ZERO TO DIGITAL SERIES:

TERMINOLOGY GUIDELINES

TRANSLATION COMMONS

2019 | INTERNATIONAL YEAR OF Indigenous Languages

Terminology Guidelines

Authors: Sue Ellen Wright, Akil Iyer, Shuto Kato
Reviewers and contributors: Craig Cornelius, Alaina Brandt, Julie Anderson, Tex Texin
Marketing: Leonidas Pappas
Project Coordination: Paula Cirilo

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

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 who 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* 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.

2. ABOUT THIS DOCUMENT

This document describes how to document terminology in an existing, undocumented, or under-documented language. It outlines procedures for recording existing words and for creating new terminology to support communication in specialized subjects for which Indigenous languages may not have particular words. This recommended methodology provides guidance about various workflows for creating or updating terminology, including planning, resource collection, terminological analysis, and documentation. This process provides a means for Indigenous communities to be able to coin terms as the need arises, for example, for subjects belonging to technological, medical, political, legal, or other fields that may otherwise be outside the historical realm of their culture or language. Expanding and updating this kind of vocabulary will help an Indigenous community access and use resources to facilitate their inclusion in a digital environment.

Some endangered languages have a long tradition of writing and others are strictly oral. Some are viewed as "family languages" that are not shared with outsiders. Language advocates need to document the words that already exist and create new words in

order to represent objects, concepts, and ideas that their people will need if they want to digitize their language.

The intended audiences of these guidelines are Indigenous communities wanting to document their vocabulary, subject experts supporting terminology work in one or more languages, and organizations supporting language communities in the digitization process. The goal is to be able to use the resulting terminology collection for oral, written, and eventually, digital resources.

The workflows described can be applied to terminology work in any domain or subject area. Examples given here rely heavily on the English language, but certain general rules and models can be modified to apply to different languages and situations, although some information will not apply to all languages. Still, the examples should give readers a good starting point to identify the parameters that will need to be considered in their language when conducting terminology work.

This document describes fundamental areas of knowledge, best practices, and resources (both technological and otherwise) that will aid Indigenous communities as they conduct terminology work in their languages to support language preservation, revitalization, and digitization efforts.

3. WHAT IS TERMINOLOGY AND TERMINOLOGY WORK?

Words, abbreviations, or even phrases that are used for special purposes in languages are considered to be part of the special language terminology of that language. Certainly, the terminology of science and technology is widely recognized, but all specialized human activities in all cultures use special terminology, whether for cooking, hunting, or expressing cultural values. Terminology work is carried out with two purposes in mind:

- Collecting words and terms: We collect and document existing words and terms. In indigenous and undocumented languages, this activity is part of language documentation and language planning, and we collect all the words and terms we can. In this environment, we don't limit ourselves to special language terms per se, but we do assign concepts to subject fields where relevant.

- Coining new terms for missing concepts: When possible, we use the existing resources of the language we have recorded to create new terms when we find that there are concepts we need to talk about that don't yet have terms in the language. Terminologists often identify these concepts by comparing similar concepts and recognizing that a concept they know in one language hasn't been created or coined in the other. The proper designation for these empty slots is "missing terms". Sometimes languages fill these slots by borrowing. English has words like *tsunami* and *Schadenfreude* that were borrowed in this way. With developing languages, it is usually considered wise to coin new terms using existing resources rather than borrowing.

Terminology work is important to cultural preservation and the promotion of equality. When languages are endangered, terminology work can be used to create a record of a language and revive its use. Welsh and modern Hebrew are examples of languages that were in danger of dying out but have become important national languages, thanks in part to systematic terminology work.

The goal of terminology work is to collect the words that people actually use and to record the shared understanding of their meanings. A collection of such words relating to the various areas of specialized activity in a community provides a basis for Indigenous language keepers to actively participate in the process of documenting their language, preserving the accuracy of terms and their appropriate uses. As noted above,

further, terminology work includes the coining of new terms when a new technology is adopted, which will be the case for many Indigenous languages that are being digitized, or made ready for use on digital devices. Digital terms, like *browser* and *tool* can be borrowed from everyday words, but terms like *menu* and *font* and even *email* will take thought because their root meanings may be unfamiliar.

4. IDEAL WORKFLOW

The following outline presents an ideal workflow for terminology work involved in digitizing languages. All the steps are important for the completion of the project, but it may not be feasible to follow them in the precise order in which they are listed. This list is an approximate guide for your work.

4.1 Are you ready to start your terminology work?

- Does your language have the features you need for digitization? For example:
- Does your language have a writing system?
- Is your writing system available in the Unicode Standard to record your terminology using computer applications?
- Has your writing system been configured for keyboard data entry and display on computers and digital devices?
- Has anyone worked out basic grammar rules for your language?

If the answer to any of these questions is “NO”, check the Translation Commons Resources tab for more information on how to address these issues. But don’t stop here – even if some of these steps are in progress, there may be many other things you can be doing in the meantime. Some of the suggestions in these Guidelines may be repetitive, but they are focused on the process of identifying and recording terms, as well as creating new ones. It is helpful to note that you can begin to make audio recordings long before a script or writing system has been finalized and you

can start to use an evolving script to record information manually before the digitization process has been completed.

4.2 Have you assembled human resources to do the job?

- Have you gathered representatives of your community to work on your project?
- Have you enlisted support from elders and other group leaders?
- Which speakers of your language know about your topics? Who is a great cook, hunter, or another expert? Make contact with that expert cook or hunter and encourage them to tell you about their work.
- Have you found some people who can provide technical support to handle the computer tasks involved in the project?
- Do you and your team know how to recognize and collect words and terms?

- It may take a while to get your team together, especially if you are gathering experts in different subject fields, and to help everyone understand your goals. This may be an ongoing process over time. Use tutorials and other materials to train people to recognize terms and record their meanings. If you are developing your grammar and standardizing spelling, for instance, and people need to learn to apply basic grammar rules, add this information to your training program.

