从零开始走向语言数字化系列：

术语指南

TRANSLATION COMMONS

2022-2032 | INTERNATIONAL DECADE OF Indigenous Languages

术语指南

主要作者：Sue Ellen Wright, Akil Iyer, Shuto Kato
其余作者和审校：Craig Cornelius, Alaina Brandt, Julie Anderson, Tex Texin
宣传：Leonidas Pappas
项目协调：Paula Cirilo

如果您对该指南有任何意见，敬请提出。
联系方式：krista@translationcommons.org

1. 前言

Translation Commons是一个非盈利性的志愿者组织，我们致力于为各类语言的数字化提供帮助、为语言专业人士提供指导，并为语言行业提供课程及资源。

语言数字化倡议（LDI）是我们进行的主要项目之一，该项目旨在帮助那些急需提高数字化能力的语言群体。全球有近六千种语言都没有被数字化，或者只被数字化了一小部分。而语言数字化倡议为这些语言群体提供了语言数字化流程指导。

我们与联合国教科文组织下属的2019国际土著语言年（2019 International Year of Indigenous Languages）行动一起，聚焦土著群体，探寻土著语言的数字化之路。土著语言使用者人数较少，因而在数字化世界中很难找寻到用土著语言所记录的内容，而这对于土著语言群体来讲是不公平的。语言数字化倡议的目标之一就是要保障这些人用土著语言获取网络信息的权利，从而确保他们能用母语参与全球网络活动，能够使用其母语版本的电脑软件，享受到现代电脑软件所带来的便利。本指南能够为这些语言群体提供数字化工具，提高其对语言数字化的理解，将土著语言带入到数字化世界，从而帮助他们加速语言数字化的进程，与此同时也保证了语言群体的自主权。除了制定本指南之外，我们还向这些语言群体提供教程以及组织开展研讨会，同时我们会向其介绍行业专家来提供标准化指导，以帮助这些语言群体完成语言的数字化。

本指南是*《从零开始走向语言数字化》*系列指南中的一本，该系列为语言的数字化实践提供了全方位的指导。本系列指南由语言技术和语言学方面的专家共同编写而成。目标受众是所有希望能在数字化系统中使用自己母语的语言群体。

语言的数字化能够扩展语言群体的交流渠道。*《语言数字化的益处》*中详细说明了语言数字化能如何造福土著群体及世界。

欲了解语言数字化详细流程，请见《从零开始走向语言数字化：如何让您使用的语言走进互联网》。Translation Commons网站中资源栏目发布了更多指南、演示和视频等语言数字化倡议的相关信息。

2. 本指南概述

本指南介绍了如何用已数字化的、未数字化的或未完全数字化语言记录术语。其概述了记录现有词汇和创建新术语的步骤，其中创建新术语可以帮助解决语义空缺现象。指南中提供了许多关于创建或更新术语的步骤指导，包括前期术语规划、术语资源收集、术语分析以及术语保存。上述流程能够帮助土著群体在必要时创造新术语，例如技术、医学、政治、法律方面的术语，或者土著文化、语言、历史中不存在的术语。扩大和更新这类术语词汇将有助于土著群体获取和利用知识资源，以便更好地融入数字化世界。

某些濒临灭绝的语言有悠久的书写记录方式，但有些语言则没有文字，仅仅只有口语一种方式。有些语言是 “族内语言”，使用群体不愿与外人共享。语言数字化倡导者需要记录已有的单词，并创建新的单词来表达数字化某一语言所需的一些事物、概念和想法。

这些指南的目标受众包括希望数字化其语言词汇的土著群体、单语种或多语种术语工作的学科专家以及支持语言数字化进程的组织。其目标是将收集和创建出的术语应用到口语、书面语以及最终的数字化层面。

本指南所记录的工作流程可以应用于任何领域或主题的术语工作。本指南给出的例子大部分为英语（部分已翻译为中文），尽管有些例子可能不适用于其他语言，但其中某些工作流程和模型是各个语言通用的，可以稍作修改以适用于不同的语言和情况。这些例子为读者提供了一个很好的起点，能帮助读者开始术语工作，以及意识到在进行术语工作时需要考虑的因素。

本指南介绍了一些语言数字化的基本知识、实践案例和资源（包括技术资源和其他资源等），来帮助土著语言群体开展术语工作，从而促进土著语言的保存、振兴和数字化工作。

3. 什么是术语和术语工作？

语言中带有特殊目的的单词、缩写甚至短语都应当算是该语言中的术语。目前科学和技术领域的术语已得到人们的广泛认可，但其实除此之外，人类所有的特定活动都会使用到术语，比如烹饪、狩猎或者是向他人阐释某文化的价值观等。开展术语工作的两个目的是：

- 为了收集词汇和术语：我们的目的之一就是为了收集和记录现有的词汇和术语。对于某些土著语言或者未数字化的语言来说，收集和记录现有的词汇和术语就是记载术语和前期术语规划的一部分，我们在这个过程中应尽可能多的收集单词和术语。在收集的过程中，我们不应只局限于术语本身，一个术语会对应多个不同的概念，我们应当把这些概念分属的主题领域写出来。

- 为了创造新术语来填补语义空缺：当我们发现想要讨论的某一概念在某一语言中还没有对应的术语时，我们应当首先尽量使用该语言中已有的词汇来创造新的术语。术语学家经常会将两个不同语言之间相似的概念进行对比，进而发现某一个语言所拥有的概念其实在另外一个语言中并不存在。而这种现象我们会将其称之为“语义空缺”。有时某一语言会通过借用外来词的方法来填补这些语义空缺。英语中*tsunami*（意为：海啸）和*Schadenfreude*（意为：幸灾乐祸），就是借用了法语的词汇来填补语义空缺。随着语言的发展，人们认为通常情况下利用某一语言现有的词汇创造新的术语比借用外来词创造新的术语会更好。

术语工作非常重要，因为它能促进文化的保护，也可以保障人们拥有平等的数字权利。当语言濒临灭绝时，术语工作可以用来记录濒临灭绝的语言并帮助其恢复。威尔士语和现代希伯来语就是很好的例子，这两种语言一度濒临灭绝，但目前已经成为两个重要的民族语言，这在一定程度上要归功于系统性的术语工作。

术语工作的目标是收集人们实际生活中会用到的词汇，并记录其含义。收集某一群体特定活动的一些术语可以给那些致力于保护土著语言的人士提供一个良好的工作基础，帮助他们积极地参与到土著语言的数字化当中，并维持术语意义和用法的准确性。上面我们说过术语工作还包括创造新术语，进一步来讲就包括为数字化系统创造新术语。对于许多正在

进行数字化进程的土著语言，或者已经准备好大量应用在电子系统上的语言来说，它们都需要为了现代数字化系统创建许多新的数字化术语。什么是数字化术语？举例来讲，例如*浏览器*和*工具*这样的为数字化系统所创建的术语就是数字化术语，这些术语可以从日常用语中借用。但像*菜单*和*字体*，甚至*电子邮件*这样的术语则无法直接从日常用语中借用，因为它们对于某一种语言来说是一种无中生有的、全新的概念。

