arXiv:2305.00606v1 [cs.CL] 1 May 2023

面向塞内加尔沃洛夫语的低资源机器翻译

Derguene Mbaye[1,2], Moussa Diallo[1], 和 Thierno Ibrahima Diop[2]

[1] 谢赫·安塔·迪奥普大学，达喀尔，塞内加尔，
derguene.mbaye@ucad.edu.sn,
moussa.diallo@esp.sn https://www.ucad.sn/[2] Baamtu，达喀尔，塞内加尔 thierno.diop@baamtu.com

摘要。 近年来，自然语言处理（NLP）研究取得了长足进展，重大突破也树立了新的基准测试。然而，这些进展主要惠及英语和法语等通常被称为资源丰富语言的一部分语言。其他资源较为匮乏的多数语言则被落在后面，沃洛夫语在内的大多数非洲语言便是如此。在这项工作中，我们构建了一个包含 123,000 个句子的沃洛夫语-法语平行语料库，并使用该语料库，在不同数据配置下对基于循环神经网络（RNN）的机器翻译模型进行了实验。我们发现，使用子词化数据训练的模型，以及使用法语-英语语言对训练的模型，性能均有所提升；在相同实验条件下，这些模型的表现优于使用法语-沃洛夫语语言对训练的模型。

关键词： 低资源、机器翻译、非洲语言、循环神经网络

1 引言

机器翻译（MT）系统可以将源语言中的文本序列（或音频源）转换为目标语言中的相同序列。长期以来，统计机器翻译（SMT）系统[1] 一直是最受欢迎的方法，直到神经机器翻译（NMT）系统 [2] 出现并取得越来越高的性能。然而，这类系统的质量始终与其设计中使用的数据量密切相关 [3]。因此，最先进的机器翻译系统采用序列到序列模型，并结合注意力机制 [4]以及 Transformer 架构 [5]进行开发。对于英语等不受这一限制的语言，人们称其为资源丰富语言，且这类语言拥有数百万个句对；大多数其他语言则属于“低资源”（LR）范畴。然而，“低资源”一词涵盖多种情况，其范围也不仅限于语言本身，还可涉及数据稀缺的领域或任务，即使相关语言本身属于资源丰富语言也是如此。 [6] 对此作了说明，其中“低资源”这一概念从三个不同方面进行界定：任务专属标签、无标注语言文本的可用性，以及

辅助数据。如 [7], 所示，大多数非洲语言都符合这一描述，这使研究人员的工作变得困难，也导致非洲语言在自然语言处理研究中的代表性不足 [8]。沃洛夫语尤其如此：除了缺乏数据之外，自然语言处理领域对这种语言的研究也很少。

在 [9] ，研究人员收集了一个涵盖 4 种非洲语言（包括沃洛夫语）的自动语音识别（ASR）数据集，并利用该数据集设计了首个针对沃洛夫语的自动语音识别系统。在 [10],，研究人员启动了首部符合 LMF[3] 标准的沃洛夫语协作式在线词典的编纂工作。作为非洲语言—法语词典（DiLAF）[4] 项目的一部分，研究人员编纂了涵盖 7 种非洲语言（包括沃洛夫语）的多部词典。然而，在本文撰写时，除沃洛夫语词典外，其他词典均已上线。作者在[11] 探讨了基于有限状态方法开发沃洛夫语形态分析器、实现并评估沃洛夫语基于词汇功能语法的句法分析器 [12], 以及创建沃洛夫语通用依存（UD）树库 [13] 。该树库是尼日尔-刚果语系语言北大西洋语支中的首个通用依存树库。在 [14], ，作者介绍了一种以词典为词库、以沃洛夫语形态分析器为基础的方法，并据此研究了沃洛夫语拼写检查器的设计。

然而，据我们所知，专门研究沃洛夫语-法语机器翻译系统的工作只有 [15] ，作者在其中介绍了一个包含 70,000 个沃洛夫语-法语平行句的语料库，并据此开发了词嵌入模型和基于长短期记忆网络的翻译模型；以及 [16] ，作者在其中将语料库扩充至 83,000 个句子，并据此训练了两个基于 Transformer 架构的神经机器翻译系统，分别用于法语→沃洛夫语和沃洛夫语→法语方向。然而，[15] 中报告的结果采用准确率指标，因此难以评估其系统的实际翻译质量。此外，也有人开发了包含沃洛夫语在内的多语种神经机器翻译系统，例如 [17] ，作者利用现有的预训练模型，为 16 种非洲语言构建了低资源翻译系统。Meta 的 No Language Left Behind 项目[5] 能够实现 200 种语言之间的互译，其中也包括沃洛夫语。尽管如此，除沃洛夫语之外，针对低资源语言神经机器翻译（LRL-NMT）的研究也已取得大量进展。Masakhane 社区[6] 提出通过采用参与式方法应对这一挑战， [18] 让所有相关资源人员参与其中，进而为 30 多种语言构建机器翻译数据集和基准测试。 [19] 对解决低资源语言神经机器翻译问题的不同方法进行了详细研究，并制定了一套指南，用于根据给定的低资源语言数据条件选择可能适用的神经机器翻译技术。 [20] 针对不同的翻译系统开展了一系列实验，涵盖

