从零开始走向语言数字化系列：

语言数据收集指南

指导您把语言带入互联网

TRANSLATION COMMONS

2019 | INTERNATIONAL YEAR OF Indigenous Languages

语言数据收集指南

主要作者：Julie Anderson、Tex Texin

其余作者：Grigory Sapunov, Jeannette Stewart, Kirti Vashee, Debbie Anderson

编辑： Andrew Owen、Akil Iyer、Shuto Kato、Paula Cirilo

插图与宣传：Leonidas Pappas

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

目录

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）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。附录 D 当中的《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社群及世界。

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 网站下资源页面发布了更多准则、演示和视频等语言数字化项目的相关信息。

1.1 本指南概述

本指南将：

- 列出语言数字化过程中所必须收集的语言数据类型
- 概述收集语言数据的过程
- 阐明各种语言数据与其技术用途的联系

要想实现语言的数字化，我们需要大量的例子来分析语言是如何使用的。语言学家和技术专家们会利用这些例子进行研究和分析，用以设计程序规则和电脑组件，这些规则和组件能够支持一门语言在数字系统当中的使用。我们需要的例子包括按照字母表排序的文字列表，文字的不同书写方式，以及带有解释的单词表等。除此之外，还需要许多其他类型的语言数据来支持语言数字化的进程。

数字系统中能够使用很多的应用程序，有些应用程序对 语言数据的要求非常简单，但有些应用对语言数据的要求则比较高。例如，Windows 上的记事本程序可能只需要支持简单的文本输入和显示就行了。但是专门处理文本的应用程序，比如 Word，可能需要提供给用户更加复杂的布局和排版功能、拼写和语法检查功能、排序功能、大纲和其他处理语言文本的功能。本指南将阐述各类应用程序对语言数据的不同需求。

2. 语言数据收集过程概述

本指南记录了数据收集过程中的一些步骤，这些步骤在语言数字化的过程中非常具有代表性和指导意义，语言数字化完成后可在数字存储库中查看相应的数据。步骤如下：

- 找到可靠的语言数据来源，并获取资料
- 接收这些语言数据
- 记录数据源并设置数据使用和发布权限
- 将非数字资料转换为数字格式
- 将语言数据和注释信息上传到语言数据存储库
- 进行数据审查（包括更正数据、添加注释信息、进行数据分类和核验）
- 发布数据（通过审查和核验的数据就可以提供给用户进行查看）

2.1. 找到语言数据来源

首先，应找出可靠的语言数据来源。任何来源不限，包括在世的语言使用者、历史文件、档案、艺术作品等。收录的语言数据不应仅限于正式文本、历史文本或文学文本，收集一些当今时代的一些非正式文本（包括口头对话或书面交流的内容）也是有益的。

2.2. 材料收集与材料数字化工作

下一步，我们就要开始收集真实的语言数据。如果这些数据还不是数字化的资料，则必须转换成适合数字存储和传输的形式（例如，转换成文本、图像、音频、视频等数字化格式）。例如，口述历史可以存为音频或视频文件或转录为文本。艺术品、书籍、文档，甚至手写的单词列表也可以对其扫描并保存成图像文件。

- 找到合适的语言数据
- 收集书面文本材料
- 向母语人士学习语言知识
- 录制音频或视频

- 转录口述资料
- 扫描或对书面材料进行拍照

收集各类语言数据的过程没有固定顺序。重点是要尽快开始数据收集工作，注释工作，以及建立语言数据存储库的工作。需要收集的语言材料包括：

- 手写信件和印刷文件
- 文本和书籍
- 单语词典，翻译词典及语法知识
- 网站数据
- 社交媒体数据
- 录音录像
- 歌曲、诗歌和表演
- 口述的历史
- 艺术品、图画和照片
- 在世语言使用者对语言群体的了解，以及他们是如何使用这门语言的

2.3.授权、所有权和出处

收集语言数据时，必须确保获得其内容的传播权。例如，如果收集的是私人信件或文档，或者某人讲话的录音，在将其上传到语言数据存储库之前，需获得作者或讲者（或其他利益相关方）的许可。您需要获得书面许可才能发布数据，并且您需要确定相关人士有权授予该许可。您需要咨询您所在辖区范围内某语言群体的长老级人士或您的相关律师，以便规避语言数据收集过程中潜在的法律风险，其中包括数据接收、数据上传、数据使用等问题。并且您需要咨询律师来帮您起草言辞得当的授权协议书。

此外，在保护知识产权、版权和数据主权方面我们虽然已有一些传统措施，但是您可能需要再三考虑这些措施，并对这些措施进行改动，以更好的保护语言群体对其语言、知识、遗产和文化的权利。解决知识产权问题有多种法律途径。当我们采用某种保护知识产权和数据主权的做法时，有必要先询问语言群体对于以上做法的观点，并将其观点融合进保护

知识产权和数据主权的工作当中。关于土著群体数据主权专题的延伸阅读，特别是与语言方面有关的阅读，详见：

- Battiste 和 Henderson 所写的，《保护土著群体的知识和遗产》
- Lovett et al 所写的，《保护土著群体数据主权的实践》
- Te Taka Keegan 所写的，《毛利人对毛利语言数据的主权》
- Christopher Hutton 所写的，《谁是语言的主人？将知识产权作为母语以及人类语言多样性的概念化》

除了数据收集的法律方面之外，记录各项数据的出处及其在语言数据存储库中的存储路径也很有必要。以上出处信息将有助于证明某项数据确实属于该语言，并且这一点将在审查过程中也会有所考虑。

2.4. 注释信息

如果可能，请尽量在收集数据的过程中将其他注释信息也记录下来。一门活跃的语言自然会随着时间的推移而产生变化。语言也会发展出地区和方言的差异。社会因素，如每个说话者的年龄、流利程度、性别或地位，以及语境 （礼仪语、正式语、非正式语等）都可能影响言语行为（口语或书面语）。

