提格利尼亚语的自然语言处理：当前状况及未来方向

Fitsum Gaim **Jong C. Park**
School of Computing
Korea Advanced Institute of Science and Technology (KAIST)
{ fitsum.gaim, jongpark } @kaist.ac.kr

Abstract

虽然有数百万人讲提格里尼亚语，但该语言在自然语言处理（NLP）研究中仍然严重不足。本研究对提格里尼亚语的 NLP 研究进行了全面调查，分析了从 2011 年至 2025 年超过十年的 40 多项研究。我们系统地回顾了当前在十个不同下游任务中的计算资源、模型和应用的现状，包括形态处理、机器翻译、语音识别和问答系统。我们的分析揭示了从基础的基于规则的系统到现代神经架构的明确发展轨迹，其中的进展不断由资源创建的里程碑解锁。我们发现根植于提格里尼亚语形态复杂性和资源稀缺性的关键挑战，同时强调了有前景的研究方向，包括形态感知建模、跨语言迁移和以社区为中心的资源开发。本研究既为研究人员提供了全面的参考，也为推动提格里尼亚语 NLP 的发展提供了一条前进的路线。调查研究和资源的精选元数据已公开提供。

1 介绍

近年来，自然语言处理（NLP）取得了显著进展，机器翻译、问答系统和语言生成领域的突破彻底改变了人类与技术互动的方式。然而，这些进步主要局限于世界上 7000 多种语言中的一小部分，导致了具有重大社会影响的数字鸿沟 [Hovy and Spruit, 2016, Joshi et al., 2020, Gaim et al., 2023] 。提格利尼亚语（ትግርኛ；ISO 639-3：tir），主要在厄立特里亚和埃塞俄比亚使用，代表了数字时代此类弱势语言所面临的挑战。

缺乏对特定低资源语言的全面调查为新研究人员设置了障碍，并阻碍了系统性进展。虽然最近的一些评论部分涉及了相关语言 Tonja et al. [2023]，但之前没有任何工作对提格里尼亚语自然语言处理研究进行了集中和全面的分析。本文通过描绘该领域从早期基于规则的形态分析器到最近大规模语言模型的发展的过程，填补了这一空白，强调了社区主导的数据集创建是进步的主要推动力。

我们对提格利尼亚语的自然语言处理研究进行了全面的调查，分析了 2011 年至 2025 年间发表的超过 40 篇研究。本文的贡献包括：

- 对涉及十个不同任务的提格里尼亚自然语言处理研究进行系统综述，分析方法和资源的时间发展。
- 对推动进展的数据集、工具和预训练模型的分析。
- 识别关键研究空白和未来工作的具体建议。
- 一种用于调查其他低资源语言进展的可重复方法。

Preprint.

2 提格里尼亚语言：特点与计算挑战

2.1 语言背景

提格里尼亚语是一种属于闪米特语族的语言，属于阿非罗-亚细亚语系，是厄立特里亚的国家语言和埃塞俄比亚提格雷地区的一种区域语言 [Negash, 2016] 。据估计，全球大约有 1000 万名以提格里尼亚语为母语的使用者 [Eberhard et al., 2025] ，该群体是一个重要但在计算机领域服务较少的语言社区。该语言属于一个语言丰富的区域，被称为东北非语言宏观区域，其特征是不同语言家庭之间共享的特征，这种情况影响了其结构和演变 Zaborski [2011] 。

2.2 书写系统和正字法

吉兹字母表 提格里尼亚语使用 Ge'ez 字母（ፊደል，fidäl），这是一种在非洲之角有着悠久历史的古老书写系统。已知最古老的该字母的例子，例如厄立特里亚 Matara 的 Hawulti 方尖碑，其历史可以追溯到公元 4 世纪 [Ullendorff, 1951] 。Ge'ez 字母是一种音节文字，其中每个字符代表一个辅音-元音音节。这些字符通过系统地修改一个基本的辅音字形来表示七个元音中的一个。提格里尼亚语的音节表包括 32 个基本辅音，每个辅音有七种元音形式（或顺序），另外还有五组唇化辅音，每组有五个元音形式，总共产生 249 个不同的字符 ($32 \times 7 + 5 \times 5 = 249$)。虽然 Ge'ez 字母的独特标点系统在现代提格里尼亚语中仍保留，但其本土数字已经被西方阿拉伯数字取代。

吉兹文字具有在计算处理上需要特别考虑的特点。首先，与基于拉丁字母的文字不同，吉兹文字没有大小写系统（即没有大写字母），这给首字母缩略词识别和命名实体识别等任务带来了复杂性。其次，该文字没有明确标记音位上区分词义的重辅音现象（辅音双写）。例如，qäräbä（"他接近"）和 qärräbä（"他提供"）在书写上都是相同的 ቀረበ 。这种词汇歧义需要从上下文推断，这对文本到语音（TTS）合成、形态分析和机器翻译等任务构成了重大障碍。

在数字时代，输入提格里尼亚语已经被专门的输入方法所简化。诸如 GeezWord、GeezIME 和 Keyman 等软件程序允许用户使用标准的 QWERTY 键盘输入基于 Ge'ez 脚本的语言，如提格里尼亚语、阿姆哈拉语和提格雷语。这些工具采用语音系统的变体，将 ASCII 符号的组合映射到 Ge'ez 字节字符。理解这些输入方法所设定的惯例对 NLP 任务至关重要，如文本规范化、拼写检查和数据预处理。

