سلسلة من الصفر إلى الرقمنة

القواعد الارشادية لجمع بيانات اللغة

دليلك لوضع لغتك على الانترنت

TRANSLATION COMMONS

2019 | INTERNATIONAL YEAR OF Indigenous Languages

إرشادات جمع بيانات اللغة

المؤلفون: جولي أندرسون وتكس تكسين
المساهمون: غريغوري سابونوف، جانيت ستيوارت، كيرتي فاشي، ديبي أندرسون
المحررين: أندرو أوين، عقيل آير، شوتو كاتو، بولا سيريلو
الرسومات والتسويق: ليونيداس باباس

نرحب بتعليقاتكم من أجل تحسين إرشاداتنا.
تواصلوا معنا krista@translationcommons.org

جدول المحتويات

1. المقدمة

ترانسلايشن كامنز هو مجتمع تطوعي غير ربحي يقدم الدعم لرقمنة اللغات ويوجه متخصصي اللغات ويوفر دورات وموارد للصناعات اللغوية.

أحد البرامج الأساسية لترانسلايشن كامنز هو مبادرة رقمنة اللغة (LDI) التي تسعى إلى توفير القدرات الرقمية للمجتمعات اللغوية التي ترغب بها. يوجد ما يقارب 6,000 لغة حول العالم بحضور رقمي قليل أو بدون حضور رقمي على الإطلاق. توفر مبادرة رقمنة اللغة خارطة طريق يمكن للمجتمع اتباعها لتحقيق رقمنة لغتهم.

عقدت ترانسلايشن كامنز شراكة مع مبادرة السنة الدولية للغات الشعوب الأصلية في عام 2019 وهي مبادرة قامت بها اليونسكو لتركيز المزيد من الاهتمام على مجتمعات الشعوب الأصليين ورقمنة لغاتهم. يُعَدُّ دعم الوصول الرقمي العادل إلى لغات الشعوب الأصلية ولغات الأقليات الأخرى جزءًا من مهمة مبادرة رقمنة اللغات لضمان أن هذه المجتمعات اللغوية قادرة على المشاركة في الأنشطة العالمية عبر الإنترنت والحصول على جميع مزايا تطبيقات الكمبيوتر الحديثة بلغتهم الأم. إن وضع الإرشادات لتزويد المجتمعات بالأدوات والفهم لرقمنة أبجديتهم وإدخال لغاتهم إلى الإنترنت يمنحهم المعرفة لتسهيل العملية مع الحفاظ على استقلاليتهم. تقدم ترانسلايشن كومنز بالإضافة إلى الإرشادات والتوجيهات دروسا تعليمية وورش عمل كما تساعد المجتمعات في رقمنة اللغة من خلال تعريفهم بالخبراء المختصين الذين يوجهونهم خلال عملية توحيد المعايير.

تُعَدُّ هذه الوثيقة واحدة من سلسلة من الإرشادات بعنوان من الصفر إلى الرقمنة، والتي تتناول ممارسات رقمنة اللغة بشكل شامل. إنَّ مؤلفي هذه الإرشادات خبراء في تكنولوجيا اللغة واللغويات والجمهور المستهدف هو أي مجتمع لغوي يبحث عن كيفية استخدام لغته في الأنظمة الرقمية.

تعمل الرقمنة على توسيع سبل التواصل في المجتمع اللغوي. راجع ملحق د: فوائد رقمنة اللغة للحصول على مزيد من التفاصيل حول فوائد رقمنة اللغة لمجتمعات الشعوب الأصليين والعالم بأسره.

لمعرفة المزيد عن عملية رقمنة اللغة، انظر من الصفر إلى الرقمنة: دليلك لنشر لغتك على الإنترنت توفر صفحة الويب الخاصة بمصادر ترانسلايشن كامنز معلومات إضافية عن مبادرة رقمنة اللغة تتضمن الإرشادات والعروض والفيديوهات ووثائق أخرى.

1.1 عن هذا المستند

أهداف هذا المستند هي:

- لإدراج أنواع بيانات اللغة التي يجب جمعها لأغراض الرقمنة.
- لوصف عملية جمع البيانات.
- لربط أنواع مختلفة من بيانات اللغة مع استخداماتها التكنولوجية.

نحتاج إلى مجموعة متنوعة من الأمثلة على استخدام اللغة لتحقيق رقمنة اللغة. يستخدم اللغويون وخبراء التكنولوجيا هذه الأمثلة للدراسة والتحليل لتصميم القواعد والمكونات لدعم اللغة في الأنظمة الرقمية. تتضمن بعض الأمثلة قائمة الحروف في الأبجدية والطرق المختلفة لكتابة الحروف وقوائم الكلمات في اللغة ومعانيها. بالإضافة إلى ذلك، هناك عدة أنواع من بيانات اللغة المطلوبة لتحقيق دعم رقمي قوي للغة.

تدعم الأنظمة الرقمية أنواعًا عديدة من التطبيقات. لبعض التطبيقات متطلبات بيانات لغة بسيطة للغاية، ولبعضها الأخر متطلبات متقدمة أكثر، فعلى سبيل المثال: قد يتطلب تطبيق المفكرة فقط إلى دعم إدخال نص بسيط وعرضه. أما تطبيق معالجة الكلمات فقد يتطلب تخطيطا معقدا وخيارات طباعة، وتدقيق إملائي ونحوي، وفرز لغوي، ومخطط تفصيلي، وميزات أخرى تعتمد على اللغة. إذا تصف هذه الوثيقة متطلبات البيانات المختلفة لمجموعة من التطبيقات.

2. نظرة عامة على عملية جمع بيانات اللغة

يصف هذا الدليل الإرشادي خطوات جمع عناصر البيانات التي تكون تمثيلية وتعليمية لرقمنة اللغة وإتاحتها للعرض في مستودع رقمي، وهذه الخطوات هي:

- دعوة المصادر الموثوقة للمساهمة في بيانات اللغة.
- الموافقة على المساهمات في بيانات اللغة.
- توثيق المصدر وحقوق الاستخدام والنشر لكل عنصر.
- تحميل المواد غير الرقمية لنُسق رقمي.
- تحميل عناصر بيانات اللغة والشروحات لمستودع بيانات اللغة.
- المراجعة (تصحيح، وشرح، وتصنيف، والتحقق من صحة البيانات)
- النشر (إتاحة العناصر التي جرى مراجعتها والموافقة عليها للاطلاع)

2.1 تحديد مصادر البيانات اللغوية

في البداية يجب على المجتمعات اللغوية تحديد المصادر الموثوقة والممكنة لبيانات اللغة، قد تتواجد تلك المصادر على نسق متعددة فتشمل: المتحدثون الحاليون، والوثائق التاريخية، والأرشيف، والأعمال الفنية وغيرها. من المفيد أيضاً تضمين نماذج لمحادثات عادية يومية (شفهية أو مكتوبة) وليس فقط استخدام رسمي أو تاريخي أو على النمط الأدبي.

2.2 جمع المواد والتنسيق الرقمي:

ومن ثم يلزم جمع المواد الفعلية. إذا لم تكن المواد في نسق رقمي بالفعل، فيجب تحويلها إلى نسق مناسب للتخزين الرقمي والنقل (على سبيل المثال: نص– صورة– صوت– فيديو– إلخ). على سبيل المثال: يمكن تسجيل التواريخ الشفوية كملفات صوتية أو ملفات فيديو أو نسخها إلى نص. ويمكن مسح الأعمال الفنية والكتب والمستندات وحتى قوائم الكلمات المكتوبة بخط اليد ضوئيًا وإرسالها كملفات صور.

- تحديد المواد المناسبة.
- تجميع المواد المكتوبة.
- الحصول على المعلومات اللغوية من المتحدثين الأصليين.
- تسجيل صوتي أو فيديو.
- نسخ السرد الشفوي.
- المسح الضوئي أو تصوير المواد المكتوبة.

يمكن لأعضاء المجتمع البدء في جمع الأنواع المختلفة من بيانات اللغة بأي ترتيب. تشمل أمثلة المواد:

- المراسلات المكتوبة بخط اليد والوثائق المطبوعة.
- النصوص والكتب.
- القواميس الأحادية اللغة والمترجمة والقواعد اللغوية.
- المواقع الشبكية.
- مواقع التواصل الاجتماعي.
- التسجيلات الصوتية والفيديو.
- الأغاني والشعر والعروض.
- التقاليد الشفوية.
- الأعمال الفنية والرسومات والصور.
- معرفة المجتمع واستخدام اللغة من المتحدثين الأحياء.

