LANGUES AUTOCHTONES :

DE ZÉRO À NUMÉRISÉE

UN GUIDE POUR METTRE VOTRE LANGUE EN LIGNE

TRANSLATION COMMONS

2022-2032 | INTERNATIONAL DECADE OF Indigenous Languages

Langues autochtones : De zéro à numérisée
Auteurs : Deborah W. Anderson, Lee Collins, Craig Cornelius, Craig Cummings
Réviseurs et contributeurs : Andrew Owen, Julia Nee, Lawrence Wolf-Sonkin, Anna Luisa Daigneault, Julie Anderson, Daniel Bogre Udell
Conception et commercialisation : Mette Attar, Johanna Behm, Leonidas Papas
Coordination du projet : Ester Perez, Jeannette Stewart

1. INTRODUCTION

Translation Commons est une organisation à but non lucratif et une communauté de bénévoles qui soutient la numérisation des langues et qui met à disposition une orientation, des cours et des ressources pour les professionnels du secteur linguistique.

L'un des principaux programmes de Translation Commons est l'Initiative de numérisation des langues (INL), dont le but est d'offrir des compétences numériques aux communautés linguistiques qui le désirent. Près de 6 000 langues dans le monde disposent d'une présence numérique limitée ou inexistante. L'INL propose à ces communautés une feuille de route qu'elles peuvent suivre pour parvenir à la numérisation de leur langue.

Translation Commons s'est associée à l'UNESCO dans le cadre de son initiative « 2019 : année internationale des langues autochtones », dans le but de ramener les communautés autochtones et la numérisation de leurs langues sur le devant de la scène. Une partie de la mission de l'INL est d'aider les locuteurs de langues autochtones et d'autres langues minoritaires à disposer d'un accès numérique équitable, pour s'assurer que de telles communautés linguistiques sont capables de participer à des activités en ligne au niveau mondial et de bénéficier de tous les avantages des applications informatiques modernes, dans leur langue natale. La création de lignes directrices pour munir ces communautés des outils et des connaissances nécessaires à la numérisation de leurs écritures et à la mise en ligne de leurs langues leur permet de se lancer dans cette entreprise tout en conservant leur autonomie. En plus de cette feuille de route, Translation Commons met à disposition des tutoriels, organise des ateliers et accompagne les communautés dans la numérisation de leur langue en les mettant en relation avec des experts du secteur qui les guident au cours du processus de standardisation.

1. 1 La série Du néant au Web

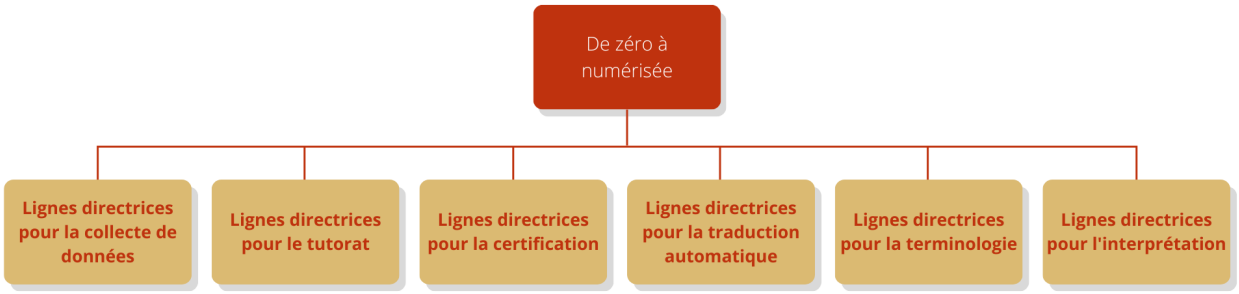
Ce document fait partie d'une série de lignes directrices intitulée Du néant au Web, qui traite des pratiques de numérisation linguistique d'une façon globale. Les auteurs de ces lignes directrices sont des experts en matière de technologies des langues et de linguistique, et ce document s'adresse à toute les communautés linguistiques souhaitant disposer de la capacité d'utiliser leur langue dans un système numérisé.

En matière de communication, la numérisation ouvre de nouvelles voies aux communautés linguistiques. Consultez l'*Annexe sur les avantages de la numérisation de la langue* pour obtenir plus d'informations sur la manière dont la numérisation d'une langue peut présenter des avantages pour les communautés autochtones comme pour le monde en général.

Pour en savoir plus sur le processus de numérisation d'une langue, consultez la page Ressources de Translation Commons, qui offre plus détails sur l'INL et d'autres éléments connexes, parmi lesquels des lignes directrices, des présentations, des vidéos et d'autres types de documentations.

Vous retrouverez dans l'image 1 ci-dessous toutes les lignes directrices créées par Translation Commons pour aider les communautés autochtones.

Image 1 : Série Du néant au Web



Ce document en particulier décrit comment faire en sorte que les logiciels pour mobile et de bureau prennent en charge une langue écrite. La mise en œuvre recommandée permet aux locuteurs natifs de communiquer en ligne, de partager des connaissances et des documents et d'utiliser des logiciels et appareils qui pourraient autrement leur être inaccessibles.

Les publics visés par ce document sont :

- Les communautés autochtones souhaitant rendre leur langue accessible sur les appareils mobiles et ordinateurs
- Les technologues soutenant la numérisation d'une ou plusieurs langues
- Les organisations souhaitant venir en aide aux communautés linguistiques

Ce document a pour but de vous aider à déterminer les outils dont vous avez besoin et comment les utiliser. Il pourrait aussi vous servir dans la recherche des outils disponibles pour utiliser votre langue en ligne.

Lorsqu'une personne demande comment utiliser sa langue sur Internet, il existe plusieurs réponses possibles. Il est utile de comprendre qu'il existe plusieurs niveaux de technologies qui doivent être pris en compte lorsque l'on commence à utiliser une langue en ligne, à la fois sur le Web et sur téléphone portable.

Ce document traite de l'utilisation des langues écrites sur Internet par ses locuteurs, lecteurs et créateurs. Cette utilisation comprend les conversations du quotidien, les SMS, les e-mails, les réseaux sociaux et les blogs. L'objectif est d'aider les personnes à développer leurs sites Internet et une grande variété de contenu, de sorte à favoriser la communication des communautés locales et des diasporas linguistiques n'importe où

dans le monde. Une langue donnée peut disposer de plusieurs écritures ou normes d'écritures. L'objectif n'est pas de dicter comment doivent être utilisées les technologies mobiles mais de permettre aux personnes d'utiliser les technologies décrites ici de manière à renforcer le prestige, l'image publique et la viabilité des langues autochtones à l'échelle mondiale. Bien que les méthodes formelles de documentation, les grammaires et les dictionnaires soient utiles pour l'étude de la langue et sa normalisation, leur utilisation n'est pas un prérequis pour utiliser votre langue en ligne.

