ZERO TO DIGITAL SERIES:

LANGUAGE DATA GATHERING GUIDELINES

A GUIDE TO BRING YOUR LANGUAGE ONLINE

TRANSLATION COMMONS

2019 | INTERNATIONAL YEAR OF Indigenous Languages

Language Data Gathering Guidelines

Authors: Julie Anderson and Tex Texin
Contributors: Grigory Sapunov, Jeannette Stewart, Kirti Vashee, Debbie Anderson
Editors: Andrew Owen, Akil Iyer, Shuto Kato, Paula Cirilo
Graphics and Marketing: Leonidas Pappas

We welcome feedback to improve our guidelines.
Contact us at krista@translationcommons.org

Table of Contents

1. INTRODUCTION

Translation Commons is a nonprofit volunteer community that supports the digitization of languages, mentors language professionals, and provides courses and resources for the language industries.

One of the principal programs at Translation Commons is the Language Digitization Initiative (LDI), which seeks to bring digital capabilities to the language communities that desire them. Nearly 6,000 languages throughout the world have a small or nonexistent digital presence. The LDI provides a roadmap that a community can follow to achieve digitization of their language.

Translation Commons partnered with the 2019 International Year of Indigenous Languages, a UNESCO initiative, to focus more attention on Indigenous communities and the digitization of their languages. Supporting equitable digital access to Indigenous and other minority languages is part of the LDI's mission to ensure that such language communities are able to participate in global online activities and have all the benefits of modern computer applications in their native language. Creating guidelines to equip communities with the tools and understanding to digitize their scripts and bring their languages to the internet gives them the knowledge to facilitate the process while maintaining their autonomy. In addition to guidelines, Translation Commons provides tutorials and workshops and assists communities with language digitization by introducing them to industry experts that guide them through the standardization process.

This document is one in a series of guidelines entitled *Zero to Digital*, which holistically addresses language digitization practices. The authors of the guidelines are experts in language technology and linguistics. The intended audience is any language community that wants the capability to use their language on digital systems.

Digitization expands a language community's avenues for communication. See the *Benefits of Language Digitization Appendix* for more detail on how digitizing a language benefits both Indigenous communities and the world at large.

To learn more about the language digitization process, see Zero to Digital: A Guide to Bring Your Language Online. The Translation Commons Resources web page provides additional LDI and related information, including guidelines, presentations, videos, and other documents.

1.1 About this document

The aims of this document are:

- To list the types of language data that must be collected for digitization purposes.
- To describe a process for gathering the data.
- To correlate the various kinds of language data with their technological uses.

A wide variety of examples of language usage is required to achieve the digitization of a language. Linguists and technology experts use these examples for study and analysis to design rules and components to support the language on digital systems. Some examples include the list of characters in an alphabet, different ways the characters are written, and lists of words in the language and their meanings. In addition to these, many other types of language data are required to achieve robust digital support for a language.

Digital systems support many kinds of applications. Some applications have very simple language data requirements. Others have more advanced requirements. For example, a notepad application may only need to support simple text entry and display. A word processing application may offer complex layout and typography options, spelling and grammar checking, linguistic sorting, outlining, and other features that are language-dependent. This document will describe the various data requirements for a range of applications.

2. PROCESS OVERVIEW FOR LANGUAGE DATA GATHERING

This guideline describes the steps for gathering data items that are representative and instructional for digitizing a language and making them available for viewing in a digital repository. These steps are:

- Invite credible sources of language data to contribute.
- Accept language data contributions.
- Document source and rights to use and publish each item.
- Convert non-digital materials into digital format.
- Upload language data items and annotations to a language data repository.
- Review (correct, annotate, categorize, and authenticate).
- Publish (make reviewed and approved items available for viewing).

2.1. Identifying Language Data Sources

First, language communities should identify potential credible sources of language data. These can exist in a variety of forms including living speakers, historic documents, archives, artwork, etc. It is beneficial to also include examples of present-day casual conversation (oral or written), not just formal, historic, or literary-style usage.

2.2. Material Collection and Digital Formatting

The actual materials then need to be gathered. If the materials are not already in digital form, they must be converted into a suitable form for digital storage and transmission (for example, text, image, audio, video, etc.). For example, oral histories can be recorded as audio or video files or transcribed to text. Artwork, books, documents, and even handwritten word lists can be scanned and submitted as image files.

- Identify suitable materials.
- Gather written materials.

- Elicit linguistic knowledge from native speakers.
- Record audio or video.
- Transcribe oral accounts.
- Scan or photograph written material.

Community members can start gathering the various types of language data in any order. The important thing is to start the process of data gathering, annotation, and repository storage. Example materials include:

- Handwritten correspondence and printed documents.
- Texts and books.
- Monolingual and translation dictionaries and grammars.
- Websites.
- Social media.
- Audio and video recordings.
- Songs, poetry, and performances.
- Oral traditions.
- Artwork, drawings, and photographs.
- Community knowledge and language use from living speakers.