2.3 الترخيص والملكية والمصدر

عند تجميع بيانات اللغة، يجب التأكد أيضا من الحصول على الحق القانوني في مشاركة المحتوى. فمثلا، إذا حصلت على رسائل أو وثائق شخصية أو تسجيل لخطاب شخص ما قد تحتاج إلى إذن من الكاتب أو المتحدث (أو أصحاب المصلحة الآخرين) قبل إتاحته للاستخدام في المستودع. وقد تحتاج إلى الحصول على إذن كتابي لنشر البيانات، فضلا عن التأكد من أن لهم الحق في منح الإذن. وقد تحتاج إلى استشارة شيوخ المجتمعات المحلية أو المحامين الموكلين لك لتحديد المسائل الخاصة بقبول البيانات ونشرها في المستودع واستخدام البيانات كأساس للرقمنة، فضلا عن الصياغة المناسبة لأي اتفاق للترخيص.

وبالإضافة إلى ذلك، قد تتطلب النهج التقليدية للملكية الفكرية وحقوق التأليف والنشر وسيادة البيانات دراسة متأنية وملائمة لحماية حقوق المجتمع في اللغة والمعرفة والتراث والثقافة. وهناك نظم وطرق مشروعة متعددة لمعالجة قضايا الملكية الفكرية. ومن الضروري العمل مع نماذج الملكية الفكرية وسيادة البيانات التي تشمل آراء مجتمع اللغات. وللاطلاع على المزيد بشأن موضوع سيادة بيانات الشعوب الأصلية، لا سيما فيما يتعلق باللغة، انظر:

- Battiste and Henderson, *"Protecting Indigenous Knowledge and Heritage"*
- Lovett et al., *"Good Practices for Indigenous Data Sovereignty"*
- Christopher Hutton, *"Who Owns Language? Mother Tongues as Intellectual Property and the Conceptualization of Human Linguistic Diversity"*

وبالإضافة إلى الجوانب القانونية لجمع البيانات، من المفيد توثيق مصدر كل عنصر ومساره إلى المستودع. واستخلاص معلومات المصدر هذه مفيد في التصديق على البيانات باعتبارها تنتمي إلى اللغة وسيُنظر فيها أثناء عملية المراجعة.

2.4 الشروحات

ومن المفيد تسجيل معلومات حول كل جزء من البيانات المجمعة عند الإمكان. تتغير اللغة الحية بطبيعة الحال مع مرور الوقت. تستحدث اللغات أيضا اختلافات إقليمية ولهجات. قد تؤثر العوامل الاجتماعية مثل: عمر كل متحدث وطلاقته وجنسه وحالته والظروف (الاحتفالي–الرسمي–غير الرسمي– إلخ.) على أفعال القول (المنطوق والمكتوب).

وتسجيل المعلومات عن المتحدثين والظروف ومتى وأين وقعت أفعال القول وغير ذلك من الشروحات تبني صورة أكثر دقة للّغة.

من أمثلة الشرح:

- تاريخ إنشاء العنصر وموقعه .
- معلومات حول المتحدث أو المؤلف والمتلقي أو الجمهور المقصود (العمر والجنس والمسمى الوظيفي وحالة المتحدث بطلاقة).
- العلاقة أو صلة القرابة بين المتحدثين.
- الظروف.
- الأسلوب الأدبي (النثر– الشعر– الكلمات– الشعائر– إلخ).

2.5 مستودع التخزين

يمكن تحميل الملفات إلى مستودع تخزين حيث يمكن عرض المعلومات ومراجعتها من قبل مجتمع اللغة واللغويين وخبراء التكنولوجيا والأطراف الأخرى التي منحتها المجتمعات حق الوصول.

قد تعتمد خطوات تحميل المحتوى إلى المستودع على التشكيل والتنفيذ المحدد للمستودع الخاص بك. راجع الوثائق والإدارة الخاصة بمستودعكم المحدد للحصول على التفاصيل.

2.6 تحكم وصول المجتمعات المحلية إلى المستودع:

يتحكم أعضاء مجتمع اللغات أو ممثليهم في الوصول إلى المستودع. يقرر مجتمع اللغات أو ممثليهم من يمكنه العمل كمسؤول عن المستودع. ويدير المسؤولون من يمكنه تحميل محتويات المستودع أو تحريرها أو عرضها أو التعامل بها بأي طريقة.

غالبا ما يُدار التحكم في الدخول إلى المستودع من خلال تحديد الأدوار أو ملفات تعريف المستخدم. كل دور يُمنح أو يُحرم له نوع من أنواع حقوق الوصول المختلفة (المسماة التصاريح). وبعد ذلك، يُضاف المستخدمين وتمكينهم من استخدام المستودع ويعهد إليهم بأدوار. يحدد تعيين الدور الخاص بهم امتيازاتهم لأداء وظائف في المستودع، بما في ذلك–على سبيل المثال–حقوق

عرض السجلات أو إنشائها أو تحريرها أو حذفها. يمكن أن تحتوي سجلات المستودع على الكثير من الحقول، بما في ذلك تلك التي تمثل بيانات وصفية عن تقوية إضافة سجل ومن قبل من ومعلومات الموقع وما إلى ذلك. لذلك قد يكون من الضروري تحديد أذونات لعرض أو تحرير كل حقل من هذه الحقول خاصة إذا كانت بعض البيانات تمثل معلومات يجب التحكم في خصوصيتها. يمكن أن يكون هناك أيضًا أذونات إدارية إما لإخفاء سجل أو نشره وإنشاء فئات أو تحريرها لتنظيم السجلات، وستعتمد الأذونات لإدارة المستخدمين (إضافة مستخدم وإزالة مستخدم وتغيير دور المستخدم وما إلى ذلك) وتفاصيل الأدوار والأذونات، إلى أخره على كيفية تكوين المستودع الخاص بك. ملاحظة: يمكن جعل التصاريح محددة لسجلات معينة أو مجموعات فرعية للمستخدم، فمثلا، يمكن منح متخصص في اللغات الهندية أذونات تحرير للسجلات المتعلقة باللغات الهندية ولكن ليس للسجلات بلغات أخرى. وقد يكون لدى مدير فريق معني بلغات النيجر والكونغو إذن بإدارة المستخدمين الذين يعملون مع تلك اللغات دون المستخدمين الذين يعملون بلغات أخرى.

ويمكن استخدام بعض التصاريح لإنشاء عملية فرز أو نوع آخر من سير العمل لمحتويات المستودع. على سبيل المثال، عند رفع المادة لأول مرة إلى المستودع، قد لا تكون مرئية سوى من قبل مجموعة من المراجعين، مما يمنح المجموعة فرصة لطرح الأسئلة والتحقق من صحة المحتويات على النحو المناسب والممثل للغة. وإذا وافق المستعرضون على ذلك، فيمكن نشر المحتوى أو توضيحه على أنه يمكن لعامة المجتمع مشاهدته. يمكن تمييز المحتوى الذي يتطلب سن الرشد لعرضه على هذا النحو. يمكن أيضًا إخفاء المحتويات التي تنتهك اللوائح الإقليمية وتمييزها على هذا النحو. تعتمد هذه الميزات على تفاصيل المستودع الخاص بك. لمزيد من التفاصيل، راجع الوثائق الخاصة بالمستودع الخاص بك.

عادةً ما يمنح مجتمع اللغة إمكانية الوصول إلى:

- أعضاء المجتمع.
- خبراء اللغة.
- خبراء الكمبيوتر.
- الأطراف المهتمة.

يمكن أن يصبح التحكم في الوصول معقدًا ويتطلب مجموعة معقدة من أدوار المستخدم وأذونات الوصول. ومع ذلك، يوضح الجدول 1 أبسط مجموعة من الأدوار والأذونات. تشير خلية جدول فارغة إلى إذن أو امتياز وصول مرفوض للأشخاص في الدور المرتبط. تُخفى السجلات الجديدة أو المحررة من النشر إلى أن يعين المراجع السجل ليكون له إعداد نشر .

الجدول 1 إدارة الوصول إلى المستودع

	الأذونات					
	العرض	التعليق	إنشاء/ تحرير	اخفاء/ نشر	إزالة	تصنيف
الأدوار						
ضيف	ممنوح					
مستخدم معتمد	ممنوح	ممنوح				
عضو مجتمعي	ممنوح	ممنوح	ممنوح			
باحث أو مهني لغوي	ممنوح	ممنوح	ممنوح			
مراجع	ممنوح	ممنوح	ممنوح	ممنوح		ممنوح
مشرف المجتمع	ممنوح	ممنوح	ممنوح	ممنوح	ممنوح	ممنوح

- الضيف هو مستخدم مجهول يرغب في استعراض محتويات المستودع.
- المستخدم المعتمد هو شخص قد يكون قدم وثائق اعتماد تبرر قدرته على المساهمة بمعرفة واحترام.
- عضو المجتمع المحلي هو عضو أصلي أو مقبول في مجتمع اللغات.
- الباحث أو اختصاصي اللغة هو خبير يُدعى لدراسة الموقع أو المساهمة فيه.
- المراجع هو شخص يتمتع بمهارات متقدمة في إدارة مناقشات المحتوى، وهو حساس لمشاعر المجتمع ومتطلباته، وكذلك للقضايا القانونية وغيرها.
- يتمتع مسؤول المجتمع بالتحكم المطلق بالمستخدمين والامتيازات وبيانات المستودع.