Ce partenariat multipartite se compose d'un comité de pilotage responsable de la coordination de la mise en œuvre, de groupes ad hoc chargés de fournir des conseils pertinents et de partenaires contributeurs.

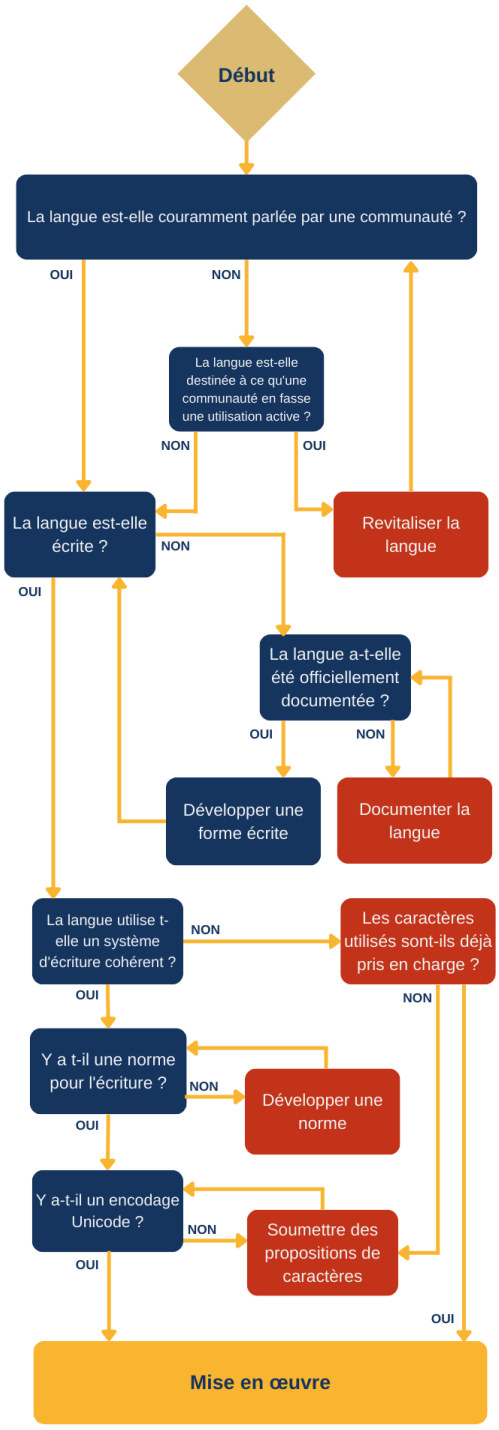
2. APERÇU DES PROCESSUS

Ce document fournit deux processus de travail. Le premier est utilisé pour déterminer le statut actuel de la langue, quand le second sert à développer des solutions technologiques pour l'usage numérique de la langue. Les étapes dans les schémas font référence aux sections du document où vous trouverez des informations plus détaillées. Les étapes sont données à titre indicatif uniquement et certaines peuvent être réalisées simultanément.

2.1. Processus de détermination du statut de la langue

Ce schéma (Image 2) décrit les étapes pour déterminer le statut actuel de la langue destinée à être utilisée en ligne.

Image 2 : Détermination du statut de la langue



2.2. Processus de mise en œuvre de la technologie

Après avoir déterminé le statut de votre langue, suivez ce processus (Image 3) pour commencer à utiliser les technologies disponibles pour mettre votre langue en ligne.

Image 3 : Mise en œuvre de la technologie

3. STATUT DE LA LANGUE

Cette section décrit les étapes du processus lié à la détermination du statut de la langue. L'objectif est de déterminer le niveau actuel de prise en charge de votre langue sur ordinateur et sur appareil mobile. Cette section contient des suggestions quant aux prochaines étapes à suivre pour vous aider à numériser votre langue. Aucune connaissance technique n'est requise pour répondre à ces questions.

3.1. La langue est-elle actuellement utilisée par une communauté ?

Oui : La langue est-elle écrite ?
Non : La langue est-elle actuellement utilisée par une communauté ?

3.2. La langue est-elle destinée à être utilisée activement par une communauté ?

Oui : Revitaliser la langue
Non : La langue est-elle écrite ?

3.2.1. Revitaliser la langue

Même les langues qui ne disposent actuellement d'aucun locuteur peuvent être ravivées grâce à l'engagement et l'organisation des communautés. Cela implique l'utilisation d'importantes ressources pour la documentation de la langue, mais également de supports pour l'enseignement, d'enseignants pour s'en servir et d'un effort important de la communauté pour maintenir ces actions sur plusieurs années, peut-être même sur des
décennies.

L'approche adoptée en Nouvelle Zélande pour la revitalisation de la langue maori (le programme Te reo) était en partie basée sur l'approche du Pays de Galles pour la revitalisation du gallois : celle-ci comprenait entre autres l'utilisation de médias en langue autochtone et d'écoles enseignant la langue. Une approche similaire a été adoptée pour l'irlandais. Voici d'autres exemples des projets entrepris à l'échelle mondiale :

- **L'hébreu** : un exemple de revitalisation linguistique après plusieurs siècles sans que personne ne parle cette langue
- **Le cherokee** : les communautés et les gouvernements consacrent des efforts pour la revitalisation du cherokee avec des cours d'immersion pour les enfants, un enseignement dispensé au niveau primaire et secondaire, ou encore des cours pour adultes.
- **Le mutsun** : la tribu Amah Mutsun en Californie dispose d'un programme de revitalisation linguistique.

- **Le chakma** : une langue du Bangladesh et de l'Est de l'Inde. La communauté a commencé à employer son ancienne écriture pour l'éducation et les efforts d'alphabétisation.
- **Le tunica :** la langue ancestrale de la Tribu Tunica-Bioloxi a été spécifiée comme dormante en 1948 et a depuis été ravivée par des locuteurs qui ont enseigné leur langue d'héritage à de nouveaux locuteurs courants et encouragé des non locuteurs à s'inscrire à des cours d'immersion.
- **Le cornique :** le cornique a été revitalisé pour la première fois au début des années 1900 mais le mouvement a pris plus d'élan dans les années 2000, quand les locuteurs du cornique ont pu utiliser des forums sur Internet pour se retrouver et parler la langue quotidiennement.

