Vuk'uzenzele 与 ZA-gov-multilingual 南非多语言语料库的构建

Richard Lastrucci[1], IsheanesuDzingirai[1],Jenalea Rajab[2],Andani Madodonga[1], Matimba Shingange[1],Daniel Njini[1],Vukosi Marivate[1,3]比勒陀利亚大学计算机科学系[2]金山大学计算机科学与应用数学学院[3]Lelapa AI richard.lastrucci@tuks.co.za, ishe.dzingirai@gmail.com,jenalea.rajab@gmail.com, andanim412@gmail.com, mrosslyns@gmail.com,vukosi.marivate@cs.up.ac.za

arXiv:2303.03750v2 [cs.CL] 5 Apr 2023

摘要

本文介绍了两个以政府为主题、涵盖多种南非语言的多语语料库。这些语料库通过收集南非政府报纸《Vuk'uzenzele》以及南非政府演讲（ZA-gov-multilingual）构建而成，内容均已翻译成全部 11 种南非官方语言。这些语料库可用于多种下游自然语言处理任务。建立这些语料库的目的是让研究人员能够研究南非政府出版物所使用的语言，重点了解南非政府官员如何与其选民沟通。

本文重点介绍语料库的收集、清理和发布过程。我们利用语言无关句子表示（LASER）嵌入，为神经机器翻译（NMT）任务创建平行句语料库。基于这些对齐句子，我们通过微调一个大规模多语言预训练语言模型，为9种原住民语言提供神经机器翻译基准。

1 引言

由于缺乏用于训练各种自然语言处理任务模型的数据（Nekoto 等，2020年），以及缺乏基准和复现方法（Mart- inus 和 Abbott，2019），非洲的自然语言处理（NLP）研究进展受到阻碍。对于许多南非语言来说，仍然难以找到易于获取的文本数据集（Marivate等人，2020年），尽管这些语言拥有众多使用者（Ranathunga 和 de Silva，2022）。因此，有必要着力开发本地语言的自然语言处理资源（Joshi 等，2020年）。

本文基于 Autshumato 项目的工作（Groenewald 和 Fourie，2009；Groenewald 和 du Plooy，2010年），创建自动对齐的南非11种官方语言的政府文本数据平行语料库。Autshumato项目专注于为五种原住民语言创建机器翻译工具，但由此生成的语料库缺少有关其来源或背景的信息，因此难以用于其他自然语言处理任务，例如分类、随时间变化的主题建模，以及其他需要内容背景信息的任务。我们的方法提供了更全面的数据，可支持更广泛的自然语言处理应用。

我们的理念是，有机会创建一个更加用户友好的数据收集流程，使其易于维护，并为他人提供提取工具。保存数据源并以支持扩展的方式组织数据至关重要。我们的目标是通过纳入所有原住民语言，并将翻译方向扩展到不再仅以英语作为源语言，来增强政府数据领域的神经机器翻译（NMT）资源。此外，我们认识到，为所有南非语言提供超越英语的对齐数据十分重要。

此外，本文介绍了使用从政府获取的文本数据创建的南非11种官方语言的平行语料库数据集。这些数据集旨在促进神经机器翻译模型的开发。语料库经过自动对齐，预计将成为从事机器学习领域的研究人员和实践者的重要资源。

这些平行语料库使用 LASER 编码器（Schwenk 和 Douze，2017）生成，从而实现了分词句子数据的一对一对齐。数据来源可靠，包括报纸和学术期刊，涵盖卫生、金融和政治等多种主题。

我们还通过对大规模多语模型（M2M100（Fan 等人，2021）进行微调，为这些平行语料库提供神经机器翻译基准，该模型基于

作品Adelani 等人（2022）。

本文结构如下。在下一节中，我们将详细介绍所编制的数据集，包括其编制方法以及所包含的信息。随后，我们将描述如何利用这些数据集进行对齐并创建平行语料库。接下来一节将介绍我们的神经机器翻译实验，并分析所得结果。最后，我们总结研究发现，并为未来研究提出建议。