2.7 معالجة الأخطاء

حتى مع وجود عملية تدقيق شاملة، فمن الحتمي أن تُدرج الأخطاء في بيانات اللغة. يمكن أن تؤدي الأخطاء المطبعية والأخطاء في النسخ والذاكرة البشرية المعرضة للخطأ وما إلى ذلك إلى حدوث أخطاء في عملية جمع البيانات الخاصة بك. يجب مراجعة البيانات قبل إتاحتها وتوثيقها كممثلة للغة. تسمح المراجعات الدورية المستمرة بتصحيح بيانات اللغة الخاطئة التي ربما أصبحت جزءًا من مجموعتك.

2.8 عملية المراجعة

تعد عملية المراجعة عملية مهمة للتأكد من أن البيانات الموجودة في المستودع موصوفة وصفا صحيحا ومصنفة ومسموح بها وجديرة بالثقة. يمكن أن تسلط المراجعات الضوء على التناقضات المحتملة في بيانات اللغة.

قد تكون عناصر البيانات غير كاملة. فعلى سبيل المثال، قد تكون العناصر مجرد أجزاء من النص. أو قد تكون مبهمة. بعض العناصر لها مساهمات مهمة ولعملية المراجعة اعتبار مفيد.

يمكن أن تأتي البيانات المساهمة في المستودع من مجموعة متنوعة من المصادر. أحيانا يقدم الأفراد ذوو النوايا الحسنة أو المؤذون معلومات تكهنية أو ملفقة.

لم تُدَرّس بعض اللغات أو تُوَثَّق من قبل، فمفرداتهم وعلم الأصوات وقواعدهم اللغوية وما إلى ذلك غير معروفة رسميًا. ستستخدم البيانات التي جُمعت لاشتقاق بنية اللغة ومصطلحاتها وما إلى ذلك. مع المزيد من أمثلة بيانات اللغة التي تمثل مجموعة كبيرة ومتنوعة من السيناريوهات زادت القدرة على رقمنة اللغة بدقة. ومع ذلك إذا كانت أي من البيانات التي جُمعت لا تمثل اللغة حقًا، فإنها تقوض الرقمنة المناسبة. يمكن أن تؤدي الأخطاء أو السهو إلى سلسلة طويلة من الاستنتاجات غير الصحيحة لذلك تقلل عملية المراجعة من احتمال حصول ذلك

لذلك من المهم أن يكون لديك عملية تُقبل فيها المساهمات في المستودع بسهولة ثم تُراجع من قبل المتحدثين الأصليين أو أعضاء المجتمع أو متخصصي اللغة أو خبراء آخرون. يمكن للمراجعين تقييم المواد والتعليق عليها وحتى الطعن في صحتها ودقتها وتفسيرها. ويمكنهم أيضًا توفير سياق إضافي والإبلاغ عن الاستخدام الصحيح وتصحيح الأخطاء. يمكن أن ينتج عن المراجعة تعديلات أو التوصية بالبحث عن أمثلة بيانات إضافية أو محددة. قد تطلب المراجعة أن يقدم المساهم معلومات المصدر أو

الترخيص. قد تؤكد المراجعة أيضًا ما إذا كانت هناك أي انتهاكات قانونية مثل الصور أو الكلام المحظور. يمكن أن تكون عملية المراجعة تكرارية.

حتى تكتمل المراجعة يكون كل عنصر مرئيًا فقط للأفراد في دور المراجع. هذا يضمن أن مستخدمي المستودع الآخرين يرون فقط المعلومات التي فُحصت. إذا كان لدى المراجعين أسئلة حول أحد العناصر فإنهم يتواصلون مع المساهم حتى تحل جميع المشكلات. ثم يمكن منح العنصر حالة النشر وجعله مرئيًا للجميع.

2.9 إنشاء عملية جمع البيانات اللغوية الخاصة بك

تتمثل الخطوة الأولى في تجميع فريق بالمهارات والأدوات اللازمة لكل مهمة من مهام جمع البيانات اللغوية. أنشئ إرشادات وحدد المسؤوليات وخطط لسير العمل الخاص بك لدعم تجميعك لبيانات اللغة. ضع في اعتبارك الأسئلة التالية:

- جمع المواد
  - هل لديك مخطط سير عمل أو عملية لدعوة المواد وقبولها ورقمنتها؟
  - ما هي المصادر المحتملة لبيانات اللغة؟
  - ما أنواع مواد البيانات اللغوية المتاحة لك؟
  - هل لديك طريقة لحفظ مواد بيانات اللغة المادية وأرشفتها؟
  - هل لديك الأدوات والمهارات اللازمة لوضع المواد في شكل رقمي؟
  - هل تعرف حقوق استخدام ونشر بيانات اللغة لمنطقتك؟
- استضافة بيانات اللغة الرقمية في مستودع
  - هل قمت ثبت وهيأت مستودع بيانات اللغة؟
  - هل لديك مهارات تكنولوجيا المعلومات لإدارة المستودع؟
- فحص بيانات اللغة
  - هل شكلت فريق من مراجعي بيانات اللغة المهرة؟

- هل لديك مبادئ توجيهية للتعليق على بيانات اللغة والتدقيق فيها؟
- هل لديك إرشادات لمتطلبات الموافقة على عناصر بيانات اللغة التي ستنشر؟
- هل لديك إرشادات لسؤال المساهمين للحصول على مزيد من المعلومات أو التشكيك باحترام في الملاءمة أو الأصالة أو حقوق النشر أو أي مشكلات أخرى تتعلق بالمواد المساهمة؟
- هل لديك إرشادات لحل الخلافات حول القضايا المتعلقة بالمواد المساهمة؟
- هل لديك مخطط سير عمل أو عملية لتحميل ومراجعة وتعليق وتصحيح بيانات اللغة الرقمية في المستودع؟

➢ إدارة فريقك

- هل يفهم أعضاء فريقك مهامهم ومسؤولياتهم؟

2.10 ملخص القسم

يُعَدّ جمع بيانات اللغة عملية مستمرة ومتكررة. تُضاف عناصر البيانات وتُراجع بمجرد توفرها. مخطط عملية جمع البيانات اللغوية هو:

- دعوة مصادر موثوقة لبيانات اللغة.
- الموافقة على المساهمات في بيانات اللغة.
- وضع المواد في شكل رقمي.
- تحميل المواد إلى المستودع.
- التحقق من حقوق الاستخدام والنشر.
- التعليق.
- المراجعة (تصويب وتعليق ومصادقة).
- النشر.

3. كيف تُستخدم بيانات اللغة في أنظمة الكمبيوتر

تُستخدم بيانات اللغة بعدة طرق لدعم رقمنة اللغة. ذكرها جميعا بالتفصيل يُعَدُّ خارج نطاق هذه الوثيقة..

ومع ذلك، فيما يلي بعض الطرق الأساسية التي يستخدم بها اللغويون وخبراء التكنولوجيا بيانات اللغة لدعم رقمنة اللغة.

3.1 عينات الكتابة

تُستخدم عينات الكتابة في البداية لإنشاء رموز الكتابة (الحروف والأرقام وعلامات التشكيل وعلامات النغمة وعلامات الترقيم والحروف الأخرى) المستخدمة في اللغة. عند تحديد هذه المجموعة من الأحرف، يمكن استخدام عينات الكتابة لاكتشاف كيفية رسم كل حرف يدويًا وإنشاء خطوط بهذه الأحرف. مجموعة الأحرف مطلوبة أيضًا لتحديد تخطيط لوحة المفاتيح أو طريقة الإدخال المستخدمة لإدخال الأحرف في نظام رقمي.

تذكر: غالبًا ما تُرسَم الحروف بطرق متعددة. فكر أنّ الأحرف يمكن تنسيقها مع أو بدون خطوط الزخرفة –مائلة– ولها العديد من الاختلافات. في بعض اللغات، تغير الأحرف شكلها بناءً على موضعها في الكلمة أو اعتمادًا على الحرف الموجود بجانبها.

أيضا، نادرا ما تستخدم بعض الأحرف يمكن استخدامها فقط في مواضع معينة أو قد تكون في الإصدارات القديمة من اللغة. لهذا السبب كلما زاد عدد العينات التي تجمعها، زادت موثوقية رقمنة لغتك واكتمالها وفائدتها.