4. 理想化的工作流程

下面的大纲为语言数字化过程中涉及的术语工作提供了一个理想的工作流程。所有步骤对于完成语言的数字化都很重要，但是请您注意，没必要严格按照下文的顺序来准备工作，因为有时候会行不通。下方的清单仅提供了一个大致合理的工作流程。

4.1 您准备好开始术语工作了吗？

- 开始之前请想一想，您的语言是否已经具备开启数字化工作所需的条件了呢？例如：
- 除了口语系统之外，这门语言是否有书写系统？
- 这门语言的书写系统是否已经囊括到Unicode标准中了呢？因为只有囊括到此标准中才能用电脑软件记录您的术语。
- 这门语言的书写系统是否已配置了对应的键盘输入方式呢？是否可以在电脑等电子设备上显示呢？
- 是否有人已经制定了这门语言的基本的语法规则呢？

如果以上任意一个条件还没有具备，那么请查看Translation Commons网站中的资源栏目，了解如何解决这些问题。可能上面的工作目前正在进行当中，目前还不具备进行语言数字化的条件，但是请您不要在此原地踏步，在此期间您还有许多其他事情可以做。尽管本指南中有些建议可能是重复的，但所有的建议都是围绕术语工作展开的。在您等待完成最初准备工作的时候，可能您的语言还没有书写系统，或者还没有完成书写系统的制定工作，但这不妨碍您开始录制音频的工作，并且您还可以在数字化过程完成之前就开始使用不断发展中的书写系统来手动记录信息。

4.2 您是否有足够的人力资源？

- 您是否召集了母语人士代表来参与您的项目？
- 您是否得到了前辈和其他团体领袖的支持？
- 在母语人士中，有哪些人对您所讨论的话题有所了解？比如说您要做有关于做饭方面的术语工作，那么您是否认识该领域的专家呢？请与这些领域专家们联系，让他们与您分享他们的工作内容，增进您对该领域的了解程度。
- 您有没有找到一些可以提供技术支持的人来处理工作过程中涉及的电脑任务？
- 您和您的团队知道如何识别和收集单词和术语吗？
- 组建您的团队可能需要一段时间，特别是如果您要召集不同学科领域的专家，而且还要帮助每个人理解您的目标，这都会花费更多时间。所以找寻队员的工作会是一个持续性的工作。建议您使用教程和其他材料来给队员提供培训，帮助队员识别术语并记录其含义。例如您正在编定一门语言的语法，或者正在对您的语言进行正字工作，那么请将语法和正字培训纳入您的培训当中，因为您的队员需要用到这些基本知识来完成工作。

当您工作的时候，一定要与母语人士一起来检查您的工作。这些母语人士是否同意您对这门语言的看法？比如在发音方面或单词表达方面是否他人有不同的见解？然后随着工作的进行，您可能就会发现越来越多的方言浮出水面。所有类似的问题都需要认真考虑，以便在您的团队中达成共识。

4.3 用您的语言创建文本并收集现有的术语

- 在做收集工作时，最好一次只做一个特定主题，但在工作过程中可能会出现其他内容，所以也请做好记录新内容的准备，最好在收集的过程中注明术语的主题领域。
- 开始之前先找找是否已经有与该主题相关的单词列表、词典或文本。如果有，可以从这些文本开始您的术语收集工作。
- 与您找到的该领域专家进行交谈，并对他们讲的故事和评论进行录音。
- 转写您的录音，必要时可人工转写录音。
- 识别重要的单词和术语，并列出清单，写出单词或术语对应的定义或解释。

- 如果您在记录某一领域的术语时，不知道您看到的一些物品、概念等对象应当叫什么，请尝试着找出它们的叫法，并把它们写进您的术语列表中。
- 您的术语是否准确？您是否针对某一术语与多个母语者进行过核对？这一术语是否有其他的变体形式？针对于这一术语，有没有其他的方言说法？
- 请您将与您正在做的主题相关的术语也都记录下来。
- 将您找到的文本和术语列表都存储起来，创建一个语料库。
- 一旦您有了庞大的语料库，有些电脑软件可以用来组织语料库或辅助您记录术语。

4.4 为语义空缺现象创造新的术语

- 是否有一些术语您需要用到，但是在您的语言中没有这些术语？
- 请您把有关的术语整理成一个概念系统，研究概念之间的关系，找到缺失的术语有哪些，在创建新术语后，请您为其撰写定义。（有关概念系统的内容详见7.3节）
- 您是否需要创建一整套术语，例如医疗术语或电脑术语？
- 请您多找几位母语人士进行核对，询问他们您创建的术语是否合理。

4.5 在共享词汇表中记录和维护术语

- 您是否已决定如何记录术语以及术语的含义了呢？
- 您是否已决定每一条术语下要记录的数据类型了呢？
- 您是否已选择好要使用的术语管理系统和记录术语的数据模型了呢？
- 是否每个术语所对应的重要的含义都记录了下来，并记录在了术语库（termbase）中呢？
- 术语的质量是否已经检查过了呢？
- 术语库是否已经和需要用到它的人共享了呢？
- 如果您的术语中是用于描述您在研究中的发现的，那么这就是所谓的*描述性术语*。
- 如果您的术语是在告诉人们应该如何正确使用术语，那么这就是所谓的*规范性术语*。
- 人们在说话的时候，如果提到某个术语，人们所指的概念是否都是术语库中给出的那个概念？

- 术语是否需要随着时间的推移而更新呢？

5. 将指南应用到实践当中

5.1 使用流程图

图1所示的流程图对第4章所提出的问题进行了一个简单的概述。请您注意，这张流程图只是一个简化版的图，您未必要严格按照流程图的顺序去做，流程图仅供参考。毕竟每个项目都不一样，都有自己的特殊性。所以请您在工作的过程中保持灵活性和创造性。

请注意，图1右列中从上到下第2项（将现有的字母和语言中不同的声音对应起来），在做对应的时候，建议您首选使用字母文字，字母文字是一种代表声音的文字，文字不代表任何实际含义，只代表声音。另外一种选择是使用语标文字，语标文字是指一种代表概念或想法的文字，其文字不代表它的发音方式。例如，英文、法文和阿拉伯文就是字母文字；而中文和日文就是语标文字。

寻找队员肯定应当是第一步要做的事情，一旦您找到了队员，请您尽快和他们一起开始工作。在您寻找队员并且获取队员信任、准备开始合作的同时，您也可以开始着手录音工作，为您的语言录制音频。即便字母系统和书写系统还没有开发完，您也可以开始着手录音工作。在语言数字化进程完成之前，您可以使用正在开发的字母系统去手动录入信息。随着越来越多的人参与到工作当中，并且开始使用手机和其他数字工具来数字化这门语言，您就可以开始采用数字媒体来收集信息了，这可以帮助您接触到更为庞大的语言群体，并从中收集信息。