2 主要数据集

2.1 Vuk'uzenzele南非多语言语料库

Vuk'uzenzele 数据集由南非政府杂志 Vuk'uzenzele[1] 的各期刊物编制而成。作为一本杂志，其文本主要聚焦时事、政治、娱乐以及其他与杂志出版相关的主题。Vuk'uzenzele 数据集全面呈现了相应时期南非的语言和讨论主题，让研究人员得以深入了解南非的历史和文化。因此，对于任何希望分析南非政治、时事和流行文化的研究人员而言，Vuk'uzenzele 数据集都是一项丰富的资源。

2.1.1 Vuk'uzenzele 的创建

Vuk'uzenzele 的原始数据是从该杂志的 PDF 版中抓取的。Vuk'uzenzele 的主要版本为英语。只有少数英语文章被翻译成另外 10 种南非官方语言（南非语、北恩德贝莱语、科萨语、祖鲁语、北索托语、南索托语、斯瓦蒂语、文达语、聪加语和茨瓦纳语）。因此，我们创建了一套处理流程，用于识别应从特定期号的各语言 PDF 文件中提取哪些文章。随后，我们提取单篇文章并将其存入文本文件。由于 PDF 提取过程的限制，提取出的文本文件仍存在一些问题。为清理文本，一名团队成员会逐一检查每个提取出的文本文件，并按以下方式进行格式整理：

- 第 1 行：文章标题（使用的语言）
- 第 2 行：空行
- 第 3 行：文章作者（如有。若无，则默认为 *Vukuzenzele* 未署名）
- 第 4 行：空行
- 第 5 行至末尾：文章正文

经过人工审阅后，数据更易于分析和可视化。这项耗费大量劳力的工作对于全面且有意义地分析杂志内容是必要的。人工审阅和提取的过程虽然耗时，但最终证明是值得的，因为它有助于更深入地了解 Vuk'uzenzele 的内容。截至本文撰写时，我们已有 *53* 期报纸，涵盖*2020*年*1*月至*2022*年*7*月。团队今后还会陆续添加更多内容。我们已建立自动化流程来下载和存档 PDF 文件，但提取并识别翻译文章仍需人工操作。数据集、代码和自动化爬虫可在 https://github.com/dsfsi/vukuzenzele-nlp 以及 Zenodo[2]（Marivate等人，2023a）获取。我们以便于其他研究人员扩展、改编和增补的格式提供这些资源（数据采用 *CC-4.0-BY-SA* 许可协议，代码采用 *MIT* 许可证）。

2.2 ZA-Gov 多语言南非语料库

ZA-Gov 多语言语料库数据集由南非政府内阁会议后的讲话构建而成。因此，该数据集涵盖了多种主题，包括能源、劳工、公共服务、犯罪、新冠肺炎、国际关系、环境以及政府事务（例如政府任命、内阁决策等）。这让我们得以一窥南非政府的运作方式，以及它如何应对各种内部和外部挑战。

2.2.1 ZA-Gov-multilingual 的创建

ZA-Gov 多语言原始数据抓取自南非政府网站（https://www.gov.za/），该网站发布了所有内阁声明及其译文。数据经过提取并整理为 JSON 格式。每篇讲话的 JSON 负载记录了以下内容：

- 日期，
- 日期时间，
- 标题（英语），

[1] https://www.vukuzenzele.gov.za/

[2] https://doi.org/10.5281/zenodo.7598539

• 网址（语音的顶部网址），

• 每种语言的语言数据（英语、南非荷兰语、南恩德贝莱语、科萨语、祖鲁语、北索托语、*sep*、茨瓦纳语、*ssw*、文达语、聪加语）。

– 标题（使用该语言）、–
文本（使用该语言）、 –
网址（翻译对应的网址）。

这种结构便于研究人员和分析师执行各种自然语言处理、数据挖掘和机器学习任务，例如情感分析、主题建模、分类、语言建模等。例如，通过情感分析和文本挖掘，分析师可以研究内阁成员声明中的观点，并追踪这些主题随时间的演变。截至撰写本文时，该数据集包含 *162* 份内阁声明，时间跨度为 *2013*年*5*月*2*日至*2022*年*12*月*1*日。当 gov.za 网站上有新的、翻译声明可用时，数据集将自动更新。数据集、代码和自动化爬虫可从以下网址获取：https://github.com/dsfsi/gov-za-multilingual 以及 Zenodo[3] （Marivate等人，2023b）。我们以便于其他研究人员扩展、混合使用和补充的方式提供该数据集（数据采用 *CC-4.0-BY-SA* 许可协议，代码采用 *MIT* 许可证）。