As you work along, be sure to check your work with members of your community. Do they agree with your decisions? Are there differences of opinion on pronunciation or the way you should represent words in your script? You may find that dialects emerge as you work along. All these issues need to be considered in order to arrive at a consensus among your group.

4.3 Create texts and collect existing terms in your language.

- Work preferably on one specific topic at a time, but be ready to document anything new that comes your way by indicating the subject field as you work along.
- Identify whether you already have any word lists, dictionaries, or texts in your language that deal with the topic and that you can use to start your term collection.
- Talk to the experts you have identified and make audio recordings of their stories and comments.
- Transcribe your recordings, starting out writing by hand if necessary.
- Identify important words and terms, and start to make lists and to write definitions or explanations.
- If you are documenting terms in your living area and you don't know the terms for some of the objects you see, try to find out those terms so you can include them in your lists.
- Are your terms accurate? Check with multiple speakers – are there variant forms? Are there dialects?
- Document related terms as you go along.
- Store your texts and word lists together to create a text corpus.
- Once you have a large corpus, there are computer tools that you may want to use to organize your corpus or to document your terms.

4.4 Create (coin) terms for missing new ideas and concepts.

- Are there terms you need that are missing in your language?
- Organize related terms in concept systems in order to study concept relations and identify missing terms and support definition writing.

- Do you need to create whole groups of terms, such as for medical or computer terms?
- Check with several or even many speakers to ask whether coined terms make sense to them.

4.5 Document and maintain terminology in shared glossaries.

- Have you decided how you will keep track of terms and their meanings?
- Have you decided on the types of data you will document about each term?
- Have you decided on the terminology management system and the data model you will use to document terms?
- Has each important concept been documented in a terminology collection, which has been organized in a terminology database (termbase)?
- Has the quality of each concept entry been verified?
- Has the termbase been shared with users who need it?
- Does your termbase describe what you have found in your research? This is called *descriptive terminology*.
- Does your termbase tell people how they should use terms? This is called *prescriptive terminology*.
- Do relevant communities of speakers refer to the concepts in the termbase consistently?
- Does the terminology need updating over time?

5. APPLYING THE GUIDELINES

5.1 Using the flowchart

The Flowchart shown in Figure 1 shows an overview of these questions. Remember this is a simplified view – you may end up doing different pieces of the project in a different

order or in parallel. Every project is different. It pays to be flexible and creative as you work along.

Note that the second item in the right-hand column preferences alphabetic scripts, where the symbols in the script represent sounds in the language without reference to meaning. An option is to work with a logographic script, where the characters in the language represent a concept or idea without reference to the way the respective word or term is pronounced. English, French, and Arabic are examples of alphabetic scripts; Chinese and Japanese Kanji are examples of logographic scripts.

Collecting people to participate should certainly be one of the first activities, and once you have identified them, you will probably want to start working with them as soon as possible. As you talk to people and gain their trust and cooperation, you can begin to make audio recordings long before a script or writing system has been finalized, and you can start to use an evolving script to record information manually before the digitization process has been completed. Over time, as more people become involved and are using cell phones and other digital tools, you can use digital media to collect information from an ever-wider group of people who know the language.

These efforts will save valuable time and provide initial information that can be used to create models for future activities as the project matures. This early work will also provide verifiable examples of text that you will need later when you apply for ISO 639 language status, script registration, or Unicode certification. As the project moves along, check back in the flowchart to see if you need to catch up on some steps you couldn't do earlier.