这将节省您宝贵的时间，并为项目提供初始信息。随着项目的成熟，这些信息可以用于创建模型，以便开展后续的活动。这些前期工作可以帮您收集可供验证的例子，可供您后续申请国际语言代码标准ISO639、为语言进行注册，或进行Unicode认证时使用。随着项目的推进，请记得拿出流程图再检查一下，是否有些以前做不了的工作现在可以开始做了，是否要赶一赶进度。

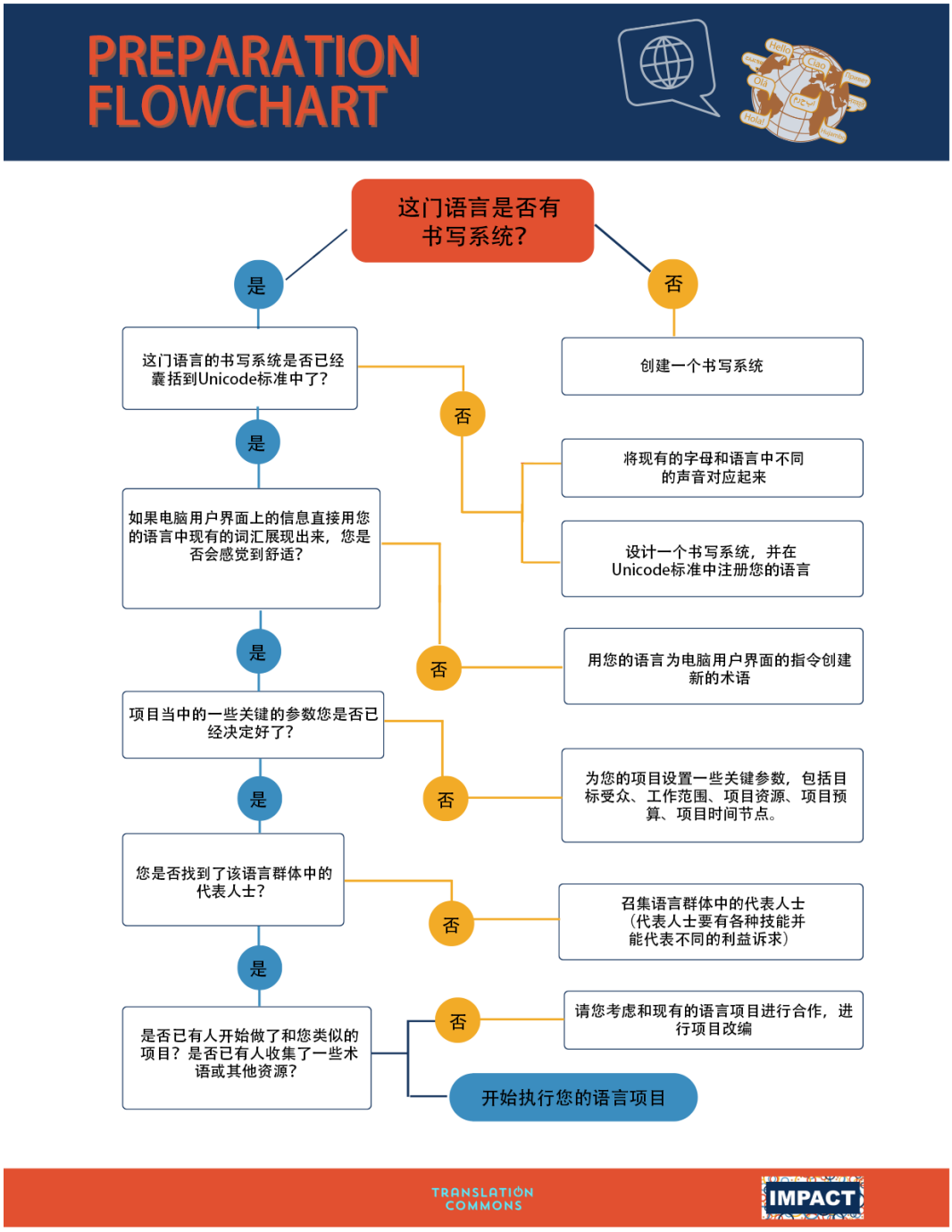


图1：项目准备流程图

5.2 传统词典学和新兴术语学

利用传统词典学所编纂的词典会将单词写在条目中，与该单词相关的所有含义（定义）都包含在同一个条目中，如图2所示。

这种传统的词典在您只使用一种语言进行工作时是非常有用的，因为您可以把有关于某个单词的所有信息记录在一起。传统词典可能还会包括语法信息，并记录许多小词（在语料库语言学中有时被称为噪音词），如冠词、介词、连词以及一些其他的日常词汇或者专业性词汇等。

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

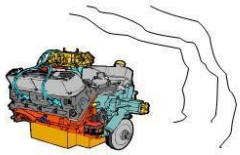


图2：传统词典中*rattle*这个词的词典条目

然而，如果您想编写一本双语词典或者多语词典，或者如果您想记录许多方言单词或同义词的时候，采用传统词典学来进行编纂工作就会出现许多问题。当您在另一种语言中寻找某个单词的对等词时，传统词典条目中所对应的每一个含义都有可能会对应一个不同的单词。并且单词不同的拼写形式或同义词在传统词典中都需要单独再重新写一个条目。这就是为什么今天大多数译者、双语作者和研究人员、从事语言规划的人员使用与“传统词典学”相对应的“新兴术语学”来记录词汇信息。

5.3 我们如何识别术语？为什么术语工作很重要？

使用术语学方法来记录语言就是指我们为一个词的每个不同含义都创建一个单独的条目。这听起来可能很复杂，举例来说如果我们按照这样的方法编纂，那么图2中的1个条目将变成11个条目。但这样做就能很方便的把一个单词的每个含义所对应的同义词、方言说法、以及外语翻译都标出来，也就是说，我们能准确地为这个单词所代表的每一个概念或意义找到对等词。事实上，这种方法是非常有必要的，这样做的好处之一是能让翻译工作变得更加准确，而除此之外，通常我们在进行语言数字化的过程中，或者讨论某一特定领域的过程中（如医学）经常会遇到语义空缺现象，使用这种方法可以帮助我们快速找到为这个领域所创建的新术语。

为了弄清楚一个特定的想法或概念与术语之间的关系，我们常常画一个语义三角形，用来展示（1）该术语在我们头脑中所反映的*概念*，（2）该术语的*所指对象*（有时把它称为*参照物*，或*现实世界中所对应的对象*），以及（3）该术语本身。如果我们在讲话或者写作的时候能够用到该术语，而且听我们讲话的人和看我们作品的读者都能立即在脑子里形成和我们脑子里一样的概念，那么我们就称这是个“有效的”术语。当我们使用一种以上的语言进行交流时，重点是要保证不同语言之间的某个词所对应的现实世界中的东西（也就是所指对象）要是相同的或者相似的，要保证讲不同语言的每个人对这个词的理解都应当和我们相同。只有这样，我们才能确定A语言中的一个术语是否和B语言中的某个术语是对等的。图3就展示了我们刚刚所说的语义三角形。