[3] 词汇标记框架[4] http://pagesperso.ls2n.fr/~enguehard-c/DiLAF/index.php[5] https://ai.facebook.com/research/no-language-left-behind/[6] https://www.masakhane.io/

基于神经网络和基于统计的方法，用于将英语翻译成冰岛语。然而，这些研究大多基于 Transformer 架构，而该架构对数据的需求非常大。在低资源环境中，较早的架构（如循环神经网络）可能表现更好，因为其所需参数较少。

本文介绍了法语-沃洛夫语平行句子数据收集工作的进展情况。截至目前，该语言对中已收集到规模最大的语料库，其中包含经过筛选并在句子层面对齐的 123,000 个句子。随后，我们拟在 [15] 的基础上进一步开展研究，通过使用双语评估替补指标（BLEU） [21]，对模型进行评估，探究循环神经网络模型在该语料库上的表现；与准确率相比，该指标更具代表性。由于已有研究表明，子词切分，即将语料库分割为词或子词，往往能够提升翻译模型的性能[22]，，因此我们进一步实验了这一方法对模型的影响。本文的组织结构如下：

– 第 2 节介绍沃洛夫语。– 第 3 节介绍数据收集和筛选过程。– 第 4 节介绍所进行的实验。– 第 5 节展示实验结果。– 第 6 节总结全文。

2 沃洛夫语

沃洛夫语是一种西大西洋语支语言，主要在塞内加尔和冈比亚使用，在毛里塔尼亚南部也有人使用。它属于尼日尔-刚果语系的大西洋语支，目前有超过700万人分布在三个西非国家并使用沃洛夫语。虽然塞内加尔人口中只有约40%是沃洛夫人，但约90%的人将该语言作为第一、第二或第三语言使用[7]。

沃洛夫语有两种主要的地域变体：一种在塞内加尔使用，另一种在冈比亚使用 [23]。尽管使用沃洛夫语的人能够相互理解，塞内加尔沃洛夫语和冈比亚沃洛夫语仍是两种不同的语言：二者分别拥有各自的 ISO 639-3 语言代码（分别为“WOL”和“WOF”）。沃洛夫语长期以来一直使用一种称为阿贾米文或沃洛法尔文的阿拉伯字母书写，后来也被改写为罗马字母。

沃洛夫语是一种黏着语 [11]，其字母表与法语字母表十分接近：除了 H、V 和 Z [24]，其字母表中的字母都与法语相同。~此外，沃洛夫语还使用字符 η （“ng”）和 N （“gn”，如西班牙语中的用法）。沃洛夫语中也有重音符号，但数量有限（$\acute{A}$、$\acute{E}$、$\tilde{A}$、$\acute{O}$）。该语言使用 29 个源自拉丁字母的罗马字母，其中大多数参与构成二合字母，用来表示长辅音和前鼻音化阻塞音。与许多其他尼日尔-刚果语系语言不同，沃洛夫语没有声调。不过，沃洛夫语音节的强度有所不同，例如，长元音的发音强度比短元音更大。书写时用重复的元音字母表示长度，而且大多数

[7] https://www.axl.cefan.ulaval.ca/afrique/senegal.htm

沃洛夫语辅音也可以延长（重复书写）。然而，沃洛夫语尚未实现标准化（有些资料甚至将“H”排除在字母表之外），因为从未有任何一种变体被认可为标准。尽管如此，达喀尔应用语言学中心（CLAD）仍负责协调沃洛夫语的正字法标准化 [9]。

3 数据收集

3.1 语料库

构建数据集是一项繁琐且耗时的工作，尤其是对于沃洛夫语这类尚未标准化的语言。学校不教授这种语言，而且很少有人遵循拼写规则，因此，社交网络等来源上的文本十分多样，难以使用。因此，我们选择收集法语数据，因为自殖民化以来，法语一直是塞内加尔的官方语言，并请专业语言学家将其翻译成沃洛夫语，从零开始构建部分数据集。

