique a fait progresser la synthèse vocale, qui donne à présent de bons résultats, en particulier dans des domaines bien circonscrits comme les annonces de noms de rues.

La synthèse vocale peut faciliter la mise en œuvre de la Charte de différentes manières. Il y aurait les **annonces bilingues dans les transports publics**. L'annonce de l'arrêt suivant est déjà souvent produite par synthèse vocale, ce qui pourrait aisément être complété par les noms de rue en langues régionales ou minoritaires. Une autre possibilité serait d'intégrer les langues régionales ou minoritaires dans les messages qui informent les piétons des moments de traversée et d'attente aux intersections (comme les messages « rouge piéton » et « vert piéton » pour aveugles et malvoyants). Certains de ces messages communiquent aussi d'autres renseignements, comme le nom de la rue ou le sens de traversée du passage clouté. Si c'est sous forme vocale, ils peuvent être formulés en langue régionale ou minoritaire.

La synthèse vocale nécessite un gros volume de données vocales enregistrées ; elle est donc plus intéressante pour des langues régionales ou minoritaires dans lesquelles il existe déjà une telle base de données, par exemple parce qu'elles sont majoritaires dans d'autres pays. Cela encouragerait les échanges transfrontaliers d'études et de recherches en vue du partage des bases de données des pays où la langue est majoritaire avec les pays dans lesquels elle est régionale ou minoritaire. Si les données existent déjà dans une langue, les annonces de noms de rues sont facilement réalisables dans d'autres pays, et peuvent compléter, voire remplacer, la production de plaques en langue régionale ou minoritaire.

3.4.4 Traduction automatique (articles 10.1.a.i-v, 10.1.b, 10.1.c, 10.2.a-f, 10.3.a-c, 10.4.a)

La traduction automatique, on l'a vu, peut être utilisée pour le **raffinage** (gisting) ou la traduction complète de documents, de textes ou de formulaires administratifs et autres pièces officielles. Ces derniers touchant à des domaines bien circonscrits et contenant beaucoup de formulations figées et répétitives, il est en outre possible de développer de nouvelles applications de traduction automatique pour les langues régionales ou minoritaires non encore prises en charge par les grands traducteurs automatiques.

3.5 Les langues régionales ou minoritaires dans les médias

Il est essentiel pour la protection et la promotion des langues régionales ou minoritaires qu'elles soient présentes dans divers types de médias. L'IA peut apporter là son aide par génération automatique de sous-titres (émissions de télévision) et extraction automatique d'informations (journaux). Ces applications peuvent également contribuer à la mise en œuvre de l'article 9.4 de la Convention-cadre.

3.5.1 Génération automatique de sous-titres (articles 11.1.a.i-iii, 11.1.c.i-ii)

La génération automatique de sous-titres est un sous-domaine de la **reconnaissance vocale** et de la **transcription automatique de texte** *(speech to text)*. Elle produit automatiquement des sous-titres à partir d'un document audio, par exemple la bande-son d'une vidéo. Les sous-titres ne transcrivent que les paroles d'un document sonore ; les sous-titres codés y ajoutent par exemple des indications sur la musique ou les bruits de fond. La reconnaissance vocale fait depuis longtemps déjà l'objet de recherches et a beaucoup progressé récemment.

La génération automatique de sous-titres requiert un modèle oral et un modèle écrit. Ces modèles sont généralement entraînés ensemble pour une meilleure précision. La reconnaissance vocale est complexe : elle doit reconnaître les variations d'accent, de prononciation, d'articulation, de hauteur, de volume ou de vitesse, ainsi que les bruits de fond, les échos, etc. Elle fonctionne mieux dans les langues où l'on dispose d'un gros volume de données d'entraînement de locuteurs différents placés dans des situations de communication variées, le logiciel pouvant ainsi apprendre à s'accommoder des chevauchements, des hésitations, etc. de la parole naturelle spontanée.