记录关于说话者、所处环境、言语行为发生的时间和地点以及其他注释信息可以让我们更准确的理解语言。

这些注释信息包括：

- 该条语言数据的日期和地点
- 关于讲话人或作者的信息、听众或者目标受众的信息（包括年龄、性别、头衔以及语言熟练级别）
- 讲话人之间的关系或相对地位
- 说话时所处的环境
- 文学文体（散文、诗歌、歌词、典礼仪式用文等）

2.5. 语言数据存储库

可以将文件上传到语言数据存储库，通过这个存储库，该语言群体、语言学家、技术专家和该语言群体有关人士授予访问权限的第三方人士可以查看和校对这些信息。

上传步骤可能会因为语言数据存储库的特定配置不同而有一定差异。有关详细信息，请参阅 语言数据存储库的记录和管理方式。

2.6. 语言数据存储库权限管理

语言数据存储库的访问控制权限由其语言群体成员或代表管理。存储库的管理员由该语言群体任命。上传、编辑、查看或以其他方式处理存储库内容的权限由管理员设置。

通常情况下，存储库中的访问控制通过设定用户的角色或通过用户配置文件来实现。每种用户角色都会有不同的访问权限。在把用户添加到存储库之后，就可以开始设定用户的角色。 用户的不同角色决定了其在存储库中的操作权限，例如，查看、创建、编辑或删除数据的权限。存储库中的数据可以有许多属性信息，包括添加时间、添加人、位置信息等元数据信息。因此，有必要为这些属性信息设定编辑和查看的权限，尤其是因为有些属性信息涉及到隐私问题。存储库还可以有隐藏或公开数据的权限、创建或编辑数据类别的权限（用于管理存储库数据），以及管理用户的权限（添加用户、删除用户、更改用户角色等）。有关设置用户角色以及用户权限的详细方案 将取决于存储库的配置情况 。注意，可以针对某一条具体的数据单独设置操作权限，或者针对某一些用户群体单独设置操作权限。例如，可以只向印度语言专家授予与印度语言相关的数据的编辑权限，不对其授予其他语言数据的编辑权限。或者可以授权负责尼日尔语—刚果语组的组长管理这两门语言相关用户的权限，但不授予他们管理其他语言相关用户的权限。

有些权限可以用于对语言数据存储库进行数据审查或者其他工作。例如，当语言材料首次上传到存储库时，可以对其设置权限，只允许一些审校人士对其进行查看。这样这些审校人士就可以对内容提出意见，审阅材料的内容是否恰当和有代表性。如果审校通过，材料

内容则可以公开发布，让更多人能够查看这些内容。对成人内容也可进行标记，设置仅成人可查看。某些违反区域规定的内容也可以进行标记，并且隐藏起来。上述有关于操作权限的设置取决于您的存储库配置情况 。更多相关细节，请参见有关语言数据存储库的文档。

通常，语言群体会将语言数据存储库的访问权限授予以下人士：

- 该语言群体成员
- 语言专家
- 电脑专家
- 其他利益相关方

语言数据存储库的访问控制可能会比较复杂，并且需要对用户的角色以及用户权限进行一系列的复杂设置。不过，下方表 1 列出了最基本的用户角色以及操作权限。下方表格空着的地方就代表该用户**没有**对应的权限。在审校人士将数据设定为*可发布*之前，新添加的或编辑过的数据将被*隐藏*起来，不对大众公开。

表 1 语言数据存储库访问权限管理

	权限					
	查看	**评论**	**创建/编辑**	**隐藏/发布**	**移除**	**分类**
用户角色						
访客	可以					
认证用户	可以	可以				
语言群体成员	可以	可以	可以			
研究人士或语言专家	可以	可以	可以			
审校人士	可以	可以	可以	可以		可以
语言群体管理员	可以	可以	可以	可以	可以	可以

- **访客：**是指想要查看存储库内容的匿名用户
- **认证用户：**是指已通过认证，证明自己有能力为该存储库做出知识性贡献的人士
- **语言群体成员：**是指该语言群体的母语人士，或得到该语言群体能力认可的有关人士
- **研究人士或语言专家：**是指受邀来做研究或建设存储库的专家
- **审校人士：**是指具有高级别语言技能的人士，能够主持对数据内容的讨论，并且对该语言群体的集体情感与诉求以及有关法律等问题敏感的人士
- **语言群体管理员：**是指拥有存储库最高权限的人士，可对用户、操作权限和存储库数据进行最终控制

2.7. 错误处理

即使审查过程十分严谨，语言数据中也不可避免地会有错误。数据收集过程中，印刷错误、转录数据错误、人类记忆错误等问题都可能会出现。在数据公开之前，以及在该数据作为代表性数据储存在存储库之前，应该对其进行审查。持续的定期审查能够对一些在存储库中的错误进行改正。

2.8. 数据审查

数据审查是非常重要的，它能确保在语言数据存储库中的数据内容、数据分类、数据权限都是正确的，并且能提升数据的可靠性。审查可以检查出语言数据中潜在的问题。

语言数据存储库中的数据条目可能并不完整。例如，某个数据可能只是文本的节选，或者数据来源不详。不过这些数据仍然很重要。所以需要数据审查为此把关。

存储库的数据来源有很多，比较繁杂。有时候人们无论是出于恶意与否，都可能会提交上来一些有争议的数据。

有些语言还尚未有人研究过或者没有被记录下来。其词汇、音韵、语法规则等人们都尚未了解。所以收集的语言数据将用于分析推测出该语言的结构以及术语等。代表各式场景的

语言例句越多，语言数字化后的准确度就越高。然而，如果所收集的语言数据不具代表性，那么语言数字化也将失败。语言数据错误或遗漏会影响随后的语言分析过程，导致最后推导出来的分析结果是完全错误的。而审查工作就能减少这种错误推导的可能性。

因此，我们需要有一整套完整的流程，让人们能够轻松地为存储库提供数据建议，并且随后应当有母语人士，语言群体成员，语言专家以及其他专家来对人们提供的数据进行审查。负责审查的人可以对数据进行评估，评论，甚至可以质疑数据的真实性，准确性和该数据的解释。他们还可以为数据加上额外的语境，指出正确的用法，并纠正错误。审查后可能会对数据进行重新编辑，或另寻一些更为准确的例子。也或许会另寻一些其他例子作为补充。审查过程中可能需要数据提供者标明数据来源或提供授权证明。审查还可以确认是否有任何违法的数据，如禁止的图像或言论。审查过程可以重复进行多次。