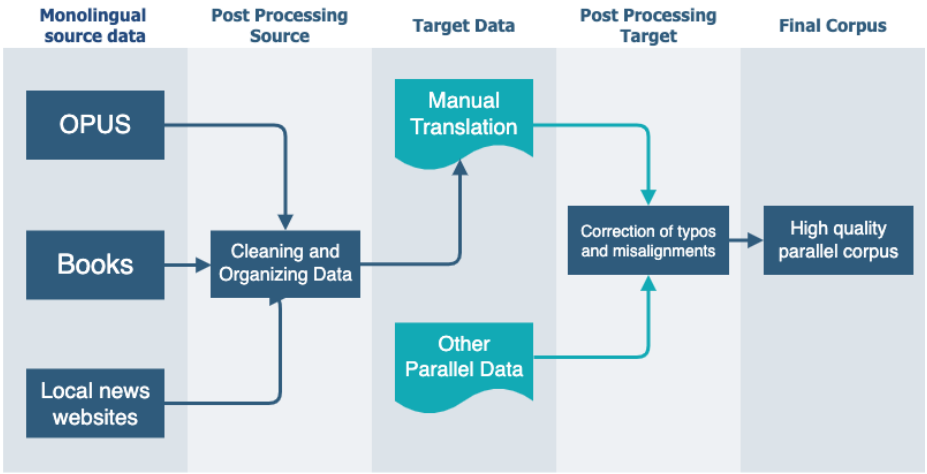


图 1. 数据收集流程

语言学家使用政府制定的官方沃洛夫语字母表[8]进行翻译。法语单语数据收集自现有资源，例如 Opus，以及从在线来源抓取的文本，其中包括新闻网站、宗教文本和博客。我们使用 Opus 收集法语单语文本数据，并收集《古兰经》和《圣经》的译文作为平行文本。我们还从离线来源收集数据，例如已被

[8] http://www.jo.gouv.sn/spip.php?article4802

翻译成沃洛夫语的法语书籍。因此，我们收集了一个包含 123,000 个法语-沃洛夫语平行句子的语料库，使我们的语料库成为迄今为止收集到的最大语料库。

出于实验目的，我们将整个数据集划分为三个子集：训练集、验证集和测试集。验证集和测试集保持固定，并与完整数据集分离，其中验证集包含 16,000 个句子，测试集包含 7,000 个句子。我们仅将训练集的规模从 10,000 个句子变更至 100,000 个句子，每次增加 10,000 个句子。

3.2 数据过滤

在根据不同的实验配置分配数据之前，我们执行了一系列后处理操作。首先，我们进行了分层抽样，以确保验证集和测试集能够代表整个数据集，从而限制抽样偏差。

由于系统的质量直接取决于数据的质量，我们借鉴了 [25] 中提出的方法，随后对数据集进行了过滤。因此，我们删除了两侧使用相同语言书写的句子、重复的句子对，以及两侧完全相同的句子。我们还删除了特殊字符和网址，并过滤掉过长以及在数据集中占比不足的句子。当句子的长度（词数）超过所考虑数据集中句子平均长度的两倍时，我们将其视为过长。

4 实验

尽管我们收集了一个包含 123,000 个句子的语料库，但对于神经机器翻译而言，我们仍处于低资源配置。因此，我们选择了一种中等数据密集型架构（相较于统计机器翻译和 Transformer模型），并利用数据处理方法来最大限度地提升模型性能。

我们使用 OpenNMT [26] 复现了与 [15]类似的架构，以便比较结果。因此，循环神经网络模型由一个长短期记忆网络层组成，该层在编码器和解码器两端均为 [27] 300 个隐藏单元，并配有一个丢弃层。丢弃率设为 0.1，嵌入大小设为 128。我们定义了一个 Adam优化器 [28] ，学习率为 0.001，批次大小设为 4096 个词元。

我们将数据集划分为不同大小的配置，在每种配置中，模型分别沿 Fr→Wo 和 Fr→En 方向进行训练，直至达到收敛。当验证集上的结果在经过 6 次检查点后不再出现改善时，即认为模型已经达到收敛。

对于数据的子词切分，我们使用了 SentencePiece [29]与字节对编码（BPE），它在沃洛夫语 [30]等黏着语中能带来显著的性能提升。随后，我们根据所考虑的尺寸配置的训练集中的所有片段生成词汇表，并使用 BLEU [21]进行自动模型评估。BLEU 是神经机器翻译中应用最广泛的指标，因为它与人工评估结果具有