当我们比较不同的语言时，语义三角形就非常有用。我们在认识一个术语的时候，首先会在脑子里想这个术语在原语言中所对应的概念。然后我们会在目标语言中寻找和它对等的概念。如果在目标语言中没有对等的概念，这种情况就是典型的语义空缺现象。要解决这种现象，我们要么创建一个新的术语，要么借用其他语言的单词来描述这个概念。理想情况下，做术语工作最好能找到目的语中的现有词汇来描述这个概念，而不是先考虑从其他语言中借用词汇。*树*这个术语就是一个很好的例子，因为美洲原住民语言切罗基语中本就有一个现成的词来表达*树*：ᏭᎬᏆ（*quigvi*[kʷigə̃i]）。然而，有很多英文词汇在切罗基语中找不到对等词汇。假设我们想为英文当中的*email*（电子邮件）这个单词所代表的概念在切罗基语中创造一个对等概念的术语。即便我们在创建术语的时候想避免借用英文词

email，我们也最好不要直接用切罗基语拼读出一个发音很像英文中“E-mail”的词汇。

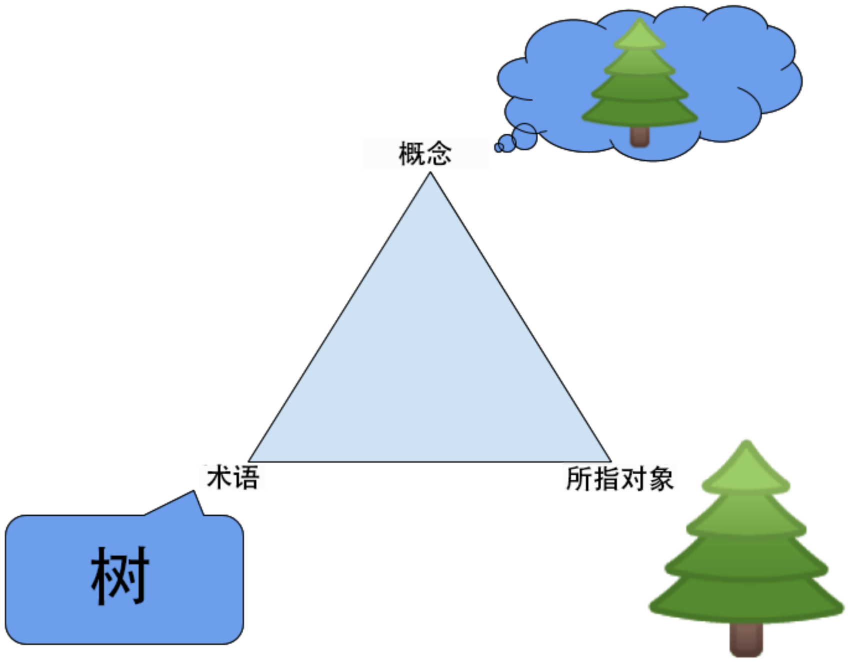


图3：“语义三角”，“树”这个单词和现实生活中的（所指对象）树相匹配，使得听到或读到这个词的人脑海中能在出现树的概念和意义。

在进行术语工作中，不论是英语使用者还是切罗基语使用者是否有使用过电子邮件的经历，*电子邮件这个词的自身*应当对两者来说是一个相同的*东西*。*电子邮件*所对应的现实世界中的东西（或称为所指对象），在两个不同语言的语义三角形中永远都不会改变。有些时候，一些已知的物品，比如树，在两个说话者的眼中可能会有一些不同（比如一棵橡树和一棵松树当然长的不一样）。但是在我们命名或定义这个已知物品的时候，我们应当关注的是它们之间相同的特征。刚刚我们说到所指对象应当是相同的，那么主要会出现不同的是术语，术语很有可能会反应说话者、读者和作者不同的经历。英文单词*email*（电子邮件）是由*electronic mail*（电子化了的邮件）缩写而来，与之相比，切罗基语中的电子邮件一词**ᎠᎾᎦᎵᏍᎬ ᎪᏪᎵ**（*anagalisgv goweli*[anagalisgə̃ goɰeli]）如果直译的话就是*lightning paper（闪电+纸）*。尽管说英语的人也见过闪电，但他们看见电子邮件中“电子”这两个字首先想到的是日常我们使用的电流，而不是先想到闪电，因为他们每天都和电力系统中流过的电流打交道（图4中用电插头表示），知道这是电子邮件能够发来的原因。当然，现代切罗基人也懂得电子邮件的原理，人们的知识已经随着现代社会的进化而进化了，但是切罗基语这门语言本身并没有参与到进化过程中。在电子邮件这个词中

，英语当中的“电流”的概念与切罗基语当中的“闪电”的概念是相对应的，英语当中email（电子化了的邮件）所体现出的电子邮件“快速”的特性，与切罗基语中ᎠᎾᎦᎵᏍᎬ ᎪᏪᎵ（闪电+纸）所体现出的电子邮件“神奇”的特性是相对应的。切罗基人选择了一个描述大自然的词（闪电）来表达电这个术语，因此切罗基语中的电子邮件这个术语就采用了*闪电+纸*。