2.3. Licensing, Ownership, and Provenance

As you gather language data, you must ensure that you also gain the legal right to share the content. For example, if you acquire personal letters or documents, or a recording of someone's speech, you may need the permission of the author or speaker (or other stakeholders) before making it available for use in the repository. You may need to get written permission to publish the data, as well as assurance that they have the right to grant permission. You may need to consult community elders or lawyers within your jurisdiction to identify the issues with accepting data, publishing it in the repository, and using the data as a basis for digitization, as well as the appropriate wording for any permission agreements.

Additionally, traditional approaches to intellectual property, copyright, and data sovereignty may require careful consideration and adaptation to protect the community's

rights to its language, knowledge, heritage, and culture. There are multiple legitimate systems and ways of addressing intellectual property issues. It is necessary to engage with the models of intellectual property and data sovereignty that include the views of the language community. For further reading on the topic of Indigenous data sovereignty, especially as it relates to language, see:

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

In addition to the legal aspects of data gathering, it is useful to document the origins of each item and its trail to the repository. Capturing this provenance information is useful in certifying the data as belonging to the language and it will be considered in the review process.

2.4. Annotations

It is useful to record information about each piece of collected data when possible. A living language naturally changes over time. Languages also develop regional and dialectal differences. Social factors such as each speaker's age, fluency, gender, or status, and the circumstance (ceremonial, formal, informal, etc.) may affect speech acts (spoken or written).

Recording information about the speakers, circumstances, when and where speech acts occurred, and other annotations builds a more accurate picture of the language.

Annotation examples include:

- The origination date and location of the item.
- Information about the speaker or author and the receiver or intended audience (age, gender, title, and native/fluent speaker status).
- Relationship or relative status between speakers.

- Circumstance.
- Literary style (prose, poetry, lyrics, ritual, etc.).

2.5. Repository Storage

Files can be uploaded to a storage repository where the information can be viewed and reviewed by the language community, linguists, technology experts, and other third parties who have been granted access by the community.

The steps for uploading content to the repository may depend on the specific configuration and implementation of your repository. Consult the documentation and administration for your specific repository for details.

2.6. Community Access Control

Language community members or their representatives control access to the repository. The community decides who can act as administrators of the repository. The administrators manage who can upload, edit, view, or otherwise work with the repository contents.

Typically, access control in a repository is managed by defining roles or user profiles. Each role is granted or denied each of the different types of access rights (called permissions). Then, as users are added and enabled to use the repository, they are assigned roles. Their role assignment determines their privileges to perform functions in the repository, including, for example, rights to view, create, edit, or delete records. Repository records can have many fields, including those that represent metadata of when a record is added, by whom, location information, etc. So it can be necessary to specify permissions to view or edit each of these fields, especially if some data represents information for which privacy must be controlled. There can also be administrative permissions to either hide or make public a record, create or edit categories for organizing records, and permissions for managing users (add user, remove user, change user role, etc.) The details of roles, permissions, etc. will depend

on how your repository is configured. Note, permissions can be made specific to particular record or user subsets. For example, a specialist in Indic languages may be granted edit permissions for records related to Indic languages but not records in other languages. A moderator of a group working on Niger-Congo languages may have permissions to manage users that are working with those languages but not users working on other languages.

Some permissions can be used to create a vetting or other type of process workflow for the repository contents. For example, when material is first uploaded to the repository it may be only viewable by a group of reviewers. This gives the group a chance to ask questions and validate the contents as appropriate and representative of the language. If reviewers approve, then the content can be published or marked as viewable by the more general community. Content that requires age of maturity to view, can be marked as such. Contents that violate regional regulations can also be hidden and marked as such. These features depend on the details of your repository. For more details, see the documentation for your repository.

Typically, the language community would grant access to:

- Community members.
- Language experts.
- Computer experts.
- Interested parties.

Access control can become complicated and require a sophisticated set of user roles and access permissions. However, Table 1 illustrates the most basic set of roles and permissions. An empty table cell indicates a permission or access privilege that is **denied** to persons in the associated role. New or edited records would be *hidden* from public view until a reviewer sets the record to have a *publish* setting.

Table 1. Repository Access Management

	Permissions					
	View	**Comment**	**Create/ Edit**	**Hide/ Publish**	**Remove**	**Categorize**
Roles						
Guest	Granted					
Approved User	Granted	Granted				
Community Member	Granted	Granted	Granted			
Researcher or Language Professional	Granted	Granted	Granted			
Reviewer	Granted	Granted	Granted	Granted		Granted
Community admin	Granted	Granted	Granted	Granted	Granted	Granted

- A **Guest** is an anonymous user wishing to view the repository contents.
- An **Approved User** is an individual who may have provided credentials justifying their ability to contribute knowledgeably and respectfully.
- A **Community Member** is a native or accepted member of the language community.
- A **Researcher** or **Language Professional** is an expert who is invited to study or contribute to the repository.
- A **Reviewer** is an individual with advanced skills for moderating content debates and is sensitive to community feelings and requirements, as well as legal and other issues.
- A **Community Administrator** has the ultimate control over users, privileges, and the repository data.