在审查完成之前，这些数据仅审查者可见。如此便能确保存储库其他的用户看到的都是审查过后的信息。如果审查者对某个数据有疑问，他们会与数据提供者进行沟通，直到解决所有疑问。然后，审查者就可以将数据标记为*公开*，使其对所有人可见。

2.9. 建立语言数据收集流程

第一步是组建一个团队，团队要拥有必要技能和工具来完成每一项语言数据收集任务。然后写一份数据收集指南、分配工作职责并规划工作流程。在工作过程中您需要考虑以下问题：

- 收集材料
  - 您是否有一套已经计划好的工作流程，用于寻找和接收数据，并将其数字化？
  - 您是否寻找到了可为您提供语言数据的人？
  - 您可以获得哪些类型的语言数据材料？
  - 您有办法保存纸质的语言数据材料吗？
  - 您是否拥有可以将语言数据数字化的工具和技能？
  - 您是否了解您所在地区发布和使用语言数据的权限？

- 在存储库中管理语言数据
  - 您是否已安装并配置了语言数据存储库？
  - 您是否具备管理存储库的 IT 技能？
- 审查语言数据
  - 您是否组建了一个由经验丰富的数据审查人员所组成的团队？
  - 您是否有数据注释信息指南和数据审查指南？
  - 您是否有数据批准和发布的指南？
  - 您是否有一本指南能够指导团队如何请求数据提供者提供更多信息，或能够指导团队对数据相关性、真实性、版权或其他有关问题进行质疑？
  - 收集的数据可能会有一些争议，您是否有关于争议处理的指南？
  - 您是否有已计划好的工作流程以便上传和审查数据、为数据添加注释和进行数据更正？
- 管理您的团队
  - 您的团队成员了解他们的任务和职责吗？

2.10. 小结

收集语言数据是一个持续不断的过程。只要能获得语言数据，就可以将其上传到语言数据存储库并进行审查。语言数据收集过程概述如下：

- 找到可靠的语言数据来源
- 接收这些语言数据
- 将材料转换为数字格式
- 上传资料到语言数据存储库
- 设置使用数据和发布数据的权限
- 为数据添加注释
- 审查数据（更正、注释和核验数据）
- 发表数据

3. 在电脑上如何使用收集到的语言数据

收集到的语言数据可以在多方面支持语言的数字化工作。有关的内容由于内容篇幅较长，故暂不详细展开说明。不过，以下是语言学家和技术专家使用语言数据的一些基本方法，这些方法可以支持语言数字化。

3.1. 手写的语言数据

手写的语言数据会最初用于确定语言中使用的书写符号（包括字母、数字、重音标记、声调标记、标点符号和其他字符）。当确定了上面一系列的书写符号后，手写的语言数据可以用于观察该语言的文字书写方式，并将其用于创建数字化字体。上面的书写符号还可以用来创建键盘布局和输入法，用于输入这门语言的文字。

请您注意，文字的书写方式并不只有一种。您需要考虑到文字可能会有衬线字体、斜体字等许多变化。在某些语言中，文字会根据它们在单词中的位置或它们旁边的字来改变文字形状。

还有，有些字很少使用。它们可能只在特定的仪式中使用，也可能是该语言文字的旧版本。正因为如此，收集的样本越多，语言的数字化就越可靠、越完整、越有用。

3.2. 句子和段落

包含完整例句的语言数据可以揭示该语言的发音、正字、排版、语法习惯等等，而在后续用电脑软件对这门语言进行文字处理的时候，这些信息就会非常有用。在这方面可用的工具包括：

- 语法规则
- 连字符用法、断字、大写、重音符号和标点符号
- 文字对齐方式和书写方向
- 发音

- 表示尊敬的形式(如敬语和姓名顺序)

通过对语言数据的分析还可以揭示语言当中用来表示日期、时间、年代、数字、百分比等独特的书写习惯和格式。

3.3. 术语表

术语表可以从手写文本数据、录音以及其他数据中收集。拼写检查器、自动更正、预测文本、光学字符识别（OCR）和其他有关文字处理的功能都需要使用术语表。

3.4. 词典

单语词典和双语词典提供释义、词性、发音、词源、翻译和其他信息。这些信息可用于在电脑上对文字进行处理，包括断字、拼写和语法检查、自动更正、机器翻译等。

词典的设计可能会很复杂。例如，词典编纂者会仔细考虑多式综合语词典的*词目*，并在词典中对词目按照单词顺序进行排序。（译者注：多式综合语的单词或句子是由许多词根组成的，详见百度百科解释；词目指词典中所汇集的每一个被注释的对象，详见百度百科解释）指导土著语言词典撰写的现有资源包括:

- Nick Thieberger：《编纂澳大利亚和太平洋地区土著语言词典》
- Antonia Cristinoi 和 François Nemo：《濒危语言词典编纂的挑战》
- Paul V. Kroskrity：《为濒危语言群体设计一本词典》
- Frawley、Hill 和 Munro：《编纂词典：保护美洲土著语言》
- Sarah Ogilvie：《语言学、词典学与濒危语言的复兴》

3.5. 录音

录音和录像可以用来研究语言的发音。还可以用于研究文本转语音、语音转文本技术。文本转语音对于视障人士、有阅读障碍的人士以及识字能力较低的人是非常有用的。语音转

文字技术可以用于开发语音控制电脑技术，从而帮助无法用键盘或者屏幕打字的残疾人士在电脑上用语音输入文字。

3.6. 双语数据

双语数据有多种用途。利用双语数据，比如翻译词典、双语字幕视频和已翻译的文档等，可以创建在线词典、语音翻译工具、机器翻译和其他工具等。此外，通过双语数据，可以对两种语言的单词进行对此，从而发现细微的语言差异。