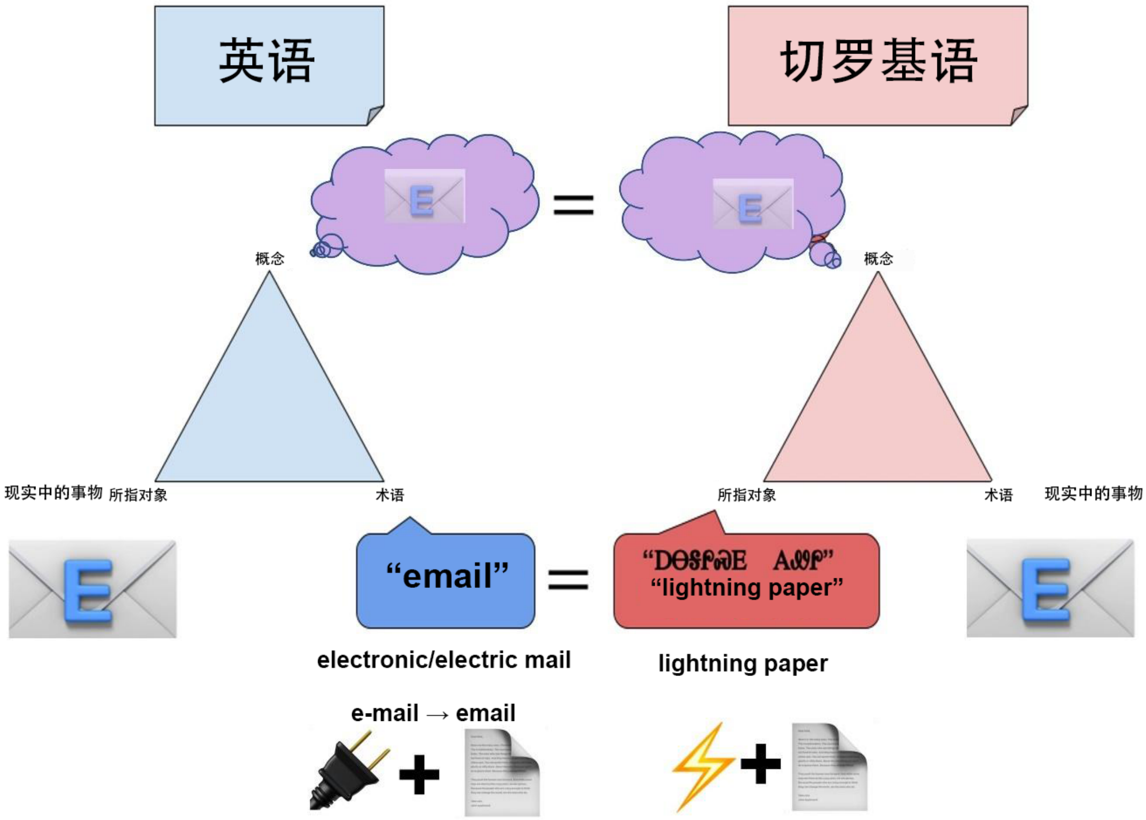


图4：两个语义三角形展现了两种语言中电子邮件这个概念的关系，电子邮件在两种人脑海里的概念都是相同的，从而我们就创造出了两个语言当中对等的术语，并且他们的所指对象也是相同的。

这样的方法，也就是把术语和想法（概念）匹配起来的方法非常重要，因为就像先前指出的，有些术语不止有一个意思，反过来一个意思也可以用多个术语表示，所以如果只把两个语言中的术语对应起来，忽略背后的所指对象和该术语在我们头脑中所反映的概念，就会造成许多错误。这样的方法能更容易把一个术语的不同意思分别与其在不同语言中的翻

译对应起来，也更能适应某一语言中不同的方言形式所带来的差异，以及更适应同一单词的不同拼写方式。

我们在上面讨论了双语术语的问题，讨论这个的主要原因是因为很多土著语言中的单词经常会和其他语言中的对等单词记录在一起，比如说切罗基语的单词经常会和英语中对应的单词记录在一起。总的来说，我们应避免借用词汇，但如果有些词已经借来用了，并且已经被广泛使用了，那么您可以直接将其保存在您的术语列表或者术语库当中，不用再创建新的术语。不过，如果有可能的话，最好还是不要借用词汇，最好用您熟悉的语言再重新创建一个术语。重新创建术语有助于其他人理解这个词汇的意思，并且还可以保证您所在的语言群体掌握语言自主权。随着更多新术语的出现，类似的好处还会在其他领域展现出来。比如在医疗领域，如果这些新术语是用熟悉的词汇创建出来的，而不是直接借用外来语的，就能够帮助医护人员快速获取信息，快速应答和交流，最终提高治疗结果。

6. 记录您的术语

在最初的阶段，用术语表来记录术语会是一个不错的选择，但是您还需要写术语的定义，有时候可能定义还不止一个。随着您的术语和术语表越来越多，找到您需要的信息也将变得越来越难。所以最后您可能需要创建*术语条目*来解决这个问题，尤其是当您想要把术语和该术语在其他多种语言中的对等词汇一起记录下来时，这将会非常有用。一个简单且完整的术语条目（*术语条目的数据模型*）示例请看下方的图5。

典型的术语条目信息包含以下几点数据，这些通常被称作*数据分类*：

- **主题领域**
- **定义**，用于记录某一个概念的定义，以及该定义的来源（如果有来源的话）
- **术语**
- **词性**
- **语境**，用于提供例句来说明某个词的典型用法，以及例句来源

主题领域： 数字通信

切罗基语术语： ᎠᎾᎦᏢᎾᎠᎬ ᎪᏯᏢ

词性： 名词

定义： ᎠᎾᎦᏢᎾᎠᎩ ᎪᏯᏢ: ᏫᏃᏢᎾᎠᎩ ᎠᏍᎪᏙᏓᎾᎠᎳᏫ ᎠᎾᎦᏢᎾᎠᎩ ᎠᏂᎩᎾᎠᏗᎾᎠᎩ ᎬᎳᏫᎩ ᎾᎾᎠᎩ ᎮᏫ ᎬᏗᎾᎠᎩ ᎠᎦᏙᏁᎵᎾᎠᏗ ᎠᎾᎢᏂᎪᏗᎾᎠᎩ ᏂᏓᏀᏓᎾᏫ ᎾᎢ ᎮᏫ ᎠᎴ ᎤᏂᎪᏗ ᏗᏂᏁᎾᎠᎩ ᎾᏆᎢ ᎢᏗᏢ ᎤᏚᏓᏃᏢ

来源: https://language.cherokee.org

语境： Ꮎ ᎠᏲᏀᏀ ᎠᎾᎦᏢᎾᎠᎩ ᎪᏯᏢ ᏫᎦᏫᏆᏏ ᎤᎢᎢ ᏫᏃᏢᎾᎠᎩ ᏫᎾᎾᎢᏴᏂᎡᎩ.

来源: Roy Boney <roy-boney@cherokee.org>

英语术语： email

词性： 名词

定义： 通过网络由一个电脑用户向一个或多个接收人以电子方式发送的信息

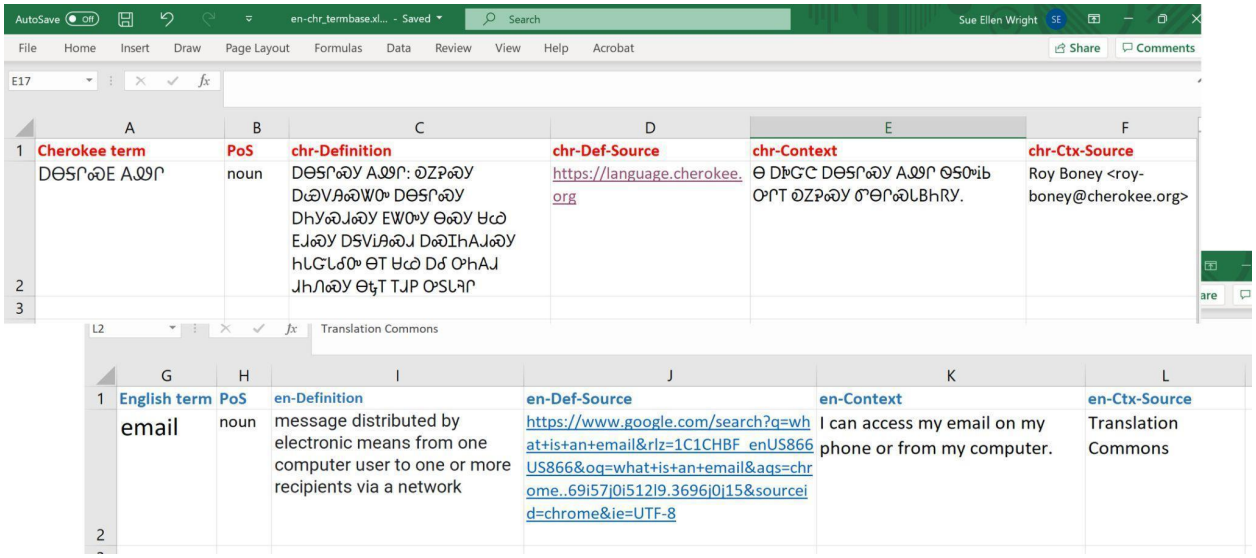
来源: https://www.google.com/search?q=define+email&rlz=1C1CHBF_enUS866US866&oq=define+email&aqs=chrome..69i57j0i512l9.2230j0j15&sourceid=chrome&ie=UTF-8