2.3 形态复杂性

提格里尼亚语展示了模板式（词根与词形）和黏着型的形态变化过程，形成了一个对计算模型来说具有挑战性的复杂系统。

模板形态学。 单词是通过将一个通常由三个辅音组成的辅音词根插入到各种元音模式中形成的。这种非连接过程允许从一个词根生成许多表面形式。从词根 s-b-r (ስብር) 派生出的一些与"打破"概念相关的例子包括：

- säbärä (ሰበረ) - "他打破了"（完成体）
- yisäbbir (ይሰብር) - "他打破"（未完成体）
- sibur (ስቡር) - "破碎的"（形容词）
- mäsbär (መስበር) - "破碎之地"（工具名词）

黏着特征 提格雷尼亚语也通过在词干上附加前缀、后缀和中缀来修饰词义。这些词缀通常是可分离的，并且在不同的词根中具有一致的语义类别。以下是一些例子：

- 前缀 ተ- (tä-) 形成被动或反身动词：säbärä（"他打破"）→ täsäbärä（ተሰበረ，"它被打破"）。
- 后缀如 -ካ（-ka）和 -ኪ（-ki）表示所有格：bet（ቤት，"房子"）→ betka（ቤትካ，"你的房子"，阳性）和 betki（ቤትኪ，"你的房子"，阴性）。

这种形态的丰富性使得单个屈折词能够表达众多语法范畴（性别、数、时态等）。由于当前语料库中类型-词例比率较高，这一过程导致词汇量迅速增长和严重的数据稀疏性，给依赖于词频和固定词汇表的统计和神经 NLP 方法带来了重大挑战。

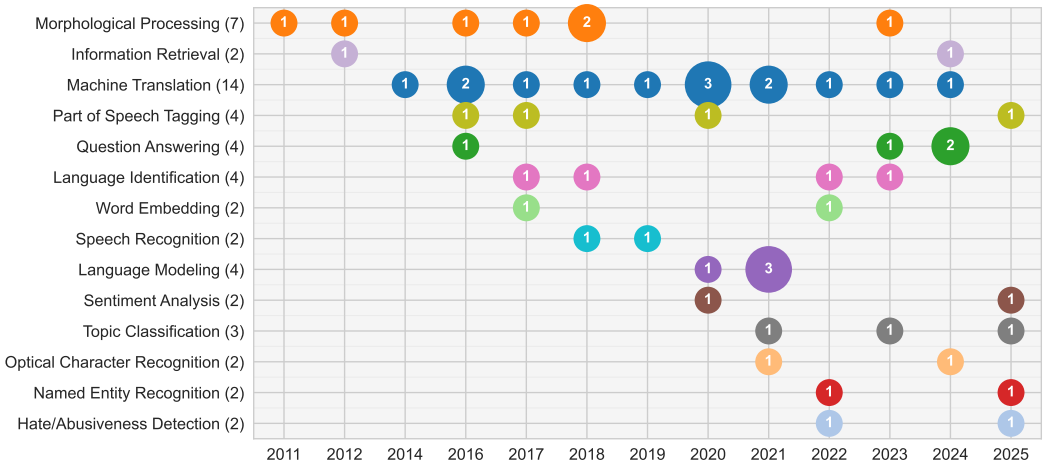

Figure 1: 提格里尼亚自然语言处理研究按任务领域的时间线和分布（2011-2025 年）。一年的出版物数量由气泡大小表示。

3 提格里尼亚自然语言处理研究的现状

我们的分析识别出一个明显的演变过程，从基础的基于规则的系统过渡到现代神经架构的采用，而这一过程是由关键资源创建的里程碑推动的。图 1 展示了出版物的时间顺序发展和分布。研究的详细情况在表 1 中有所描述，该表按任务总结了关键的方法论贡献，而表 2 列出了主要的公开可获取的数据集。

鉴于提格里尼亚语的复杂性，形态分析一直是一个基础的研究领域。早期的工作包括用于信息检索的基于规则的词干提取器 Osman and Mikami [2012] 和像 HornMorpho 这样的综合有限状态转换器系统用于分析和生成 Gasser [2011] 。其他方法已经探索了解析器组合器 Littell et al. [2018] ，使用 LSTM 的神经方法用于语素边界检测 Tedla and Yamamoto [2018] ，并引入一个新的形态分割数据集和分析以进一步支持这一核心任务 Gebremeskel et al. [2023] 。

在绝大多数研究中（约 30%），重点集中在机器翻译上，因为它在提高提格里尼亚语社群信息获取的潜力巨大。早期的工作集中于统计机器翻译（SMT），通过应用形态学分割来减轻数据稀疏的问题、探索因子模型以及创建平行语料库。结合 SMT 与句法重排序规则的混合方法获得了高达 32.64 的 BLEU 分数。最近的研究转向了神经机器翻译（NMT），研究 Transformer 架构、从其他 Ge'ez 文字语言中进行迁移学习，以及通过回译进行数据增强。尽管取得了这些进展，性能仍受限于有限的平行语料库，大多数研究使用的句对不到 20,000 个。错误分析证实，误译和遗漏仍然是经常出现的问题。