2.7. Error Handling

Even with a thorough vetting process, it is inevitable that mistakes will be included in the language data. Typographical mistakes, errors in transcription, fallible human memory, etc., can introduce errors into your data collection. Data should be reviewed before it is

made available and documented as representative of the language. Ongoing periodic reviews allow for the correction of erroneous language data that may have become part of your collection.

2.8. Review Process

A review process is important to ensure that the data in the repository is correctly described, categorized, permitted and trustworthy. Reviews can bring to light potential discrepancies in the language data.

Data items may be incomplete. For example, items may just be fragments of text. Or their origins may be ambiguous. The items may still be important contributions. The review process provides useful consideration for their relevance.

Data contributed to the repository can come from a variety of sources. Sometimes, either well-meaning or malicious individuals submit information that is speculative or contrived.

Some languages have not been studied or documented previously. Their vocabulary, phonology, grammar rules, etc. are not formally known. The collected data will be used to derive the language's structure, terminology, etc. With more language data examples representing a large variety of scenarios, the greater the ability to accurately digitize the language. However, if any of the collected data is not truly representative of the language, it undermines proper digitization. Errors or omissions can lead to a long chain of incorrect conclusions. A review process reduces this likelihood.

Therefore, it is important to have a process where contributions to the repository are easily accepted and then reviewed by native speakers, community members, language professionals, or other experts. Reviewers can assess, comment, and even challenge the materials as to authenticity, accuracy, and interpretation. They may also supply additional context, inform about proper usage, and correct errors. The review can result in edits or recommend a search for additional or specific data examples. The review might request that the contributor supply provenance or licensing information. The review might also confirm whether there are any legal violations, such as prohibited images or speech. The review process can be iterative.

Until a review is complete, each item is visible only to individuals in the reviewer role. This ensures that other repository users see only information that has been vetted. If reviewers have questions about an item, they communicate with the contributor until all issues are resolved. Then the item can be given a *publish* status and made visible to everyone.

2.9. Creating Your Language Data Gathering Process

The first step is to assemble a team with the necessary skills and tools for each of the language data gathering tasks. Create guidelines, define responsibilities, and plan your workflow to support your gathering of the language data. Consider the following questions:

Gathering Materials

- Do you have a planned workflow or process for inviting, accepting, and digitizing materials?
- Who are potential sources of language data?
- What types of language data materials are available to you?
- Do you have a way to preserve and archive physical language data materials?
- Do you have the tools and skills to put materials into digital format?
- Do you know the rights to use and publish language data for your region?

Hosting Digital Language Data in a Repository

- Have you installed and configured a language data repository?
- Do you have the IT skills to manage the repository?

Vetting Language Data

- Have you formed a team of skilled language data reviewers?
- Do you have guidelines for annotating and vetting language data?
- Do you have guidelines for requirements to approve language data items to be published?

- Do you have guidelines for asking contributors for further information or respectfully questioning relevancy, authenticity, publishing rights, or other issues with contributed materials?
- Do you have guidelines for resolving disputes over issues with contributed materials?
- Do you have a planned workflow or process for uploading, reviewing, annotating, and correcting digital language data in the repository?

Managing Your Team

- Do your team members understand their assignments and responsibilities?

2.10. Section Summary

Gathering language data is an ongoing and iterative process. Data items are added and reviewed as they become available. An outline of the language data gathering process is:

- Invite credible sources of language data.
- Accept language data contributions.
- Put materials into digital format.
- Upload materials to the repository.
- Verify rights to use and publish.
- Annotate.
- Review (correct, annotate, and authenticate).
- Publish.

3. HOW LANGUAGE DATA IS USED IN COMPUTER SYSTEMS

Language data is used in many ways to support the digitization of a language. It is beyond the scope of this document to detail them all. However, here are some of the most basic ways in which language data is used by linguists and technology experts to support the digitization of a language.

3.1. Writing Samples

Writing samples are used initially to establish the writing symbols (letters, digits, accent marks, tone marks, punctuation, and other characters) used in a language. As this set of characters is determined, the writing samples can be used to discover how each character is hand-drawn and to create fonts with those characters. The character set is also required to define the keyboard layout or input method that is used to enter characters into a digital system.

Remember, characters are often drawn in multiple ways. Consider that characters can be stylized with and without serifs, italicized, and have many variations. In some languages, characters change shape based on their position in a word or depending on the character they are alongside.

Also, some characters are rarely used. They may be used only in certain ceremonies or may be in old versions of the language. Because of this, the more samples you collect, the more reliable, complete, and useful the digitization of your language will be.