语境： I can access my email on my phone or from my computer.

来源: Translation Commons

图5：典型的术语条目布局

您可以在开始的阶段使用支持txt格式的编辑器（例如Windows上的记事本），或者文字处理软件（例如Word或WPS）来收集术语条目信息，但我们更推荐您使用电子表格软件（例如Excel或WPS）来收集术语条目信息。电子表格文件请见图6。图6中的两个图像实际上是电子表格中的一整行，只不过该行很长，所以截了两张图。这样创建出来的表格横向会很长，有一些方法可以避免这个问题，但是如果您决定后续要构建术语库，那么把某个单词对应的所有不同语种的翻译放在同一行将有助于您更容易地把数据导出到其他系统当中。电子表格适用于收集数量较少的术语，但随着您收集的术语越来越多，电子表格也会越来越冗长，不便于使用。但如果您能将电子表格布局设计成适合术语库模型的布局，相对来说将来在导出和导入数据时候就会容易很多。



	A	B	C	D	E	F
1	Cherokee term	PoS	chr-Definition	chr-Def-Source	chr-Context	chr-Ctx-Source
2	ᎠᎾᏕᎶᏍᎬ ᎠᏯᏁ	noun	ᎠᎾᏕᎶᏍᎩ ᎠᏯᏁ: ᏫᏃᏣᏍᎩ ᎠᏍᏙᏱᏍᏍᏙᏯ ᎠᎾᏕᎶᏍᎩ ᎠᏂᎩᏍᏍᏙᏯ ᎬᏩᏍᏍᏙᏯ ᏱᏩᏍ ᎬᏩᏂᎯᏍᏗᏍᏙᏯ ᏂᏓᏩᏓᏱᏍ ᎾᎢ ᏱᏩᏍ ᎠᎴ ᎤᏂᎪᏩ ᏗᏂᏁᎾᎩ ᎾᎿᎢ ᎢᏗᏢ ᎤᏕᏓᎾᏁ	https://language.cherokee.org	Θ ᎠᏴᏩᏣ ᎠᎾᏕᎶᏍᎩ ᎠᏯᏁ ᎾᏕᎤᏰ ᎤᎢ ᏫᏃᏣᏍᎩ ᎤᎾᎡᏍᏓᏞᏴᏂᎡᎩ.	Roy Boney <roy-boney@cherokee.org>
3						

	G	H	I	J	K	L
1	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

图 6：在电子表格程序中创建的一整行的术语条目。

主题领域：可以帮助您对术语按照话题进行分类管理。在写主题领域的时候，要保持每次使用的单词是一致的（例如：有关于医疗的词汇要在主题领域始终写*医疗*两字，不要一会儿写 *医疗* 一会儿又写 *医药* ）。保持主题领域用语的一致性可以保证后续筛选医疗术语会非常容易。

定义：好的定义通常尽可能短，但要尽可能完整。假设术语A所描述的是事物，那么对该术语A的定义通常会是：术语A“*等同于*”术语B（此时术语B是术语A的上层概念），但术语A带有一些自己独有的特征【例如，*狼*（是）犬科的*大型食肉动物*】。在上述的例子中，犬科对于狼来说就是上层概念，而大型食肉动物则是狼带有的一些自己独有的特征。对于其他一些术语，比如术语A，我们可能会将其定义为：术语A是术语B的*“一部分”*【例如，*叶子是植物的*（一部分），*它附着在植物上并通过光合作用为植物的生长提供养分*】。动词的定义通常会用到解释性的动词短语（例如，*走路：意为以固定的速度，依次抬起每只脚向前移动*）。7.3和7.4小节讨论了如何创建概念系统图的问题。一旦您收集了大量的术语，您就可以开始用概念系统图来表示术语和术语之间的关系。这个过程可以帮助您重新调整您对术语的定义，以便反映不同概念之间的关系。

术语：在您的项目开始时，一旦听到术语，您可能就想把它们记录下来。随着您对语法了解的加深，请尝试弄清楚您的术语是否有类似*基本形式*的东西。对于许多语言来说，所谓

的基本形式就是名词最常见的单数形式，其复数形式可能和单数形式不同。并且在某些语言中，某个术语的不同形式会在句子中以不同的方式发挥作用。如果您所记录的术语是包括不同形式的，您就需要考虑一下以何种流程来记录这些不同形式会比较好。某个术语最简单的基本形式通常称为*词目*。英文中词目的例子包括“run”（不是“runs”或“running”）、“book”（不是“books”或“booked”）、“red”（不是“redden”或“redness”，每一个单词后面跟着的两个单词和原单词是实际上不同的概念）。

词性：上面我们只讲了名词和动词，但还有一些小的功能词，比如介词和连词，通常用来描述名词的形容词，以及用来描述动词、名词和非名词短语的副词。上面这些被称为“词性”。词性分类因语言而异，所以对您的语言进行语法分析非常重要，语法分析可以把所有其他的词性分类情况都展现出来。如果您在编写的是单语词典，那么词性以及其他的语法信息最好都要写进去，而如果您要编写的是术语库，那么无论该术语是什么词性，都需要将其记录在您的术语库中。比如术语库中虽然常见的是名词和动词，但是某些形容词和副词，如果有比较特殊的含义，也可以记录在您的术语资源中。

需要注意的一点是，有时一种语言的名词可能在另外一种语言中是动词，反之亦然。英语经常将描述动作的名词转换为动词，而其他语言（例如美洲原住民语言帕塔瓦米语）可能会把有生命的或者会动的名词当作动词来用（例如，*河流*这个词，本身是个名词，但是河流会流动，所以帕塔瓦米语会把河流这个词当作动词使用）。您需要认真思考*词性*在您的语言中到底分为几种，并且是如何起到语法作用的，并根据您这门语言的词性情况调整您在术语库中记录信息的方式。