3 语料库作为其他自然语言处理任务和进一步研究的基础

除了支持神经机器翻译模型的构建（见下一节）之外，我们的数据集还有望成为翻译以外诸多其他自然语言处理任务的基础。我们相信，这些数据集将为研究南非政府的沟通方式提供宝贵资源，也可直接用于构建多语言文档分类系统（Schwenk和Li，2018）、简化系统（Lu等人，2021；Siddharthan，2014；Martin等人，2022）、实体抽取（Tedeschi等人，2021；Chen等人，2018；Pappu等人，2017；Emelyanov和Artemova，2019）以及其他自然语言处理任务。为进一步拓展数据集的用途，我们建议参考平行意义库（Abzianidze等人，2017）等研究。这类资源可以为跨语言知识迁移提供启发，并为南部非洲语言提供可能有用的新基准

[3]https://doi.org/10.5281/zenodo.7635167

以及南非以外的语言。我们设想，这些数据集将成为进一步研究南非语言和非洲语言多语言自然语言处理的起点。

4 处理与汇编方法

这些数据集是由南非 11 种官方语言构成的双向平行语料库，这些语言列于表1 中，并附有相应的 ISO 639-2 代码。数据集包含以南非官方语言书写的文本，包括南非语、英语、北恩德贝莱语、科萨语、祖鲁语、斯瓦蒂语、北索托语、聪加语、斯瓦蒂语、文达语和茨瓦纳语。因此，这些语言可以组成 55 种配对方式，在每个数据集中形成 55 个不同的语料库。数据集采用 ISO 639-2 语言代码命名，例如“aligned-afr-zul.csv”。由于汇编方式不同，一些数据集的观测值多于其他数据集，这可能导致结果有所差异。例如，如果用于神经机器翻译，那么观测值更多的数据集可能比观测值较少的数据集更有助于为两种语言生成更好的模型。这些数据的汇编有助于进一步探索南非语言与话语的复杂性，从多维角度呈现语言在南非的使用和解读方式。通过这些数据集，可以探索南非语言使用的多样性，并深入了解不同语言之间的互动。

表 1：语言列表及 ISO 639-2 代码

Name	Code
祖鲁语	zul
科萨语	xho
南非语	afr
英语	eng
北索托语	nso
茨瓦纳语	tsn
聪加语	tso
南索托语	sot
斯瓦蒂语	ssw
文达语	ven
北恩德贝莱语	nbl

4.1 预处理

在进行LASER编码和对齐之前，需要对原始抓取数据进行预处理，以加以整理。

预处理步骤如下所列，并且会略有不同，因为源数据和抓取方法会影响所获得的数据。例如，ZA-Gov-Multilinguals包含许多嵌套编号，即*2.1.2*，需要将其移除；相比之下，Vuk'uzenzele数据使用项目符号进行列举。

4.1.1 Vuk'uzenzele

收集到的原始文本按以下方式进行了预处理：

- 文本已转换为小写。
- 连字符和项目符号已移除。
- 双空格、制表符和换行符均替换为单个空格。
- 标准撇号（即 ’）取代了 Unicode 撇号。

4.1.2 ZA-Gov-multilingual

收集的数据中的原始文本按以下方式进行了预处理：

- 移除行首的点号（或一位数或多位数的数字）
- 在格式为 $x \cdot y$ 的一串数字后插入句号。
- 在格式为 x的一串数字后添加句号。
- 将一系列标点符号（例如句号、冒号、分号或这些符号的组合）后跟字母的情况替换为单个句号。

4.2 语料库对齐

预处理完成后，文本被传递给 NLTK 分词器“punkt”，该工具会返回一个由句子词元组成的向量。n 个分词后的句子被发送到 LASER 编码器，该编码器将其编码为n 个句子向量，每个向量长度为 1024。对这些句子向量进行比较，并使用余弦相似度算法计算两个向量之间的相似度得分（0 到 1），具体如第4.2.1节所述。