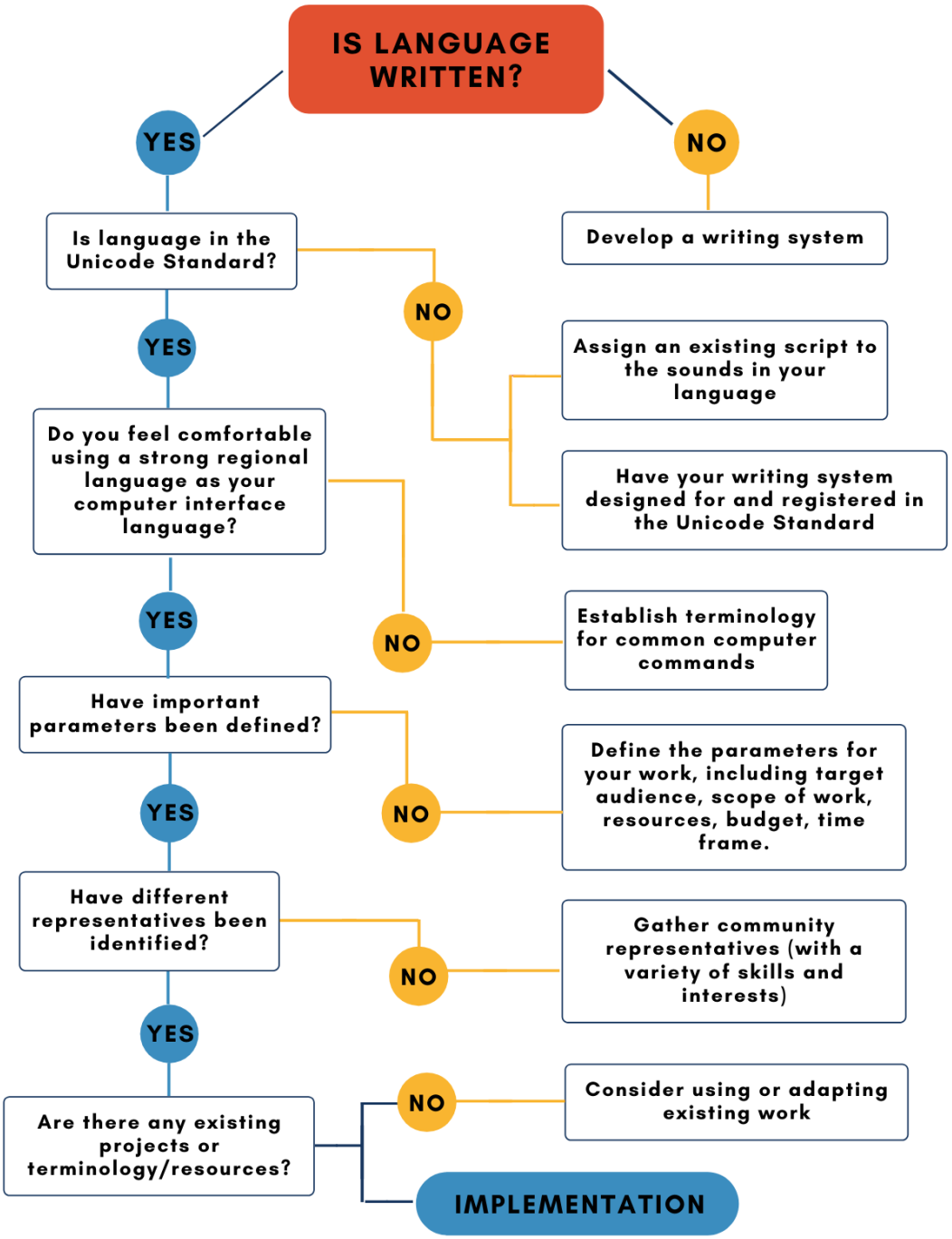


Figure 1: Project preparation flowchart

5.2 Lexicography (familiar dictionaries) and terminology (terminography)

Traditional dictionaries list words in entries where all the meanings (definitions) associated with that word are included in the same entry, as shown in Figure 2.
Lexicographical dictionaries are very useful when you work in just one language because you can record all your information about a single word together in one place. They also may include grammatical information, and they are used to document many small words (sometimes called noise words in corpus collections) like articles, prepositions, conjunctions, and other language particles that are used throughout both everyday and specialized language.

rattle (rat'l) vi. -tled, -tling, [ME …]
1. to make a series of sharp, short sounds in quick succession
2. to go or move with such sounds [a wagon rattling over the stones]
3. to talk rapidly and incessantly; chatter [often with on: rattle on]
-vt.
1. to cause to rattle [to rattle the handle of a door]
2. to utter or perform rapidly
3. to confuse or upset; disconcert [to rattle a speaker with catcalls]
-n.
1.quick succession of sharp, short sounds
2. a rattling noise made by air passing through the mucous of a partially closed throat: cf. DEATH RATTLE
3. a noisy uproar; load chatter; engine rattle
4. a series of horny rings at the end of a rattlesnake's tail, used to produce a rattling sound
5. a device, as a baby's toy or a percussion instrument, made to rattle when shaken
Collocation: to rattle around in a house that is too big for one's needs

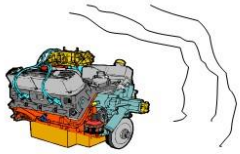


Figure 2: Dictionary (lexicographical) entry for the word *rattle*

Lexicographical entries become problematic, however, when you try to use them as the basis for bilingual or multilingual dictionaries, or if you want to document

many dialect words or synonyms. Each one of the meanings in the entry has the potential to be a different word when you look for equivalents in another language, and different spellings or synonyms will end up in multiple entries. This is why today most translators and other bilingual authors and researchers, as well as language planners, use what is called the “terminological approach” to recording lexical information.

5.3 How do we identify terminology and why does terminology work matter?

Using the terminological approach to language documentation, we create a separate entry for each one of the different meanings of a word. That may sound complicated because our one entry in Figure 2 becomes eleven entries if we create terminology entries. If we do that, however, it is easy to assign a term, its synonyms, and equivalents in multiple languages or dialects to each one of the meanings, that is, to accurately assign equivalents for each idea or concept it represents. This approach is essential, in fact, not only for accurate translation but also if we are to navigate the coining of any new terms for concepts we need for the digitization process or for other missing concepts, such as in medicine.

To visualize the relationship between a given idea or concept and the term that represents it, we often draw a triangle, showing (1) the *concept* the term creates in our minds, (2) the *object* the term refers to (sometimes called the *referent*, or the *object in the real world*), and (3) the term itself. If a term “works”, we can use it when we speak or write, and the listener or reader can immediately form the same idea in their mind. When we communicate with more than one language, it is important that the object (the referent) be the same or very similar, and that everyone in both languages understands the same concept we have in mind. Only then can we decide that a term in Language A is adequately equivalent to a term in Language B. An example of a semantic triangle, which is used to represent meaning, is shown in Figure 3.

The triangle arrangement becomes very useful, then, when we compare languages. We start by recognizing that a word represents a given concept. Then we look for the equivalent concept in the target language. If there is no equivalent word for that concept, we have a typical situation of a “missing” word and we may create a new word or borrow a word. Ideally, terminology work sets out to identify existing terms in a language community rather than borrowing words from outside languages. The term *tree* is a good example because the Native American language Cherokee has a word for *tree*: ᏭᎬᎢ (*quigvi* [kʷigə̃i]). However, there are many other English words for which there was originally no Cherokee equivalent. Suppose we wanted to coin a good term for the concept represented by the English term *email*. If we want to avoid borrowing terms, we wouldn’t want to just spell out a word in Cherokee that sounds like “E-MAIL”.