3.3. La langue figure-t-elle dans un registre public ?

Si la langue est sur un inscrite sur un registre ou dans une liste publiques de langues.

Oui : La langue est-elle écrite ?
Non : Existe-t-il un code de langue normalisé ?

3.4. La langue est-elle écrite ?

Est-ce que votre langue dispose d'une forme écrite ?

Oui : Documenter la langue
Non : Développer une forme écrite

3.4.1. Développer une forme écrite

Les langues non écrites dépassent du cadre de ce document. Cependant, vous pouvez quand même utiliser votre langue en ligne avec des ressources audio et vidéo. Il existe des organisations qui proposent des conseils et des outils pour les langues principalement orales. Les départements de linguistique des universités peuvent constituer de bonnes ressources.

Les ressources audio et vidéo sont plus utiles lorsqu'elles peuvent être retrouvées par des utilisateurs et des chercheurs. Tous les enregistrements devraient inclure un attribut ou une étiquette de langue pour pouvoir être identifiés par indexation automatique. Utilisez des codes de langue normalisés tels que l'IETF BCP-47.

3.4.2. Documenter la langue

Votre langue a-t-elle été officiellement documentée, avec par exemple une grammaire, un dictionnaire ou une étude linguistique ?

Oui : La langue est documentée
Non : La langue n’est pas documentée

3.4.2.1. La langue est documentée

Si des dictionnaires, grammaires et autres informations linguistiques existent, sont-ils accessibles aux utilisateurs de la langue ? Pensez à des manières de rendre ces informations plus accessibles et utiles pour les communautés concernées. Cela pourrait passer par la création de ressources en ligne, de supports pédagogiques adaptés à l’enseignement en immersion et dans le primaire. Faites en sorte d’attribuer les droits d’auteurs de ces ressources aux communautés linguistiques. Obtenez ou créez des formats numériques de ces informations pour que ces communautés puissent les mettre en ligne comme elles le souhaitent.

Lorsque ces ressources ne sont pas accessibles, prenez des mesures pour partager des livres, des bases de données et d’autres informations sur la langue avec les communautés linguistiques.

3.4.2.2. La langue n’est pas documentée

Si la langue n’est pas encore documentée, il pourrait quand même être possible d’utiliser la langue sous forme orale ou écrite. Cependant, l’assistance technologique sous forme de suggestions d’orthographe, de recherche et de prédictions de texte sera limitée.

3.5. La langue utilise-t-elle un système d’écriture cohérent ?

Un système d’écriture est un ensemble de règles d'utilisation pour une ou plusieurs écritures dans une langue spécifique. Beaucoup de langues disposent de plusieurs formes d’écriture, par exemple le serbo-croate peut être écrit en caractères cyrilliques ou latins en fonction de la communauté. De même, la plupart des systèmes d’écriture sont utilisés dans plus d’une langue. Par exemple, le birman, le shan, le môn et encore d’autres langues peuvent être écrits en écriture birmane.

La langue utilise-t-elle au moins un système d’écriture de manière cohérente, y compris pour l’orthographe ?

Oui : L'écriture est-elle supportée par une norme ?
Non : Les caractères utilisés sont-ils déjà pris en charge ?

3.5.1. Les caractères utilisés sont-ils déjà pris en charge ?

S’il n’existe pas de système d’écriture cohérent, l’écriture peut être informelle, sans orthographe ou grammaire cohérente (ou même utilisant plus d’une écriture). Bien qu’un texte de ce genre puisse être créé et affiché par des outils existants, les outils

d'orthographe et de grammaire ne disposeront que d'une utilité limitée. Lorsque plusieurs systèmes d'écriture sont en concurrence, les personnes extérieures à ces communautés doivent comprendre que des systèmes différents peuvent présenter différents intérêts au sein de la communauté linguistique.

Vous pourriez avoir besoin de l'aide de linguistes et de personnes du milieu technique :

- Beaucoup de communautés utilisent plusieurs orthographes et jeux de caractères pour l'écriture des langues, qui varient aussi entre les dialectes. Les concepteurs d'outils linguistiques tels que des claviers doivent en être conscients et fournir plusieurs jeux de caractères, signes diacritiques et suggestions d'orthographe pour tenter de répondre aux besoins de tous les groupes communautaires.
- La communauté technique devrait penser à des méthodes pour identifier ces variations, les rendre visibles et peut être même développer des moyens pour faciliter la conversion entre les variantes, en fonction des besoins des membres de la communauté.
- Si les caractères utilisés pour écrire sont déjà pris en charge sur les ordinateurs sous une forme ou une autre, la communauté technique peut travailler avec la communauté linguistique pour identifier et concevoir des polices de caractères et des claviers.
- Si le manque de cohérence représente une barrière à l'utilisation de la langue, les enseignants locaux, les décideurs politiques, les leaders de la communauté linguistique, les linguistes, etc., peuvent aider à adopter une forme plus cohérente pour son utilisation sur Internet.

3.6. L'écriture est-elle supportée par une norme ?

Le système d'écriture dispose-t-il d'une orthographe et d'un ensemble de règles grammaticales communément acceptées, qu'elles soient formelles ou informelles ?

Même sans normes, la rédaction et la communication informelle restent possibles. Cependant, une orthographe, une grammaire et un vocabulaire non standardisés ou des variations régionales importantes peuvent rendre difficiles des formes d'utilisation plus poussées, comme pour les réseaux sociaux, le partage de documents et la recherche de sites web et d'informations avec les moteurs de recherche.

Oui : Soumettre des propositions de caractères
Non : Développer une norme

3.6.1. Soumettre des propositions de caractères

Lorsque les différents aspects d'une langue écrite (comme l'orthographe, la grammaire et la ponctuation) sont les mêmes, les membres de la communauté auront bien plus de

facilités à utiliser les services disponibles en lignes et les outils tels que la suggestions de saisie, les prédictions de texte, la recherche en ligne et les outils d'écriture.

La norme Unicode régit les écritures plutôt que les langues. Une écriture, par exemple les caractères latins peut être utilisée pour plus d'une langue, par exemple pour l'anglais, le swahili, l'indonésien, etc. Si les caractères de votre langue sont déjà pris en charge, vous pouvez continuer le processus de mise en œuvre technologique de la langue.