3.7. 数据收集量

通常来讲，收集的数据越多，语言数字化的质量就越好。同时该语言的语法、单词义项等会变得更加准确、细节更加丰富。通过收集大量数据，一些习语和罕见的用词也可以保存下来。

有些词语只在特定主题或领域内使用。例如有些词只在卫生、农业、监管等领域使用。收集的数据量越大，就能覆盖越多的领域。

此外，有些以语言为基础的功能只有在掌握了大量的语言数据之后才能够创造出来。例如，想要创造机器翻译引擎，就必须要掌握大量的语言数据用于训练电脑的机器翻译引擎。

3.8. 保持语言数据的多样性

所有类型的语言数据都可以为语言的数字化提供强有力的支持。所以在数据收集的过程中，不要排除那些看起来已经过时的用语、过于正式/不正式的用语、过于官方的用语、只针对年轻人或者文化水平较低的人的用语，夸张的用语（例如广告），以及带有迷信色彩的用语（比如传说或者历史故事）。以下的这些语言数据都是有用的：

- 儿童故事和图画书
- 教育材料

- 通讯信息、私人信件、便笺和短信消息
- 法律文件（出生证明、结婚证、死亡证明等）
- 词典（单语词典和双语词典）
- 书籍
- 报纸
- 路标、广告牌
- 海报
- 古代的文字
- 口述历史和传统
- 图画和其他艺术品
- 日常会话（口头的、书面的、正式和非正式的）

3.9. 创造新术语

人们需要新的术语来表示新奇的想法、发明和活动。尤其是当我们需要使一门语言数字化时，我们就更　需要新的术语。例如，用户在软件和硬件的界面中看到的术语必须是用其母语创造的或改编的术语。这些术语就包括了数字系统中常见的术语，例如菜单、按钮、下拉、文件、编辑、帮助、退出、确定、取消、单击、下载等。

可能您的母语中并没有这些术语，而想要为这些术语找到合适的翻译，或者用您的母语为这些词汇创造新的术语，不总是那么容易。而很多时候，用您的母语直接对这些术语进行直译可能并不是非常好。例如，在英语当中的“home page”就不能用中文直接翻译成“家页”，而中文当中我们会将其翻译成“首页”或“起始页”，很多其他的语言也是如此。此外，一个术语可能有多种用途，需要根据语境进行不同的翻译。例如，英语中的“cancel”在中文里实际上有两种意思，一种是“取消”，一种是“撤销”。到底是完全取消操作，还是撤回到上一步的操作？这时候就需要看语境来抉择。又例如，英语中的“orange”和中文的“橙”字一样，它既代表一种颜色，又代表一种水果，但在其他语言中，可能并不是这样。

语言群体可能需要创建一个工作流程，用于为电脑系统创造新的术语，而且要保证整个语言群体都能在术语使用上达成一致。有关创造新术语的信息，请参阅《从零开始走向语言数字化：术语指南》

3.10. 小结

收集语言数据的数量和范围越大，语言数字化就越准确和全面。下表说明了语言数据是如何应用在电脑系统中的。

表 2 语言数据在电脑系统中的应用

语言数据类型	语言学用途	电脑系统用途
手写的语言数据	分析书写符号和书写习惯	确定字符集、字体、键盘布局、输入法等
电子化文本（包含句子或段落）	分析语音、正字法、排版、语法和其他的语言习惯	用电脑进行文字和语音处理
术语表	分析语言中的词汇	用于拼写检查、自动更正、预测文本和 OCR 文字识别
词典	分析词汇的定义和词性	用于拼写和语法检查器、文字处理、文字转语音和机器翻译
音频和视频	分析发音规则和口语习惯	文字转语音和语音识别
双语数据	用于翻译、分析和其他语言之间的关联以及分析该语言的书写系统	制作双语词典和带字幕翻译的视频、翻译文档和软件界面、制作语音翻译工具和机器翻译软件
创造出的新术语	为数字化系统创建出新术语	用于软件的菜单、命令、对话等界面的翻译

4. 专业语言实用工具

4.1. Unicode 通用语言环境数据信息库

通用语言环境数据信息库（CLDR）是 Unicode（统一码）下设立的一个数据信息库，为软件的国际化和本地化提供了关键的数据模块。顾名思义，通用语言环境数据信息库就是一个存储了大量的语言环境数据的信息库。（译者注：举例来说，苹果公司是一家美国公司，所以苹果公司的软件是基于英语开发的，最开始只会有英文版本，而后续在全球推广苹果系统的时候，需要将其翻译为不同的语言，适合当地的语言习惯。）很多公司都会利用通用语言环境数据信息库来翻译其开发的软件

通用语言环境数据信息库中包含翻译软件的过程中所需的许多专有名词的翻译，如月份、星期、国家及其分区、语言名称、计量单位和货币信息等。通用语言环境数据信息库还提供了编码表达式，以便在开发软件时根据当地的日期、时间、数字、计量单位、货币和其他惯例来编排数据。

如果您使用的语言尚未在包含在通用语言环境数据信息库中，请考虑联系有关人员在信息库中添加上您的语言数据。这将使软件开发者能更容易地把其开发的软件翻译成您的语言版本。查阅通用语言环境数据信息库的内容也可以帮助您确定数字系统所需要信息和术语，并且有助于您创建新术语。

4.2. 机器翻译

机器翻译是一种越来越重要的语言工具，因为它可以快速翻译大量的文本。

然而，机器翻译引擎的建立需要大量的双语数据。理想情况下，最好是用已经对齐了的双语数据，也就是说最好用平行语料。平行语料是指源语言的每个句子片段都已经与目标语言的句子片段匹配了起来。关于机器翻译的重要性及其具体要求的更多信息，详见附录 C：机器翻译数据要求。其中包含机器翻译概述以及建立机器翻译引擎的数据要求。

4.3. 自然语言处理工具

本节介绍了有助于分析语言和生成语言数据的工具。举个例子来讲，这样的工具可能是一个能够扫描文件并提取语言单词列表的工具。在数字化语言过程的最初，您需要确定您所需的工具，来完成较为紧迫的任务。这些工具可能需要改造一下以便处理您的语言。