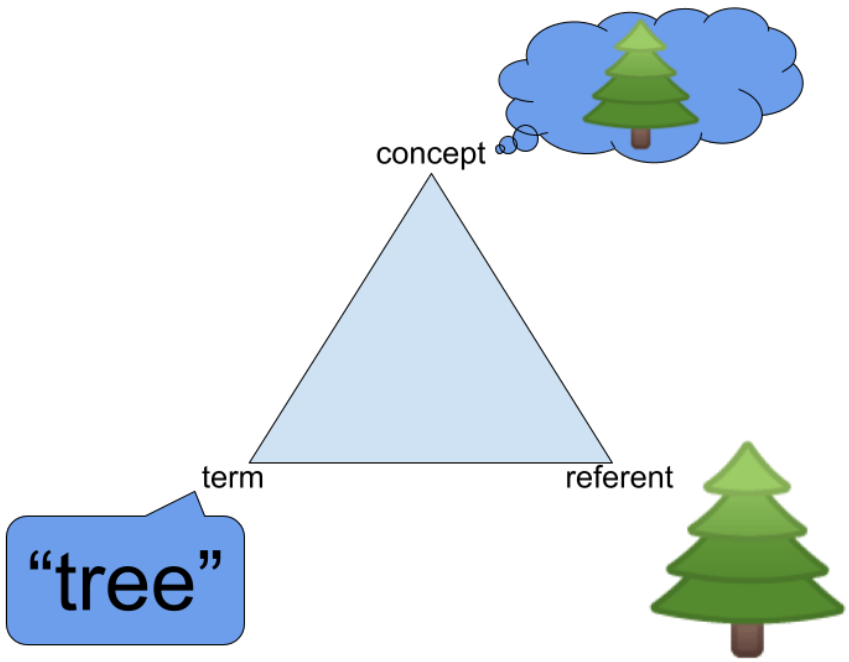


Figure 3: “Semantic triangle”, where the term “tree” matches the object (referent) tree, which triggers the idea or concept of the tree in the mind of the person who hears or reads the term.

In terminology work, we consider that regardless of whether the English speaker and the Cherokee speaker have had the same experience with emails at the time the terms were created, the *email itself* is the same *thing* for both of them. The object or referent

email doesn’t ever change in this equation. Sometimes known objects, like trees, will look a little different to the two speakers (for instance, an oak tree looks different from a baobab tree), but we look at the common characteristics of the two objects when we name or define our object. It is primarily the term node that is likely to be different and likely to reflect the different experiences of the speakers, readers, and writers. Compare the English word *email*, which is shortened from *electronic mail*, and the Cherokee word ᎠᎾᎦᎵᏍᎬ ᎪᏪᎵ (*anagalisgv goweli* [anagalisgə̃ goɰeli]) which directly translates as *lightning paper*. The English speakers, although they have seen lightning, deal daily with electricity that flows through an electrical system (represented here by an electrical plug), and they know that is how the email reaches them. Of course, modern Cherokee speakers share this experience, but the language itself didn’t participate in the same evolution. Here we have the convergence of the concept of electricity that travels through a wire and electricity in the form of lightning, together with the speed and magical properties of the email. And the Cherokee chose to use a natural word for electricity in this term, hence *lightning paper*.

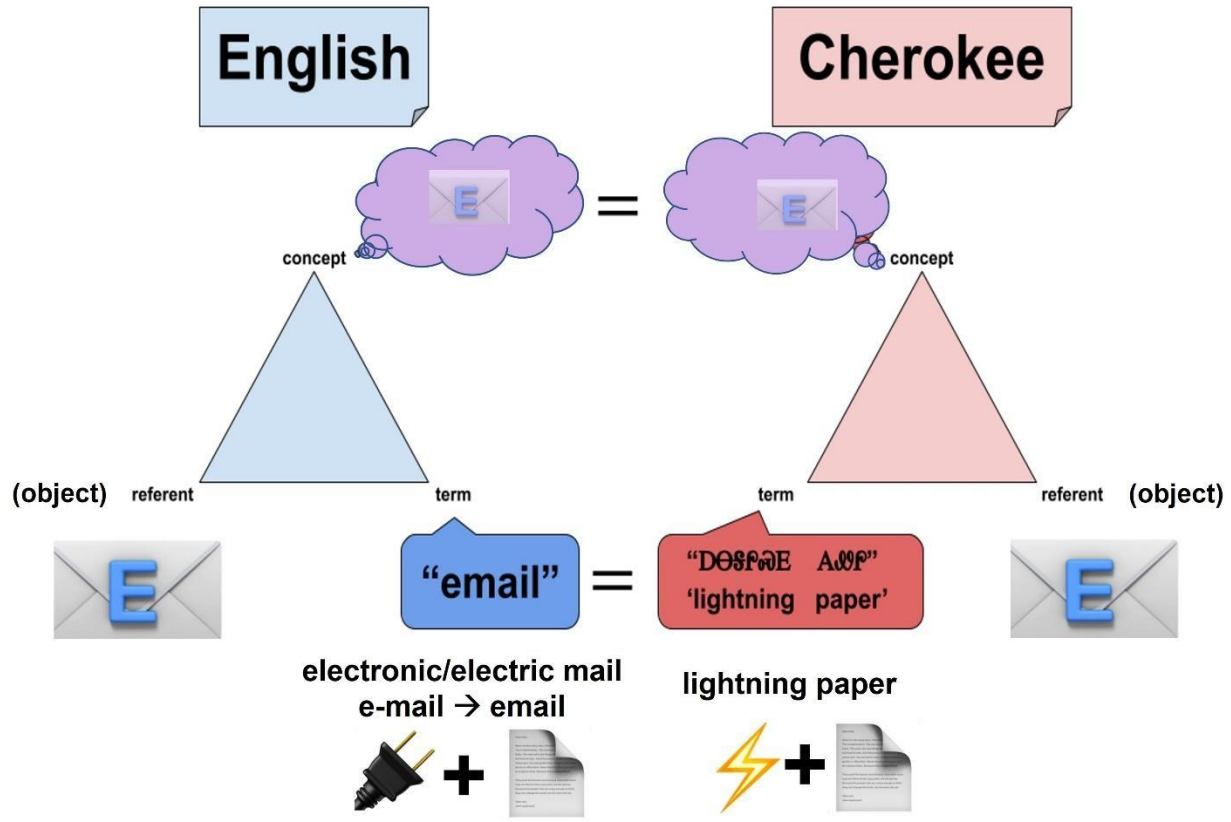


Figure 4: Triangle showing the relationship between concepts in two languages, where matching concepts support equivalent terms for the same referent object**.**

This approach to match term + idea (concept) is important because, as noted, some terms mean more than one thing, and some ideas can be represented by more than one term, which makes it dangerous to just try to match term to term independently. This approach makes it easier to match concepts in two or more languages, accommodate different forms in multiple dialects, and report alternative spellings.