语境：如果针对于某一个术语您有对应的例句，那么请写在语境这里。最好能找到一个例句能够表明该术语的定义，或者至少能暗示出该术语的定义。并且最好能找到包含这个术语的一些习语，或者包含这个术语的常用短语表达。我们用“书”这个词来举个例子。“小明每次去镇上都会买书。”这句话就没能很好暗示出“书”的定义。再看后面这句话“他们坐在一起轮流大声朗读书上的故事，这个故事是有关于一个渔夫的故事。”这句话就很好的暗示出了“书”的定义，至少让人们明白了，书是一个上面记录着故事的东西。

请注意：图6的电子表格中没有设计*注释*部分，但如果您对于某一条术语有其他的额外信息补充，您可以自行添加注释一栏。

7. 其他需要做的事情

7.1 创建文本语料库

您这门语言可能会有一些现存的文本作品，是否已经有人收集过这些现存的作品了呢？一些语言有着更为古老的文字形式，而直到现在这些古文才被数字化。不过，这可以更好的保护古文，并且数字化后人们获取到古文资料会更容易。有一些语言群体可能会有丰富的“口头文学”，包括故事、历史、音乐、诗歌等，这些作品在数个世纪以来以口头的方式传承到一代又一代人身上，他们没有文字来记载这些作品。这样的作品会面临很大的失传风险，因为语言群体的数量在不断减少，年轻一代的人们不再记得这些作品。正如前文所指出的，您可以先对这些类型的“文本”进行录音，然后将其转录成文字，以便后续数字化。无论是以上哪种情况，您都会通过文本作品和口头作品收集到大量的文本。把以上收集的大量的文本集合在一起，就可以形成一个*文本语料库*。如果您想了解有关于如何收集语料库，保护您的版权，以及对文本进行归档，您可以参考在Translation Commons网站中 资源栏目中的有关信息。

目前已有各式各样的电脑软件可以用来处理文本语料库的有关工作，其中也包括术语提取工作。术语提取工具是专门用来识别术语、定义、语境的工具。该工具一般支持处理一些世界上主要的语言。如果您已经形成了一个体量较大的语料库，那么您可以寻找一些编程人员来帮助您制定一款或者多款能够处理您这门语言的电脑软件，来帮助您处理语料库。这些文本往往是您所在语言群体的文化遗产，在处理这些文本的时候，保护和尊重群体的信仰与传统始终是非常重要的。请您在与团队进行合作的时候更加认真一些，确保一些神圣的文本材料或者机密的文本材料能够得到有效的保护，因为您的这项工作会为您的子孙后代留下宝贵的遗产。

7.2 添加已翻译的文本以创建平行文本语料库

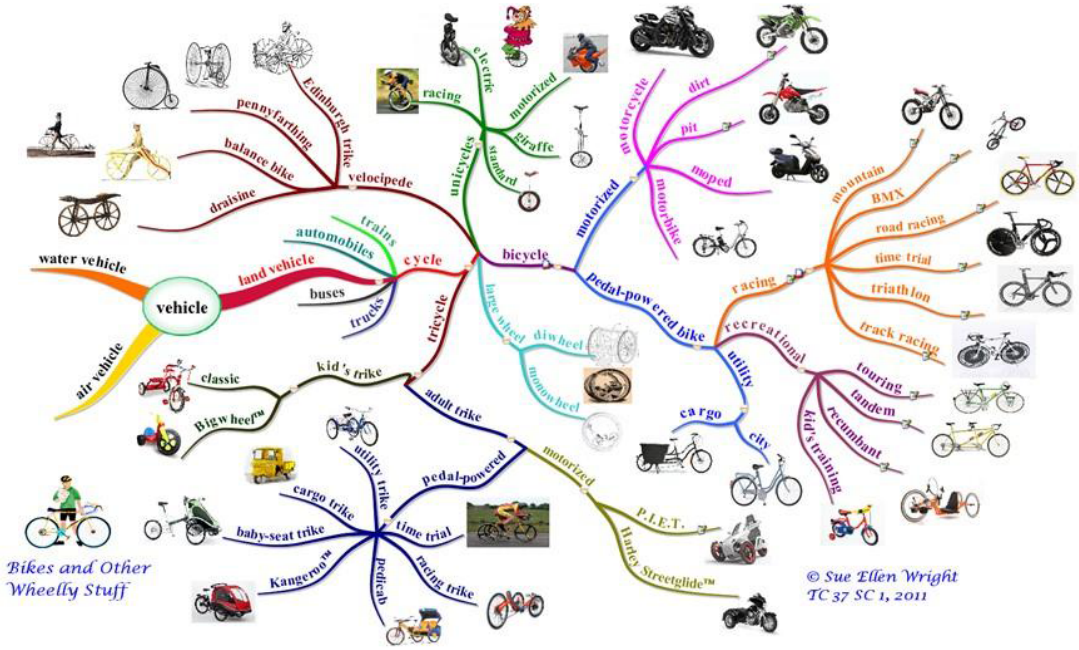
如果您可以把您的语料库做成双语的语料库，那么您的语料库将会有更大的价值。双语语料库可以拓展您获取术语的渠道，如此您便可以通过人类译员或者机器翻译（MT）来获取术语。若您想创建双语语料库，具体来说就需要把以您语言记录的文本逐句翻译成另一种

语言，然后您就能得到原文和译文匹配非常精准的语料，这种匹配精准的语料实际上就是翻译记忆库，这个记忆库当中包含许多*句段*。这样的双语语料库也叫做“平行文本语料库”。若您想使用机器翻译来辅助您进行语言数字化，首先必须要有大量已经翻译好的双语文件，这些双语文件会用来“训练”机器对您的语言进行学习，以便后续机器能够帮您提取出翻译好的术语和句段（有关训练机器学习的内容详见《从零开始走向语言数字化，机器翻译指南》）。

着手解决您的语言没有机器翻译问题的第一步就是要改造现有的计算机辅助翻译软件（CAT）来处理您的语言并且能在电脑屏幕上正确显示出您的语言。一旦人工译员可以使用这些计算机辅助翻译软件来创建翻译记忆库（TM），您就可以把原文和译文对齐了的平行文本保存在您的翻译记忆库当中了。您可以使用这样句段对齐的方式来保存您以后所有的文本语料，或者您也可以把现有的文本语料全部转化成这种句段对齐的格式。如此，您可以建立起一个带有原文和译文的平行语料库。随着您的语料库数据变得越来越多，您就可以开始寻找市面上一些主流的机器翻译公司来帮您创建适用于您的语言的机器翻译引擎。拥有数据量相当大的翻译记忆库不仅仅能够帮助您提取术语，并且可以帮您提取出用于创建一个完整术语条目所需的信息。