3.2 الجمل والفقرات

يمكن أن تكشف البيانات التي تحتوي على جمل وفقرات كاملة عن المصطلحات الصوتية أو الإملائية أو المطبعية أو النحوية أو غيرها من المصطلحات اللغوية الضرورية لمعالجة الكلمات. وتشمل الأمثلة:

- القواعد اللغوية.
- الشرطة وفواصل الكلمات والكتابة بالأحرف الكبيرة والتأكيد وعلامات الترقيم.
- التبرير واتجاه الكتابة.

- النُطق.
- أشكال المناداة (على سبيل المثال: التكريم وترتيب الاسم).

يمكن أن تكشف البيانات أيضًا عن اصطلاحات كتابة فريدة وتنسيقات تستخدم لتمثيل التواريخ والأوقات والعصور والأرقام والنسب المئوية وما إلى ذلك.

3.3 قوائم المصطلحات

يمكن استخلاص قوائم المصطلحات من كتابة العينات والتسجيلات الصوتية والبيانات الأخرى. يمكن استخلاص قوائم المصطلحات من عينات الكتابة والتسجيلات الصوتية والبيانات الأخرى. تُستخدم هذه القوائم من الكلمات والعبارات بواسطة المدققين الإملائيين والتصحيح التلقائي والنص التنبؤي والتعرف الضوئي على الحروف وغيرها من الوظائف الرقمية.

3.4 القواميس

توفر القواميس الأحادية اللغة والثنائية اللغة التعريفات وأجزاء الكلام والنطق وأصول اللغة والترجمات وغيرها من المعلومات. يمكن أن تكون هذه المعلومات مفيدة لمعالجة الكلمات والشرطة والتدقيق الإملائي والنحوي والتصحيح التلقائي والترجمة الآلية والجوانب الأخرى للرقمنة.

يمكن أن يكون تصميم القاموس وتنسيقه أمرًا معقدًا. يولي مؤلفو القواميس اهتماما دقيقا على سبيل المثال لمقاربات الكلمات الرئيسية في اللغات متعددة التركيبات، حيث تُبنى الكلمات أو الجمل المعقدة من أجزاء كثيرة. وترتيب الكلمات أبجديا.

تتضمن بعض الموارد المتاحة لإنشاء قواميس للغات السكان الأصليين ما يلي :

- Nick Thieberger, “The lexicography of Indigenous languages in Australia and the Pacific”

- Antonia Cristinoi and François Nemo, "Challenges in endangered language lexicography"
- Paul V. Kroskrity, "Designing a Dictionary for an Endangered Language Community"
- Frawley, Hill, and Munro, *Making Dictionaries: Preserving Indigenous Languages of the Americas*
- Sarah Ogilvie, "Linguistics, Lexicography, and the Revitalization of Endangered Language"

3.5 التسجيلات

يمكن استخدام التسجيلات الصوتية والمرئية لاكتشاف قواعد النطق. تتيح هذه المعلومات إمكانات تحويل النص إلى كلام وتحويل الصوت إلى نص. يُعَدُّ تحويل النص إلى كلام مفيدٍ للأشخاص الذين يعانون من إعاقات في الرؤية أو القراءة ومن يعانون من ضعف الإلمام بالقراءة والكتابة. يتيح التعرف على الكلام إمكانات الأوامر الصوتية وهو مفيد أيضًا للأشخاص ذوي الإعاقة الذين لا يستطيعون الكتابة على لوحة المفاتيح أو الشاشة التي تعمل باللمس.

3.6 بيانات ثنائية اللغة

تخدم البيانات ثنائية اللغة العديد من الأغراض. على سبيل المثال تتيح قواميس الترجمة ومقاطع الفيديو المترجمة والمستندات المترجمة إنشاء قواميس البحث عن الكلمات عبر الإنترنت وأدوات الترجمة الصوتية والترجمة الآلية وأدوات أخرى. كما تكشف مقارنة المصطلحات بين اللغات عن اختلافات دقيقة.

3.7 الصوت

عادة، كلما جُمِعَت المزيد من البيانات، كانت جودة الرقمنة أفضل. تصبح القواعد النحوية وتعريفات الكلمات وما إلى ذلك أكثر دقة ونوعية. يمكن توثيق العبارات والمصطلحات نادرة الاستخدام..

تُستخدم بعض المصطلحات فقط بالاقتران مع مواضيع أو مجالات معينة. تشمل الأمثلة الصحة والزراعة والتنظيم وما إلى ذلك.

كلما زاد حجم البيانات التي يتم جمعها زادت احتمالية تغطية المزيد من المجالات.

أيضًا بعض التطبيقات اللغوية–على سبيل المثال– الترجمة الآلية لا تنجح إلا عندما يكون هناك حجم كبير من بيانات اللغة المتاحة لتدريب نظام الترجمة.

3.8 متنوع

جميع أنواع بيانات اللغة مفيدة لإنشاء دعم قوي لرقمنة لغتك. لا تستبعد مصادر بيانات اللغة التي تبدو قديمة أو غير رسمية أو رسمية للشباب أو غير المتعلمين أو المبالغ فيها (إعلانات) أو غير المعقولة (قصص أسطورية أو تاريخية). سيكون أي مما يلي مفيدًا:

- قصص الأطفال والكتب المصورة.
- المواد التعليمية.
- التواصل والرسائل والملاحظات.
- الوثائق التنظيمية (شهادات الولادة والزواج والوفاة).
- المعاجم (أحادية اللغة وثنائية اللغة)، الكتب.
- الجرائد.
- اللافتات.
- الملصقات.
- كتابة قديمة.
- التواريخ والتقاليد الشفوية.
- الرسومات والأعمال الفنية.
- محادثة يومية (شفوية أو كتابية أو رسمية أو غير رسمية).

3.9 صياغة مصطلحات جديدة

من الطبيعي أن تكون هناك حاجة إلى مصطلحات جديدة للدلالة على الأفكار والاختراعات والأنشطة الجديدة. ويصح هذا بوجه خاص لأن اللغة جاهزة للرقمنة. فعلى سبيل المثال، يجب تحديد المصطلحات المستخدمة في واجهة المستخدم الخاصة بالبرامج والأجهزة أو تكييفها باللغة الأصلية. مصطلحات مثل قائمة وزر وقائمة منسدلة وملف وتحرير ومساعدة وخروج وموافق وإلغاء وانقر وتنزيل، إلخ.

ليس من السهل دائمًا اختيار أو اختراع معادلات اللغة الأصلية. قد لا تكون الترجمات الحرفية هي الخيار الأفضل. على سبيل المثال المصطلح "الصفحة الرئيسية" يترجم في بعض اللغات إلى "صفحة البداية". أيضًا قد يكون للمصطلح استخدامات متعددة ويتطلب ترجمات مختلفة وفقًا للسياق. على سبيل المثال، أحيانًا يكون لكلمة "إلغاء" أن تعني "تراجع". ضع في اعتبارك أن كلمة "برتقال" في اللغة الإنجليزية هي لون وفاكهة، ولكن في لغات الأخرى توجد مصطلحات منفصلة لكل منهما.

قد يحتاج المجتمع اللغوي إلى إنشاء عملية لإنشاء المصطلحات الخاصة برقمنة الكمبيوتر والاتفاق عليها. أنظر إلى قواعد المصطلحات من الصفر إلى الرقمنة للمزيد من المعلومات لكيفية إنشاء مصطلح جديد.

3.10 ملخص القسم

كلما زاد حجم ونطاق بيانات اللغة في مجموعتك، زادت دقة وشمولية رقمنة لغتك. يوضح الجدول التالي بعض الطرق الأساسية لاستخدام بيانات اللغة في أنظمة الكمبيوتر.

الجدول 2. استخدام بيانات اللغة في أنظمة الكمبيوتر

نوع بيانات اللغة	القيمة اللغوية	استخدامات البرمجيات
عينات الكتابة	كشف رموز الكتابة والأعراف	تحديد مجموعة الأحرف والخط وتخطيط لوحة المفاتيح وطريقة الإدخال وما إلى ذلك.
بيانات الجملة والفقرات	كشف المصطلحات الصوتية والإملائية والطبوغرافية والنحوية وغيرها من المصطلحات اللغوية	معالجة الكلمات والخطاب
قوائم المصطلحات	كشف المفردات	المدققات الإملائية والتصحيح التلقائي والنص التنبؤي والتعرف البصري على الحروف
القواميس	قدم تعريفات وأجزاء من الكلام	المدققات الإملائية والنحوية ومعالجة النصوص وتحويل النص إلى الصوت والترجمة الآلية
تسجيلات الصوت والفيديو	الكشف عن قواعد النطق وأنماط الكلام العامية	تحويل النص إلى صوت والتعرف على الكلام
بيانات ثنائية اللغة	يوفر الترجمات والارتباطات عبر اللغات وأنظمة الكتابة	قواميس ثنائية اللغة ومقاطع فيديو مترجمة ومستندات وبرامج مترجمة وأدوات الترجمة الصوتية والترجمة الآلية
مصطلحات مستحدثة.	يحدد المصطلحات التي تحتاجها الأنظمة الرقمية	قوائم البرامج والأوامر والحوارات وما إلى ذلك.