许多软件应用依靠自然语言处理（NLP）组件来处理语言数据。在您完成语言数据收集工作的关键部分之后，您可以开始投资进行语言建模，并且更新您的软件库。许多软件界面可能还不支持您的语言，所以需要对其进行翻译，而预先训练好的语言模型可以加速软件界面翻译的过程。

尽管语言的内容是丰富多样的，但许多句子可以通过类似的句法、语法等模式归为一类。能识别语法或者句法模式的工具可以帮助您处理您的语言数据。例如，有一些工具可以分析某些语言的文本并创建单词列表。还有一些更复杂的工具能够使用语言模型来处理语言数据。语言数据也可用于训练和创建更为完善的语言模型。有一些工具是专门设计用来处理数字化程度较低的语言的。

有几个开源网站能托管自然语言处理和语言建模工具。有些软件可以处理与您的语言在语法上非常类似的语言，请您考虑使用这些软件，并对其加以定制改造，以便能够处理您的语言。成功后，您也可以把您训练出的语言模型添加到这些网站。如此，软件开发者就能更加容易地开发支持您的语言的软件，然后您可以邀请主流的软件公司将您的语言加入他们的开发列表中。

以上提到的工具包括：

- 词向量工具，如 word2vec（从大型文本语料库中识别单词关联的模型）
- 已训练好的模型，例如 BERT 和 GPT（能实现分析语法功能、含义和单词关联）
- 命名实体识别（NER）模型（包括姓名、地理、日期等）
- 词性标注模型（包括名词、动词等）
- 句法/依存关系分析模型（主语/宾语/...，动词和名词短语，等等）

- Huggingface 公司的“Transformers”（支持信息提取、翻译和其他自然语言处理功能）
- SpaCy 高级自然语言处理开源软件库（支持信息提取和其他自然语言处理功能）
- 自然语言工具包（是一个文本处理工具库）

许多应用程序要依靠语言识别库来识别当前使用的语言，为程序正常运行做出调整。（译者注：比如 Word 文档首先要识别输入的文字是什么语言，才能对应的功能，比如拼写检查功能）这些应用程序无法支持语言识别库里面没有的语言。您可以向语言识别库中提供信息，帮助它们识别您的语言，使得软件能尽早地支持您的语言。

在这方面可用的工具包括：

- 谷歌的 CLD3（紧凑型语言检测器 v3）
- Facebook 的 fastText

4.4. 小结

下面这些工具可以帮助分析您的语言，支持其数字化，帮助软件开发者支持您的语言，并加快译入和译出您的语言的速度。

- Unicode 通用语言环境数据信息库
- 机器翻译
- 自然语言处理工具

5. 本指南英语原文所涉及的专业性词汇解释

专业术语	解释
Character	翻译为：书写符号　　解释：一门语言的字母、记号、标志、标记或其他书写符号等（译者注：Character 这个词的意思有很多，在做这本指南的翻译过程中，为保证读者能理解，译者已经对这个词进行了灵活的翻译处理，并未将其全部都翻译成“书写符号”，有些根据语境翻译成了“文字”等其他中文词）
Corpus	翻译为：语料库　　百度百科解释：语料库指经科学取样和加工的大规模电子文本库，其中存放的是在语言的实际使用中真实出现过的语言材料
Diachronic	翻译为：有关于演化的　　解释：有关于事物发展变化的，尤其是语言的发展变化的【百度百科：演化语言学（也称历时语言学）是语言研究的一个分支，是从纵向发展的角度研究某种语言从一个时代到另一个时代的发展变化的语言学分支】
Dialectal	翻译为：方言的　　解释：和方言有关的
Digitize	翻译为：数字化/电子化　　解释：将数据转换为可由电脑处理的数字形式
File format	翻译为：文件格式　　解释：是指电脑文件所存储的方式，该存储方式会根据存储的数据类型发生变化，一般可以通过文件扩展名来识别（如：.html 代表一个网页）
Font	翻译为：字体　　解释：文本的图形化表示，或是一组具有相似图形设计的书写符号
Glossary	翻译为：单词表　　解释：按字母顺序排列的与某一特定主题有关的词汇及其定义
Indigenous	翻译为：土著的　　解释：来源于某一特定地区的
Lexicography	翻译为：词典编纂　　解释：指词典编纂的过程（译者注：本文的英语原文经常用这个单词来代替“词典学”，词典学是有关于如何编纂词典的学问，在本文的翻译过程中译者已经对这个单词的翻译进行了灵活地处理）

Locale Data	翻译为：语言环境数据　　解释：用于为一个地区的特定语言和文化定制软件的用户界面的数据
Media	1. 翻译为：媒体　　解释：是一种大众传播手段（包括广播、出版和互联网等） 2. 翻译为：存储介质　　解释：用于存储数据的设备
MT	翻译为：机器翻译　　解释：利用电脑进行自动化的翻译工作
NER	翻译为：命名实体识别　　百度百科解释：又称作“专名识别”，是指识别文本中具有特定意义的实体，主要包括人名、地名、机构名、专有名词等
NLP	翻译为：自然语言处理　　百度百科解释：是研究能实现人与计算机之间用自然语言进行有效通信的各种理论和方法的学问
NLTK	翻译为：自然语言工具包　　解释：用于处理自然语言的工具的集合
Orthographic	翻译为：正字法：　　解释：是关于文字使用的规范性法则，包括拼写、连字符、大小写、断字、重音符号和标点符号的规范
Phonological	翻译为：语音学的　　解释：与语言中的声音研究有关的。
Polysynthetic	翻译为：多式综合语　　解释：其特点是这种语言中会有由几个语素组成的复杂词，其中一个词就相当于一整个句子
Repository	翻译为：数据存储库　　解释：储存和管理数据的一个库
Scan	翻译为：扫描　　解释：以数字手段对信息进行复制和存储的方式
Segment	翻译为：句段　　解释：一个独立的、有意义的文本或口语单位句段分割就是将书面文本划分为有意义的语义单位（如单词、句子或主题）的过程。语音分割是识别口语中不同的单词、音节或音素之间的界限的过程
Serif	翻译为：衬线　　解释：部分印刷体的西文字母顶端或底部的短线
Text-to-speech	翻译为：文本转语音　　解释：向用户大声读出电子文本的辅助技术