3.1 词性标注

初始词性标注研究是基于 Nagaoka Tigrinya 语料库 Tedla et al. [2016] 完成的，这是一个手动标注的包含 4,656 句子的数据集。该资源为 CRF 和 SVM 模型的应用提供了便利，之后的工作分析了词嵌入在该任务中的有效性 Tedla and Yamamoto [2017] 。深度学习模型，特别是 BiLSTM，后来被证明是有效的 Tesfagergish and Kapociute-Dzikiene [2020] 。目前最先进的技术是通过微调基于 transformer 的单语预训练语言模型建立的，这些模型实现了 95.49 % 的准确率 Gaim et al. [2021b] 。

3.2 命名实体识别 (NER)

关于提格利尼亚命名实体识别的研究已经从使用小数据集和标签集的调查发展到开发大规模、综合的资源，例如 TiNC24 Berhane et al. [2025] 。TiNC24 数据集拥有超过 200K 的标注实体，这使得能够训练出强大的序列标注模型，且通过微调预训练的提格利尼亚语言模型，实现了 90.18 的 F1 分数% 。

3.3 文本分类

文本分类的研究已涉及各类下游任务，如主题分类、情感分析、仇恨言论和滥用语言检测。早期的工作主要集中在使用 CNN 进行新闻分类 Fesseha et al. [2021] 。最近的努力则针对用户生成的内容，包括通过迁移学习对社交媒体评论进行情感分析 Tela et al. [2020] 和在 Facebook 上检测仇恨言论 Bahre [2022] 。近期，一个全面的多任务基准被引入用于滥用语言检测，并提供了关于滥用性、情感和主题分类的联合注释 Gaim et al. [2025] 。

自从早期使用 SMT 技术探索事实性问答开始，

3.4 问答 (QA)

问答系统取得了显著进展。随着几个专用数据集的发布，该领域的水平大幅提升：TiQuAD，这是一个大规模本地注释的阅读理解基准 Gaim et al. [2023] ；TIGQA，这是一个专家注释的教育数据集 Teklehaymanot et al. [2024] ；还有 Belebele，这是一个包含提格里尼亚语部分的多语言多项选择问答数据集 Bandarkar et al. [2024] 。这些资源促进了对跨语言和多语言模型的稳健实验，其中在 TiQuAD 上达到了高达 85.12 % 的 F1 分数。

3.5 语言建模

单语语料库的发展，例如 TLMD Gaim et al. [2021a] ，是至关重要的。该资源支持了第一个提格利尼亚语言模型（TiRoBERTa、TiBERT、TiELECTRA）Gaim et al. [2021b] 的预训练，并且显示出在各种下游自然语言理解任务中显著优于多语种替代模型的表现。与此同时，其他研究探索了多语种语言模型 AfriBERTa 的开发，该模型在包括提格利尼亚在内的 11 种非洲语言的集合上进行训练，在多种低资源语言中取得了可喜的成果。[Ogueji et al., 2021]

3.6 其他下游任务

初步工作已在几个其他领域进行。在语音识别方面，工作集中在语料库设计 Abera and H/Mariam [2018] 和深度神经网络的应用 Abera and H/mariam [2019] 。在光学字符识别（OCR）方面，GLOCR 数据集提供了关键资源 Gaim [2021] ，使基于 CRNN 的模型能够在印刷文本上达到高精度 Hailu et al. [2024] 。语言和方言识别研究包括区分密切相关的埃塞-闪米特语言 Asfaw [2018], Gaim et al. [2022] ，识别提格里尼亚方言 Gedamu and Hadgu [2023] ，并研究与阿姆哈拉语的互通性 Feleke [2017] 。为了方便评估词嵌入模型，还通过翻译和优化英语数据集为提格里尼亚开发了一套专用类比测试集 Gaim and Park [2022] 。信息检索（IR）是提格里尼亚中研究较少的领域之一。早期的工作探索了基于 Lucene 的词干提取词汇检索系统 Osman and Mikami [2012] 。最近，Gaim [2024] 提供了提格里尼亚文本检索和表示的单语双编码器模型，这些模型通过问答数据集进行训练。然而，要实现从大规模提格里尼亚文本集合中高效的语义信息检索，仍然存在研究缺口。最后，对机器翻译系统中性别偏见的初步评估发现，80 % 的句子表现出性别偏见 Sewunetie et al. [2024] ，这是由于提格里尼亚的语法性别系统的典型挑战，因为该系统通常在未指定时默认为男性。

3.7 数据集和资源

所审阅文献中的大部分都通过同时开发语言资源如数据集、基准和模型来进行。这些贡献在推进提格里尼亚自然语言处理领域方面发挥了关键作用。表 2 总结了一些主要资源，突出了社区在构建提格里尼亚的稳健数据基础设施方面的重点。