4. الأدوات اللغوية المتخصصة

4.1 مستودع البيانات الخاص بالترميز الموحد (Unicode):

يقدم مستودع البيانات المحلي العام الخاص بالترميز الموحد أدوات ووسائل مفيدة لتدويل البرمجيات وتعريبها لدعم لغات العالم. كما يوحي اسمه، يمثل مستودع البيانات المحلي العام. مجموعة كبيرة من البيانات المحلية. تستخدمه مجموعة واسعة من الشركات لتكييف برامجها لدعم اصطلاحات اللغات المختلفة.

على سبيل المثال، يحتوي مستودع البيانات المحلي العام على ترجمات مقبولة للعديد من أسماء العلَم التي تحتاجها تطبيقات البرامج، مثل الشهور وأيام الأسبوع والبلدان والأقسام الفرعية وأسماء اللغات ووحدات القياس والعملات. يوفر مستودع البيانات المحلي العام أيضًا على رموز تشفير يمكن للبرنامج استخدامها لتنسيق البيانات وفقًا للاتفاقيات المحلية للتاريخ والوقت والأرقام والقياس والعملة وغير ذلك.

إذا لم تكن لغتك ممثلة بالفعل في يحتوي مستودع البيانات المحلي العام ، ففكر في إرسال معلومات لملئها. سيمكن ذلك مطوري تطبيقات البرامج من إضافة لغتك بسهولة أكبر إلى أنظمتهم. يمكن أن تساعدك مراجعة محتويات يحتوي مستودع البيانات المحلي العام أيضًا في تحديد المعلومات والمصطلحات التي تحتاجها الأنظمة الرقمية ويجب صياغتها للغتك.

4.2 الترجمة الآلية

تُعَد الترجمة الآلية من التطبيقات اللغوية التي تزداد أهميتها لأنها يمكن أن تترجم كميات كبيرة من النص بسرعة وكفاءة.

ومع ذلك، لإنشاء أنظمة الترجمة الآلية، يلزم وجود كميات كبيرة من البيانات ثنائية اللغة. من الأفضل أن تكون البيانات محاذية والمجسمات موازية إذ يقرن كل جزء او جملة من اللغة المصدر مع الجزء المطابق في اللغة الهدف. لمزيد من المعلومات حول أهمية الترجمة الآلية ومتطلباتها المحددة، راجع الملحق ج: متطلبات الترجمة الآلية. يوفر نظرة عامة على الترجمة الآلية ومتطلبات البيانات لبناء نظام ترجمة آلية.

4.3 أدوات معالجة اللغة الطبيعية

يقدم هذا القسم موضوع الأدوات التي تساعد في تحليل لغتك وإنشاء بيانات اللغة. مثال على ذلك أداة تقوم بمسح مستند ضوئي وتستخرج قائمة من الكلمات بلغتك. في البداية، يجب عليك تحديد الأدوات التي يمكن أن تساعد في المهام الفورية. قد تتطلب هذه الأدوات بعض التخصيص للعمل مع لغتك.

تعتمد الكثير من تطبيقات البرامج على مكونات معالجة اللغة الطبيعية (NLP) للعمل مع بيانات اللغة. بعد أن تكون قد حققت إنجازًا في تجميع اللغة، قد ترغب في الاستثمار في تدريب نماذج اللغة وتحديث مكتبات البرامج المستخدمة في الصناعة. يمكن أن يؤدي توفير نماذج مُدربة مسبقًا للغتك إلى تسريع اعتماد لغتك من خلال العديد من التطبيقات.

على الرغم من تنوع اللغات، يمكن تجميع العديد من اللغات معًا بواسطة أنماط متشابهة من النحو والقواعد وما إلى ذلك. يمكن أن تساعد الأدوات التي تتعرف على الأنماط التي تستخدمها لغتك في معالجة بيانات اللغة. على سبيل المثال، هناك أدوات يمكنها تحليل النص لأنواع لغات معينة وإنشاء قوائم كلمات. هناك أدوات أكثر تعقيدًا تستخدم نماذج اللغة لمعالجة بيانات اللغة. يمكن أيضًا استخدام بيانات اللغة للتدريب وإنشاء نماذج أكثر دقة للغة تدريجيًا. هناك بعض الأدوات المصممة خصيصًا للعمل مع اللغات غير الموثقة.

هناك الكثير من مواقع الويب مفتوحة المصدر التي تستضيف أدوات البرمجة اللغوية العصبية ونمذجة اللغة. فكر في اختيار أداة تدعم اللغات المشابهة لغتك نحويًا ثم خصصها ودربها على بيانات لغتك. عندما تحقق النجاح، يمكنك إضافة نموذج مدرب لغتك إلى هذه المواقع. يمكن لمطوري البرامج دعم لغتك بسهولة أكبر ، ويمكنك بعد ذلك دعوة شركات البرامج الكبرى لإضافة لغتك إلى مكتباتهم.

تشمل الأدوات ما يلي:

- متجهات الكلمات مثل word2vec (نموذج لتحديد اقتران الكلمات من مجموعة كبيرة من النص).
- نماذج مدربة مثل BERT وGPT (تلتقط الوظائف النحوية والمعاني وارتباطات الكلمات).

- نماذج للتعرف على الكيانات المسماة (NER) (الأسماء والجغرافيا والتواريخ وما إلى ذلك).
- نماذج لوضع علامات على جزء من الكلام (اسم، فعل، إلخ).
- نماذج للتحليل النحوي / التبعية (الموضوع / الكائن / ...، عبارات الفعل والاسم، إلخ).
- "Huggingface's Transformers" (استخراج المعلومات والترجمة ووظائف البرمجة اللغوية العصبية الأخرى).
- مكتبة SpaCy (استخراج المعلومات ووظائف البرمجة اللغوية العصبية الأخرى).
- مجموعة أدوات اللغة الطبيعية (NLTK) (مكتبات معالجة النصوص).

تعتمد الكثير من التطبيقات على مكتبات تعريف اللغة لتكوين نفسها للغة الحالية. لا يمكن لهذه التطبيقات دعم اللغات التي لا تتعرف عليها هذه المكتبات. يمكن للمساهمة بالمعلومات في هذه المكتبات، حتى يتعرفوا على لغتك، أن تسرع من قبول لغتك ودعمها.

الأمثلة تشمل:

- 3CLD من Google (كاشف اللغة المضغوط 3v).
- Facebook's fastText.

4.4 ملخص القسم

يمكن أن تساعد هذه الأدوات في تحليل لغتك ودعم رقمنتها وتعزيز اعتماد لغتك من قبل مطوري البرامج وتسريع ترجمة المواد من وإلى لغتك.

- المستودع البيانات المحلي العام للترميز الموحد.
- الترجمة الآلية.
- أدوات معالجة اللغة الطبيعية.

5. معجم

المصطلح	وصف
حرف	حرف أو رسم بياني أو اشارة أو علامة أو رمز مستخدم في الكتابة.
مجموعة	مجموعة كبيرة أو كاملة من الكتابات.
اللسانيَّات الدايكرانيَّة	يهتم بالطريقة التي يتطور بها شيء ما وخاصة اللغة مع مرور الوقت.
لهجي	الانتماء إلى لهجة لغة أو مرتبطة بها.
رقمنة	حوِّل إلى نموذج رقمي يمكن معالجته بواسطة الكمبيوتر.
تنسيق الملف	طريقة لتخزين المعلومات في ملف الكمبيوتر. تختلف الطريقة باختلاف نوع البيانات التي تُخزن ويمكن تحديدها بشكل عام بواسطة امتداد الملف. (على سبيل المثال،.html لصفحة الويب).
الخط	تمثيل رسومي للنص. مجموعة من رموز الكتابة بتصميم رسومي مماثل.
معجم	قائمة أبجدية بالكلمات وتعريفاتها المتعلقة بموضوع معين.
السكان الأصليين	موطن منطقة معينة.
علم اللغة	ممارسة تجميع القواميس.
البيانات المحلية	معلومات تستخدم لتخصيص واجهات المستخدم للغة منطقة معينة وثقافتها.
وسائل الإعلام	1- وسائل الاتصال الجماهيري (البث والنشر على الانترنت) مجتمعة. 2- أجهزة تخزين البيانات.
MT	الترجمة الآلية.
NER	التعرف على الكيان المحدد.
NLP	معالجة اللغة الطبيعية.
NLTK	مجموعة أدوات اللغة الطبيعية.