Typographic	翻译为：排版　　解释：一种安排字体的艺术或技术，使书面语言在显示时更为清晰可读，能吸引读者
Unicode	翻译为：统一码　　解释：统一码，也叫万国码、单一码，是计算机科学领域里的一项业界标准，包括字符集、编码方案等。是为了解决传统的字符编码方案的局限而产生的，它为每种语言中的每个字符设定了统一并且唯一的二进制编码，以满足跨语言、跨平台进行文本转换、处理的要求（译者注：本文在翻译的过程中没有翻译 Unicode，直接采用的英文原文，因为通常中文语境下都不翻译这个单词。）
Upload	翻译为：上传　　解释：将数据从一台电脑转移到另一台电脑上，通常是转移到更大的，或远离用户的，或作为服务器使用的电脑上
URL	翻译为：网址　解释：网络上一个网页或其他信息的地址
UTF-8	翻译为：UTF-8　百度百科解释：针对 Unicode 的可变长度字符编码
Voice-to-text	翻译为：语音转文字　　解释：语音识别程序，可将口头语言转换为书面文字
XLIFF	翻译为：XLIFF　　解释：XLIFF 是由软件开发商、本地化服务提供商、本地化工具提供商等团体共同倡议和设计，由 OASIS 标准组织发布的用于本地化数据交换的格式标准。它基于 XML 技术制定软件资源文件格式的转换规格，其目的在于提高软件的本地化作业效率

附录 A：数字系统中语言应用实例

在数字系统中，您的语言可能会在如下的一些常见的活动中使用。当然您也可以根据人们的需要创建新的应用程序。

通讯

- 收发短信
- 收发电子邮件
- 收发媒体文件（图像、音频和视频）
- 自动翻译文本，机器翻译
- 自动转换语音信息为文本，或将文本转换为语音

出版、存储信息和文字处理

- 在网站上发布和获取信息
- 创建和分享文件、书籍、新闻媒体、标牌、海报、教育材料
- 编纂印刷或在线词典
- 扫描文件转换成电子文本
- 为您的文本创建一个字体
- 在线购买和销售物品
- 将网站和应用程序翻译为您的语言
- 创建您的语言的应用程序
- 拼写检查
- 语法检查和自动更正

用户界面和残障人士支持

- 语音识别（适用于残障人士）
- 使用语音命令来控制设备

- 文字转语音和屏幕阅读器（适用于视力障碍人士或低识字率地区）
- 语音转文字和实时字幕（适用于听障人士）

附录 B：软件应用的数据要求

下方的软件应用数据要求表是一个可视化的指南，能帮助您更好地开始收集语言数据，最终将您的语言应用到这些软件功能上。

该表能从两个不同的角度帮助语言群体：一、语言群体可以先找到他们希望使用的软件功能，然后再看下表，找到需要哪些语言数据来创造这个功能。二、语言群体也可以反过来先看看自己收集了哪些语言数据，然后对着下表看看近期能够实现什么样的软件功能。

该表与本指南中所概述的大部分信息相同，能帮助用户建立一个合理的数字化目标预期。

表 3 软件应用数据要求表

数据类型	书写符号	术语	翻译文本	讲话	语言惯例和排版设计		大量的语言数据
举例内容	字符、字母、数字、标点符号、字体等	术语表、单语词典	双语词典、词网、语料库	语音、发音规则、音频和视频	正字法、语法规则、连字法、大写字母、标点符号、书写方向、对齐方式	写作习惯：日期、时间、年代、数字、百分比等	句子，段落，单语文本语料库
软件应用							
通讯							
发送/接收文本信息	x						
发送/接收电子邮件	x						
自动翻译文本，机器翻译	x	x	x		x	x	x
数据类型	书写符号	术语	翻译文本	讲话	语言惯例和排版设计		大量的语言数据
自动转换语音为文字，或将文字转换为语音	x	x		x	x	x	x
出版、存储、文字处理							

在网站上发布和获取信息	x				x	x	
创建和分享文件、书籍、新闻媒体、标牌、海报、教育材料	x				x	x	
编写纸质/在线词典	x	x	x		x	x	
扫描文件转换为数字文本（OCR）	x						x
为您的文字设计字体	x				x		x
在线购买和销售物品	x				x	x	
将网站和应用程序翻译为您的语言	x	x	x		x	x	
用您的语言创建应用程序	x				x	x	
拼写检查	x	x					
语法检查、自动更正	x	x			x	x	
用户界面和残障人士支持							
数据类型	书写符号	术语	翻译文本	讲话	语言惯例和排版设计		大量的语言数据
语音识别（适用于身体残疾人士，也可以用于和智能手机的语音助手交谈，例如苹果手机的Siri）		x		x			

文字转语音和屏幕阅读器（适用于视力障碍人士或低识字率地区）	x	x		x	x	x	x
语音文字转换和实时字幕（适用于听障人士）	x	x		x	x	x	x
图例	x 带有 x 的表格表示实现某种软件应用功能所需要的数据类型。						
注意	本表格所列举的数字类型和软件应用功能可能并不全面。						

附录 C：机器翻译数据要求

什么是机器翻译？

机器翻译（MT）是指使用软件将文本或语音从一种语言翻译成另一种语言。

为了实现高质量的翻译，机器翻译不仅仅是简单的字对字的翻译。机器翻译引擎使用先进的算法，需要大量的语言数据来训练翻译引擎。

机器翻译通常与人工翻译的质量无法相比。为了提高质量，机器翻译引擎通常按领域或专业进行定制，以限制内容的范围，使其翻译的更为准确。

机器翻译作为一种辅助人类翻译的工具是很有用的，而且在某些特定的情况下，机器翻译的译文可直接拿来使用。

建立机器翻译引擎

为了实现语言和语言之间的自动翻译，必须建立一个专门针对这两门语言的机器翻译引擎。这就要求先选择一个机器翻译所使用的技术，然后用两种语言去训练翻译引擎。

要想为两门语言创建机器翻译引擎，就必须要确保这两门语言都有强大的数字基础。也就是说，对于一门刚刚被数字化的语言来说，它的语言资源必须要不断的增长，这样才能保证有足够的数据来训练机器翻译引擎。（译者注：例如训练中文-英文机器翻译引擎需要两个条件。一是数字系统首先要支持这两门语言。二是在网络上有人用中英文发表过大量的文本，因为只有这样才能产生源源不断的语言资源供训练机器翻译引擎使用。对于刚刚被数字化的语言来说，可能只满足条件一，不满足条件二，所以无法训练机器翻译引擎。）