Si le système d'écriture est disponible et utilisé, mais que ses caractères ne figurent pas encore dans la norme Unicode, la démarche recommandée est de normaliser les caractères pour que la langue puisse être utilisée sur les appareils connectés tels que les appareils mobiles, les ordinateurs portables et les ordinateurs de bureau. Même sans normalisation, il est quand même possible pour des utilisateurs de partager des fichiers, du moment qu'ils disposent des mêmes polices de caractères et méthodes de saisie. Cependant, les possibilités des outils et services en ligne seront limitées.

Pour parvenir à la normalisation Unicode de votre langue :

1. Vérifiez si les caractères sont déjà inclus ou pris en charge en codage Unicode.
2. S'il manque des caractères, préparez et soumettez une proposition.

3.6.2. Développer une norme

Même si ce n'est pas une étape obligatoire et que les langues peuvent présenter différentes formes d'écriture, lorsque la formalisation est conforme aux objectifs de la communauté, il est recommandé de créer un ensemble de règles d'orthographe, de ponctuation et de grammaire qui permettent de communiquer des idées sous format numérique. Cela améliore aussi les suggestions de mots et la correction orthographique. En plus, cela permet aux moteurs de recherche et aux autres services et outils en ligne de donner des résultats plus utiles et pertinents.

3.7. Procéder à la mise en œuvre

La prise en charge des caractères est disponible en Unicode. Les caractères du système d'écriture sont trouvés dans les écritures déjà prises en charge. Vous pouvez maintenant passer à la mise en œuvre technologique de la langue.

4. PROCÉDÉ DE MISE EN ŒUVRE TECHNOLOGIQUE DE LA LANGUE

4.1. Remarque sur les technologies appliquées aux textes au sein des systèmes numériques

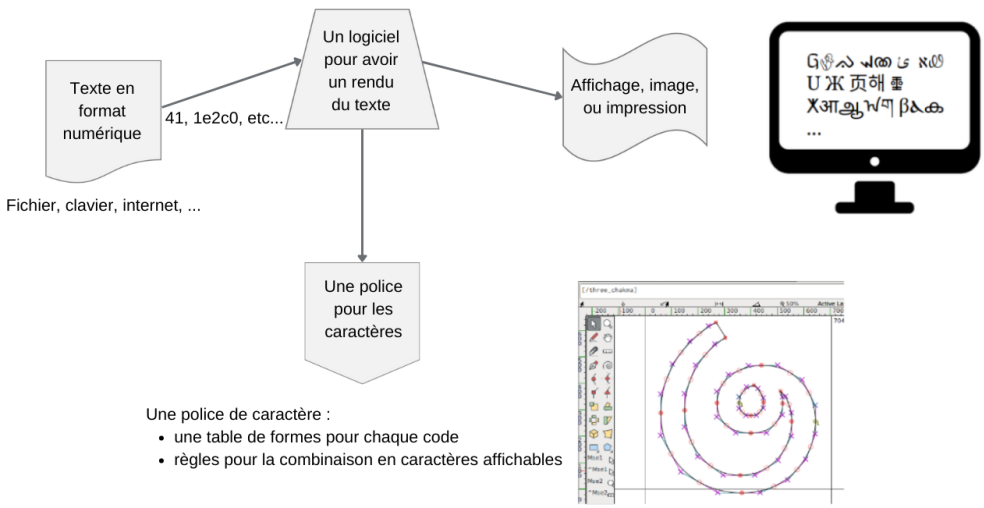
Sur un appareil numérique, le texte est stocké sous forme de séquences de bits servant de *points de code*. Par exemple, le point de code U+0041 pourrait représenter le

caractère « A ». Ce code peut être créé par un clavier, stocké dans un fichier ou transmis à une autre application. Il est affiché ou *rendu* par des systèmes qui comprennent le système de codage utilisé pour les caractères.

Une norme de codage attribue une signification à chaque code. Par exemple, la norme ISO-8859-1 définit les caractères des valeurs 0 à 255, donnant une signification à chaque point de code. L'Unicode est une norme pour les caractères qui attribue une valeur unique à chaque caractère de plus de 150 écritures différentes, parmi lesquelles le latin, le cyrillique, le chinois, l'arabe, l'hébreu, la devanagari, le tamoul, le birman et bien d'autres.

Un texte associé à un codage défini peut être créé avec une méthode de saisie compatible et être affiché dans une police compatible. Ce processus est illustré dans le diagramme ci-dessous (Image 4) qui montre qu'un système comme l'Unicode est capable d'afficher les caractères de nombreuses langues.

Image 4 : le système Unicode est capable d'afficher les caractères de nombreuses langues



Plusieurs systèmes d'écriture utilisent des caractères de base avec des signes diacritiques optionnels ou des lettres modificatives, telles qu'un « e » accent aigu, soit un « é ». Un codage peut par exemple définir que pour produire le caractère combiné, les caractères de base peuvent être suivis du code d'une diacritique. Dans certains cas, un codage peut inclure des caractères précomposés, utilisant des points de code uniques qui comprennent à la fois les points de code de base et combinés.

Toute langue écrite destinée à être utilisée sur des appareils numériques doit donc inclure :

- une norme de codage des caractères, par exemple l'Unicode ;
- des systèmes de rendus et des polices pour les caractères utilisés dans la langue et ;
- des méthodes ou des applications pour obtenir un rendu des points de code sur les supports désirés.

Dans ce document, nous mettons en avant la norme Unicode car elle est utilisée sur tous les appareils numériques modernes. C'est également le codage par défaut pour de nombreuses sources de texte. L'Unicode simplifie une grande partie des efforts requis pour rendre une langue disponible en ligne.

4.2. Définitions pour la mise en œuvre d'une prise en charge technologique

Cette section a pour but de vous aider à mettre en place une prise en charge technologique pour votre langue écrite. Bon nombre de langues utilisent des polices et écritures non standard. Cela n'empêche pas l'utilisation de la langue en ligne mais cela peut limiter le degré de prise en charge disponible. Les étapes suivantes décrivent les aspects de la prise en charge numérique pour les langues écrites.