较高的相关性。我们使用 SacreBLEU[31] 实现[9] BLUE 指标来评估模型。

5 结果

我们在相同的数据规模配置下比较相同的架构：(i) 将原始形式的数据（即未进行子词切分）提供给模型，与训练前进行子词切分后再提供给模型的情况；(ii) 使用不同语言对进行训练，即比较 Fr→En 与 Fr→Wo。第一种情况可用于衡量子词切分对翻译质量的影响；第二种情况则可用于观察语言间的语言特性如何促进或妨碍翻译性能。

表 1 和表 2 展示了翻译实验的结果，所有 BLEU 分数均在测试集上计算。

表 1. 法语-沃洛夫语实验

训练规模	未进行子词切分	进行了子词切分
100k	15.22	16.71
90k	14.41	15.28
80k	15.12	16.09
70k	12.76	14.85
60k	12.11	14.23
50k	10.45	12.14
40k	9.35	11.03
30k	7.33	9.73
20k	5.58	7.45
10k	3.94	4.84

在表 1 中，我们观察到，针对 Fr→Wo 数据，原始语料库与子词化语料库的 BLEU 分数相差约 1.6 分。这可以解释为子词出现频率更高，因此模型能更好地学习它们。图 2 更清楚地展现了这一提升：在每个训练检查点上，模型在子词化数据上的表现都更好。

在表 2 中，我们观察到 Fr→En 数据也呈现类似趋势，此次 BLEU 分数提升约 4 分。比较这两种语言对的实验结果时，我们还注意到，在相同的实验条件下（语料库大小和子词切分），Fr→En 数据的 BLEU 分数比 Fr→Wo 高约 3.5 分。

图 3 清楚地展示了 Fr→En 模型在不同数据集格式下的表现，并显示其在 30k 检查点处急剧下降。这是因为

[9] 版本 2.0.0

表 2. 法语—英语实验

训练规模	无子词化	有子词化
100k	18.88	22.19
90k	18.52	21.11
80k	18.05	20.79
70k	17.82	20.57
60k	16.70	19.28
50k	15.17	18.94
40k	14.18	17.52
30k	4.68	16.22
20k	10.5	14.8
10k	3.34	10.1

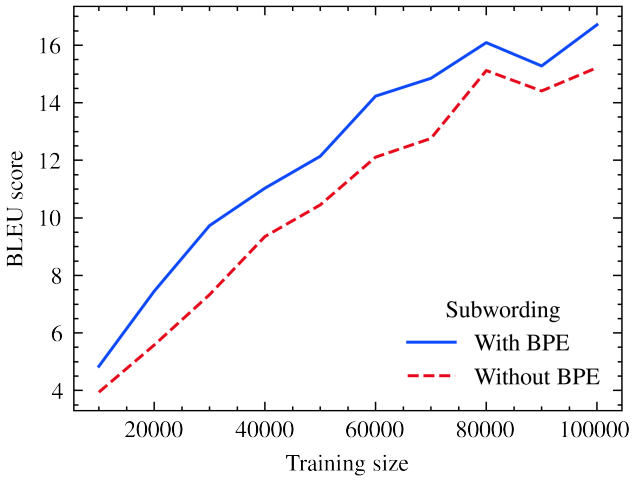


图 2. 在相同数据规模配置下，Fr→Wo 神经机器翻译模型在子词化与非子词化数据上的性能变化

所添加数据片段的质量较低，其中包含大量伪影，说明并非所有数据点都对训练有用。

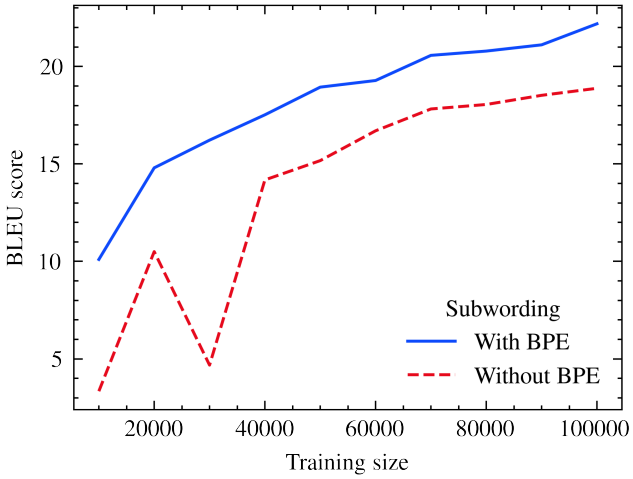


图3。 在相同数据规模配置下，基于子词化和非子词化数据的 Fr→En 神经机器翻译模型的性能变化

除子词切分外，我们还希望观察两种语言之间共享的语言特性能否影响翻译性能。图4和图5展示了两个模型在相同配置（架构和数据）下分别用于 Fr→Wo 和 Fr→En 语言对时的性能。