4.2.1 LASER 编码

为了比较文本的相似度，我们使用了 LASER 编码器。LASER 是 Language-Agnostic Sentence Representations（语言无关句子表示）的缩写，是 Facebook AI Research 的一个研究项目。LASER 通过将句子编码为向量来生成句子表示。这些向量充当句子的机器表示。可以使用余弦相似度比较这些向量，从而得出一个介于 0 和 1 之间的分数。余弦分数越接近 1，表示相似度越高。该分数记录在 LASER 数据集（可在数据集仓库中获取）中。以下表格列出了每个数据集中得分高于 0.65（即相似度为 65%）的观测值2 和3。所有数据集中观测值数量的完整表格收录于数据集仓库的 README 文件中：
h
ttps://github.com/dsfsi/gov-za-multilingual（ZA-Gov-Multilingual）以及 https://github.com/dsfsi/vukuzenzele-nlp（Vuk'uzenzele）。

表 2：Vuk'uzenzele 中余弦分数大于或等于 0.65 的观测值最多的前十个数据集。

语言 pair	No. 观测值占 Vuk'uzenzele
ssw-科萨语	2,202
ssw-祖鲁语	2,183
科萨语-祖鲁语	2,102
北索托语-科萨语	2,081
北索托语-茨瓦纳语	2,071
斯瓦蒂语-聪加语	2,034
nso-ssw	2,021
茨瓦纳语-聪加语	2,020
茨瓦纳语-科萨语	2,009
聪加语-科萨语	2,009

4.3 后处理

对于 LASER 数据集，对齐数据的源句、目标句和余弦分数被写入一个 CSV 文件，文件名遵循以下格式：’aligned-{src_语言_代码}- {tgt_语言_代码}.csv’，例如 ’aligned-afr-zul.csv’。有关数据集命名中使用的语言代码，请参阅1 中的语言列表。

对于简单对齐数据集，源句和目标句被写入一个 CSV 文件，其命名结构与 LASER 数据集相同。

表3：余弦分数大于或等于0.65且观测值最多的前十个数据集

语言 pair	观测值数量 在 ZAgov 多语言语料库中 gual
nbl-ven	18,984
nso-ssw	18,697
zul-ssw	18,563
xho-ssw	18,387
科萨语-祖鲁语	18,145
科萨语-北索托语	18,110
科萨语-聪加语	17,954
斯瓦蒂语-聪加语	17,880
祖鲁语-聪加语	17,789
祖鲁语-北索托语	17,630

5 神经机器翻译基准

对于资源匮乏的非洲语言而言，对齐句子语料库数量有限，这阻碍了从头开始训练的神经机器翻译模型的质量（Martinus 和 Abbott, 2019；Nekoto 等，2020年；Adelani 等，2022）。最近，Adelani 等 (2022) 通过在少量对齐句子上微调大规模多语模型（包括 M2M100 模型（Fan 等，2021）来解决这一问题。M2M100 模型是一种非英语中心的多对多语言模型，经过训练，可在 100 种语言之间直接翻译，其中包括五种南非官方语言（Fan 等，2021）。Adelani 等(2022) 展示了如何有效利用该模型，以少量数据为预训练中未涵盖的语言和领域创建神经机器翻译系统。

在他们工作的基础上，我们通过微调 M2M100 模型，为政府出版物领域的 Vuk'uzenzele 和 ZA-gov-multilingual 数据集创建基准翻译。为将结果置于具体背景中并便于比较，我们还使用从南非数字语言资源中心（McKellar，(2021，2，2，2，0，2）(https://sadilar.org）获取的现有 Autshumato 平行语料库子集，对 M2M100 模型进行微调。我们着重为数据集中的低资源非洲语言提供神经机器翻译基准；由于南非语的数字数据集相对丰富，因此不纳入南非语译文，并将其留待未来工作处理。