有些语言有着千万级别的使用人数，但是目前这些语言还是没有可用的翻译引擎。这种情况有两个原因，要么是人们付出的努力不够，要么就是语言数据不足，无法训练翻译引擎。虽然随着核心机器翻译技术的不断改进，人们用较少的语言数据也能建立机器翻译引擎，但我们需要时刻谨记，建立一个***好的***机器翻译引擎是一定需要大量的语言数据的。

用于建立机器翻译引擎的语言数据也被称为训练数据。

假使目前只有一些基础的训练数据，我们就可以快速地创建一个翻译引擎。随着时间的推移，我们可以继续向现有的机器翻译引擎中添加额外的训练数据以不断优化机器翻译引擎。

随着新训练数据的添加，机器翻译技术的不断革新，以及人们给机器翻译的不断纠错，机器翻译引擎也会不断的发展优化。并且我们应当对机器翻译引擎进行定期评估和更新。

机器翻译在哪些地方有用处？

在需要以相对较低的成本快速提供大量的信息和知识资源时，机器翻译是非常有用的。然而，我们也应该明白，目前机器翻译无法达到一名优秀人工译员的水平。

但是，机器翻译的好处在于翻译速度比人快，而且通常产出的译文质量还算可用，并且在机器翻译引擎建立后可以随意进行部署，可供数百万人在网上使用。机器翻译的存在使数以百万计的人能够获取他们原本无法获取的信息。虽然关于机器翻译造成误译的情况比比皆是，但对许多人来说，更加重要的是要学会如何使用机器翻译，以及如何对机器翻译技术进行扩展。虽然在产出高质量译文方面，机器翻译还是不太可能取代人类，但有越来越多的案例表明机器翻译可以适用于以下方面：

- 重复性高的内容
- 没有人工译员能够提供翻译的情况
- 支付不起人工翻译费用的情况
- 内容价值含量高，但是每个小时或每天在不断发生变化的内容
- 能够促进关键知识在全球范围内传播的内容
- 能够与全球客户加强沟通的内容，通常这种客户都更喜欢自助式服务

- 不求多高的翻译质量，大概能理解就行的内容

开发一个机器翻译引擎需要什么样的数据？

所有的现代机器翻译开发技术都是需要由数据驱动的，也就是说，电脑通过分析大量的已经翻译好的翻译语言数据来*学习*如何从一种语言翻译到另一种语言。这种用于开发机器翻译引擎的语言数据被称为训练数据。目前使用最为广泛的机器翻译技术是神经机器翻译技术，它正在慢慢取代许多被称为统计机器翻译的旧技术。这两种技术都是用于开发机器翻译引擎的技术，这两种技术都需要以下这些语言数据：

- 双语文本
- 翻译词汇表
- 目标语言的单语数据
- 源语言或源语言相关语言的单语数据

双语数据和目标语言

逐句翻译的大型文本集被称为平行语料库。创建一个初始的翻译引擎至少要用到十万个双语翻译句段（句子）。一个翻译句段可以是一个完整的句子，也可以是一组翻译好的关键术语和短语。如果是理想情况下，至少要用到一百万个翻译句段。而有些机器翻译引擎是用数十亿个翻译句段建立的。一般来讲，用的高质量双语句段越多，训练出的机器翻译引擎就越好。

但许多语言群体没有如此大量的数据。数据采集阶段往往需要政府机构、教育机构和整个语言群体之间的一致努力和长期合作。目前，现有的技术已经可以使用更少的句段来初步建立一个翻译引擎，随着人工不断对机器翻译引擎进行纠正，机器翻译引擎可以不断得到优化。

用于训练机器翻译引擎的语料库文本通常是从大量的平行文本中提取出来的，比如对同样一个事件进行了类似报道的新闻双语文本。

然而，提取出的片段可能是有噪音的，即在每个语料库中都有一些多余的数据，这些数据只在一个语料库中存在，而另外一个语料库没有。（译者注：举例来说，两份分别为中文和英文的报纸都对甲事件做了类似的报道，但可能中文报纸中有的内容英文报纸中却没有，反之亦然，而这样的数据是不利于机器翻译引擎训练的，这些就是所谓的“数据噪音”，需要想办法剔除这些“噪音”，以训练更好的机器翻译引擎。）提取技术可以区分在两个语料库中都存在双语数据和只在一个语料库存在的单语数据，通过这样的对比就可以剔除那些只在一个语料库中存在的数据，减小数据的噪音，提取更纯净的双语数据。比较语料库则可以用于直接获取翻译的有关知识。（译者注：比较语料库的具体介绍详见：https://zhuanlan.zhihu.com/p/59514775）然而，高质量的平行语料很难获得，尤其是对于一些记载资料不足的语言。

用于创建机器翻译引擎的训练数据通常来自于翻译记忆库（翻译记忆库就好像一个仓库，存储着过去已经翻译好的文本）或一段时间以来收集的其他翻译资产。这些训练数据就决定了将来训练出的翻译引擎更加适合翻译什么主题的内容。通常情况下，可用的训练数据量是有限的。在这种情况下，我们需要努力让机器翻译引擎学习的内容要尽可能贴合它将来最有可能翻译的领域。请记住，您给机器翻译引擎训练什么主题的材料，它将来就会在什么主题上翻译的最好。因此，想要用于翻译医学内容的机器翻译引擎就最好用医学主题的翻译记忆库和词汇表来训练。