Plusieurs applications génèrent automatiquement des sous-titres. L'une d'entre elles, gratuite, largement répandue et facile d'utilisation, est l'**algorithme de reconnaissance vocale de Google** utilisé par YouTube. Il prend actuellement en charge neuf langues d'Europe (l'allemand, l'anglais, le français, l'italien, le néerlandais, le portugais, le russe, l'espagnol et le turc). On télécharge une vidéo dans l'éditeur YouTube, puis on active le sous-titrage automatique. La page d'aide de YouTube contient des indications sur la manière de procéder, en fonction du type de vidéo. Une fois les sous-titres générés, on relit ce qui a été reconnu puis on télécharge le fichier texte de sous-titres. Cet algorithme étant auto-apprenant, il s'améliore et intègre constamment de nouvelles langues. Autre outil utile de YouTube : la **synchronisation automatique des sous-titres**. On fournit un document texte contenant toutes les paroles d'une vidéo. L'algorithme trouve ensuite les moments exacts de la vidéo où le texte apparaît. La page d'aide de YouTube contient des instructions détaillées. L'outil fonctionne pour les mêmes langues que la reconnaissance vocale et peut se révéler précieux, la transcription d'une vidéo demandant beaucoup moins de temps que la synchronisation des sous-titres.

La génération automatique de sous-titres en langue régionale ou minoritaire peut contribuer à rendre les émissions de télévision dans ces langues accessibles aux malentendants. Elle rehausse par ailleurs la visibilité des langues régionales ou minoritaires. Et elle peut déboucher sur la traduction automatique d'émissions de télévision. Il serait envisageable, par exemple, de sous-titrer en langue régionale ou minoritaire des émissions importantes, ou de sous-titrer en langue majoritaire des émissions produites en langue régionale ou minoritaire. Plusieurs chaînes travaillent déjà à la génération automatique de sous-titres par IA, comme le réseau de télévision franco-allemand Arte ou la chaîne de télévision catalane Betevé. Là encore s'offrent des possibilités d'échanges transfrontaliers entre chercheurs.

3.5.2 Extraction automatique d'informations (article 11.1.e.i-ii)

L'extraction automatique d'informations consiste à retrouver automatiquement des informations structurées dans des documents non structurés lisibles par une machine, le plus souvent des textes en langage naturel. Elle convient particulièrement bien aux séries de documents construits sur le même modèle, où chaque document présente les choses ou événements de façon identique, mais avec des détails différents. L'extraction commence par un prétraitement, comme l'apposition d'une **étiquette morpho-syntaxique** (verbe, adjectif, préposition, etc.) à chaque mot, ou encore la détermination du radical de chaque verbe du document. Diverses sous-tâches (comme la reconnaissance d'entités nommées ou la résolution de coréférences) permettent ensuite d'extraire l'information demandée.

Le résumé automatique est l'une des grandes applications de l'extraction automatique d'informations. Un ensemble de données est condensé informatiquement pour créer un sous-ensemble contenant les informations les plus utiles de l'original. Cela peut se faire par extraction ou par abstraction. Dans l'extraction, on recherche les expressions clés du document et on extrait les phrases les plus pertinentes. Il s'agit donc d'un classement des phrases. Dans l'abstraction, un algorithme d'autoapprentissage génère d'abord une représentation sémantique interne de l'original, puis une paraphrase raccourcie, précise et fluide. Cette technique, évidemment beaucoup plus délicate sur le plan informatique, a fait l'objet de beaucoup de recherches et fournit de bons résultats.

Il existe plusieurs ressources en ligne de création de résumés automatiques par extraction. Elles aident à publier des condensés d'articles de presse en langue régionale ou minoritaire sur d'autres pages, ou d'articles de presse en langue majoritaire pour les rendre rapidement et plus facilement traduisibles. Ces deux méthodes peuvent contribuer à simplifier la **publication** régulière d'**articles de journaux** en langues régionales ou minoritaires.