5.1 预处理

对齐后的数据集经过处理，删除了重复和冲突的译文（包括源句和目标句中的译文），随后打乱顺序以消除顺序偏差，然后创建训练集、测试集和开发集。数据划分比例为：70% 用于训练，20% 用于测试，10% 用于开发。为便于比较，所有模型均使用“xxx-eng”翻译方向进行微调，其中“xxx”代表非洲本土源语言，“eng”代表英语目标译文。

现有的 Aushumato 平行语料库（从各种政府资源和网络爬取中提取（Groenewald 和 Fourie，2009 ））在领域上与所创建的 ZA-gov-multilingual 平行语料库相当，但目前数据集规模大得多。因此，为了直接比较神经机器翻译结果，我们在“xxx-英语”翻译方向上，提取与 ZA-gov-multilingual 语料库相同数量的预处理后对齐句子。用于神经机器翻译基准测试的语料库句子数量和词元数量见附录 A.1 表5、6 和7。

5.2 结果

Vuk'uzenzele、ZA-gov-multilingual 以及可用 Autshumato平行语料库子集的 M2M100 微调基准测试结果见表4。表中列出了微调的翻译方向，并为便于参考，特别标出了未纳入 M2M100 原始预训练的源语言。此外，各数据集中取得的最高 BLEU 分数结果以粗体显示。对于某种语言，如果 Autshumato平行语料库无法获取或不存在，则以“-”符号表示。

在 Vuk'uzenzele 数据集上进行微调所取得的整体 BLEU 分数最低，这符合预期，因为与 ZA-gov-multilingual 和 Autshumato 数据集相比，其对齐数据集规模较小。最高的 BLEU 分数分布于 ZA-gov-multilingual 和 Autshumato 神经机器翻译模型之间，但分布并不一致；其中，ZA-gov-multilingual 模型在茨瓦纳语、聪加语和祖鲁语方面取得了更高的分数。这可能是由于从 Autshumato 数据集中随机抽取子集所致，也可能是 Autshumato 项目在创建对齐语料库时，所采用的源组合和清理方法存在差异所致。值得注意的是，这些差异

表 4：使用 Vuk'uzenzele（Vuk.）、ZA-gov-multilingual（Gov.）以及可用 Autshumato 数据集（Aut.）子集进行 xxx-英语翻译的超大规模多语言迁移 BLEU 分数

翻译方向	BLEU		
	Vuk.	Gov.	Aut.
南恩德贝莱语→英语	7.33	8.04	**12.24**
北索托语→英语	9.29	**26.50**	-
ssw→英语	4.80	**28.72**	-
南索托语→英语	4.55	10.21	**14.83**
茨瓦纳语→英语	2.80	**29.68**	28.04
聪加语→英语	13.86	**35.40**	32.10
文达语→英语	2.32	9.68	**17.24**
科萨语→英语	6.05	26.81	-
祖鲁语→英语	9.97	**30.03**	25.90

子集的选取会导致结果不同，深入分析则留待未来工作。

在所有数据集中，我们在聪加语（'tso-eng'）上取得了最高的基准测试结果。M2M100 模型此前并未针对这种语言进行预训练，这表明迁移学习对于新的低资源语言数据集是有效的。

另需指出的是，现有针对 Autshumato 数据集的神经机器翻译资源，仅涵盖聪加语、茨瓦纳语、塞佩迪语、南索托语和祖鲁语的「英语-xxx」翻译方向（Skosana 和 Mlambo，2021）。因此，我们的贡献扩展了基准翻译资源（在政府数据领域），将其涵盖范围拓展至北恩德贝莱语、科萨语、斯瓦蒂语和文达语；同时也拓宽了翻译方向，不再局限于以英语为源语言。

6 结论

最后，本文介绍了两个自动对齐的多语种语料库，以促进不同语言文本之间的翻译。这些数据集丰富了用于训练非洲语言神经机器翻译模型的语料库资源，而这些语言在当代自然语言处理研究中往往被忽视或资源匮乏。作者希望这些数据集能有助于为低资源非洲语言创建神经机器翻译模型，并让这些模型用于促进人们获取翻译服务，帮助非洲语使用者和作者更有效地使用自己的母语交流。作者还希望这些数据集能进一步推动关于非洲语言多语性的自然语言处理研究，并增进人们对非洲存在的各种方言的认识。