总体而言，无论数据是否经过子词切分，我们都注意到，在所有训练检查点上，使用 Fr→En 语言对训练的模型性能都优于使用 Fr→Wo 语言对训练的模型，只有 30k 检查点例外，该处观察到性能急剧下降。这可以通过法语和英语之间的语言相似性来解释：尽管两者属于不同语系，但使用相同的字母表。两者还具有 27% 的词汇相似性[23] ，并且一种语言中的词语可以在另一种语言中找到，或源自另一种语言。我们推测，语言对之间的形态学差异会影响模型将一种语言翻译成另一种语言的能力。沃洛夫语字母表的字母数量多于法语字母表，且沃洛夫语的形态更为丰富，这可能会妨碍模型捕捉该语言的特性。

6 结论

本文介绍了一个包含 123,000 个句子的法语—沃洛夫语平行语料库。该语料库大部分是从零开始收集的，因为公开可访问的资源

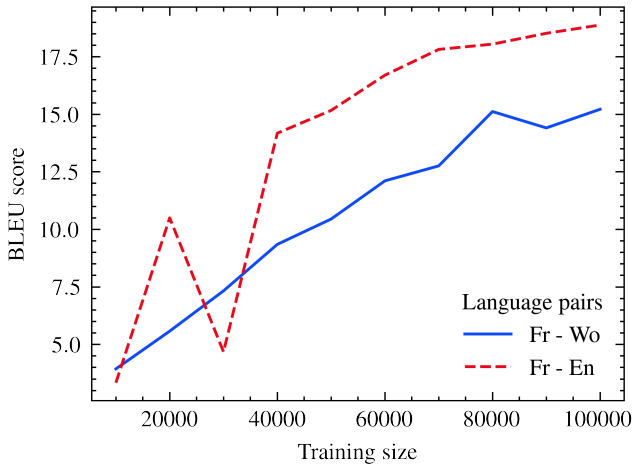


图 4. 在相同数据规模配置下，Fr→Wo 和 Fr→En 神经机器翻译模型在原始数据上的性能变化

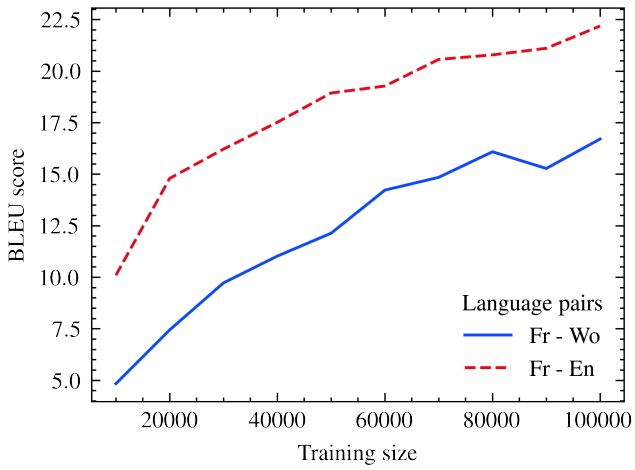


图 5. 在相同数据规模配置下，Fr→Wo 和 Fr→En 神经机器翻译模型在子词化数据上的性能变化

关于这一语言对的研究还很少。由于语料收集项目仍在进行中，数据集尚未开放。随后，我们对基于长短期记忆网络和全局注意力的神经机器翻译模型的多种架构进行了实验，结果表明，这些系统在子词化数据上的效率更高。进一步的实验试图通过在相同实验条件下比较两个不同语言对上的系统，研究语言对之间的语言相似性对翻译性能的影响：Fr→Wo 和 Fr→En。

据我们所知，我们的语料库是迄今为止为这一语言对收集的最大语料库，也是首项针对 Fr↔Wo 语言对展示基于长短期记忆网络的机器翻译系统性能、并采用 BLEU 指标以便更好地评估神经机器翻译模型性能的研究。