A major reason for talking about bilingual terminology entries is that indigenous terms are often documented along with their equivalents in a second language, such as Cherokee and English. Although we generally avoid borrowing, if some terms have already been borrowed and are well established, you may want to keep them in your terminology list or add them later to your termbase. If possible, however, it is wise to

create your own terms using familiar words instead of borrowing. This approach helps empower your speakers to understand and to take control of the digitization of your language. The benefits also spill over into other areas as new materials are created. For instance, in the medical field, using familiar words helps health care workers to access information and respond and communicate quickly and supports successful outcomes.

6. DOCUMENTING YOUR TERMS

Term lists are useful to start with, but you will also need definitions, sometimes multiple definitions, for your terms, and as your list or lists grow, it becomes increasingly difficult to find information in your lists when you need it. You will eventually want to create *terminology entries*, especially if you want to document your terms together with equivalents in more than one language. A simple complete term entry (*term entry data model*) might look something like Figure 5.

Typical information to include in a term entry includes the following data items, usually called *data categories*:

- **Subject field**
- **Definition** of the concept plus a source, if you have one
- **Term**
- **Part of speech**
- **Context** in a sample sentence where the term is used in a typical way, plus a source

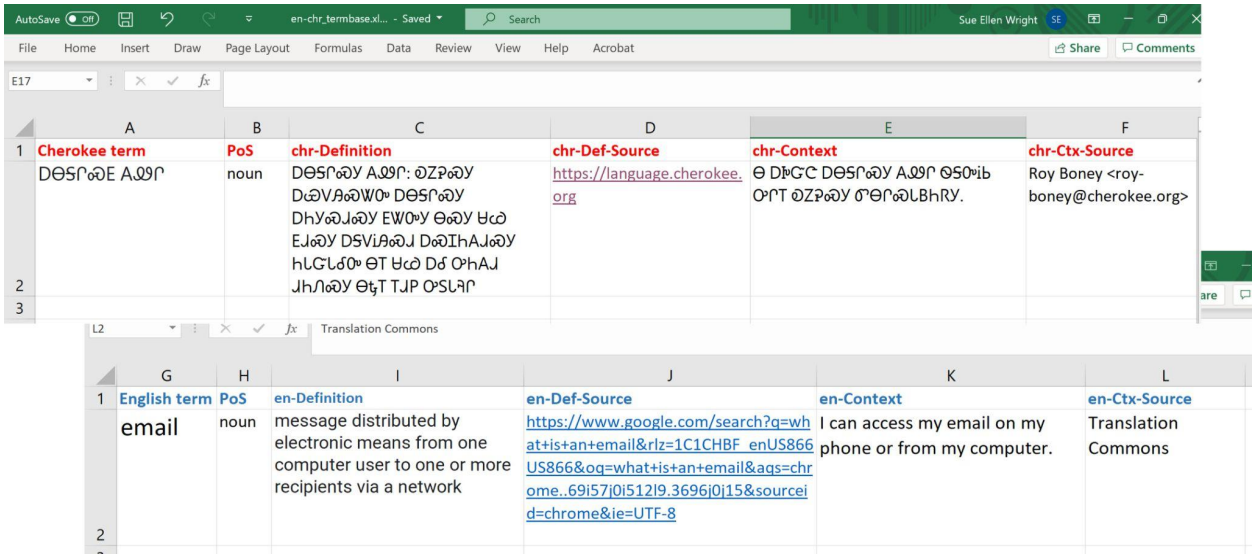
Subject field: digital communications

Cherokee term: ᎠᎾᎦᏞᏍᎬ ᎪᏪᏞ
Part of speech: noun
Definition: ᎠᎾᎦᏞᏍᎩ ᎪᏪᏞ: ᏫᏃᏞᏍᎩ ᎠᏍᏙᏗᏫᏍᎳᏅ ᎠᎾᎦᏞᏍᎩ ᎠᏂᎩᏍᎫᏍᎩ ᎬᎳᏅᎩ ᎾᏫᎩ ᏌᏫ ᎬᎫᏍᎩ ᎠᎦᏙᎥᏗᏫᏍᎫ ᎠᏫᎢᏂᎪᎫᏍᎩ ᏂᏓᏀᏓᎴᏅ ᎾᎢ ᏌᏫ ᎠᎴ ᎤᏂᎪᎫ ᎫᏂᏁᏍᎩ ᎾᏆᎢ ᎢᎫᏢ ᎤᏚᏓᏢ
Source: https://language.cherokee.org/
Context: Ꮎ ᎠᏲᏀᏳ ᎠᎾᎦᏞᏍᎩ ᎪᏪᏞ ᏧᎦᏅᎥᏏ ᎤᏢᎢ ᏫᏃᏞᏍᎩ ᏅᎤᏞᏍᏫᏓᏂᏒᎩ.
Source: Roy Boney <roy-boney@cherokee.org>

English term: email
Part of speech: noun
Definition: message distributed by electronic means from one computer user to one or more recipients via a network
Source:
https://www.google.com/search?q=define+email&rlz=1C1CHBF_enUS866US866&oq=define+email&aqs=chrome..69i57j0i512l9.2230j0j15&sourceid=chrome&ie=UTF-8
Context: I can access my email on my phone or from my computer.
Source: Translation Commons

Figure 5: Typical term entry layout

You can start by entering this information in a text file or a word-processing file, but it is smarter to use some sort of spreadsheet program to collect term entries. A spreadsheet file might look something like Figure 6. The two images in this figure are actually one long row in the spreadsheet. There are ways to create narrower spreadsheets, but putting each concept for both (or all) your languages in a single row makes it very easy to export your data to other systems later if you decide to build a database. Spreadsheets are good for collecting smaller numbers of terms, but they, too, become cumbersome as your collection grows. If you design your spreadsheet to match a termbase model, it will be relatively easy to export and import your data.