Les exigences de base pour un système d'écriture numérique sont :

- **Les caractères** : l'identification de l'ensemble des lettres, diacritiques, ligatures, ponctuations, nombres, idéogrammes et autres symboles utilisés dans l'orthographe de la langue.
- **Le code ISO** : pour la langue et pour l'écriture. Peut comprendre des paramètres régionaux optionnels, par exemple « es-MX » pour l'espagnol utilisé au Mexique.
- **Une norme de codage** : qu'il s'agisse d'une norme formelle (Unicode avec collation, tri et combinaisons de caractères) ou d'une norme informelle (codage d'une police).
- **La prise en charge de la police** : les informations nécessaires pour créer une police de caractères comprenant toutes les formes de caractères nécessaires :
    - Les détails sur la mise en page d'un texte, tels que les règles de fusions, de ligatures, de caractères combinatoires (qui peuvent déjà avoir été fournis dans les documents de la norme Unicode)
    - Un échantillon de texte que les concepteurs et développeurs de polices peuvent utiliser pour réaliser des tests.
      Des mises à jour peuvent être requises pour les moteurs de rendu ou d'autres logiciels.
- **La méthode de saisie :** méthode employée pour entrer les caractères sur tous les appareils (généralement à l'aide d'un clavier physique ou tactile sur un écran).

Des ressources supplémentaires pour les langues sont disponibles :

- Le projet Common Locale Data Repository (CLDR, répertoire de données de paramètres régionaux classiques) propose des informations supplémentaires sur les langues, par exemple pour les calendriers.
- Les étiquettes de langue, par exemple, celles de l'IETF BCP-47.
- Des outils pour la segmentation des mots par un logiciel :
    - Des dictionnaires pour les mots écrits sans césure explicite
    - Des règles, par exemple pour les espaces et la ponctuation

4.3. Existe-t-il un code de langue normalisé ?

Est-ce qu'un code de langue normalisé est disponible et approuvé de tous ?

Oui : Une police Unicode est-elle disponible ?
Non : Faire la demande d'un code de langue

4.3.1. Faire la demande d'un code de langue

Si aucun code de langue n'est disponible, vous aurez besoin d'établir un code de langue normalisé avec optionnellement une variante régionale, une écriture ou les deux. Les étiquettes et codes de langue non standard peuvent prêter à confusion et mener à des erreurs d'identification des ressources linguistiques. Suivez les directives des normes ISO-639 et IETF BCP-47.

4.4. Une police Unicode est-elle disponible ?

Existe-t-il une police compatible avec le système Unicode qui soit accessible par les membres de la communauté sur un ordinateur de bureau ou un ordinateur portable ?

Oui : La police est-elle disponible sur les appareils mobiles et ordinateurs ?
Non : Créer une police de caractère

4.4.1. Créer une police de caractère

Quand une écriture est ajoutée au système Unicode, il est possible qu'il n'existe pas encore de polices capables de prendre en charge les caractères. Dans ce cas, vous pouvez consulter des typographes et des experts techniques pour créer des polices qui prennent en charge les textes dans votre écriture.

4.5. La police est-elle disponible sur les appareils ?

La police est-elle disponible sur les téléphones et les autres appareils accessibles par la communauté ? Si c'est le cas, les textes sur les sites web, les réseaux sociaux et les autres applications seront lisibles.

Il est important de noter que la plupart des ordinateurs et appareils mobiles fonctionneront mieux avec des textes Unicode et polices compatibles, en plus de méthodes de saisie de texte Unicode.
Vérifiez si la police fait partie des polices de caractères Noto ou si elle est disponible en téléchargement sur d'autres sites proposant des polices Unicode. Notez aussi que certains appareils mobiles peuvent ne pas prendre directement en charge l'installation ou le téléchargement de polices. Les ordinateurs de bureau comme les ordinateurs portables permettent l'installation de polices téléchargeables.

Vus aurez peut-être besoin de configurer certains logiciels comme les traitements de texte et les navigateurs avant que la police s'affiche dans votre document ou sur la page web.

Les polices web peuvent être utilisées par les sites pour présenter un texte à l'aide de la police fournie par le site web. Cela permet aux créateurs de contenu de préciser la police à utiliser et de rendre le site lisible sur les appareils mobiles et les ordinateurs, même si la police n'est pas déjà installée sur l'appareil. Notez bien qu'une police web fonctionne uniquement dans les pages qui la définissent expressément et qu'une telle utilisation n'installe pas de manière permanente une police sur un appareil.
Les dernières polices Noto sont disponibles en tant que polices web, ce qui est utile si l'appareil ne dispose pas déjà de la bonne version de la police ou si celle-ci ne peut pas être installée.

Oui : L'appareil prend-il en charge la saisie ?
Non : Installer manuellement ou demander aux fabricants sa prise en charge.

4.5.1. Installer manuellement ou demander aux fabricants sa prise en charge

Les systèmes d'écriture récemment normalisés peuvent ne pas encore être proposés sur les appareils mobiles, même lorsqu'une police est disponible. Insistez pour que les fabricants des appareils intègrent les polices requises pour votre langue.

Il est également possible d'installer manuellement une police Unicode sur un appareil mobile. Cette solution peut être efficace mais demande souvent des connaissances spécialisées.

Attention : installer une police sur un appareil mobile à partir d'Internet peut entraîner des problèmes de sécurité et de confidentialité. De même, une telle installation est susceptible d'invalider la garantie de l'appareil ou les droits aux services après-vente.

4.6. L’appareil prend-il en charge la saisie ?

Les utilisateurs d’une langue auront besoin d’une méthode de saisie pour pouvoir écrire des messages, des e-mails, des articles de blog ou d’autres contenus dans cette langue. L'appareil (mobile ou autre) prend-il en charge de façon native une méthode de saisie de la langue ?

Oui : La saisie est-elle prise en charge par des tiers (applications ou appareils) ?
Non : Développer une méthode de saisie

4.7. La saisie est-elle prise en charge par des tiers (applications ou appareils) ?

Les claviers normalisés du système intègrent-ils déjà la prise en charge de la saisie ? Si oui, apprenez à activer ou installer la prise en charge pour le clavier. Cette procédure est spécifique à chaque appareil, mais vous pouvez vous aider de ressources en ligne.

Les fabricants des appareils mobiles proposent beaucoup de claviers existant déjà sur iOS et Android. Par exemple :

- Gboard, le clavier Google pour téléphone portable (https://support.google.com/gboard/answer/6380730?hl=fr&co=GENIE.Platform%3DAndroid) peut être utilisé pour beaucoup de langues, à la fois sur iOS et sur Android.
- Les Outils de saisie Google (https://www.google.com/intl/fr/inputtools/) possèdent des claviers virtuels pour une utilisation sur les pages web dans le navigateur Chrome (uniquement sur ordinateur).
- Plusieurs applications tierces pour la saisie de texte sont disponibles sur l’App Store d’Apple et sur Google Play.