3.2. Sentences and Paragraphs

Data containing complete sentences and paragraphs can reveal the phonological, orthographic, typographic, grammatical, or other linguistic conventions of a language necessary for word processing. Examples include:

- Grammar rules.
- Hyphenation, word breaks, capitalization, emphasis, and punctuation.

- Justification and writing direction.
- Pronunciation.
- Forms of address (e.g., honorifics and name order).

The data can also reveal unique writing conventions and formatting used to represent dates, times, eras, numbers, percentages, etc.

3.3. Term Lists

Terminology lists can be gleaned from writing samples, audio recordings, and other data. These lists of words and phrases are used by spell checkers, autocorrect, predictive text, optical character recognition (OCR), and other digital functions.

3.4. Dictionaries

Monolingual and bilingual dictionaries provide definitions, parts of speech, pronunciation, etymologies, translations, and other information. This information can be useful for word processing, hyphenation, spell and grammar checking, autocorrect, machine translation, and other aspects of digitization.

Designing and formatting a dictionary can be complicated. Lexicographers give careful consideration, for example, to approaches to *head words* in polysynthetic languages, where complex words or sentences are built from many parts, and to alphabetizing words. Some of the resources available on making dictionaries for Indigenous languages include:

- 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. Recordings

Audio and video recordings can be used to discover the norms of pronunciation. This information enables text-to-speech and voice-to-text capabilities. Text-to-speech is useful for people with sight or reading disabilities and low literacy. Speech recognition enables voice command capabilities and is also useful for people with disabilities who cannot type on a keyboard or touchscreen.

3.6. Bilingual Data

Bilingual data serves many purposes. For example, translation dictionaries, subtitled videos, and translated documents enable the creation of online word lookup dictionaries, voice translation tools, machine translation, and other tools. Also, comparing terminology between languages reveals nuanced differences.

3.7. Volume

Typically, the more data collected, the better the quality of digitization. Grammar, word definitions, etc. become more accurate and nuanced. Idioms and rarely used terms can be documented.

Some terminology is only used in association with particular topics or domains. Examples include health, agriculture, regulatory, etc. The greater the volume of data gathered, the higher the likelihood that more domains are covered.

Also, some language applications, for example, machine translation, are successful only when there is a significant volume of language data available for training the translation system.

3.8. Variety

All kinds of language data are useful to establish robust support for digitizing your language. Do not exclude sources of language data that seem archaic, informal, formal, official, for young or uneducated people, hyperbolic (advertising), or implausible (legendary or historical stories). Any of the following will be useful:

- Children's story and picture books.
- Educational materials.
- Correspondence, personal letters, notes, and messages.
- Regulatory documents (certificates of birth, marriage, death, etc.).
- Dictionaries (monolingual and bilingual).
- Books.
- Newspapers.
- Signage.
- Posters.
- Archaic writing.
- Oral histories and traditions.
- Drawings and other artwork.
- Everyday conversation (oral, written, formal, and informal).

3.9. Coining new terminology

New terminology is naturally needed to denote novel ideas, inventions, and activities. This is especially true as a language is being readied for digitization. For example, terminology used in the user interface of software and hardware must be defined or adapted in the native language. Terms such as menu, button, drop-down, file, edit, help, exit, ok, cancel, click, download, etc.

Native language equivalents are not always easy or obvious to choose or invent. Literal translations may not be the best choice. For example, the term "home page" is translated as "start page" in some languages. Also, a term may have multiple uses and require different translations according to context. For example, sometimes "cancel" can

have the meaning “abort” and other times it can mean “undo”. Consider that in English “orange” is both a color and a fruit, but in other languages, there are separate terms for each.

A language community may need to create a process to generate and agree on their terminology for computer digitization. See the ***Zero to Digital Terminology Guidelines*** for information on coining new terms.

3.10. Section Summary

The greater the volume and range of language data in your collection, the more accurate and comprehensive the digitization of your language will be. The following table illustrates some of the basic ways that language data is used in computer systems.

Table 2. Language Data Usage in Computer Systems

Language Data Type	Linguistic Value	Software Uses
Writing Samples	Reveal writing symbols and conventions	Determining character set, font, keyboard layout, input method, etc.
Sentence Data, Paragraphs	Reveal phonological, orthographic, typographic, grammatical, and other linguistic conventions	Word and speech processing
Term Lists	Reveal vocabulary	Spell-checkers, autocorrect, predictive text, and optical character recognition
Dictionaries	Provide definitions and parts of speech	Spell and grammar checkers, word processing, text-to-voice, and machine translation
Audio, Video Recordings	Reveal norms of pronunciation and colloquial speech patterns	Text-to-voice and speech recognition

Bilingual Data	Provides translations, cross-language correlations, and writing systems	Bilingual dictionaries, subtitled videos, translated documents and software, voice translation tools, and machine translation
Coined Terms	Establishes terms that are needed by digital systems	Software menus, commands, dialogues, etc.