	Cherokee term	PoS	chr-Definition	chr-Def-Source	chr-Context	chr-Ctx-Source
2	ᎠᎾᏕᎵᏍᎬ ᎯᏊᏁ	noun	ᎠᎾᏕᎵᏍᎩ ᎯᏊᏁ: ᏫᏃᏊᏍᎩ ᎠᏍᏙᏱᏍᎩ ᎠᎾᏕᎵᏍᎩ ᎠᏂᎩᏍᎩᏁᏗ ᎤᏍᏆᏂᎪᏁᏍᎩ ᎬᏩᏍᎩ ᎤᏪᏅ ᎬᏩᏂᏍᎩ ᎤᏂᏣᏘᏳᏍᏗ ᎾᏍᎩᏯ ᏫᏃᏊᏍᎩ ᎤᎾᏕᎵᏍᎩ ᏗᏥᏂᎨᏒᎾ.	https://language.cherokee.org	ᎾᏂᏥᏣᏛᏅ ᎠᎾᏕᎵᏍᎩ ᎯᏊᏁ ᏫᏃᏊᏍᎩ ᎤᎾᏕᎵᏍᎩ ᏗᎾᏂᏥᏣᏛᏅ.	Roy Boney <roy-boney@cherokee.org>

	English term	PoS	en-Definition	en-Def-Source	en-Context	en-Ctx-Source
2	email	noun	message distributed by electronic means from one computer user to one or more recipients via a network	https://www.google.com/search?q=what+is+an+email&rlz=1C1CHBF_enUS866US866&oq=what+is+an+email&aqs=chrome..69i57j0i512l9.3696j0j15&sourceid=chrome&ie=UTF-8	I can access my email on my phone or from my computer.	Translation Commons

Figure 6: Term entry created as a single row in a spreadsheet program.

Subject field: The subject field category in your entry helps you organize your terms according to the topic. Always use stable values here (for instance: always use *Medical*, don't vary back and forth between *Medical* and *Medicine*). If you are consistent, it will be easy to filter your resource for all your medical terms.

Definition: Good definitions are usually as short as possible, but as complete as necessary. If the term names a thing, the definition usually indicates that the term "*is a*" broader concept with distinguishing characteristics (for instance, *wolf* (is a) *large carnivorous animal of the dog family)*. For some other terms, you might indicate "*is a part of*" a broader concept (for instance, *leaf* (is a) *part of a plant that is attached to the plant and supports the plant through a process called photosynthesis*). Verbs are often defined with an explanatory verb phrase (for instance, *walk: to move at a regular pace lifting each foot in turn*). Sections 7.3 and 7.4 discuss creating concept system diagrams. Once you have collected a sizable number of terms, you can begin to represent their relations to one another using these kinds of graphs. This process can lead you back to fine-tune your definitions to reflect relations among your concepts.

Term: At the beginning of your project, you may want to just start recording terms as you hear them. As you learn more about the grammar, try to figure out if there is something like a *base form* for your terms. In many languages, this will be the most common singular form of a noun, but the plural may be different, and in some languages, many forms function in different ways in the sentence. If there are many forms, you will want to decide on the best procedures for recording the forms in your language. The simple base form is often called a *lemma*. Examples of lemmas in English include “run” (not “runs” or “running”), “book” (not “books” or “booked”), “red” (not “redden” or “redness”-—which are actually different concepts).

Part of speech: The discussion in the last paragraph just talked about nouns and verbs, and there are small functional words, such as prepositions and conjunctions, adjectives that usually describe nouns, and adverbs that describe verbs, words, and phrases other than nouns. These categories are called “parts of speech.” This situation varies from language to language, so the grammar analysis of your language will cover these variations and others. A term’s part of speech and other grammatical notes will probably best be recorded in your monolingual dictionary, while the terms themselves, no matter what part of speech, will be recorded in your termbase. Unique words, particularly adjectives and adverbs, can be recorded in your terminology resource, however.

One thing to note is that sometimes nouns in one language may be verbs in another or vice versa. English often converts verbal nouns to action verbs, and other languages (Potawatomi, a Native American language, for instance) may treat animate nouns that are always in motion as if they were verbs (the word for *river*, for instance). You may need to give serious thought to how *part of speech* works in your language and adjust the way you record the information in your termbase.

Context: If you have a good example sentence using your term, record it in this field. It is especially useful to find contexts that indicate or at least hint at, the definition of the

concept, and that use any idioms or turns of phrase that are commonly used with the term. Using the term “book”, for example, the sentence “John buys books every time he comes to town.” does not hint at the meaning of “book” as well as “They sat together and took turns reading aloud from a book of stories about the fishermen of the South Java Sea.”

Note: The sample entry doesn’t show a *Note* field, but you can add one if you have additional information about any part of the entry.