Notez que plusieurs langues possèdent déjà des dispositions de claviers qui sont disponibles publiquement, par exemple dans le référentiel de claviers CLDR.

4.7.1. Développer une méthode de saisie

Il existe beaucoup d’outils pour installer des méthodes de saisie pour de nombreuses langues ainsi que pour développer de nouveaux claviers. En voici quelques exemples :

- Les applications de clavier disponibles dans l’App Store ou Google Play, ou à d’autres endroits
- Keyman : un outil pour claviers multilingues (https://keyman.com/)
- L’application Microsoft Keyboard Layout Creator (https://www.microsoft.com/en-us/download/details.aspx?id=102134)
- L’application Ukelele de SIL (https://scripts.sil.org/cms/scripts/page.php?site_id=nrsi&id=ukelele)

Attention : l'installation d'un clavier ou d'une autre méthode de saisie à partir d'Internet peut entraîner des problèmes de sécurité et de confidentialité.

4.8. L'appareil prend-il en charge les données Unicode ?

Les données Unicode sont-elles disponibles ?

Oui : continuez à utiliser votre langue.
Non : si nécessaire, vous pouvez penser à ajouter les données linguistiques nécessaires au CLDR. Ces informations comprennent entre autres l'étiquette de la langue et son nom. Pour en savoir plus sur le CLDR consultez la section Prise en charge supplémentaire de la langue.

5. PRISE EN CHARGE SUPPLÉMENTAIRE DE LA LANGUE

La normalisation Unicode reprend de nombreux aspects du traitement de texte d'un système d'écriture et d'une langue. Grâce à la prise en charge de la police et de la saisie, plusieurs éléments fonctionneront déjà, parmi lesquels la plupart des fonctions de base d'un traitement de texte, d'un tableur et des e-mails.

Mais ce n'est que le début. Le référentiel de données localisées communes (CLDR, Common Locale Data Repository) fournit « des éléments fondamentaux pour la prise en charge des langues par les logiciels » en collectant des informations utiles pour différentes régions (langues et pays). Ces données peuvent inclure les noms des langues, des pays, des mois, des jours et d'autres informations utiles. Elles permet d'adapter le formatage en fonction des paramètres régionaux de date, d'heure, de nombres et d'autres informations souvent formatées de façon définie.

Même si les données du CLDR ne sont pas essentielles à la communication textuelle de base dans les langues autochtones, ces informations renforcent la fonctionnalité de la langue. Presque tous les outils tels que les e-mails, les SMS, les réseaux sociaux, etc. fonctionneront bien avec des polices et un clavier adaptés.

Ces informations sont utilisées par les programmeurs pour créer des contenus pour les applications en ligne dans des contextes spécifiques, par exemple pour les versions localisées des calendriers, des tableurs, des contenus numériques, des options de menus et dans d'autres contextes d'interface utilisateur.

Le CLDR stocke également des informations supplémentaires pour les langues, telles que les caractères utilisés pour les écrire et les dispositions de clavier pour la saisie de texte. Consultez la section Claviers CLDR pour plus d'informations.

Cependant, pour une prise en charge plus complète des langues sur les applications web et mobiles, des données et outils de traitement de texte supplémentaires

pourraient être nécessaires. Les éléments suivants sont nécessaires pour atteindre la parité entre les langues autochtones et les langues entièrement prises en charge :

- **La segmentation et la césure des mots** : la segmentation du texte est nécessaire pour une mise en page correcte ainsi que pour que l'utilisateur puisse sélectionner des groupements de graphèmes, de mots et de phrases. Beaucoup de langues ne disposent d'aucun signal explicite pour indiquer les séparations entre les mots, par exemple des espaces ou des signes de ponctuation. Dans ces cas, vous aurez besoin de données issues d'un dictionnaire ou d'algorithmes pour fournir ces informations. Pour plus d'informations, consultez :https://unicode-org.github.io/icu/userguide/boundaryanalysis/.
- **Les sauts de ligne** : les langues disposent de nombreuses règles différents concernant l'endroit où le texte peut être interrompu pour un changement de ligne. Par exemple, les citations sont entourées de plusieurs caractères spécifiques à chaque langue, qui suivent différentes règles pour séparer les mots et les phrases. En outre, l'endroit où une ligne de texte peut être coupée dépend des propriétés des caractères Unicode dans le système d'écriture. De la même manière, un nombre et le signe de la devise, par exemple « 10$ », peuvent avoir besoin d'être collés ensemble pour prévenir toute méprise quant à leur signification. Les règles de ponctuation varient à la fois selon la langue et la région. Le saut de ligne est un des cas étudiés dans l'Unicode Boundary Analysis (https://unicode-org.github.io/icu/userguide/boundaryanalysis/#line-break-boundary).
- **L'identification de la langue d'un contenu textuel :** les documents doivent être clairement étiquetés avec un code de langue ou un autre élément identificateur qui décrit le langage humain du document. Lorsque de telles informations sont disponibles dans un document, des outils peuvent les utiliser pour trouver plus efficacement les informations appropriées pour les utilisateurs. Dans un document multilingue, des sections individuelles ou même des paragraphes peuvent être étiquetés avec la langue du texte. L'utilisation d'étiquettes normalisées comme celles de la norme IETF BCP-47 (https://tools.ietf.org/html/bcp47) est particulièrement importante :
    - Le mécanisme d'identification varie selon les applications et l'utilisateur a besoin d'apprendre à le faire. Notez que les étiquettes prises en charge peuvent être limitées et fournies à partir d'une liste.
    - Pour les documents en ligne et les sites web, l'HTML donne l'attribut *lang* (https://www.w3schools.com/tags/att_global_lang.asp) pour étiqueter de manière explicite la langue de tout composant HTML. La valeur de cet attribut doit figurer parmi celles données dans les listes standard d'identifiants de langues, et ne pas faire l'objet d'un choix arbitraire. Par exemple, utilisez rs pour le serbe, de pour l'allemand, zh-Hans ou zh-CN (ou simplement zh) pour le chinois en caractères simplifiés.
- **La détection de la langue** : les services en ligne et les autres applications peuvent souvent donner des résultats plus pertinents et utiles lorsque la langue du texte est connue. Pour les textes qui ne sont pas clairement étiquetés, des détecteurs de

langue tels que cld2 (https://github.com/optimaize/language-detector) ont été conçus. Ces outils font généralement une analyse statistique des caractères du texte et identifient la langue humaine la plus probable. Cette étape est nécessaire car la plupart des systèmes d'écriture sont utilisés pour plusieurs langues, par exemple les lettres latines pour le swahili, le lakota, le warlpiri et le finnois, le cyrillique pour le russe, l'ukrainien et le kazakh, le birman pour la langue birmane, le shan et le môn, etc.