الهجاء	مجموعة من الطرق الثابتة لكتابة اللغة. يتضمن معايير التهجئة، الشرطة والكتابة بالأحرف الكبيرة وفواصل الكلمات والتشديد وعلامات الترقيم.
صوتي	تتعلق بالأصوات في لغة معينة أو في اللغات أو بدراسة ذلك.
الدلالة	دلالة أو تتعلق بلغة تتميز بكلمات معقدة تتكون من عدة صيغ، حيث يمكن أن تعمل كلمة واحدة ككلمة كاملة.
مخزن	موقع مركزي تخزن فيها البيانات وإدارتها.
مسح ضوئي	نسخ وتخزين المعلومات في شكل رقمي.
جزء	وحدة منفصلة وذات مغزى من النص أو اللغة المنطوقة. تجزئة النص هي عملية تقسيم النص المكتوب إلى وحدات ذات معنى، مثل الكلمات أو الجمل أو الموضوعات. تجزئة الكلام هي عملية تحديد الحدود بين الكلمات أو المقاطع أو الصوتيات في اللغات الطبيعية المنطوقة.
زخرفة	إسقاط طفيف ينتهي بضربة حرف بحروف معينة.
النص إلى الكلام	تقنية مساعدة تقرأ النص الرقمي بصوت عالٍ للمستخدم.
مطبعي	فن وتقنية ترتيب الكتابة لجعل اللغة المكتوبة مقروءة وجذابة عند عرضها.
الترميز المُوحد	معيار ترميز دولي يدعم النص الرقمي بلغات ونصوص مختلفة. تعين قيمة رقمية فريدة لكل حرف أو رقم أو رمز أو أي حرف آخر ووظائفها باتساق عبر الأنظمة الأساسية والبرامج المختلفة.
رفع	نقل (البيانات) من كمبيوتر إلى آخر، عادةً إلى كمبيوتر أكبر، أو بعيدًا عن المستخدم، أو يعمل كخادم.
URL	محدد موقع المعلومات عنوان صفحة أو معلومات أخرى على الويب.
UTF-8	ترميز الأحرف المستندة إلى الرمز الموحد متغير العرض المستخدمة للاتصال الإلكتروني للنص
صوت إلى نص	برنامج التعرف على الكلام الذي يحول اللغة المنطوقة إلى نص مكتوب.
XLIFF	تنسيق ملف تبادل تعريب XML تنسيق ملف مستند إلى XML لتبادل البيانات القابلة للترجمة.

الملحق أ: مثال على التطبيقات الرقمية

قد تكون هذه الأنشطة المشتركة متاحة بلغتك الأم على الأنظمة الرقمية. ومن الممكن أيضًا إنشاء تطبيقات جديدة خاصة باحتياجات مجتمعك.

التواصل

- إرسال واستقبال الرسائل النصية.
- إرسال واستقبال رسائل البريد الإلكتروني.
- إرسال واستقبال الوسائط (الصور والصوت والفيديو).
- ترجمة النص تلقائيًا والترجمة الآلية.
- تحويل تلقائي للرسائل الصوتية لعرضها كنص والعكس صحيح.

النشر والتوثيق ومعالجة الكلمات

- النشر والوصول إلى المعلومات على المواقع.
- إنشاء المستندات والكتب ووسائل الإعلام الإخبارية واللافتات والملصقات والمواد التعليمية ومشاركتها.
- إنشاء قواميس مطبوعة وعبر الإنترنت.
- مسح المستندات ضوئيًا لتحويلها إلى نص رقمي.
- إنشاء خط للبرنامج النصي الخاص بك.
- شراء وبيع الأشياء عبر الإنترنت.
- توطين مواقع الويب والتطبيقات بلغتك.

- إنشاء تطبيقات اللغة الأم.
- التدقيق الإملائي.
- التدقيق النحوي والتصحيح التلقائي.

واجهات المستخدم ودعم الإعاقة

- التعرف على الكلام (مفيد للأشخاص ذوي الإعاقات الجسدية).
- استخدام الأوامر الصوتية للتحكم في الأجهزة.
- تحويل النص إلى كلام وقارئات الشاشة (مفيدة لضعاف البصر أو استخدام منخفضي القراءة والكتابة).
- تحويل الكلام إلى نص والتعليق في الوقت الفعلي (مفيد لضعاف السمع).

الملحق ب: متطلبات البيانات للتطبيقات التكنولوجية

جدول متطلبات البيانات للتطبيقات التكنولوجية هو دليل مرئي لمساعدتك في تحديد نقطة بداية جيدة في جمع بيانات اللغة فيما يتعلق بأهدافك لتمكين التطبيقات التكنولوجية بلغتك.

يساعد الجدول المجتمعات من زاويتين: يمكن للمجتمع البحث عن تطبيقات الكمبيوتر المطلوبة واكتشاف أنواع البيانات اللغوية اللازمة لدعم إنشائها. أو بناءً على أنواع بيانات اللغة المتاحة أو التي يمكن الحصول عليها، يمكن للمجتمع تحديد التطبيقات التي يمكن تحقيقها على المدى القريب.

يعرض الجدول الكثير من نفس المعلومات الموضحة في هذا المستند ويساعد المستخدمين على الحصول على توقعات واقعية لتحقيق أهداف الرقمنة الخاصة بهم.

جدول 3 متطلبات البيانات للتطبيقات التكنولوجية

أنواع البيانات	كتابة الرموز	المصطلحات	الترجمة	كلام	اصطلاحات اللغة والتخطيط		بيانات اللغة المجمعة
مثال على المحتوى	الأحرف والأرقام وعلامات الترقيم والخطوط وما إلى ذلك.	قوائم المصطلحات، قواميس أحادية اللغة	قواميس ثنائية، قاعدة بيانات لربط الكلمات على الإنترنت (وردنت)، مجاميع	علم الأصوات وقواعد النطق والتسجيلات الصوتية والمرئية	الهجاء، القواعد النحوية، الواصلة، الكتابة بالأحرف الكبيرة، علامات الترقيم، اتجاه الكتابة، التبرير	اصطلاحات الكتابة: التواريخ، الأوقات، العصور، الأرقام، النسب المئوية، إلخ.	جمل، فقرات، مجموعة نصية أحادية اللغة
التطبيقات الرقمية							
التواصل							
إرسال / استقبال الرسائل النصية	x						
إرسال / استقبال رسائل البريد الإلكتروني	x						
ترجمة النص تلقائيًا والترجمة الآلية	x	x	x		x	x	x
أنواع البيانات	كتابة الرموز	المصطلحات	الترجمة	كلام	اصطلاحات اللغة والتخطيط		بيانات اللغة المجمعة
تحويل تلقائي للرسائل الصوتية	x	x		x	x	x	x

لعرضها كنص والعكس صحيح							
النشر والتوثيق ومعالجة الكلمات							
نشر والوصول إلى المعلومات على المواقع	x				x	x	
إنشاء ومشاركة المستندات والكتب ووسائل الإعلام واللافتات والملصقات والمواد التعليمية	x				x	x	
إنشاء قواميس مطبوعة وعبر الإنترنت	x	x			x	x	
مسح المستندات ضوئيًا لتحويلها إلى نص رقمي (OCR)	x						x
إنشاء خط للبرنامج النصي الخاص بك	x				x		x
شراء وبيع الأشياء عبر الإنترنت	x				x	x	

مواقع الويب والتطبيقات المترجمة إلى لغتك	x	x			x	x	
إنشاء تطبيقات اللغة الأم	x				x	x	
التدقيق الإملائي	x	x					
تدقيق نحوي، تصحيح تلقائي	x	x			x	x	
واجهات المستخدم ودعم الإعاقة							
أنواع البيانات	كتابة الرموز	المصطلحات	الترجمة	كلام	اصطلاحات اللغة والتخطيط		بيانات اللغة المجمعة
التعرف على الكلام (مفيد للأشخاص ذوي الإعاقات الجسدية، والتحدث أيضا إلى سيري أو اليكسا)		x		x			
تحويل النص إلى كلام وقارئات الشاشة (مفيدة لضعاف البصر أو استخدام ضعيفي القراءة والكتابة)	x	x		x	x	x	x

نقل الكلام إلى نص وإدراج شرح مكتوب في الوقت الحقيقي (مفيد لضعاف السمع)	x	x		x	x	x	x
مفتاح الرموز	X	تشير علامة الاختيار إلى نوع البيانات الأكثر احتمالاً لدعم إنشاء هذا النوع من التطبيقات.					
ملاحظة	هذا الجدول ليس قائمة كاملة بأنواع البيانات أو التطبيقات.						

ملحق ج: متطلبات بيانات الترجمة الآلية

ما هي الترجمة الآلية؟

الترجمة الآلية (MT) هي استخدام برنامج لترجمة النص أو الكلام من لغة إلى أخرى.

لإنتاج ترجمات عالية الجودة، تُعَد الترجمة الآلية أكثر من مجرد استبدال ميكانيكي كلمة بكلمة. تستخدم الترجمة الآلية خوارزميات متقدمة ويتطلب كميات كبيرة من بيانات اللغة لتكوين نظام إنتاجي.