然而，BLUE 指标可能会带来偏差，因此不足以全面评估我们系统的实际质量 [32]。子词切分也带来了显著提升，但 SentencePiece 方法不依赖特定语言，未必适用于所有语言。另一方面，即使使用长短期记忆网络或门控循环单元[27] ，循环神经网络系统仍无法处理长序列。如今，最先进的系统主要基于 Transformer 架构，它更善于处理较长序列，并且由于不进行顺序处理而支持并行化。跨语言迁移学习方法在解决低资源语言的机器翻译问题上也展现出很大潜力，因此值得进一步探索。

在未来的工作中，我们计划进一步扩展数据集，并探索基于 Transformer 的模型。尽管这类模型对数据需求较高，但可以针对有限的资源配置进行优化 [33]。我们还将对多语言模型进行比较分析，以选择在沃洛夫语上迁移学习性能更佳的模型，并在该模型上进行迁移学习。

参考文献

[1]Koehn P（2009）《统计机器翻译》。剑桥大学出版社，https://doi.org/10.1017/CBO9780511815829[2] Bahdanau D、Cho K、Bengio Y（2014）《通过联合学习对齐与翻译实现神经机器翻译》。ArXiv 1409[3]Koehn P、Knowles R（2017）《神经机器翻译面临的六项挑战》。载于：第一届神经机器翻译研讨会论文集，计算语言学协会，温哥华，第 28–39 页，https://doi.org/10.18653/v1/W17-3204，网址：https://aclanthology.org/W17-3204[4]Luong T、Pham H、Manning CD（2015）《基于注意力的神经机器翻译的有效方法》。载于：2015 年自然语言处理实证方法会议论文集，计算语言学协会，里斯本，葡萄牙，第 1412–1421 页，https://doi.org/10.18653/v1/D15-1166，网址：https://aclanthology.org/D15-1166[5] Vaswani A、Shazeer N、Parmar N、Uszkoreit J、Jones L、Gomez AN、Kaiser Lu、Polosukhin I（2017）《注意力就是你所需要的一切》。载于：Guyon I、Luxburg UV、Bengio S、Wallach H、Fergus R、Vishwanathan S、Garnett R（编）《神经信息处理系统进展》，Curran Associates, Inc.，第 30 卷，网址：
h
t
t
p
s://
p
r
o
ceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf
[6] Hedderich M、Lange L、Adel H、Str¨otgen J、Klakow D（2021）《低资源场景下自然语言处理近期方法综述》。第 2545–2568 页，https://doi.org/10.18653/v1/2021.naacl-main.201[7] Adebara I、Abdul-Mageed M（2022）《迈向面向非洲语言的非洲中心主义自然语言处理：我们的现状与未来方向》。载于：第 60 届计算语言学协会年会论文集（第 1 卷：长篇论文），计算语言学协会，都柏林，爱尔兰，第 3814–3841 页，https://doi.org/10.18653/v1/2022.acl-long.265，网址：https://aclanthology.org/2022.acl-long.265[8] van Esch D、Lucassen T、Ruder S、Caswell I、Rivera CE（2022）《面向 2,800+ 种语言变体的书写系统与说话人元数据》。载于：语言资源与评测会议论文集，马赛，法国，第 5035–5046 页，网址：https://aclanthology.org/2022.lrec-1.538[9] Gauthier E、Besacier L、VoisinS、Melese M、Elingui UP（2016）《为自动语音识别收集撒哈拉以南非洲语言资源：沃洛夫语案例研究》。载于：第十届国际语言资源与评测会议（LREC'16）论文集，欧洲语言资源协会（ELRA），波尔托罗日，斯洛文尼亚，第 3863–3867 页，网址：https://aclanthology.org/L16-1611