- **Les dictionnaires pour le traitement de texte** : la plupart des applications de traitement de texte prennent en charge les fonctions de base de création, d'édition, de partage et d'impression de texte. Certains outils, comme ceux pour la prédiction de texte, la correction orthographique et les suggestions de grammaire utilisent des listes de mots avec des données de fréquence et d'occurrence, ainsi que des dictionnaires et d'autres données linguistiques. De plus, les synonymes et les expressions fréquemment utilisées sont utiles pour certains outils, comme pour la recherche en ligne.
- **Les chiffres non-ASCII** : plusieurs écritures utilisent des chiffres différents des chiffres occidentaux. C'est le cas du birman, de l'adlam, de l'arabe, et du farsi. Cependant, beaucoup d'applications, par exemple les tableurs, n'interprètent pas ces chiffres comme des valeurs numériques mais comme des valeurs textuelles. Les concepteurs de ces applications prennent parfois en compte les propriétés Unicode de ces caractères pour les traiter comme des chiffres, mais cette prise en charge n'est pas systématique (https://en.wikipedia.org/wiki/Numerals_in_Unicode).
- **Les interfaces utilisateur traduites** : dans certaines applications, en particulier celles avec un contenu pédagogique ou des informations dans la langue concernée, il peut être utile de traduire le texte qui apparaît dans l'interface utilisateur (UI). Par exemple, les éléments d'un menu, tels que « Démarrer » ou « Ouvrir le fichier », pour les fonctions d'un système d'exploitation peuvent être traduits. Mais bien souvent, fournir des traductions pour les petites communautés linguistiques n'est pas faisable pour le propriétaire de l'application. Lorsque l'interface utilisateur est disponible dans au moins une des langues comprises par l'utilisateur, il apparaît comme moins urgent de traduire l'interface.
- **La reconnaissance optique de caractères (RCO) :** beaucoup de langues disposent d'une importante littérature écrite, dans des livres et dans d'autres documents. La RCO peut être utilisée pour convertir le texte de ces documents en format numérique. Des projets de RCO sont disponibles en open source (https://pdf.iskysoft.com/ocr-pdf/open-source-ocr.html) et il est possible de les entraîner à reconnaître de nouveaux systèmes d'écriture. Notez bien qu'un modèle de langue doté de listes de mots fréquents améliore grandement la précision des méthodes de RCO.

5.1. Les ressources linguistiques grand public

Il existe un grand nombre de ressources disponibles, gratuites et sur abonnement, que les communautés peuvent utiliser :

5.1.1. Les ressources linguistiques

- Panlex (https://panlex.org/)
- Wikitongues (https://wikitongues.org/)

5.1.2. Utilitaires

- SIL - Keyman
- Outils de police de caractères
- Outils open source pour les dictionnaires

5.2. Letchnologies linguistiques avancées

Les fonctionnalités suivantes requièrent de grandes quantités de données afin d'entraîner des systèmes utilisant l'apprentissage automatique. Des logiciels libres et en open source commencent à être disponibles mais la plupart des avancées réalisées dans ce domaine se sont cantonnées à la recherche universitaire ou à la R&D des entreprises. Il est peu probable que la majorité de ces fonctionnalités deviennent disponibles pour la plupart des langues dans un proche avenir.

- **La reconnaissance vocale** : C'est la conversion des sons d'un message oral en texte écrit, à travers la reconnaissance des mots d'une personne qui parle. Ce texte peut alors être transcrit dans des documents ou utilisé pour contrôler des applications et des appareils mobiles ou ordinateurs.
- **La synthèse vocale** : C'est la production d'un discours dans une voix naturelle, à partir d'un texte. Elle est donc utile pour les interfaces mains-libres et pour la lecture automatique à partir de sources textuelles.
- **La transcription des sources audio pour la recherche** : C'est important pour la documentation de langue, particulièrement pour les études linguistiques. Certains projets open source commencent à répondre à ce besoin, y compris :
- **La transcription accélérée pour les linguistes** : https://github.com/CoEDL/elpis
- **La traduction automatique** : La traduction d'un texte d'une langue humaine à une autre est l'une des tâches les plus difficiles pour un ordinateur. Les systèmes actuels arrivent à traduire parmi un nombre de langues limité. Toutefois, ces systèmes ne comprennent généralement pas le contexte et ne sont pas au même niveau que la traduction humaine. Cependant, grâce aux nouvelles techniques d'apprentissage automatique, la qualité et la fiabilité de la traduction automatique augmente rapidement pour ceux possédant de grands corpus. La traduction automatique des langues autochtones n'est généralement pas disponible, mais on commence à voir des projets open source et des universités qui s'y efforcent. Par exemple, Apertium (https://www.apertium.org/index.fra.html) est un outil en ligne qui soutient les efforts de traduction automatique des langues non dominantes.

6. GLOSSAIRE

ASCII (American Standard Code for Information Interchange) : Le code américain normalisé pour l'échange d'information ; un code de caractères utilisé pour la

communication électronique

BCP-47 : étiquettes de l'IETF pour l'identification des langues

CLDR (Common Locale Data Repository) : référentiel de données localisées communes ; contient des informations supplémentaires sur les langues

Caractère : plus petit composant de la langue écrite avec une valeur sémantique ; il fait référence à la forme et/ou à la signification abstraite, plutôt qu'à une forme précise.
Réf : Glossaire Unicode

Point de Code : nombre attribué à un caractère ou à un formatage spécifique

Diacritique : caractère ajouté à une lettre ou à un glyphe, généralement pour en modifier le son ou la signification

Langue dormante : langue actuellement sans locuteurs

Police de caractères : ensemble de glyphes utilisés pour la représentation visuelle des données sous forme de caractères

Glyphe : élément d'un ensemble de symboles représentant un caractère lisible utilisé pour l'écriture

Grammaire : règles régissant la composition d'un *langage naturel*

IETF (Internet Engineering Task Force) : groupe de travail pour l'élaboration de normes Internet