4. SPECIALIZED LANGUAGE UTILITIES

4.1. Unicode Common Locale Data Repository

The Unicode Common Locale Data Repository (CLDR) provides key building blocks for software internationalization and localization in support of the world's languages. As its name implies, CLDR represents a large collection of locale data. It is used by a wide spectrum of companies to adapt their software to support the conventions of different languages.

For example, CLDR contains accepted translations of many proper names that are needed by software applications, such as months, days of the week, countries and their subdivisions, language names, measurement units, and currencies. CLDR also provides coding expressions that software can use to format data according to local conventions for date, time, number, measurement, currency, and others.

If your language is not already represented in CLDR, consider submitting information to populate it. This will enable software application developers to more easily add your language to their systems. Reviewing CLDR contents can also help you determine information and terminology that is needed by digital systems and must be coined for your language.

4.2. Machine Translation

Machine translation is an increasingly important language application because it can translate large volumes of text quickly and efficiently.

However, to establish machine translation systems, large quantities of bilingual data are required. Ideally, the data should be aligned, parallel corpora where each segment (or sentence) of the source language is paired with a matching target language segment. For more information about the importance of machine translation and its specific requirements, see *Appendix C: Machine Translation Requirements*. It provides an overview of machine translation and the data requirements for building a machine translation system.

4.3. Natural Language Processing Tools

This section introduces the subject of tools that help with the analysis of your language and the generation of language data. An example is a tool that scans a document and extracts a list of words in your language. Initially, you should identify tools that can help with immediate tasks. These tools may require some customization to work with your language.

Many software applications rely on Natural Language Processing (NLP) components for working with language data. After you have achieved milestones for language gathering, you may want to invest in training language models and updating software libraries used by the industry. Providing pre-trained models for your language can accelerate the adoption of your language by many applications.

Although languages are diverse, many languages can be grouped together by similar patterns of syntax, grammar, etc. Tools that recognize patterns used by your language can help process language data. For example, there are tools that can parse text for certain language types and create word lists. There are more sophisticated tools that use models of the language to process language data. Language data can also be used to train and create increasingly more refined models of the language. There are some tools that are specifically designed to work with under-documented languages.

There are several open-source websites that host NLP and language modeling tools. Consider selecting a tool that supports languages grammatically similar to yours and then customize and train it on your language data. As you are successful you can add your language's trained model to these websites. Software developers can support your language more easily, and you can then invite major software companies to add your language to their libraries.

Example tools include:

- Word vectors like word2vec (model to identify word associations from a large corpus of text).
- Trained models like BERT and GPT (capturing grammatical functions, meanings, and word associations).
- Models for named entity recognition (NER) (names, geography, dates, etc.).
- Models for part-of-speech tagging (noun, verb, etc.).
- Models for syntactic/dependency parsing (subject/object/..., verb and noun phrases, etc.).
- Huggingface's "Transformers" (information extractions, translation, and other NLP functions).
- SpaCy library (information extraction and other NLP functions).
- Natural Language ToolKit (NLTK) (text processing libraries).

Many applications rely on language identification libraries to configure themselves for the current language. They cannot support languages that these libraries do not recognize. Contributing information to these libraries, so they recognize your language, can accelerate acceptance and support for your language.

Examples include:

- Google's CLD3 (Compact Language Detector v3).
- Facebook's fastText.

4.4. Section Summary

These utilities can help with the analysis of your language, support its digitization, promote the adoption of your language by software developers, and accelerate the translation of materials to and from your language.

- Unicode Common Locale Data Repository.
- Machine Translation.
- Natural Language Processing Tools.

5. GLOSSARY

Term	Description
Character	A letter, logograph, sign, mark, or symbol used in writing.
Corpus	A large or complete collection of writings.
Diachronic	Concerned with the way something, especially language, evolves over time.
Dialectal	Belonging to or related to a dialect of a language.
Digitize	Convert into a digital form that can be processed by a computer.
File format	A method for storing information in a computer file. The method varies with the type of data that is being stored and can generally be identified by the file extension. (e.g., .html for a web page).
Font	Graphical representation of text. A collection of writing symbols with a similar graphical design.
Glossary	Alphabetical list of words and their definitions relating to a specific subject.
Indigenous	Native to a specific region.
Lexicography	The practice of compiling dictionaries.
Locale Data	Information used to customize user interfaces for a given language and culture of a region.
Media	1. Means of mass communication (broadcasting, publishing, and the internet) regarded collectively. 2. Data storage devices.
MT	Machine Translation.
NER	Named Entity Recognition.
NLP	Natural Language Processing.
NLTK	Natural Language Toolkit.