7. OTHER ACTIVITIES

7.1 Creating a text corpus

Does your language have collections of existing texts? Some languages have ancient scripts that are only now being digitized, which will make them accessible and preserve them. Other communities have a rich tradition of oral “literature” – stories, histories, music, and poetry – that have been handed down over the centuries orally but are not written, and that may be in danger of loss if the numbers of speakers are dwindling, and younger generations no longer memorize the texts. As noted earlier in this document, you can start by audio recording these kinds of “texts” and then transcribe them for digitization. In either case, you will gradually collect a body of texts by recording and collecting oral and written information. Taken as a whole, your collection can comprise a *text corpus*. There is information on collecting corpora, preserving copyright, and creating archives in the Translation Commons Resources.

There is a wide variety of computer tools designed to manipulate and work with text corpora, including extracting terms. Term extraction tools are designed particularly to identify term candidates and possible definitions or contexts, working with major languages. If you have amassed a significant corpus, you may be able to enlist

programming help to customize one or more of these tools to use with your language. It is always important when working with texts that represent the cultural heritage of your group to protect and respect beliefs and traditions. Work carefully with your team to ensure that sacred or confidential materials are protected even as you work to preserve them for future generations.

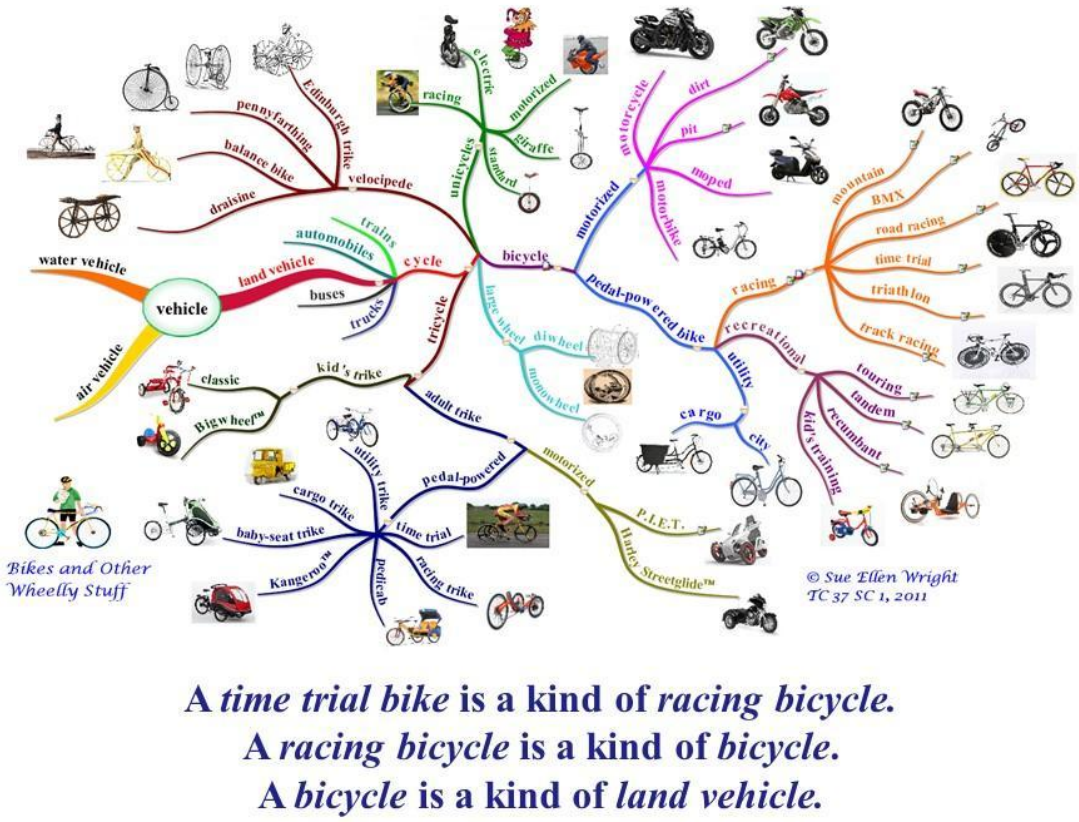
7.2 Adding translations to create a parallel corpus

The value of your corpus is significantly amplified if you can add a bilingual layer because this will allow you to expand your terminology resources through both human and machine translation (MT). Creating this bilingual layer, specifically by translating text from your language sentence-by-sentence into another language, so there is a strict match-up of translated sentences and chunks of sentences, results in the creation of a translation memory consisting of so-called *text segments*. The resulting texts then comprise a “parallel corpus”. To add MT to your set of digital resources, it is first essential to create a substantial set of translated documents that can be used to “train” your automated system to discover and extract consistently translated terms and segments. (See Zero to Digital, Machine Translation Guidelines.)

The first major step to implementing translation solutions is to adapt existing Computer-Aided Translation (CAT) tools to process your script and correctly display your language. Once human translators can use these tools to produce a Translation Memory (TM), you can begin to save matching segments consisting of source language sentences or phrases translated into another language. You can use these segments to save or reconstruct whole documents, which will enable you to build a parallel corpus of matching translated texts. As your corpus grows, you can enlist the assistance of major MT providers to add your language to the languages covered by their computer systems. Significant TMs can also be used to extract not just term lists, but also much of the data you need for full, meaningful terminological entries.

7.3 Creating generic concept systems

Terminologists strongly recommend that you organize your terms as you work along into concept systems. In describing definitions, we talked about "*is a*" and "*is a part of*" relations. Ask yourself: what is term A? If the answer is *Term A is a kind of Term B*, you can start building a generic concept system. For instance, many people are familiar with bicycles. If you were to create a concept system showing all the kinds of cycles you can think of, it might look something like Figure 7. This kind of concept system is a good way to learn to define concepts. If you look at the drawing, you can start a definition by writing: *a bicycle is a kind of land vehicle with two wheels that is frequently human-powered using foot pedals*.