لا تحقق أنظمة الترجمة الآلية مستويات جودة الترجمة البشرية. لتحسين الجودة، غالبًا ما تخصص أنظمة الترجمة الآلية حسب المجال أو المهنة لتحديد نطاق المحتوى.

الترجمة الآلية مفيدة كأداة لمساعدة المترجمين البشريين ولأغراض معينة يمكن أن تنتج مخرجات يمكن استخدامها كما هي.

بناء أنظمة الترجمة الآلية:

للترجمة بين لغتين يجب بناء نظام الترجمة الآلية خصيصًا لهذا الزوج. يتضمن ذلك اختيار تقنية الترجمة الآلية واستخدام بيانات اللغة لكلتا اللغتين لتكوين النظام.

لا يمكن بناء أنظمة الترجمة الآلية لمجموعات اللغات الجديدة إلا بعد أن يكون لكل لغة أساس رقمي آمن. يجب أن يكون هناك أيضًا تكاثر طبيعي ومتزايد للموارد اللغوية في لغة رقمية حديثة.

هناك الكثير من اللغات التي بها عشرات الملايين من المتحدثين الذين ليس لديهم أنظمة الترجمة الآلية القابلة للاستخدام اليوم، إما بسبب عدم توفر موارد بيانات كافية، أو عدم بذل جهد كافٍ. بينما تستمر تقنية تطوير الترجمة الآلية الأساسية في التحسن،

ومن الأسهل على نحو متزايد بناء أنظمة جديدة ببيانات أقل، يجب أن يُفهم في البداية أن كميات كبيرة من البيانات مطلوبة لإنتاج أنظمة ترجمة آلية جيدة.

البيانات المستخدمة لبناء أنظمة الترجمة الآلية تسمى بيانات التدريب.

في حين أنه من الممكن تطوير نظام أساسي بسرعة، بافتراض توفر بعض بيانات التدريب التأسيسي، يمكن إضافة بيانات إضافية بمرور الوقت إلى نظام الترجمة الآلية الحالي لدفع التحسينات والأداء المستمر.

تتطور أنظمة الترجمة الآلية وتتحسن باستمرار مع دمج البيانات الجديدة والتقنيات الجديدة وردود الفعل التصحيحية المستمرة. يجب التخطيط للتقييم والتحديث الدوريين.

ما فائدة الترجمة الآلية؟

تعد الترجمة الآلية مفيدة في إتاحة كميات كبيرة من المعلومات وموارد المعرفة بسرعة وبتكلفة منخفضة نسبيًا. ولكن يجب أن نفهم أيضًا أن الترجمة الآلية حاليًا لا ترقى إلى مستوى الترجمة البشرية المختصة.

ومع ذلك، فإن الترجمة الآلية سريعة، وغالبًا ما تكون جيدة بما فيه الكفاية، ويمكن نشرها حسب الرغبة للملايين لاستخدامها على شبكة الانترنت بعد إنشاء نظام الترجمة الآلية. يتيح وجود الترجمة الآلية للملايين من الأشخاص الوصول إلى المعلومات التي لن يتمكنوا من الوصول إليها لولا ذلك. في حين أن قصص حوادث الترجمة المكتسبة والترجمات الخاطئة تكثر، أصبح أصبح شديد الوضوح للكثيرين أنه من الضروري تعلم كيفية استخدام وتوسيع قدرات هذه التكنولوجيا بنجاح. بينما من غير المحتمل أن تحل الترجمة الآلية محل البشر في أي تطبيق تكون فيه الجودة أمرًا بالغ الأهمية، إلا أن هناك عددًا متزايدًا من الحالات التي تُظهر أن الترجمة الآلية مناسبة لما يلي:

- محتوى متكرر للغاية.
- المحتوى الذي لن تتم ترجمته بطريقة أخرى.
- محتوى الذي لا يمكنه تحمل تكلفة الترجمة البشرية.

- محتوى عالي القيمة يتغير كل ساعة وكل يوم.
- محتوى معرفي يسهل الانتشار العالمي للمعرفة النقدية.
- المحتوى الذي ينشأ لتسريع التواصل مع العملاء العالميين الذين يفضلون نموذج الخدمة الذاتية.
- المحتوى الذي لا يلزم أن يكون مثاليًا ولكنه مفهوم تقريبًا.

ما نوع البيانات اللازمة لتطوير نظام الترجمة الآلية؟

تعتمد جميع تقنيات تطوير الترجمة الآلية على البيانات، أي أن أجهزة الكمبيوتر تحلل كميات كبيرة من بيانات الترجمة المتراكمة لتتعلم كيفية الترجمة من لغة إلى أخرى. تسمى هذه البيانات اللغوية المستخدمة لتطوير أنظمة الترجمة الآلية ببيانات التدريب.

تكنولوجيا الترجمة الآلية الحالية التي تَنشر على نطاق واسع هي الترجمة الآلية العصبية، وهي تحل ببطء محل الكثير من المنشآت من نهج أقدم يسمى الترجمة الآلية الإحصائية .

كلاهما نهج لتطوير أنظمة الترجمة الآلية باستخدام الأنواع التالية من البيانات:

- نص ثنائي اللغة.
- مسارد الترجمة.
- بيانات أحادية اللغة في اللغة الهدف.
- بيانات أحادية اللغة في لغة المصدر أو اللغات وثيقة الصلة .

ثنائي اللغة المصدر واللغة الهدف:

مجموعات كبيرة من النصوص المترجمة جملة بجملة تسمى المجاميع الموازية. يمكن إنشاء محرك ترجمة أولي بحد أدنى 100000 مقطع مترجم ثنائي اللغة (جمل). يمكن أن يكون مقطع الترجمة جملة كاملة أو مجموعة كلمات تترجم المصطلحات والعبارات المهمة. من الناحية المثالية، يجب أن يكون هناك ما لا يقل عن 1000000 شريحة. بعض أنظمة الترجمة الآلية مبنية

بمليارات القطاعات. بشكل عام، يمكننا القول أن الكميات الأكبر من المقاطع ثنائية اللغة عالية الجودة ستوفر مخرجات الترجمة الآلية عالية الجودة.

الكثير من المجتمعات اللغوية ليس لديها كميات كبيرة من هذه البيانات. غالبًا ما تتطلب مرحلة الحصول على البيانات جهدًا منسقًا وطويل الأمد وتعاونًا بين الوكالات الحكومية والمؤسسات التعليمية والمجتمع ككل. في غضون ذلك، توجد تقنية تسمح باستخدام عدد أقل من المقاطع، مع توفير التغذية الراجعة البشرية تحسينات تدريجية أثناء ترجمة البيانات.

عادةً ما يتم استخراج Corpora المجاميع كمجموعات بيانات تدريبية لخوارزميات الترجمة الآلية من مجموعات كبيرة من مصادر مماثلة، مثل قواعد بيانات المقالات الإخبارية المكتوبة بلغات المصدر والهدف التي تصف أحداثًا مماثلة.

ومع ذلك، قد تكون الأجزاء المستخرجة مشتتة مع إدخال عناصر إضافية في كل مجموعة. يمكن لتقنيات الاستخراج التفريق بين العناصر ثنائية اللغة الممثلة في كل من العناصر الجماعية والعناصر أحادية اللغة الممثلة في مجموعة واحدة فقط لاستخراج أجزاء متوازية أكثر نقاءً من العناصر ثنائية اللغة. تُستخدم المجموعات المماثلة للحصول على المعرفة مباشرة لأغراض الترجمة. من الصعب الحصول على بيانات متوازية عالية الجودة، خاصة بالنسبة للغات قليلة الموارد.

غالبًا ما تكون بيانات التدريب المستخدمة لبناء نظام الترجمة الآلية عبارة عن ذاكرة ترجمة (الترجمة الآلية، أرشيف للترجمات السابقة) أو غيرها من أصول الترجمة القديمة التي جمعت على مدار بعض الوقت. ستحدد هذه المعلومات ما سيتعلمه نظام الترجمة الآلية ترجمته بشكل أفضل. غالبًا ما تكون هناك حدود لحجم البيانات المتاحة. في مثل هذه الحالات، يجب بذل جهود خاصة لتعليم النظام كيفية تعلم المادة التي من المرجح أن تركز على الترجمة. تذكر أن ما تتدرب عليه هو أفضل ما سيترجمه نظامك. وبالتالي، فإن نظام الترجمة الآلية الذي سوف يستخدم لترجمة المحتوى الطبي يدرب بشكل أفضل باستخدام ذاكرة الترجمة الطبية والمسارد.

عندما تُوفر بيانات ثنائية اللغة للتدريب، يجب محاذاة هذه البيانات:

يجب أن يكون المصدر والهدف ترجمات مباشرة لبعضهما البعض. النصوص المترجمة التي هي خلاصات أو مقتطفات أو تعليقات على النص الأصلي غير مناسبة. يجب فحص البيانات بعناية قبل استخدامها، للتأكد من أنها ستكون مفيدة لأغراض التدريب .