当提供双语数据进行训练时，数据必须是**对齐的**，源语和目标语必须是彼此的直接翻译。翻译的文本如果是对原始文本的总结、摘要或评论，则不适合进行训练。在使用前必须仔细检查数据，以确保能够用于机器翻译引擎训练。

如果在翻译的过程中能提供一个包含关键词汇的词汇表，并且包含对应的词汇翻译，这将会提升机器翻译引擎的翻译质量。

为您所收集的语言数据制定一个全面元数据战略将具有长期价值。在数据采集的初始阶段，重点通常是尽可能地寻找更多的数据，以满足开始阶段对于数据量的需求。然而，随着机器翻译引擎的成熟，使用正确的语言数据类型来训练引擎，可以让机器翻译引擎有更好的表现。因此，我们需要对机器翻译引擎针对特定的主题进行优化，比如针对医疗领域或者

电脑技术领域进行优化，而不是创建一个万能的机器翻译引擎。这种专业化的机器翻译引擎往往会表现更加出色。

双语数据格式

作为机器翻译引擎训练数据的双语数据需要具备三个重要特征。它应该是机器翻译引擎可以导入的文件格式。双语文本数据应采用 Unicode UTF-8 字符编码。而且，双语数据应该是对齐的，也就是说每个源语句段需要都与一个目标语句段进行匹配。

双语数据可以为下列文件格式（大致按照最好到最差的格式列出）：

- 翻译记忆库格式为最好的格式（TMX、TBX、XLIFF、CSV）
- 纯文本 （TXT）
- 网页格式（HTML）
- 结构化文本（XML）
- 微软 Office 格式（DOC、DOCX、PPT、PPTX、XLS、XLSX）
- 出版或 DTP 格式（TTX、PDF、FrameMaker)
- 光学字符识别（OCR），（TIFF、PNG、JPEG 等）

语言数据在交付的时候可以是对齐好的语言数据，也可以用原始形式交付，例如，Word 文档或 HTML 文档。如果语言数据没有对齐，则有必要使用软件来精准对齐语言数据。

单语数据

虽然翻译记忆库和双语文本可能是建立机器翻译引擎最关键的数据，但高质量的目标语言单语数据也是必不可少的。这些数据在机器翻译引擎训练中用来学习正确的语法结构，并从统计学角度影响引擎输出，使其实现所需的写作风格。在创建统计机器翻译引擎时尤为如此。（译者注：统计机器翻译与神经机器翻译相对，详情请查阅互联网）

单语数据比双语数据更容易获取。收集具有类似主题或语法风格的网站的网址就可以。我们可以从这些网站中挖掘语言数据，利用其包含的语言知识。

通常情况下，土著语言的单语数据比较难以获得。对于大多数语言来说，都有许多单语数据的来源。单语数据很重要的一个例子就是辅助建立医学机器翻译引擎。对富含这种特定主题内容的医学和相关网站进行数据抓取，可以找到关键性的语言信息，用于构建词汇表和翻译记忆库。

双语文本数据应采用 Unicode UTF-8 字符编码。双语数据可以为下列文件格式（大致按照最好到最差的格式列出）：

- 纯文本
- 翻译记忆库（TMX、TBX、XLIFF、CSV）
- 网站的网址（URL）
- 网页格式（HTML）
- 结构式文本（XML）
- 微软 Office 格式（DOC、DOCX、PPT、PPTX、XLS、XLSX）
- 出版或 DTP 格式（TTX、PDF、FrameMaker）

附录 D：语言数字化的益处

我们希望《从零开始走向语言数字化》系列指南可以为有兴趣的语言群体提供一条清晰的语言数字化之路，帮助他们能够在电脑上使用其母语版本的软件系统。

语言数字化后的具体益处将取决于该语言使用群体的目标。语言群体的目标可能包括欣赏语言之美，保存知识体系，传播价值观，创造应用程序和产品，分享民族故事和历史，促进环境管理和思想领导的发展，以及扩大贸易、教育、就业，促进娱乐、健康和安全。语言的数字化可使一个语言群体能够利用电脑软件工具来保护、振兴和教授语言，而且这些软件工具的数量会随着时间不断增加。当语言能够以数字化的形式存在于数字系统当中时，该语言的知名度将得到提高，从而使政府制定出更多支持土著群体的政策，并使企业变得更为包容。鉴于如今的年轻人普遍都有智能手机，语言数字化之后可能这些年轻人自然而然地就会使用他们的母语进行交流了。（译者注：年轻人不使用母语在网络中交流是因为智能手机不支持他们的母语。）在一门语言经过数字平台的曝光之后，就可以给讲这门语言的人带来更多的机会，对其母语人士的需求也会增加。

当该语言群体通过数字平台在全球舞台上占有一席之地时，它将使全世界都受益。因为语言群体的经验、知识和独特的世界观将会对世界其他地区做出重大贡献，由此产生的协同作用也许能为许多世界性问题带来新的解决方案。语言的数字化促进了这些信息的保存和出版，使世界了解到人类语言的广度和性质，并且以一种人们目前未知的方式保护了人类的利益。

总而言之，通过语言数字化我们能获得的一些好处是：

- 可以让只会一种语言的人能够轻松地访问、创建和交换母语内容，包括跨越长距离发送母语内容，向个人或大型团体发送信息
- 增加人们获取医疗保健信息的途径

- 支持救生应急工作和灾难通信
- 扩大本地商业和电商
- 创造新途径用于分享艺术、思想领导力和哲学
- 编写母语教育材料
- 改善与邻里的关系，加强与邻里的沟通
- 让人们可以更好地解决争端
- 能够用母语进行宣传，用母语了解法律和政府规章制度
- 可以增强人们用母语，或者利用机器翻译获取网络信息的能力，包括获取教育、商业信息和一些其他机会信息
- 让更多人能认可这门语言、以及其文化和智慧
- 在全球范围内扩大土著群体的知名度，提高其地位
- 让边缘化群体或少数群体即使在主流群体的影响下，也能够保护或振兴他们的语言
- 保护本土的知识体系、文化、历史、艺术、医药、智慧、价值观和世界观