7.3 创建通用的概念系统

根据术语专家的建议，我们强烈建议您在工作中把您所创建的术语组织成一个概念系统。在第6小节的开头部分，我们在谈论术语条目的定义时，我们谈到了两种常见的定义描述方式，即术语A*“等同于”*术语B，以及术语A是术语B的*“一部分”*。请您问问自己：术语A是什么？如果答案是*术语A等同于术语B*，那么您就可以开始建立一个通用的概念系统。例如，许多人都很熟悉自行车。假如您要创建一个自行车的概念系统，把所有您能想到的自行车都囊括进去，那么这个概念系统看起来就会像是图7一样。这种概念系统是学习如何来定义概念的一个好方法。如果您看着图7，您就可以尝试写出一个自行车的定义，比如：*自行车，是一种有两个轮子的陆地交通工具，通常由人用脚踏板驱动*。



计时赛公路自行车是公路自行车的一种。
公路自行车是自行车的一种。
自行车是一种陆地交通工具。

图 7：通用概念系统

7.4 部分—整体概念系统

您也可以按照事物划分成不同部分的方式来创建概念系统。例如，在图8中，*踏板*是*自行车传动系统*的一部分。

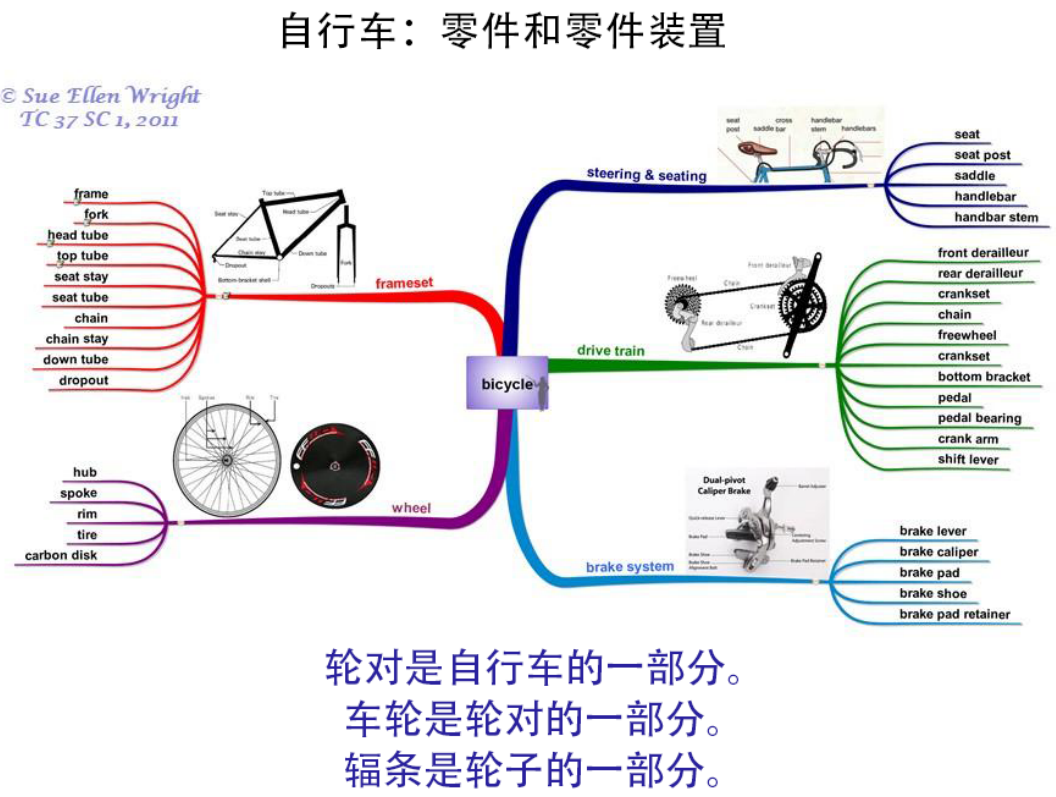


图 8：部分—整体概念系统

当您在收集术语或者是翻译术语的时候，画出这样一个概念系统图也有助于帮您找到漏掉的术语。画出概念系统图也常常可以帮助您意识到一个术语的多种不同含义，这些不同的含义一开始您可能并没有想到。

8. 分享和宣传您的术语资源

如上所述，快速增加您的术语数量并能让更多人用上您的术语的最好方法就是用术语管理系统，将您的数据存到该系统中，并且要保证该系统可随时在线上使用。例如，*Terminologue*软件，它最初由爱尔兰都柏林城市大学Fiontar & Scoil na Gaeilge学院为Foras na Gaeilge开发，用于管理爱尔兰国家术语数据库*Téarma*（网址：téarma.ie）。Terminologue网站提供了如何建立术语库并且录入数据的教程。

请您与您的组员协商决定谁来主要负责编辑和管理术语库。并且要确保感兴趣的用户可以通过电脑、手机等其他电子设备访问您的术语库。请记得在术语库中提供一个反馈和编辑的渠道，方便术语库访客给您提出新术语的建议，并且能够编辑优化现有术语条目。但要记得对您的术语数据做好保护，以防有人故意乱改术语等不好的情况发生。

参考资料

语言数字化项目的例子：
澳大利亚土著语言教育网站Patyegarang
http://www.indigoz.com.au/language/gaps.html

上述网页提供了一个非常好的创建术语的教程，教程是根据澳大利亚一个语言社区经验所编写的，该社区拥有大量创造新术语的经验。

工具：

管理文本语料库的工具

https://www.Corpus-Analysis.com 提供了用于语料库分析的 261 种工具集合。

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

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

这个网站收集了许多工具，并且这个网站的收录的工具还会不断更新。许多工具是免费提供的，有些工具尽管不是免费的，但可以为您的项目提供特殊的帮助。

术语管理工具：

Terminologue 是一个开源的术语管理工具。该软件由都柏林城市大学Fiontar & Scoil na Gaeilge学院的Gaois研究小组开发和维护。该软件的版权归都柏林城市大学所有，可在开源 MIT 许可下使用。首席开发人员是Michal Boleslav Měchura。如果您想安装*Terminologue*软件，请从这个网站下载软件。

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

术语提取工具：

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

术语条目中的数据类别：

还有许多数据类别可用于术语条目中。您可以在https://www.datcatinfo.net找到这些数据类别，以及它们的定义和用法示例。

计算机辅助翻译工具（CAT 工具），包括免费和付费两种：

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

Good Firms网站提供了十大免费和开源计算机辅助翻译软件推荐集合
https://www.goodfirms.co/blog/the-top-10-free-and-open-source-computer-assisted-translation-software

Translate5开源翻译工具。
https://www.translate5.net/en/translate5-open-source-translation-system-2

中文译员 Linguists
Han Liu
Mengzhen Wang
Liu Newman

本地化项目管理 Localization PMs
Chen Yao　Yutong Du
Sophie Liu　Han Liu
Liuyi Yang　Jiaqian Wang

本地化工程师 Localization Engineer
Liuyi Yang