7 局限性

本文讨论的神经机器翻译模型是为所述数据集的基准测试而创建的，尚未针对生产用途进行全面的质量测试。我们也仅测试了以英语为目标语言对 M2M100 模型进行微调的效果，并未将神经机器翻译系统扩展至南非本土语言之间的翻译，因此仍需开展进一步的基准测试。我们希望将（McKellar 和 Puttkammer，2020）创建的评估集纳入测试。该评估集包含所有 11 种南非官方语言的数据（这些数据未纳入 Autshumato 语料库），可用于更准确地比较神经机器翻译模型。需要指出的是，尽管 BLEU 分数结果令人鼓舞，但仍需对翻译模型进行定性语言学分析，以确定某些语言翻译（即「tso-eng」）的 BLEU 分数是否与我们领域内的翻译准确性相符。作为未来工作，我们希望与相关语言学家合作，提升此类南非语言神经机器翻译系统的质量和有效性（Skosana 和 Mlambo，2021）。

8 伦理声明

用于翻译模型基准的数据集完全取自南非政府资源。因此需要强调的是，如果这些模型用于生产环境，它们可能会忽视某些社会结构，并且会体现数据集收集时的主导政党的立场（Bender 等人，2021）。我们还注意到，这些基准模型和数据集并未经过整理，以确定其中存在的任何偏见。因此，系统中现有的任何偏见在用于自然语言处理下游生产任务时，都可能对特定群体造成伤害（Bender 等人，2021）。

9 致谢

我们谨此致谢以下资助：ABSA 数据科学讲席、GoogleResearch 学者计划、TensorFlow 机器学习奖助金以及 NVIDIA 公司硬件奖助金。衷心感谢所有匿名的 RAIL 审稿人提出宝贵意见。

参考文献

Lasha Abzianidze、Johannes Bjerva、Kilian Evang、Hessel Haagsma、Rik van Noord、Pierre Ludmann、Duc-Duy Nguyen 和 Johan Bos。2017。《平行意义库：构建一个以组合式意义表示标注的多语言翻译语料库》。载于计算语言学协会欧洲分会第 *15* 届会议论文集：第 *2* 卷，短篇论文，第 242–247 页。

David Adelani、Jesujoba Alabi、Angela Fan、Julia Kreutzer、Xiaoyu Shen、Machel Reid、Dana Ruiter、Dietrich Klakow、Peter Nabende、Ernie Chang、Tajuddeen Gwadabe、Freshia Sackey、Bonaventure F. P. Dossou、Chris Emezue、Colin Leong、Michael Beukman、Shamsuddeen Muhammad、Guyo Jarso、Oreen Yousuf、Andre Niyongabo Rubungo、Gilles Hacheme、Eric Peter Wairagala、Muhammad Umair Nasir、Benjamin Ajibade、Tunde Ajayi、Yvonne Gitau、Jade Abbott、Mohamed Ahmed、Milli- cent Ochieng、Anuoluwapo Aremu、Perez Ogayo、Jonathan Mukiibi、Fatoumata Ouoba Kabore、God- son Kalipe、Derguene Mbaye、Allahsera Au- guste Tapo、Victoire Memdjokam Koagne、Edwin Munkoh-Buabeng、Valencia Wagner、Idris Abdul- mumin、Ayodele Awokoya、Happy Buzaaba、Bless- ing Sibanda、Andiswa Bukula 和 Sam Manthalu。2022。几千条翻译大有裨益！利用预训练模型进行非洲新闻翻译。见*2022* 年北美计算语言学协会分会会议论文集：人类语言技术，第 3053–3070 页，美国西雅图。计算语言学协会。

Emily M. Bender、Timnit Gebru、Angelina McMillan- Major 和 Shmargaret Shmitchell。2021。论随机鹦鹉的危害：语言模型会不会过大？载于*2021* 年 *ACM* 公平、问责与透明度会议论文集，FAccT '21，第 610–623 页，美国纽约州纽约市。计算机协会。