4 挑战和未来方向

我们对提格利尼亚自然语言处理领域的分析揭示了一系列相互关联的挑战以及未来进展的明确路线图。我们将在下文中把挑战、机遇和研究空白的发现整合为一个统一的讨论。

4.1 挑战与方法学机遇

提格里尼亚语自然语言处理的主要障碍类似于许多低资源语言的特点 [Joshi et al., 2020] ，但由于特定的语言特征而被放大。

Table 1: 提格利尼亚自然语言处理研究中按任务领域划分的方法论贡献

Task Area	Methodological Contributions & Milestones
Morphological Processing	Rule-based stemming; Finite-State Transducers (FST); Parser combinators; Neural boundary detection (LSTM); Hybrid rule-neural analyzers. Gasser [2011], Osman and Mikami [2012], Littell et al. [2018], Tedla and Yamamoto [2018], Gebremeskel et al. [2023]
Machine Translation	Factored SMT; Hybrid SMT-rule systems; NMT (Transformer); Transfer learning from related languages; Data augmentation (back-translation). Tsegaye [2014], Gaim [2017], Berihu et al. [2020], Adhanom [2021], Kidane et al. [2021], Hadgu et al. [2022]
Part-of-Speech Tagging	CRF & SVM models; Analysis of word embeddings; Bi-LSTM architectures; Fine-tuning of Pre-trained Language Models (PLMs). Tedla et al. [2016], Tedla and Yamamoto [2017], Tesfagergish and Kapociute-Dzikiene [2020], Gaim et al. [2021b]
Text Classification	Sentiment analysis for social media text; CNNs for news classification; Multi-task learning for abusive language detection. Tela et al. [2020], Fesseha et al. [2021], Bahre [2022], Gaim et al. [2025]
Language/Dialect ID	Mutual intelligibility studies; Dialect detection; Benchmark for typologically related languages; N-gram models; Fine-tuning PLMs for language identification. Feleke [2017], Asfaw [2018], Gaim et al. [2022], Gedamu and Hadgu [2023]
Named Entity Recognition	Small-scale and large-scale annotated datasets; Fine-tuning of PLMs for sequence labeling; Joint dataset with POS tagging. Yohannes and Amagasa [2022], Berhane et al. [2025]
Question-Answering	Factoid QA via SMT; Large-scale native extractive QA datasets; Expert-annotated educational datasets; Cross-lingual transfer. Amare [2016], Gaim et al. [2023], Teklehaymanot et al. [2024], Bandarkar et al. [2024]
Language Modeling	Construction of large monolingual corpora; Pre-training of monolingual Transformer-based models (e.g., TiRoBERTa); Multilingual modeling for low-resource African languages. Gaim et al. [2021a,b], Ogueji et al. [2021]
Optical Character & Text Recognition	Development of large-scale labeled text image dataset for Ge'ez script languages; Application of CRNN-based models for text recognition. Gaim [2021], Hailu et al. [2024]
Speech Recognition	Design and construction of speech corpus; Application of Deep Neural Networks (DNNs) and LSTMs for recognition. Abera and H/Mariam [2018], Abera and H/mariam [2019]

数据稀缺性和形态复杂性。 最显著的障碍是这两个因素的结合。大规模、带注释的数据集的缺乏由于提格利尼亚语的复杂形态而严重加剧，这导致了高出词汇表（OOV）率和极端的数据稀疏性。这种协同作用挑战了标准的标记和建模技术，使得在没有海量数据集的情况下效果不佳。这突显出一些正在成为低资源 NLP 研究核心的方法论机会。数据高效学习提供了一种关键策略，通过迁移学习 [Ruder, 2019] 和跨语言方法利用多语言模型来提升性能。类似地，诸如机器翻译的银标准数据集等合成数据生成也可以有效地增强本地标注的语料库。另一个重要的机会在于设计形态感知的架构。未来的工作可以研究混合分词方案（例如，与形态分割结合的 BPE [Sennrich et al., 2016] ）或本质上更适合提格利尼亚语非连音结构的字符级模型。

有限的标准化资源。 该领域缺乏稳健的开源预处理工具，并且缺乏多领域的评估基准，这使得直接的模型比较变得困难。一个关键的机会在于以社区为中心的资源开发。通过参与式设计，与提格里尼亚语的侨民和当地机构合作，不仅可以扩大数据收集规模，还可以确保文化上的相关性，并从根本上帮助缓解偏差问题 [Bird, 2020] 。

与许多语言一样，基于历史或网络爬取数据训练的模型可能会加剧社会偏见。初步研究已经确认了机器翻译系统中存在显著的性别偏见。这设定了一个关键的未来方向：不仅仅关注性能指标，还要关注责任和公平的自然语言处理。这不仅包括对性别偏见的审计，还包括对方言偏见（鉴于已知的方言差异）以及以新闻为中心的语料库中潜在的政治或社会偏见。