Orthographic	Set of conventions for writing a language. It includes norms of spelling, hyphenation, capitalization, word breaks, emphasis, and punctuation.
Phonological	Relating to the sounds in a particular language or in languages, or to the study of this.
Polysynthetic	Denoting or relating to a language characterized by complex words consisting of several morphemes, in which a single word may function as a whole sentence.
Repository	A central location in which data is stored and managed.
Scan	Copy and store information in digital form.
Segment	A discrete meaningful unit of text or spoken language. Text segmentation is the process of dividing written text into meaningful units, such as words, sentences, or topics. Speech segmentation is the process of identifying the boundaries between words, syllables, or phonemes in spoken natural languages.
Serif	A slight projection finishing off a stroke of a letter in certain typefaces.
Text-to-speech	Assistive technology that reads digital text out loud to the user.
Typographic	Art and technique of arranging type to make written language legible, readable, and appealing when displayed.
Unicode	An international encoding standard supporting digital text in different languages and scripts. Each letter, digit, symbol, or other character is assigned a unique numeric value and functions consistently across different platforms and programs.
Upload	Transfer (data) from one computer to another, typically to one that is larger, or remote from the user, or functioning as a server.
URL	Uniform Resource Locator, the address of a page or other information on the Web.
UTF-8	Variable-width Unicode-based character encoding used for electronic communication of text.
Voice-to-text	A speech recognition program that converts spoken language to written text.
XLIFF	XML Localization Interchange File Format, an XML-based file format for exchanging localizable data.

Appendix A: Example Digital Applications

These common activities may be available in your native language on digital systems. It is also possible to create new applications that are specific to the needs of your community.

Communication

- Send and receive text messages.
- Send and receive emails.
- Send and receive media (images, audio, and video).
- Automatically translate text, machine translate.
- Auto-convert voice messages to display as text and vice-versa.

Publishing, Documentation, and Word Processing

- Publish and access information on websites.
- Create and share documents, books, news media, signage, posters, and educational materials.
- Create print and online dictionaries.
- Scan documents to convert to digital text.
- Create a font for your script.
- Buy and sell things online.
- Localize websites and applications into your language.
- Create native language applications.
- Spellcheck.
- Grammar check and autocorrect.

User Interfaces and Disability Support

- Speech recognition (useful for people with physical disabilities).
- Use voice commands to control devices.

- Text-to-speech and screen readers (useful for visually impaired or low-literacy use).
- Speech-to-text and real-time captioning (useful for the hearing impaired).

Appendix B: Data Requirements for Technological Applications

The Data Requirements for Technological Applications table is a visual guide to help you determine a good starting place in the collection of language data with regard to your goals for enabling technological applications in your language.

The table helps communities from two angles: a community can look up the desired computer applications and discover the kinds of language data that are needed to support their creation. Or, based on the kinds of language data that are available or able to be acquired, a community can determine the applications that may be achieved in the near term.

The table displays much of the same information outlined in this document and helps users have realistic expectations for achieving their digitization goals.

Table 3. Data Requirements for Technological Applications

Data Types	Writing Symbols	Terminology	Translation	Speech	Language and Layout Conventions		Bulk Language Data
Example Content	characters, letters, digits, punctuation marks, fonts, etc.	term lists, monolingual dictionaries	bilingual dictionaries, wordnets, corpora	phonological, pronunciation rules, audio and video recordings	orthographic, grammar rules, hyphenation, capitalization, punctuation, writing direction, justification	writing conventions: dates, times, eras, numbers, percentages, etc.	sentences, paragraphs, monolingual textual corpora
Digital Applications							
Communication							
Send/receive text messages	x						
Send/receive emails	x						
Automatically translate text, machine translate	x	x	x		x	x	x

Data Types	Writing Symbols	Terminology	Translation	Speech	Language and Layout Conventions		Bulk Language Data
Auto-convert voice messages to display as text and vice-versa	x	x		x	x	x	x
Publishing, Documentation, Word Processing							
Publish and access information on websites	x				x	x	
Create and share documents, books, news media, signage, posters, educational materials	x				x	x	
Create print and online dictionaries	x	x	x		x	x	
Scan documents to convert to digital text (OCR)	x						x
Create font for your script	x				x		x
Buy and sell things online	x				x	x	
Web sites and applications localized into your language	x	x	x		x	x	
Create native language applications	x				x	x	
Spellcheck	x	x					
Grammar check, autocorrect	x	x			x	x	

User Interfaces and Disability Support							
Data Types	Writing Symbols	Terminology	Translation	Speech	Language and Layout Conventions		Bulk Language Data
Speech recognition (useful for people with physical disabilities, also talk to Siri or Alexa)		x		x			
Text-to-speech and screen readers (useful for visually impaired or low-literacy use)	x	x		x	x	x	x
Speech-to-text and real-time captioning (useful for the hearing impaired)	x	x		x	x	x	x
Legend	**x**	The checkmark indicates the type of data most likely to support the creation of this type of application.					
Note		This table is not a complete list of data types or applications.					