Muhao Chen、Yingtao Tian、Kai-Wei Chang、Steven Skiena 和 Carlo Zaniolo。2018。协同训练知识图谱与实体描述的嵌入，以实现跨语言实体对齐。*arXiv* 预印本，*arXiv: Arxiv-1806.06478*。

Anton A. Emelyanov 和 E. Artemova。2019。使用预训练嵌入、注意力机制和 NCRF 进行多语言命名实体识别。*BSNLP@ACL*。

Angela Fan、Shruti Bhosale、Holger Schwenk、Zhiyi Ma、Ahmed El-Kishky、Siddharth Goyal、Mandeep Baines、Onur Celebi、Guillaume Wenzek、Vishrav Chaudhary 等。2021年。超越英语中心主义的多语言机器翻译。《机器学习研究杂志》，22(107):1–48。

Hendrik J Groenewald 和 Liza du Plooy。2010年。为三个南非语种处理平行文本语料库

l语言对 irs in th[e auts] humato pro ject. *AfLaT 2010*年，第27页。Hendrik Johannes Groenewald 和 Wildrich Fourie。2009年。介绍 Autshumato 集成翻译环境。载于第*13*届欧洲机器翻译协会年会论文集。Pratik Joshi、Sebastin Santy、Amar Budhiraja、Kalika Bali 和 Monojit Choudhury。2020年。自然语言处理领域语言多样性与包容性的现状与未来。载于第*58*届计算语言学协会年会论文集，第6282–6293页，线上。计算语言学协会。

Xinyu Lu、Jipeng Qiang、Yun Li、Yunhao Yuan 和 Yi Zhu。2021年。一种用于构建多种语言句子简化语料库的无监督方法。实证自然语言处理会议。

Vukosi Marivate、Daniel Njini、Andani Madodonga、Richard Lastrucci、Isheanesu Dzingirai 和 Jenalea Rajab。2023a。Vuk'uzenzele 南非多语言语料库。

Vukosi Marivate、Tshephisho Sefara、Vongani Chabalala、KeamogetsweMakhaya、Tumisho Mokgonyane、Rethabile Mokoena 和 Abiodun Modupe。2020年。探讨低资源语言数据集创建、整理和分类方法：茨瓦纳语和北索托语。载于首届非洲本土语言资源研讨会论文集，第 15–20 页。

Vukosi Marivate、Matimba Shingange、Richard Las- trucci、Isheanesu Dzingirai 和 Jenalea Rajab。2023b。南非政府 ZA 多语言语料库。

Louis Martin、Angela Fan、Eric Villemonte de la Clerg- erie、Antoine Bordes 和 Benoît Sagot。2022年。MUSS：通过挖掘释义进行多语无监督句子简化。国际语言资源与评测会议。

Laura Martinus 和 Jade Z Abbott。2019。聚焦非洲语言的神经机器翻译。*arXiv*预印本 *arXiv:1906.05685*。

Cindy McKellar。（2020年）。Autshumato 英语-文达语平行语料库，版本1。0, [多语言文本语料库：对齐]。检索自https://hdl.handle.net/20.500.12185/569。

Cindy McKellar。（2021）。Autshumato 英语-北恩德贝莱语平行语料库，版本 1。0, [多语言文本语料库：对齐]。检索自https://hdl.handle.net/20.500.12185/572。

Cindy McKellar。（2022）a。Autshumato 英语-祖鲁语平行语料库，第 2 版。0, [多语言文本语料库：对齐]。检索自 https://hdl.handle.net/20.500.12185/575。

Cindy McKellar。（2022）b。Autshumato 英语-南索托语平行语料库，第 1 版。0, [多语言文本语料库：对齐]。检索自https://hdl.handle.net/20.500.12185/577。

Cindy McKellar. (2022)c. Autshumato 英语-茨瓦纳语平行语料库，版本 2。0, [多语言文本语料库：对齐]。检索自https://hdl.handle.net/20.500.12185/578。

Cindy McKellar. (2022)d. Autshumato 英语-聪加语平行语料库，版本 2。0, [多语言文本语料库：对齐]。检索自https://hdl.handle.net/20.500.12185/579。

Cindy A McKellar 和 Martin J Puttkammer. 2020年。用于评估 11 种南非语言之间机器翻译可比性的*DataBrief*，29:105146。