3.6 Les langues régionales ou minoritaires dans les activités et équipements culturels

Les langues régionales ou minoritaires constituent un riche patrimoine culturel, historique et identitaire. La préservation et la promotion de leur profil culturel sont un volet important du soutien à leur accorder. Des actions culturelles modernes peuvent améliorer leur image, notamment auprès des jeunes, et jouent un rôle important dans le développement d'une langue vivante. L'importance de l'IA dans ce domaine n'est peut-être pas aussi évidente que dans d'autres, mais les applications de NLP (structuration des données, traduction automatique et génération automatique de sous-titres, par exemple) peuvent s'y révéler utiles. Elles peuvent par ailleurs contribuer à la mise en œuvre de l'article 5.1 de la Convention-cadre.

3.6.1 Structuration des données (articles 12.1.g, 12.1.h)

La structuration des données consiste à structurer ou à regrouper automatiquement de gros volumes de données à l'aide de l'IA. Le **classement de documents** par domaines est l'une des principales applications du traitement du langage naturel.

Plusieurs algorithmes et techniques (Naive Bayes, tf-idf[5], Support Vector Machines, etc.) repèrent les régularités dans les bases de données. Ils rangent chaque document ajouté à la base dans une catégorie, à côté des documents déjà classés.

[5] *Term frequency-inverse document frequency* (fréquence du mot — fréquence inverse de document*).*

L'entraînement des algorithmes se fait par **apprentissage supervisé ou non supervisé**. Dans le premier cas, un réseau neuronal est entraîné sur un ensemble de données étiquetées. Dans le second, le réseau mesure les écarts de contenu entre documents en les plaçant dans un espace mathématique et en calculant leur proximité. Cette méthode, décrite plus en détail dans la section consacrée à la traduction automatique, est appelée **regroupement des données**.

La structuration des données se prête notamment à de gros volumes de données non structurées, mais aussi à l'organisation de données structurées.

Dans le cadre de la mise en œuvre de l'article 12, elle peut servir à **classer des œuvres publiées** en langues régionales ou minoritaires et à **structurer les bases de données terminologiques** de ces langues. Le recours à ces méthodes intelligentes de structuration des données pour la création et la gestion des bases de données de ce type peut aider les institutions culturelles à réduire notablement le temps et les coûts qu'elles leur consacrent. On a le choix entre plusieurs prestataires et didacticiels, en fonction de la taille de la base, de la langue et du sujet, ainsi que du pays concerné. Les projets de structuration de données pourraient par ailleurs fournir des sujets d'étude et de recherche à des étudiants de plusieurs filières d'études en gestion et traitement de données.

3.6.2 Traduction automatique (articles 12.1.a, 12.1.b, 12.1.c)

La traduction automatique convient mal aux œuvres culturelles. Une œuvre littéraire est délicate à traduire, même pour un traducteur professionnel. Et la traduction de certains textes très liés à la langue, comme la poésie, peut se révéler impossible. La traduction automatique d'œuvres littéraires est évidemment très difficile. Elle peut néanmoins avoir son utilité, par exemple pour **des résumés ou des sous-titres**. Les résumés traduits pourraient éveiller l'intérêt pour des œuvres en langues régionales ou minoritaires, les locuteurs d'autres langues étant plus enclins à les lire si l'intrigue est déjà traduite dans ses grandes lignes. Il va sans dire que cette méthode fonctionne également dans l'autre sens, pour faciliter l'accès des locuteurs de langues régionales ou minoritaires à des œuvres littéraires d'autres langues. Les œuvres dont les résumés traduits sont très demandés pourraient en outre faire l'objet d'une traduction humaine, sachant que l'ouvrage intéresse un large public.

3.6.3 Génération automatique de sous-titres (articles 12.1.b, 12.1.c)

L'article 12 mentionne les « activités de sous-titrage » qui, on l'a vu, peuvent être automatisées par reconnaissance vocale. Cela réduit l'investissement en temps et financier à leur consacrer puisqu'il suffit de vérifier les sous-titres, il n'y a plus à les faire intégralement traduire par un humain. Pour des exemples déjà fonctionnels de génération automatique de sous-titres et les options possibles, se reporter ci-dessus à la présentation plus détaillée de la section sur l'article 11.