Table 2: 提格利尼亚语自然语言处理研究中的公开可用资源概览，显示所涉及任务的多样性和社区贡献的语言资产的规模。

Resource	Task Area	Size	Reference
NTC	POS Tagging Dataset	4.6K sents, 72K tokens	Tedla et al. [2016]
Sentiment	Sentiment Analysis Dataset	50K samples	Tela et al. [2020]
TLMD	Language Modeling Dataset	2M sents, 40M tokens	Gaim et al. [2021a]
TiQuAD	Question-Answering Dataset	10.6K QA pairs	Gaim et al. [2023]
TIGQA	Question-Answering Dataset	2.6K QA pairs	Teklehaymanot et al. [2024]
NER Corpus	Named Entity Recognition	3.6K sents, 69K entities	Yohannes and Amagasa [2022]
TiNC24	Named Entity Recognition	13K sents, 200K entities	Berhane et al. [2025]
Analogy Test	Word Embedding Evaluation	18.5K entries	Gaim and Park [2022]
GeezSwitch	Language Identification	15K samples	Gaim et al. [2022]
MasakhaNEWS	Topic Classification Dataset	3K articles	Adelani et al. [2023]
TiALD	Abusive Language Dataset	13.7K comments	Gaim et al. [2025]
TiPLMs	Pre-trained Language Models	TiRoBERTa (125M) TiBERT (110M) TiELECTRA (14M)	Gaim et al. [2021b]
AfriBERTa	Pre-trained Language Models	AfriBERTa large (126M) base (11M), and small (97M)	Ogueji et al. [2021]
TiBiEncoders	Information Retrieval Models	TiELECTRA-bi-encoder (14M) TiRoBERTa-bi-encoder (125M)	Gaim [2024]
Speech Corpus	Speech Recognition Dataset	10K labeled utterances	Abera and H/Mariam [2018]
GLOCR	OCR Dataset	710K text-image pairs	Gaim [2021]

4.2 研究空白和建议

基于这些挑战和机遇，我们的分析揭示了几个未充分探索的领域，未来研究在这些领域可能产生重要影响。这些空白包括在会话 AI（例如，对话系统）和多模态应用（例如，图像描述）方面完全缺乏工作。此外，在高级语音技术领域的研究仅限于初步的语音识别（ASR）原型，文本到语音（TTS）合成则完全未被探索。最后，现有资源集中在新闻文本上，迫切需要对医疗和法律等关键领域进行领域适应，而有关偏见和公平性的工作才刚刚开始触及表面。

为了弥补这些空白，我们建议采用一种策略，侧重于技术重点和社区参与。社区应优先开发基础资源和工具，例如扩大平行语料库，创建多领域评估基准，并标准化开源预处理库。同时，研究人员应研究高级语言感知模型，包括形态感知神经架构和全面审计，以减轻性别、文化和其他社会偏见。然而，持久进步取决于与当地大学和文化机构建立社区合作伙伴关系，以建立当地研究能力并优先考虑解决直接社区需求的应用，例如教育工具、卫生信息获取。最后，所有创建的资源——数据集、模型和工具——必须开放且可访问，并有详细的文档，以支持下一波研究人员和开发人员。

5 调查方法

5.1 文献检索

我们在包括 Google Scholar、Semantic Scholar、ACL Anthology、JSTOR、IEEE Xplore 和 ACM Digital Library 在内的多个数据库中进行了全面搜索。搜索词包括“提格里尼亚语”、“自然语言处理”、“计算语言学”、“机器学习”和具体任务名称的组合。我们还包括已发表的语言资源和灰色文献，如论文和技术报告，以确保全面覆盖。

5.2 纳入和排除标准

为了将一项研究纳入本次调查，我们规定其主要焦点必须是自然语言处理或计算语言学，并对提格利尼亚语进行了实质性处理。符合条件的研究作品进一步限制为原始研究文章和系统综述，以确保专注于新颖的贡献和严格的综述。因此，本次调查不包括在搜索时未公开访

问的研究，例如大学内部项目。此外，如果广泛的多语言出版物缺乏足够的技术细节或有意义的语言特定分析，则被省略，因为这将无法对其方法论和对提格利尼亚自然语言处理的影响进行适当评估。

5.3 数据提取与分析

对于每篇被纳入的论文，我们提取了：作者、年份、目标、方法、数据集、主要发现以及开发的资源。论文是根据主要的自然语言处理任务进行分类，但我们也允许每篇论文中有多个任务，当论文对主题进行充分讨论时，这样可以系统地分析每个领域的进展。

6 结论

这项调查对提格里尼亚语自然语言处理研究进行了全面分析，揭示了一个既具有显著基础进展又面临持续挑战的领域。虽然现在已经有一些核心任务的资源，但在数据的可用性、工具的发展和高级应用方面仍存在显著缺口。近年来的加速进展主要由社区努力和利用跨语言方法的针对性研究计划推动。未来的发展路径需要将技术创新与社区参与相结合，以构建强健且公平系统所需的文化相关的大规模数据集。跨语言迁移学习和关注形态学的模型为克服数据稀缺提供了有希望的途径，而继续投资于特定语言、经过社区验证的资源仍然是必要的。随着自然语言处理技术日益塑造全球信息获取，对如提格里尼亚语等语言的进步保障不仅是一个技术挑战，还是一个语言平等和数字包容的迫切要求。这项调查为研究人员提供了对当前能力的理解和对未来工作具体方向的基础。

我们真诚地感谢所有审阅本文草稿并提供宝贵反馈和建议的研究人员。我们邀请未来以修正或添加论文和资源的形式为开源 提格里尼亚语自然语言处理文集 项目做出贡献。

References

Dirk Hovy and Shannon L. Spruit. The social impact of natural language processing. In Katrin Erk and Noah A. Smith, editors, Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) , pages 591–598, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-2096. URL https://aclanthology.org/P16-2096/.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. The state and fate of linguistic diversity and inclusion in the NLP world. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics , pages 6282–6293, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.560. URL https://aclanthology.org/2020.acl-main.560/.

Fitsum Gaim, Wonsuk Yang, Hancheol Park, and Jong C. Park. Question-answering in a low-resourced language: Benchmark dataset and models for Tigrinya. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) , pages 11857–11870, Toronto, Canada, July 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.acl-long.661.