Appendix C: Machine Translation Data Requirements

What is Machine Translation?

Machine translation (MT) is the use of software to translate text or speech from one language to another.

To produce high-quality translations, MT is more than mechanical word-for-word substitution. MT uses advanced algorithms and requires large quantities of example language data to configure a productive system.

MT systems do not achieve the quality levels of human translations. To improve quality, MT systems are often customized by domain or profession to limit the scope of the content.

MT is useful as a tool to assist human translators and for some purposes can produce output that can be used as-is.

Building Machine Translation Systems

To automate translation between a language pair, an MT system must be built specific to that pair. This involves choosing an MT technology and using language data for both languages to configure the system.

Building MT systems for new language combinations is only possible after each language has a secure digital foundation. There must also be a natural and growing proliferation of linguistic resources in a newly digitized language.

There are many languages with tens of millions of speakers that do not have usable MT systems today, because either adequate data resources are not available, or enough effort has not been expended. While the core MT development technology continues to

improve, and it is increasingly easier to build new systems with less data, it should be understood at the outset that significant volumes of data are required to produce *good* MT systems.

The data used to build MT systems is called training data.

While it is possible to develop a basic system quickly, assuming some foundational training data is available, additional data can be added over time to an existing MT system to drive ongoing improvements and performance.

MT systems are also continually evolving and improving as new data, new techniques, and ongoing corrective feedback are incorporated. Periodic evaluation and updating should be planned for.

Where is MT useful?

MT is useful for making large volumes of information and knowledge resources available quickly at a relatively low cost. However, we should also understand that, currently, MT falls short of a competent human translation.

However, MT is quick, often a good enough approximation, and can be deployed at will for millions to use on the web after an MT system has been built. The existence of MT enables millions of people to access information that they would not otherwise be able to. While stories of MT mishaps and mistranslations abound, it is becoming increasingly apparent to many that it is crucial to learn how to use and extend the capabilities of this technology successfully. While MT is unlikely to replace human beings in any application where quality is paramount, there are a growing number of cases that show that MT is suitable for:

- Highly repetitive content.
- Content that would just not get translated otherwise.
- Content that cannot afford human translation.
- High-value content that is changing every hour and every day.

- Knowledge content that facilitates and enhances the global spread of critical knowledge.
- Content that is created to enhance and accelerate communication with global customers who prefer a self-service model.
- Content that does not need to be perfect but just approximately understandable.

What kind of data is needed to develop an MT system?

All modern MT development technologies are data-driven, that is, computers analyze large quantities of accumulated translation data to *learn* how to translate from one language to another. This linguistic data that is used to develop MT systems is called training data. The current MT technology that is most widely deployed is Neural MT, and this is slowly replacing the many installations of an older approach called Statistical MT. They are both approaches to developing MT systems using the following types of data:

- Bilingual text.
- Translation glossaries.
- Monolingual data in the target language.
- Monolingual data in the source language or closely related languages**.**

Bilingual Source and Target Language

Large collections of texts translated sentence-by-sentence are called parallel corpora. An initial translation engine can be created with a minimum of 100,000 bilingual translated segments (sentences). A translation segment can be a full sentence or a group of words that translate crucial terms and phrases. Ideally, there should be at least 1,000,000 segments. Some MT systems are built with billions of segments. Generally, we can say that larger volumes of high-quality bilingual segments will deliver higher quality MT output.

Many language communities do not have large volumes of such data. The data acquisition phase then often requires a concerted and long-term effort and collaboration between government agencies, educational establishments, and the community at large.

In the interim, technology exists that allows a smaller number of segments to be used, with human feedback providing incremental improvements as data is translated.

Corpora used as training data sets for MT algorithms are usually extracted from large bodies of similar sources, such as databases of news articles written in the source and target languages describing similar events.

However, extracted fragments may be noisy, with extra elements inserted in each corpus. Extraction techniques can differentiate between bilingual elements represented in both corpora and monolingual elements represented in only one corpus to extract cleaner parallel fragments of bilingual elements. Comparable corpora are used to directly obtain knowledge for translation purposes. High-quality parallel data is difficult to obtain, however, especially for under-resourced languages.

The training data used to build an MT system is most often translation memory (TM, an archive of past translations) or other legacy translation assets that have been collected over some time. This information will define what the MT system will learn to translate best. Often there are limits to the volume of data available. In such cases, special efforts need to be made to teach the system to learn the material it is most likely going to be focused on translating. Remember that what you train on is what your system will translate best. Thus, an MT system that will be used to translate medical content is best trained with medical TM and glossaries.

When bilingual data is provided for training, the data must be **aligned**: the source and target must be direct translations of each other. Translated texts that are summaries, abstracts, or commentaries on the original text are not suitable. Data must be carefully examined before use, to ensure that it will be useful for training purposes.