A *time trial bike* is a kind of *racing bicycle.*
A *racing bicycle* is a kind of *bicycle.*
A *bicycle* is a kind of *land vehicle.*

Figure 7: Generic concept systems

7.4 Part-whole concept systems

You can also diagram concepts according to the way that the thing being defined is either divided up into parts. In the example in Figure 9, a *pedal* is part of the *drive train* of a *bicycle*.

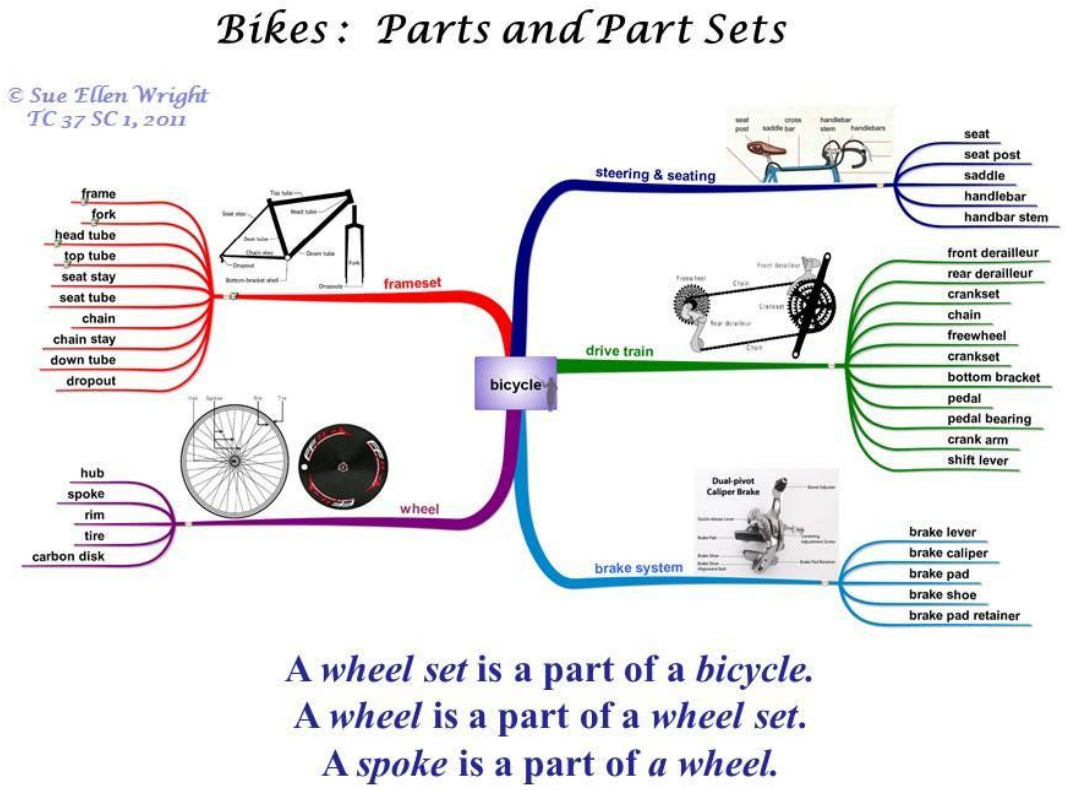


Figure 8: Part-whole concept systems

Drawing concept systems is also a good way to identify missing terms when you are collecting terms or creating term equivalents. Concept systems also often reveal that there are really multiple meanings for a term that you haven't thought of initially.

8. SHARING AND PUBLICIZING YOUR TERM RESOURCES

As noted above, the best way to grow your terminology collection and to make it accessible to a larger audience is to archive your data using a terminology management system that will be readily available on the web, for instance, the *Terminologue* software, which was originally developed by Fiontar & Scoil na Gaeilge for Foras na Gaeilge to manage *Téarma*, the National Terminology Database for Irish (téarma.ie). The Terminologue website provides tutorial information on how to set up termbases and start entering data.

Coordinate with your group to decide who will be primarily responsible for editing and administering your termbase. Make the termbase itself available to interested users to access and read via computers, phones, and other digital devices. Provide means for guests and interested parties to provide suggestions for new entries and edits for existing ones, but protect your data from unwanted editing or unintended changes.

REFERENCES

Example project:
Patyegarang, the Indigenous Australian languages education website
http://www.indigoz.com.au/language/gaps.html

This webpage provides excellent information on how to coin terms, based on the experience of an Australian community that has a great deal of experience coining new terms.

Tools:

Tools for managing text corpora

https://www.Corpus-Analysis.com provides a comprehensive list of 261 tools for use in corpus analysis.

https://tesolpeter.wordpress.com/a-brief-guide-to-corpus-analysis-tools/

http://inmyownterms.com/readings-tools-and-useful-links-for-corpus-analysis/

There are numerous collections of tools and the selection is constantly changing. Many are available for free, and some may make special allowances for projects like these.

Terminology management:

Terminologue is an open-source terminology management tool. The software is developed and maintained by the Gaois research group in Fiontar & Scoil na Gaeilge, Dublin City University. The software is copyright of Dublin City University and is available under the open-source MIT license. The lead developer is Michal Boleslav Měchura. Download the software from this website if you wish to install your own instance of *Terminologue*.

https://www.terminologue.org/docs/info.cs/

Term extraction tools:

https://termcoord.eu/free-term-extractors/

Terminological data categories:

There are many more data categories that can be used in terminology entries. You can find a collection of them, together with definitions and examples for use at https://www.datcatinfo.net.

Computer Aids for Translators (CAT Tools), both free and paid:

https://www.marstranslation.com/blog/top-free-and-paid-cat-tools

Good Firms. “The Top 10 Free and Open Source Computer-Assisted Translation Software”
https://www.goodfirms.co/blog/the-top-10-free-and-open-source-computer-assisted-translation-software

Translate5 Open Source Translation Tool.
https://www.translate5.net/en/translate5-open-source-translation-system-2