3.7 Les langues régionales ou minoritaires dans la vie économique et sociale

La pleine fonctionnalité des langues régionales ou minoritaires impose qu'elles soient omniprésentes dans la vie économique et sociale. L'IA peut aider à atteindre ce but par l'analyse des sentiments dans ces langues et, là encore, la traduction automatique.

3.7.1 Analyse des sentiments (article 13.1.c, 13.1.d, 13.2.b)

Dans l'analyse des sentiments, le NLP **classe un document d'information subjective** (comme une critique ou une réponse à une enquête) dans une catégorie d'attitude (positive/satisfaite, neutre, négative/irritée, etc.). Cela permet d'explorer rapidement de gros volumes de données textuelles.

La classification par analyse des sentiments est semblable à la classification des documents décrite ci-dessus à propos de la structuration des données. Le réseau neuronal est d'abord entraîné, avec ou sans supervision, et apprend à associer une entrée (par exemple, une critique de film) à une étiquette (« positif », par exemple). Cet apprentissage peut être entièrement automatique ou s'appuyer sur des règles définies manuellement, comme « génial = positif ». En phase de travail, le réseau reçoit du texte qu'il n'a jamais vu et le transforme en vecteurs de mots. Cette **extraction de caractéristiques** (ou d'éléments) est décrite plus en détail ci-dessus à la section consacrée à la traduction automatique. Ces vecteurs sont ensuite traités par un algorithme de classification, qui attribue une étiquette au document, indiquant par exemple que le tweet « Le film était vraiment génial ! » est un retour positif.

L'analyse des sentiments sert surtout à l'**analyse des réactions des clients**. Développer des logiciels d'analyse des sentiments en langues régionales ou minoritaires permettrait d'inclure leurs locuteurs dans les enquêtes de satisfaction de la clientèle et les consultations sur les moyens d'améliorer les services d'une entreprise.

Les prestataires de services d'analyse des sentiments varient en fonction du pays d'origine et de la langue, ainsi que de la taille et du type d'analyse des sentiments souhaités. Il y aurait là encore des possibilités d'encouragement de recherches et d'études sur les langues régionales ou minoritaires. L'analyse des sentiments est très utilisée et d'une grande actualité, avec de nombreux tutoriels en ligne ; elle figure au programme de toutes les formations initiales au NLP et à l'IA. Compte tenu des données disponibles, le développement d'un tel système est tout à fait envisageable, par exemple pour des équipes pluridisciplinaires d'étudiants en programmation et informatique.

3.7.2 Traduction automatique (articles 13.1.a, 13.1.d, 13.2.a, 13.2.b, 13.2.d, 13.2.e)

Plusieurs dispositions de l'article 13 demandent directement ou indirectement la traduction de documents en langues régionales ou minoritaires : consignes de sécurité et informations sur les droits des consommateurs, contrats, documentation technique, ordres de paiement ou autres documents financiers par exemple. La traduction automatique permet de mettre plus rapidement et à moindres frais ces documents à disposition dans la langue requise. Les documents de ce type couvrant toujours des domaines spécifiques et recourant beaucoup à des formules toutes faites, la traduction automatique est possible, qu'il s'agisse de raffinage ou de traduction automatique complète. Il est également possible de développer de nouvelles applications de traduction automatique pour les langues régionales ou minoritaires non prises en charge par les applications existantes. On a également ici des pistes d'études et de recherches intéressantes. Surtout s'il s'agit de consignes de sécurité, le **contrôle** est toutefois plus important qu'ailleurs, les petites erreurs de traduction que commet encore la traduction automatique pouvant avoir des conséquences graves.

3.8 Les langues régionales ou minoritaires dans les échanges transfrontaliers (articles 7.1.i, 14)

La coopération transfrontalière facilite considérablement la promotion des langues régionales ou minoritaires présentes comme langues officielles ou minoritaires dans d'autres pays, eu égard à la possibilité de reprendre ou d'adapter l'infrastructure existante dans des domaines comme l'éducation (matériel pédagogique, formation des enseignants, etc.) ou les médias.