Glossaries and dictionaries of key terminology with their translations enable greater translation accuracy.

There is also long-term value in developing a comprehensive metadata strategy for the linguistic data that is gathered. In the initial phases of the data acquisition, the focus is usually on finding data wherever possible to address the need for the critical data mass

needed to begin. However, as the MT engines mature there can be significant performance benefits if the right type of data is used to build the engines. Thus, an MT system can be optimized for medical domain-related content or for computer technology-related content, rather than having a single system that is doing everything. This specialization will often result in better-performing MT systems.

Bilingual Data Formats

Bilingual data has three important characteristics to be useful as machine translation training data. It should be in a file format that MT systems can import. Text data should be in Unicode UTF-8 character encoding. And, the parallel data should be aligned, each segment of source language paired with a matching target language segment.

Data can be delivered in the following file formats and are listed in rough order of preference:

- Translation Memory (TMX, TBX, XLIFF, CSV): preferred format.
- Plain text (TXT).
- Web content (HTML).
- Structured (XML).
- Microsoft Office (DOC, DOCX, PPT, PPTX, XLS, XLSX).
- Publishing or DTP formats (TTX, PDF, FrameMaker).
- Optical Character Recognition (OCR), (TIFF, PNG, JPEG, etc.).

Data can be already aligned and matched between source and target language or delivered in its raw forms as, for example, Microsoft Word or HTML documents. If the data is not aligned, it will be necessary to use tools that will enable the alignment of data with high levels of accuracy.

Monolingual Language Data

While archive translations and bilingual text are perhaps the most critical data to building an MT engine, it is also imperative to have good quality monolingual data in the target language. This data is used to learn the correct grammatical structure during translation

and statistically influences the output to read in the desired style of writing. This is especially true when building statistical MT systems.

Monolingual data is much easier to acquire than bilingual data. It can be useful to collect URLs of websites that have a similar domain or grammatical style. Their language data can be mined, and the linguistic knowledge contained in this data can be leveraged.

Typically, it is harder to get monolingual language data for Indigenous languages. For majority languages, there are many sources of monolingual language data. However, a use case where monolingual language data can be important is the building of a medical MT system. Crawling medical and related websites rich in this specific content can identify language information critical for constructing glossaries and translation memories.

Text data should be in Unicode UTF-8 character encoding. Data can be delivered in the following formats and are listed in rough order of preference:

- Plain text (TXT).
- Translation memory (TMX, TBX, XLIFF, CSV).
- URLs of web content.
- Web content (HTML).
- Structured (XML).
- Microsoft Office (DOC, DOCX, PPT, PPTX, XLS, XLSX).
- Publishing or DTP formats (TTX, PDF, FrameMaker).

Appendix D: Benefits of Digitization of a Language

We hope the Zero To Digital series of guidelines will offer any interested language community a clear path to digitization if they wish to enjoy full computer capabilities in their native language.

The benefits of rendering a language in digital form will depend on the goals of its speaker community. These may include celebrating the beauty of the language, maintaining knowledge systems, disseminating values, creating apps and products, sharing stories and histories, facilitating environmental stewardship and thought leadership, and expanding trade, education, employment, entertainment, health, and safety. Digitization allows a community to avail itself of an ever-expanding set of computer-based tools for language maintenance, revitalization, and education. A strong digital presence online, and therefore increased visibility, may help influence government policies in support of Indigenous communities and tip businesses toward inclusion. Given the ubiquity of smartphones among youth, digitization may be a natural avenue to engage them with their native language. Greater exposure via digital platforms can lead to more opportunities and demand for native speakers.

When language communities take their place on the global stage via digital platforms, it benefits the broader world. Their experiences, knowledge, and unique worldviews will be a significant contribution to the rest of the world, and the resulting synergies may bring new solutions to the world's problems. Digitization facilitates the preservation and publication of this information, making the breadth and nature of human language available to the world and preserved for humanity's benefit in ways we cannot yet predict.

To summarize, some of the benefits gained by language digitization are:

- Enable monolingual speakers to easily access, create, and exchange native language content, including across long distances, and to individuals or large groups.
- Increase access to medical and healthcare information.
- Support lifesaving emergency and disaster communications.
- Expand both local and e-commerce.
- Create new avenues for sharing art, thought leadership, and philosophies.
- Develop native-language educational materials.
- Improve relations and communications with neighbors.
- Improve dispute resolution.
- Enable advocacy and access to legal and governmental procedures in the native language.
- Increase access to information on the Internet for education, commerce, and participation, whether in the native language or other languages as translation tools develop.
- Bring recognition of native language, culture, and wisdom to others.
- Expand the role and visibility of Indigenous communities globally.
- Enable marginalized or minoritized groups to sustain or revitalize their language even though engulfed by dominant groups.
- Preserve native knowledge systems, culture, history, art, medicine, wisdom, values, and worldview.