[10] Nguer EH、Khoul´e M、Thiar´e O、Ciss´e MT、Mangeot M（2016）《在线沃洛夫语词典：现状与展望》，网址 https://hal.archives-ouvertes.fr/hal-01311413，工作论文或预印本[11] Dione CMB（2012）《运用有限状态技术构建沃洛夫语形态分析器》。载于：第八届语言资源与评测国际会议（LREC'12）论文集，欧洲语言资源协会（ELRA），土耳其伊斯坦布尔，第894–901页，网址 http: // www.lrec-conf.org/proceedings/lrec2012/pdf/572_Paper.pdf[12] Dione CMB（2020）《沃洛夫语基于词汇功能语法的句法分析器的实现与评测》。载于：第12届语言资源与评测会议论文集，欧洲语言资源协会，法国马赛，第5128–5136页，网址 https://aclanthology.org/2020.lrec-1.631[13] Dione CB（2019）《为沃洛夫语开发通用依存》。载于：第三届通用依存研讨会（UDW，SyntaxFest 2019）论文集，计算语言学协会，法国巴黎，第12–23页，https://doi.org/10.18653/v1/W19-8003，网址 https: // aclanthology.org/W19-8003[14] Lo A、Nguer EHM、Abdoulaye N、Dione CB、Mangeot M、Khoule M、Bao-Diop S、Ciss´e MT（2016）《沃洛夫语拼写纠错：现状与展望》。载于：JEP-TALN-RECITAL 2016：非洲语言自动处理 TALAF 2016，法国巴黎，网址 https://hal.archives-ouvertes.fr/hal-02054917[15] Nguer EM、Lo A、Dione CMB、Ba SO、Lo M（2020）SENCORPUS：法语-沃洛夫语平行语料库。载于：第12届语言资源与评测会议论文集，欧洲语言资源协会，法国马赛，第2803–2811页，网址 https://aclanthology.org/ 2020.lrec-1.341[16]Dione CMB、Lo A、Nguer EM、Ba S（2022）《低资源神经机器翻译：沃洛夫语最先进 Transformer 的基准测试》<->法语。载于：第十三届语言资源与评测会议论文集，欧洲语言资源协会，法国马赛，第6654–6661页，网址 https://aclanthology.org/2022.lrec-1. 717[17]Adelani D、Alabi J、Fan A、Kreutzer J、Shen X、Reid M、Ruiter D、Klakow D、Nabende P、Chang E、Gwadabe T、Sackey F、Dossou BFP、Emezue C、Leong C、Beukman M、Muhammad S、Jarso G、Yousuf O、Niyongabo Rubungo A、Hacheme G、Wairagala EP、Nasir MU、Ajibade B、Ajayi T、Gitau Y、Abbott J、Ahmed M、Ochieng M、Aremu A、Ogayo P、Mukiibi J、Ouoba Kabore F、Kalipe G、Mbaye D、Tapo AA、Memd-jokam Koagne V、Munkoh-Buabeng E、WagnerV、Abdulmumin I、Awokoya A、Buzaaba H、Sibanda B、Bukula A、Manthalu S（2022）《数千条翻译，大有裨益！利用预训练模型翻译非洲新闻》。载于：2022年北美计算语言学协会分会会议：人类语言技术论文集，计算语言学协会，Seat-

tle，美国，第3053–3070页，https://doi.org/10.18653/v1/2022.naacl-main.223，网址 https://aclanthology.org/2022.naacl-main.223[18] Nekoto W、Marivate V、Matsila T、Fasubaa T、Fagbohungbe T、Akinola SO、Muhammad S、Kabongo Kabenamualu S、Osei S、Sackey F、Niyongabo RA、Macharm R、Ogayo P、Ahia O、Berhe MM、Adeyemi M、Mokgesi-Selinga M、Okegbemi L、Martinus L、Tajudeen K、Degila K、Ogueji K、Siminyu K、Kreutzer J、Webster J、Ali JT、Abbott J、Orife I、Ezeani I、Dangana IA、Kamper H、Elsahar H、Duru G、Kioko G、Espoir M、van Biljon E、White-nack D、Onyefuluchi C、Emezue CC、Dossou BFP、Sibanda B、Bassey B、¨Olabiyi A、Ramkilowan A、Oktem A、Akinfaderin A、Bashir A（2020）《面向低资源机器翻译的参与式研究：非洲语言案例研究》。载于：计算语言学协会论文集：EMNLP 2020，计算语言学协会，在线，第2144–2160页，https://doi.org/10.18653/v1/2020.findings-emnlp.195，网址 https://aclanthology.org/2020.findings-emnlp.195[19] Ranathunga S、Lee ESA、Skenduli MP、Shekhar R、Alam M、Kaur R（2021）《低资源语言神经机器翻译：综述》。ArXiv abs/2106.15115[20] J´onsson HP、S´ımonarson HB、Snæbjarnarson V、Steingr´ımsson S、Lofts-son H（2020）《在中等资源环境中尝试不同的机器翻译模型》。载于：Sojka P、Kopeˇcek I、Pala K、Hor´ak A（编），《文本、语音与对话》，施普林格国际出版，Cham，第95–103页[21] Papineni K、Roukos S、Ward T、Zhu WJ（2002）《BLEU：一种机器翻译自动评测方法》。载于：计算语言学协会第40届年会论文集，计算语言学协会，美国宾夕法尼亚州费城，第311–318页，https://doi.org/10.3115/1073083.1073135，网址 https://aclanthology. org/P02-1040[22] Domingo M、Garcıa-Martınez M、Helle A、Casacuberta F、Herranz M（2018）《分词对神经机器翻译的影响有多大？》arXiv 预印本 arXiv:1812.08621，1812.08621[23] Eberhard D、Simons G、Fennig C（2019）《民族语：世界语言》，第22版[24] Adelani DI、Abbott J、Neubig G、D'souza D、Kreutzer J、Lignos C、Palen-Michel C、Buzaaba H、Rijhwani S、Ruder S、Mayhew S、Azime IA、Muhammad S H、Emezue CC、Nakatumba-Nabende J、Ogayo P、Anuoluwapo A、Gitau C、Mbaye D、Alabi J、Yimam SM、Gwadabe TR、Ezeani I、Niy-ongabo RA、Mukiibi J、Otiende V、Orife I、David D、Ngom S、Adewumi T、Rayson P、Adeyemi M、Muriuki G、Anebi E、Chukwuneke C、Odu N、Waira-gala EP、Oyerinde S、Siro C、Bateesa TS、Oloyede T、Wambui Y、Akinode V、Nabagereka D、Katusiime M、Awokoya A、MBOUP M、Gebreyohannes D、Tilaye H、Nwaike K、Wolde D、Faye A、Sibanda B、Ahia O、Dossou BFP、Ogueji K、DIOP TI、Diallo A、Akinfaderin A、Marengereke T、Osei S（2021）MasakhaNER：非洲语言命名实体识别。《计算语言学协会汇刊》9：1116–1131，