Wilhelmina Nekoto、Vukosi Marivate、Tshinondiwa Matsila、Timi Fasubaa、Taiwo Fagbohungbe、Solomon Oluwole Akinola、Shamsuddeen Muhammad、Salomon Kabongo Kabenamualu、Salomey Osei、Freshia Sackey 等。2020年。非洲语言低资源机器翻译的参与式研究：一项案例研究。载于计算语言学协会 *Findings*：*EMNLP 2020*，第 2144–2160 页。

Aasish Pappu、Roi Blanco、Yashar Mehdad、Amanda Stent 和 Kapil Thadani。2017。轻量级多语言实体抽取与链接。载于第十届 *ACM* 国际网络搜索与数据挖掘会议论文集，第 365–374 页。

Surangika Ranathunga 和 Nisansa de Silva。2022。有些语言比其他语言更平等：深入探究自然语言处理领域中的语言差异。载于亚太计算语言学协会第二届会议暨自然语言处理国际联合会议第十二届会议论文集（第一卷：长篇论文），第 823–848 页，仅在线发布。计算语言学协会。

Holger Schwenk 和 Matthijs Douze。2017。利用神经机器翻译学习联合多语言句子表示。*ACL2017*，第 157 页。

Holger Schwenk 和 Xian Li。2018。一个涵盖八种语言的多语言文档分类语料库。国际语言资源与评测会议。

Advaith Siddharthan。2014。文本简化研究综述。*ITL*-国际应用语言学杂志，165(2):259–298。

Nomsa Skosana 和 Respect Mlambo。2021。南非语言 Autshumato 机器翻译网络服务简述。*Literator*，42(1):7。

Simone Tedeschi、Valentino Maiorca、Niccolò Campolungo、Francesco Cecconi 和 Roberto Navigli。2021。WikiNEuRal：融合神经网络与面向多语言命名实体识别的基于知识的银标数据创建。载于计算语言学协会论文集：*EMNLP 2021* 研究成果，第 2521–2533 页，蓬塔卡纳，多米尼加共和国。计算语言学协会。

A 附录

附录A.1 数据统计

用于神经机器翻译基准测试的数据集的数据统计见表5、6和7。

表5：Vuk'uzenzele（Vuk.）数据集的翻译数据特征

翻译方向	Size #句子（#源语言词元 / #目标语言词元）
南恩德贝莱语→英语	136（3.4k / 3.9k）
北索托语→英语	1715（53.7k / 41.6k）
ssw→英语	1588（29.9k / 37.6k）
南索托语→英语	260（9.9k / 7.5k）
茨瓦纳语→英语	1366（49.8k / 31.7k）
聪加语→英语	1998（58.9k/ 46.6k）
文达语→英语	230（9.1k / 7k）
科萨语→英语	1338（25.8k / 31.5k）
祖鲁语→英语	1874（34.1k / 43k）

表 6：ZA-gov-multilingual（政府）数据集中翻译数据的特征

翻译方向	Size #句子数（#源语言词元 / #目标语言词元）
南恩德贝莱语→英语	3513（63.9k / 107k）
北索托语→英语	14742（460.9k / 375k）
ssw→英语	15139（291k/ 377.8k）
南索托语→英语	4995（145.9k / 153.5k）
茨瓦纳语→英语	14068（493.1k / 362.2k）
聪加语→英语	15393（466.4k / 381.2k）
文达语→英语	3404（68.2k / 96.6k）
科萨语→英语	15853（318.2k / 389.5k）
祖鲁语→英语	15503（327.5k / 384.1k）

表7：现有 Autshumato数据集子集的翻译数据特征（Aut.）

翻译方向	Size #句子（#源语言词元 / #目标语言词元）
南恩德贝莱语→英语	3513（48.4k / 65k）
南索托语→英语	4995（118.3k / 100.4k）
茨瓦纳语→英语	14068（335.1k / 274.7k）
聪加语→英语	15393（193.7k/166k）
文达语→英语	3404（80.6k / 65.9k）
祖鲁语→英语	15503（231.8k / 307.6k）