Langue autochtone : langue native d'un territoire précis

IYIL2019 (2019 International Year of Indigenous Languages) : Année internationale des langues autochtones 2019

Revitalisation de la langue : Inverser processus d'inversion du déclin d'une langue ou de revitalisation d'une *langue dormante*

Ligature : combinaison d'au moins deux glyphes pour former un glyphe unique

Langue naturelle : langue qui a évolué naturellement chez les humains et qui se distingue des langues formelles, par exemple de celles utilisées par les ordinateurs. Les langues naturelles peuvent être orales, visuelles, visuo-gestuelles (signées) et écrites

Noto : famille de polices comprenant plus de 100 polices individuelles, qui dans leur ensemble sont destinées à servir pour toutes les écritures codées en Unicode (toutes celles comprises actuellement dans la version Unicode 6.0 et les versions précédentes)

Orthographe : ensemble des conventions d'écriture d'une langue

PUA (Private Use Area) : Zone d'utilisation privée ; plage de *points de code* dans la norme Unicode qui ne seront jamais associés à des caractères

Ponctuation : espacements et signes qui n'ont pas de son mais qui aident à la compréhension du texte

Écriture : ensemble des lettres et des autres signes écrits utilisés pour représenter des informations textuelles dans un ou plusieurs systèmes d'écriture

Translation Commons : communauté et plateforme en ligne pour le libre partage des connaissances linguistiques

UNESCO : Organisation des Nations Unies pour l'éducation, la science et la culture.

Unicode : norme la plus couramment utilisée pour le codage des caractères des systèmes d'écritures du monde et leur numérisation

Système d'écriture : ensemble des règles d'utilisation d'une ou de plusieurs écritures pour une langue donnée

7. RÉFÉRENCES

7.1. Revitalisation de la langue

Routledge Handbook of Language Revitalization (Hinton, Huss, et Roche 2018)
The Green Book of Language Revitalization in Practice (Hinton et Hale 2001)
Language Documentation and Revitalization in Latin American Contexts (Perez-Baez, Rogers, et Roses Labrada 2016)
Developing Orthographies for Unwritten Languages (Cahill et Rice 2014)
http://cherokeepreservation.org/what-we-do/cultural-preservation/cherokee-language/
https://language.cherokee.org/
http://amahmutsun.org/language
https://rising.globalvoices.org/blog/2011/11/29/languages-online-activism-to-save-chakma-language/
https://www.languageconservancy.org/programs/indigenous-language-program-support/

7.2. Registres de langues

https://www.ethnologue.com/
https://glottolog.org/

7.3. Codage Unicode et codage de la police de caractères

https://unicode.org/main.html
https://unicode.org/standard/supported.html
https://unicode.org/standard/where/
https://unicode.org/pending/proposals.html

https://unicode.org/glossary/
https://linguistics.berkeley.edu/sei/

Les polices spécialisées peuvent utiliser des approches non standard pour les caractères, telles que les zones d'utilisation privée (PUA) ou d'autres types de caractères (l'ASCII ou l'arabe) avec des glyphes personnalisés pour les points de code. Ces spécifications sont comprises dans le codage de la police. Ces approches non standard permettent aux utilisateurs de voir les caractères qu'ils saisissent même si les autres ne les voient pas, à moins que ces derniers n'utilisent les mêmes polices. Les outils et services en ligne ne seront pas capables non plus d'interpréter le texte avec ces codages de police parce que le codage associé ne contient pas les informations sur la signification voulue pour le caractère.

Même si les textes en codage Unicode disposent d'avantages par rapport aux textes avec un codage de police, il peut être nécessaire d'utiliser des polices spécialisées jusqu'à ce que les caractères d'un système d'écriture soient inclus dans la norme Unicode. Dans ce cas, la police ne devrait utiliser que des valeurs comprises dans les PUA Unicode, sans réutiliser de valeurs déjà réservées à d'autres écritures. Cette précaution évite que les valeurs de code se chevauchent et facilite l'utilisation des écritures existantes. Convertir cette police en codage Unicode (une fois l'écriture normalisée) est relativement facile pour les codes en PUA, à partir du moment ou ces derniers sont utilisés de manière cohérente.

7.4. Codes de langues

https://en.wikipedia.org/wiki/List_of_ISO_639-1_codes

7.5. Polices de caractères

https://www.google.com/get/noto/

Exemples d'outils pour concevoir des polices de caractères :
- FontForge
- FontLab
- Glyphs
- BirdFont (https://birdfont.org/)

L'université de Reading propose un master en conception de polices (typefacedesign.net/) au cours duquel les étudiants peuvent participer à la création d'une nouvelle police Unicode.

Les développeurs de polices commerciales peuvent les inscrire parmi les nouvelles polices Unicode.

8. REMARQUES

Le cadre de ce document se limite aux langues écrites. Les formes de communications peuvent également comprendre les éléments suivants :

- Les émojis
- Les langues orales
- Les langues visuo-gestuelles (signées)
- Les langues visuelles

Les langues dotées de leur propre écriture peuvent aussi être écrites dans d'autres écritures. Par exemple :

- À l'origine, le turc était écrit avec l'écriture arabe mais il est dorénavant écrit dans l'écriture latine.
- Le chinois peut être écrit avec l'écriture latine (pinyin).

Ce document n'aborde pas la question des dialectes mais il vaut la peine de prendre ces derniers en compte lors de la proposition de normes pour les langues.

La recherche d'informations de haute qualité dans une langue présente des exigences spécifiques :

- L'identification de la langue (à partir du texte)
- La segmentation : couper les textes en mots
- La détermination de la racine des mots, dans une langue avec plusieurs variantes : les mots *dentiste* et *dentaire* ont pour racine le mot *dent*.

Même si les langues orales dépassent le cadre de ce document, les ressources suivantes peuvent être utiles :

- Google Earth (https://docs.google.com/forms/d/e/1FAIpQLSdphaDaz33syPoUDyTOTwwkaLWZx90zopUklha4uadfkUKG8A/viewform)
- https://www.blog.google/products/earth/indigenous-speakers-share-their-languages-google-earth/
- https://www.gerlingo.com/
- Xtrans (https://www.ldc.upenn.edu/language-resources/tools/xtrans) est un outil de transcription multi-plateforme, multilingue et multicanal de dernière génération, prenant en charge la transcription et l'annotation manuelle des enregistrements audios.