https://doi.org/10.1162/tacl_a_00416，URL https://aclanthology.org/2021.tacl-1.66[25] Pinnis M（2018）Tilde 为 WMT 2018 提供的平行语料过滤方法。载于：第三届机器翻译会议论文集：共享任务论文，计算语言学协会，比利时布鲁塞尔，第 939–945 页，https://doi.org/10.18653/v1/W18-6486，URL https:// aclanthology.org/W18-6486[26] Klein G、Kim Y、Deng Y、Senellart J、Rush A（2017）OpenNMT：用于神经机器翻译的开源工具包。载于：ACL 2017 论文集，系统演示，计算语言学协会，加拿大温哥华，第 67–72 页，URL https://www.aclweb.org/anthology/ P17-4012[27] Hochreiter S、Schmidhuber J（1997）长短期记忆。《神经计算》9：1735–80，https://doi.org/10.1162/neco.1997.9.8.1735[28] Kingma DP、Ba J（2015）Adam：一种随机优化方法。CoRR abs/1412.6980[29] Kudo T、Richardson J（2018）SentencePiece：用于神经文本处理的简单、独立于语言的子词分词器和反分词器。载于：2018 年自然语言处理经验方法会议论文集：系统演示，计算语言学协会，比利时布鲁塞尔，第 66–71 页，https://doi.org/10.18653/v1/D18- 2012，URL https://aclanthology.org/D18-2012[30]Sennrich R、Haddow B、Birch A（2016）使用子词单元进行罕见词的神经机器翻译。载于：计算语言学协会第54届年会论文集（第 1 卷：长篇论文），计算语言学协会，德国柏林，第 1715–1725 页，https://doi.org/10.18653/v1/P16-1162，URL https: // aclanthology.org/P16-1162[31]Post M（2018）呼吁明确报告 BLEU 分数。载于：第三届机器翻译会议论文集：研究论文，计算语言学协会，比利时布鲁塞尔，第 186–191 页，https://doi.org/10.18653/v1/W18-6319，URL https: // aclanthology.org/W18-6319[32]Wieting J、Berg-Kirkpatrick T、Gimpel K、Neubig G（2019）超越 BLEU：使用语义相似度训练神经机器翻译。载于：计算语言学协会第57届年会论文集，计算语言学协会，意大利佛罗伦萨，第 4344–4355 页，https://doi.org/10.18653/v1/P19-1427，URL https: // aclanthology.org/P19-1427[33]Araabi A、Monz C（2020）优化 Transformer 以实现低资源神经机器翻译。载于：第28届国际计算语言学会议论文集，国际计算语言学委员会，西班牙巴塞罗那（在线），第 3429–3435 页，https://doi.org/10.18653/v1/2020.coling-main.304，URL https:// aclanthology.org/2020.coling-main.304