Atnafu Lambebo Tonja, Tadesse Destaw Belay, Israel Abebe Azime, Abinew Ali Ayele, Moges Ahmed Mehamed, Olga Kolesnikova, and Seid Muhie Yimam. Natural language processing in Ethiopian languages: Current state, challenges, and opportunities. In Proceedings of the Fourth workshop on Resources for African Indigenous Languages (RAIL 2023) , pages 126–139, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.rail-1.14.

Abraham Negash. The origin and development of tigrinya language publications (1886 - 1991) volume one. University Library, Santa Clara University , Paper 131, 2016. URL http://scholarcommons.scu.edu/library/131.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig, editors. Ethnologue: Languages of the World . Ethnologue Series. SIL International, Dallas, Texas, 28 edition, 2025. Online version: www.ethnologue.com.

Andrzej Zaborski. Language subareas in ethiopia reconsidered. Lingua Posnaniensis , 52(2):99–110, Feb. 2011. doi: 10.2478/v10122-010-0017-7. URL https://pressto.amu.edu.pl/index.php/linpo/article/view/v10122-010-0017-7.

Edward Ullendorff. The obelisk of maara. Journal of the Royal Asiatic Society of Great Britain and Ireland , (1/2):26–32, 1951. ISSN 0035869X. URL http://www.jstor.org/stable/25222457.

Yitna Firdyiwek and Daniel Yaqob. The system for ethiopic representation in ascii. Citeseer , 2000.

Fitsum Gaim, Wonsuk Yang, and Jong C. Park. Tlmd: Tigrinya language modeling dataset. Zenodo , July 2021a. doi: 10.5281/zenodo.5139094. URL https://doi.org/10.5281/zenodo.5139094.

Omer Osman and Yoshiki Mikami. Stemming Tigrinya words for information retrieval. In Proceedings of COLING 2012: Demonstration Papers , pages 345–352, Mumbai, India, December 2012. The COLING 2012 Organizing Committee. URL https://aclanthology.org/C12-3043.

Michael Gasser. Hornmorpho: a system for morphological processing of amharic, oromo, and tigrinya. In Conference on Human Language Technology for Development, Alexandria, Egypt , 2011.

Patrick Littell, Tom McCoy, Na-Rae Han, Shruti Rijhwani, Zaid Sheikh, David Mortensen, Teruko Mitamura, and Lori Levin. Parser combinators for Tigrinya and Oromo morphology. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018) , Miyazaki, Japan, May 2018. European Language Resources Association (ELRA). URL https://aclanthology.org/L18-1611.

Yemane Tedla and Kazuhide Yamamoto. Morphological segmentation with lstm neural networks for tigrinya. In Intenational Journal on Natural Language Computing (JNLC) , volume 7, 2018.

Hagos Gebremedhin Gebremeskel, Feng Chong, and Huang Heyan. Unlock tigrigna nlp: Design and development of morphological analyzer for tigrigna verbs using hybrid approach. SSRN Electronic Journal , 2023. URL https://api.semanticscholar.org/CorpusID:265633133.

Yemane Tedla and Kazuhide Yamamoto. The effect of shallow segmentation on english-tigrinya statistical machine translation. In 2016 International Conference on Asian Language Processing (IALP) , pages 79–82. IEEE, 2016.

Fitsum Gaim. Applying Morphological Segmentation to Machine Translation of Low-Resourced and Morphologically Complex Languages: The case of Tigrinya. Master's thesis, School of Computing, Korea Advanced Institute of Science and Technology (KAIST), July 2017.

Tarik Tsegaye. English-tigrigna factored statistical machine translation. Master's thesis, Addis Ababa University, 2014.

Solomon Teferra Abate, Michael Melese Woldeyohannis, Martha Yifiru Tachbelie, Million Meshesha, Solomon Atinafu, Wondwossen Mulugeta, Yaregal Assabie, Hafte Abera, Binyam Ephrem Seyoum, Tewodros Abebe, Wondimagegnhue Tsegaye, Amanuel Lemma, Tsegaye Andargie, and Seifedin Shifaw. Parallel corpora for bi-lingual english-ethiopian languages statistical machine translation. In International Conference on Computational Linguistics , 2018.

Solomon Teferra Abate, Michael Melese, Martha Yifiru Tachbelie, Million Meshesha, Solomon Atinafu, Wondwossen Mulugeta, Yaregal Assabie, Hafte Abera, Biniyam Ephrem, Tewodros Gebreselassie, Wondimagegnhue Tsegaye Tufa, Amanuel Lemma, Tsegaye Andargie, and Seifedin Shifaw. English-ethiopian languages statistical machine translation. In WNLP Workshop at ACL , 2019.

Zemicheal Berihu, Gebremariam Mesfin, Mulugeta Atsibaha, Tor-Morten, and Grønli. Enhancing bi-directional english-tigrigna machine translation using hybrid approach. 2020.

Isayas Berhe Adhanom. A first look into neural machine translation for tigrinya. arXiv , 2021.

Alp Öktem, Mirko Plitt, and Grace Tang. Tigrinya neural machine translation with transfer learning for humanitarian response. arXiv preprint arXiv:2003.11523 , 2020.

Lidia Kidane, Sachin Kumar, and Yulia Tsvetkov. An exploration of data augmentation techniques for improving english to tigrinya translation. ArXiv , abs/2103.16789, 2021.