L'IA, on l'a vu, peut faciliter la mise en œuvre de la Charte en encourageant les **échanges transfrontaliers de recherches et d'études** sur les applications et les jeux de données. Tout cela ayant par nature une dimension internationale, les langues régionales ou minoritaires peuvent bénéficier d'applications NLP d'autres pays, notamment ceux où la langue est majoritaire. Une application de ce

type est aisément adaptable à d'autres langues régionales ou minoritaires, par apport de données dans la langue concernée, dès lors que le code source est mis à la disposition d'autres équipes de recherche. Un dialogueur créé par l'administration d'un pays pour l'un de ses services est facile à transférer dans un autre pays où se parle la même langue régionale ou minoritaire, ou à convertir dans une autre langue régionale ou minoritaire. L'IA non seulement facilite ainsi la mise en œuvre de la Charte, mais favorise et encourage aussi les échanges transnationaux. Elle peut alors contribuer à la mise en œuvre de l'article 18 de la Convention-cadre.

Perspectives

Les applications de l'IA étant très nombreuses et la situation évoluant très rapidement dans ce domaine, le présent rapport ne peut donner qu'un premier aperçu et quelques exemples des façons dont des applications NLP peuvent aider les États parties à mettre en œuvre la Charte européenne des langues régionales ou minoritaires ainsi que, dans une certaine mesure, la Convention-cadre pour la protection des minorités nationales. Le document ne prétend nullement à l'exhaustivité : nombre d'autres ressources, applications et possibilités permettraient de mettre l'IA au service de la mise en œuvre de la Charte. Les pistes ouvertes ici peuvent par ailleurs toutes déboucher sur de nouveaux débats et projets d'étude.

Le présent rapport a montré que l'intelligence artificielle et le traitement du langage naturel offrent plusieurs possibilités nouvelles de protection et de promotion des langues régionales ou minoritaires, ainsi que de recherches à leur sujet. La position dominante actuelle de l'anglais dans l'IA menace toutefois d'« extinction numérique » des langues régionales ou minoritaires. Ce risque appelle une réaction vigoureuse, notamment dans le domaine de la **collecte de données**. Des volumes suffisants de données, comme des corpus parallèles alignés en langue régionale ou minoritaire et en langue officielle ou majoritaire, ouvriraient aux langues régionales ou minoritaires de multiples possibilités pour s'affirmer comme des langues modernes, utiles et vivantes. Il faudrait ainsi commencer par la collecte de données en langage naturel pour presque toutes les activités proposées, et le faire aussi rapidement que possible pour garantir la présence de la langue dans les technologies modernes de traitement du langage naturel.

Les recherches en IA et NLP avancent très rapidement, et toutes les applications présentées ici s'améliorent constamment. Bien des traitements semi-automatiques d'aujourd'hui seront entièrement automatisés dans les années qui viennent, et des techniques actuellement rudimentaires produiront un jour des résultats comparables à ceux des humains. L'aide que l'IA peut déjà apporter aux États parties dans la mise en œuvre la Charte ne fera très probablement que s'amplifier ces prochaines années. Investir dans le NLP appliqué aux langues régionales ou minoritaires permettrait d'en bénéficier aujourd'hui, et bien davantage encore demain.

Prems 063222

FRA

www.coe.int

Le Conseil de l'Europe est la principale organisation de défense des droits de l'homme du continent. Il comprend 46 États membres, dont l'ensemble des membres de l'Union européenne. Tous les États membres du Conseil de l'Europe ont signé la Convention européenne des droits de l'homme, un traité visant à protéger les droits de l'homme, la démocratie et l'État de droit. La Cour européenne des droits de l'homme contrôle la mise en œuvre de la Convention dans les États membres.

COUNCIL OF EUROPE

CONSEIL DE L'EUROPE