تمكّن مسارد وقواميس المصطلحات الرئيسية بترجمة دقيقة أكثر.

وهناك أيضا قيمة طويلة الأجل في تطوير إستراتيجية بيانات وصفية شاملة للبيانات اللغوية المجمعة. في المراحل الأولية من الحصول على البيانات، ينصب التركيز عادةً على العثور على البيانات حيثما أمكن ذلك لتلبية الحاجة إلى كتلة البيانات الهامة اللازمة للبدء. ومع ذلك تصب محركات الترجمة الآلية، يمكن أن تكون هناك فوائد كبيرة في الأداء إذا تم استخدام النوع الصحيح من البيانات لبناء المحركات. وبالتالي، يمكن تحسين نظام الترجمة الآلية للمحتوى المتعلق بالمجال الطبي أو المحتوى المرتبط بتكنولوجيا الكمبيوتر، بدلاً من امتلاك نظام واحد يقوم بكل شيء. غالبًا ما يؤدي هذا التخصص إلى تحسين أداء أنظمة الترجمة الآلية.

تنسيقات البيانات ثنائية اللغة

البيانات ثنائية اللغة لها ثلاث خصائص مهمة لتكون مفيدة كبيانات تدريب على الترجمة الآلية. يجب أن تكون بتنسيق ملف يمكن لأنظمة الترجمة الآلية استيراده. يجب أن تكون بيانات النص بترميز أحرف Unicode UTF-8. ويجب أن تحاذى البيانات المتوازية، بحيث يُقرَن كل جزء من لغة المصدر بجزء مطابق للغة الهدف.

يمكن تسليم البيانات بتنسيقات الملفات التالية ويتم سردها بترتيب تقريبي للأفضلية:

- ذاكرة الترجمة (TMX ، TBX ، XLIFF ، CSV): التنسيق المفضل.
- نص عادي (TXT).
- محتوى موقع الانترنت (HTML).
- منظم (XML).
- مايكروسوفت أوفيس (DOC, DOCX, PPT, PPTX, XLS, XLSX).

- تنسيقات النشر أو ال DTP (TTX, PDF, FrameMaker)
- التعرف البصري على الحروف (JPEG ، PNG ، TIFF) ، (OCR ، الخ).

يمكن بالفعل محاذاة البيانات ومطابقتها بين لغة المصدر والهدف أو تسليمها في أشكالها الأولية، على سبيل المثال، مستندات Microsoft Word أو HTML. إذا لم تتم محاذاة البيانات، فسيكون من الضروري استخدام الأدوات التي ستمكن من محاذاة البيانات بمستويات عالية من الدقة.

بيانات أحادية اللغة

قد تكون الترجمات الأرشيفية والنصوص الثنائية اللغة البيانات الأكثر أهمية لبناء محرك الترجمة الآلية لكن من الضروري أيضا الحصول على بيانات أحادية اللغة جيدة الجودة في اللغة الهدف. تُستخدم هذه البيانات لمعرفة البنية النحوية الصحيحة أثناء الترجمة وتؤثر إحصائيًا على الناتج بأسلوب الكتابة المطلوب. هذا صحيح بشكل خاص عند بناء أنظمة الترجمة الآلية الإحصائية.

يعد الحصول على البيانات أحادية اللغة أسهل بكثير من الحصول على البيانات ثنائية اللغة. قد يكون من المفيد جمع عناوين مواقع الويب التي لها نفس المجال أو الأسلوب النحوي. يمكن استخراج بيانات لغتهم، ويمكن الاستفادة من المعرفة اللغوية الواردة في هذه البيانات.

عادةً ما يكون من الصعب الحصول على بيانات أحادية اللغة للغات السكان الأصليين. بالنسبة للغات الأغلبية، هناك العديد من مصادر بيانات اللغة أحادية اللغة. ومع ذلك ، فإن حالة الاستخدام التي يمكن أن تكون فيها بيانات اللغة أحادية اللغة مهمة هي بناء نظام الترجمة الآلية الطبية. يمكن أن يؤدي تتبع الارتباطات في المواقع الطبية ومواقع ويب ذات الصلة الغنية بهذا المحتوى المحدد إلى تحديد المعلومات اللغوية المهمة لإنشاء المسارد وذاكرات الترجمة.

يجب أن تكون بيانات النص بترميز أحرف Unicode UTF-8. يمكن تسليم البيانات بالتنسيقات التالية ويتم سردها بترتيب تقريبي للأفضلية:

- نص عادي (TXT).

- ذاكرة الترجمة (CSV, XLIFF,TBX,TMX).
- عناوين URL لمحتوى الويب.
- محتوى ويب (HTML).
- مهيكل (XML).
- مايكروسوفت أوفيس (DOC, DOCX, PPT, PPTX, XLS, XLSX).
- تنسيقات النشر DTP أو(TTX,PDF,FrameMaker)

الملحق د: فوائد رقمنة اللغة

نأمل أن تقدم سلسلة إرشادات من الصفر إلى الرقمنة لأي مجتمع لغة مهتم مسارًا واضحًا للرقمنة إذا كانوا يرغبون في الاستمتاع بإمكانيات الكمبيوتر الكاملة بلغتهم الأم.

ستعتمد فوائد تحويل اللغة في شكل رقمي على أهداف مجتمع المتحدثين بها. قد يشمل ذلك الاحتفال بجمال اللغة، والحفاظ على أنظمة المعرفة ونشر القيم وإنشاء تطبيقات ومنتجات ومشاركة القصص والتاريخ وتسهيل الإشراف البيئي وقيادة الفكر وتوسيع التجارة والتعليم والتوظيف والترفيه والصحة والسلامة. تسمح الرقمنة للمجتمع بالاستفادة من مجموعة دائمة التوسع من الأدوات المعتمدة على الكمبيوتر لصيانة اللغة وتنشيطها وتعليمها. قد يساعد الوجود الرقمي القوي على الإنترنت، وبالتالي زيادة الرؤية، في التأثير على السياسات الحكومية لدعم مجتمعات السكان الأصليين وتوجيه الشركات نحو الاندماج. نظرًا لوجود الهواتف الذكية في كل مكان بين الشباب، قد تكون الرقمنة وسيلة طبيعية لإشراكهم في لغتهم الأم. يمكن أن يؤدي التعرض الأكبر عبر المنصات الرقمية إلى المزيد من الفرص والطلب على المتحدثين الأصليين.

عندما تأخذ المجتمعات اللغوية مكانها على المسرح العالمي عبر المنصات الرقمية، فإنها تفيد العالم الأوسع. ستشكل خبراتهم ومعرفتهم ووجهات نظرهم الفريدة للعالم مساهمة كبيرة في بقية العالم، وقد يؤدي التآزر الناتج إلى تقديم حلول جديدة لمشاكل العالم. تسهل الرقمنة الحفاظ على هذه المعلومات ونشرها، مما يجعل اتساع وطبيعة اللغة البشرية متاحة للعالم ويتم الحفاظ عليها لصالح البشرية بطرق لا يمكننا التنبؤ بها حتى الآن.

للتلخيص، بعض الفوائد المكتسبة من رقمنة اللغة هي:

- تمكّين المتحدثين أحاديي اللغة من الوصول بسهولة إلى محتوى اللغة الأم وإنشائه وتبادله، بما في ذلك عبر المسافات الطويلة، وللأفراد أو المجموعات الكبيرة.
- زيادة الوصول إلى المعلومات الطبية والرعاية الصحية.

- دعم اتصالات الطوارئ والكوارث المنقذة للحياة.
- توسيع نطاق التجارة المحلية والإلكترونية.
- إنشاء طرق جديدة لمشاركة الفن وقيادة الفكر والفلسفات.
- تطوير مواد تعليمية باللغة الأم.
- تحسين العلاقات والتواصل مع الجيران.
- تحسين تسوية المنازعات.
- تمكين المناصرة والوصول إلى الإجراءات القانونية والحكومية باللغة الأم.
- زيادة الوصول إلى المعلومات على الإنترنت من أجل التعليم والتجارة والمشاركة، سواء باللغة الأم أو باللغات الأخرى مع تطور أدوات الترجمة.
- جعل الآخرين يعرفون اللغة الأم والثقافة والحكمة.
- توسيع دور وإبراز مجتمعات السكان الأصليين على الصعيد العالمي.
- تمكين الفئات المهمشة أو الأقلية من الحفاظ على لغتهم أو تنشيطها على الرغم من غمرها من قبل الجماعات المهيمنة.
- الحفاظ على أنظمة المعرفة والثقافة والتاريخ والفن والطب والحكمة والقيم والنظرة العالمية.