Asmelash Teka Hadgu, Abel Aregawi, and Adam Beaudoin. Lesan–machine translation for low resource languages. In NeurIPS 2021 Competitions and Demonstrations Track , pages 297–301. PMLR, 2022. URL https://proceedings.mlr.press/v176/hadgu22a/hadgu22a.pdf.

Nuredin Ali Abdelkadir, Negasi Haile Abadi, and Asmelash Teka Hadgu. Error analysis of Tigrinya - English machine translation systems. In Proceedings of the 4th Workshop on African Natural Language Processing, AfricaNLP@ICLR 2023, Kigali, Rwanda, May 1, 2023 , 2023. URL https://openreview.net/pdf?id=BQVqNyzCxx.

Yemane Tedla, Kazuhide Yamamoto, and A. Marasinghe. Tigrinya part-of-speech tagging with morphological patterns and the new nagaoka tigrinya corpus. International Journal of Computer Applications , 146:33–41, 2016.

Yemane Tedla and Kazuhide Yamamoto. Analyzing word embeddings and improving pos tagger of tigrinya. 2017 International Conference on Asian Language Processing (IALP) , pages 115–118, 2017.

Senait Gebremichael Tesfagergish and Jurgita Kapociute-Dzikiene. Part-of-speech tagging via deep neural networks for northern-ethiopic languages. Information Technology and Control , 49:482–494, 2020.

Fitsum Gaim, Wonsuk Yang, and Jong C. Park. Monolingual pre-trained language models for tigrinya. In 5th Widening NLP (WiNLP2021) workshop, co-located with the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP) , 2021b. URL http://www.winlp.org/wp-content/uploads/2021/11/winlp2021_62_Paper.pdf.

Hailemariam Mehari Yohannes and Toshiyuki Amagasa. Named-entity recognition for a low-resource language using pre-trained language model. In Proceedings of the 37th SIGAPP Symposium on Applied Computing , SAC '22, page 837–844, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450387132. doi: 10.1145/3477314.3507066. URL https://doi.org/10.1145/3477314.3507066.

Sham K. Berhane, Simon M. Beyene, Yoel G. Teklit, Ibrahim A. Ibrahim, Natnael A. Teklu, Sirak A. Bereketeab, and Fitsum Gaim. Towards Neural Named Entity Recognition System in Tigrinya with Large-scale Dataset. Language Resources and Evaluation , June 2025. ISSN 1574-0218. doi: 10.1007/s10579-025-09825-4. URL https://doi.org/10.1007/s10579-025-09825-4.

Awet Fesseha, Shengwu Xiong, Eshete Derb Emiru, Moussa Diallo, and Abdelghani Dahou. Text classification based on convolutional neural networks and word embedding for low-resource languages: Tigrinya. Inf. , 12:52, 2021.

Abrhalei Tela, Abraham Woubie, and Ville Hautamaki. Transferring monolingual model to low-resource language: The case of tigrinya, 2020.

Weldemariam Bahre. Hate speech detection from facebook social media posts and comments in tigrigna language. Master's thesis, The faculty of informatics, St. Mary's University, June 2022. URL http://repository.smuc.edu.et/handle/123456789/6929.

Fitsum Gaim, Hoyun Song, Huije Lee, Changgeon Ko, Eui Jun Hwang, and Jong C. Park. A multi-task benchmark for abusive language detection in low-resource settings, 2025. URL https://arxiv.org/abs/2505.12116.

Kibrom Haftu Amare. Tigrigna question answering system for factoid questions. Master's thesis, Addis Ababa University, June 2016. URL http://etd.aau.edu.et/handle/123456789/2375.

Hailay Kidu Teklehaymanot, Dren Fazlija, Niloy Ganguly, Gourab Kumar Patro, and Wolfgang Nejdl. TIGQA: An expert-annotated question-answering dataset in Tigrinya. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024) , pages 16142–16161, Torino, Italia, May 2024. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.1404/.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. The belebele benchmark: a parallel reading comprehension dataset in 122 language variants. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) , pages 749–775, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.44. URL https://aclanthology.org/2024.acl-long.44/.

Kelechi Ogueji, Yuxin Zhu, and Jimmy Lin. Small data? no problem! exploring the viability of pretrained multilingual language models for low-resourced languages. In Duygu Ataman, Alexandra Birch, Alexis Conneau, Orhan Firat, Sebastian Ruder, and Gozde Gul Sahin, editors, Proceedings of the 1st Workshop on Multilingual Representation Learning , pages 116–126, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.mrl-1.11. URL https://aclanthology.org/2021.mrl-1.11/.

Hafte Abera and Sebsibe H/Mariam. Design of a Tigrinya language speech corpus for speech recognition. In Proceedings of the First Workshop on Linguistic Resources for Natural Language Processing , pages 78–82, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics. URL https://aclanthology.org/W18-3811.

Hafte Abera and Sebsibe H/mariam. Speech recognition for Tigrinya language using deep neural network approach. In Proceedings of the 2019 Workshop on Widening NLP , pages 7–9, Florence, Italy, August 2019. Association for Computational Linguistics. URL https://aclanthology.org/W19-3603.

Fitsum Gaim. GLOCR: GeezLab OCR Dataset, June 2021. URL https://doi.org/10.7910/DVN/RQTSD2.

Aaron Afewerki Hailu, Abiel Tesfamichael Hayleslassie, Danait Weldu Gebresilasie, Robel Estifanos Haile, Tesfana Tekeste Ghebremedhin, and Yemane Keleta Tedla. Tigrinya ocr: Applying crnn for text recognition. In Biao Luo, Long Cheng, Zheng-Guang Wu, Hongyi Li, and Chaojie Li, editors, Neural Information Processing , pages 456–467, Singapore, 2024. Springer Nature Singapore. ISBN 978-981-99-8184-7.

Rediat Bekele Asfaw. A comparative study of automatic language identification on ethio-semitic languages. Master's thesis, Addis Ababa University, June 2018.

Fitsum Gaim, Wonsuk Yang, and Jong C. Park. GeezSwitch: Language identification in typologically related low-resourced East African languages. In Proceedings of the Thirteenth Language Resources and Evaluation Conference , pages 6578–6584, Marseille, France, June 2022. European Language Resources Association. URL https://aclanthology.org/2022.lrec-1.707.

Asfaw Gedamu and Asmelash Teka Hadgu. Tigrinya dialect identification. In AfricaNLP workshop at ICLR2023 , 2023. URL https://openreview.net/pdf?id=kiG8qiUFm2u.

Tekabe Legesse Feleke. The similarity and mutual intelligibility between Amharic and Tigrigna varieties. In Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial) , pages 47–54, Valencia, Spain, April 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-1206. URL https://aclanthology.org/W17-1206.

Fitsum Gaim and Jong C. Park. Tigrinya Analogy Test for evaluating Word Embeddings, May 2022. URL https://doi.org/10.5281/zenodo.7089051.

Fitsum Gaim. Semantic search models for tigrinya, 2024. URL https://huggingface.co/fgaim/tiroberta-bi-encoder.

Walelign Sewunetie, Atnafu Tonja, Tadesse Belay, Hellina Hailu Nigatu, Gashaw Gebremeskel, Zewdie Mossie, Hussien Seid, and Seid Yimam. Gender bias evaluation in machine translation for Amharic, Tigrigna, and afaan oromoo. In Beatrice Savoldi, Janiça Hackenbuchner, Luisa Bentivogli, Joke Daems, Eva Vanmassenhove, and Jasmijn Bastings, editors, Proceedings of the 2nd International Workshop on Gender-Inclusive Translation Technologies , pages 1–11, Sheffield, United Kingdom, June 2024. European Association for Machine Translation (EAMT). URL https://aclanthology.org/2024.gitt-1.1/.

David Ifeoluwa Adelani, Marek Masiak, Israel Abebe Azime, Jesujoba Alabi, Atnafu Lambebo Tonja, Christine Mwase, Odunayo Ogundepo, Bonaventure F. P. Dossou, Akintunde Oladipo, Doreen Nixdorf, Chris Chinenye Emezue, Sana Al-azzawi, Blessing Sibanda, Davis David, Lolwethu Ndolela, Jonathan Mukiibi, Tunde Ajayi, Tatiana Moteu, Brian Odhiambo, Abraham Owodunni, Nnaemeka Obiefuna, Muhidin Mohamed, Shamsuddeen Hassan Muhammad, Teshome Mulugeta Ababu, Saheed Abdullahi Salahudeen, Mesay Gemeda Yigezu, Tajuddeen Gwadabe, Idris Abdulmumin, Mahlet Taye, Oluwabusayo Awoyomi, Iyanuoluwa Shode, Tolulope Adelani, Habiba Abdulganiyu, Abdul-Hakeem Omotayo, Adetola Adeeko, Abeeb Afolabi, Anuoluwapo Aremu, Olanrewaju Samuel, Clemencia Siro, Wangari Kimotho, Onyekachi Ogbu, Chinedu Mbonu, Chiamaka Chukwuneke, Samuel Fanijo, Jessica Ojo, Oyinkansola Awosan, Tadesse Kebede, Toadoum Sari Sakayo, Pamela Nyatsine, Freedmore Sidume, Oreen Yousuf, Mardiyyah Oduwole, Kanda Tshinu, Ussen Kimanuka, Thina Diko, Siyanda Nxakama, Sinodos Nigusse, Abdulmejid Johar, Shafie Mohamed, Fuad Mire Hassan, Moges Ahmed Mehamed, Evrard Ngabire, Jules Jules, Ivan Ssenkungu, and Pontus Stenetorp. MasakhaNEWS: News topic classification for African languages. In Jong C. Park, Yuki Arase, Baotian Hu, Wei Lu, Derry Wijaya, Ayu Purwarianti, and Adila Alfa Krisnadhi, editors, Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers) , pages 144–159, Nusa Dua, Bali, November 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.ijcnlp-main.10. URL https://aclanthology.org/2023.ijcnlp-main.10/.

Sebastian Ruder. Neural Transfer Learning for Natural Language Processing . PhD thesis, National University of Ireland, Galway, 2019.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) , pages 1715–1725, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1162. URL https://aclanthology.org/P16-1162.

Steven Bird. Decolonising speech and language technology. In Donia Scott, Nuria Bel, and Chengqing Zong, editors, Proceedings of the 28th International Conference on Computational Linguistics , pages 3504–3519, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.313. URL https://aclanthology.org/2020.coling-main.313/.