The Rise of AfricaNLP: Contributions, Contributors, and Community Impact (2005–2025)

Tadesse Destaw Belay[1,7,12], **Kedir Yassin Hussen**[2], **Sukairaj Hafiz Imam**[3], **Ibrahim Said Ahmad**[4], **Isa Inuwa-Dutse**[5], **Abrham Belete Haile**[6], **Grigori Sidorov**[1], **Iqra Ameer**[7], **Idris Abdulmumin**[8], **Tajuddeen Gwadabe**[9], **Vukosi Marivate**[8], **Seid Muhie Yimam**[10], **Shamsuddeen Hassan Muhammad**[11]

[1]Instituto Politécnico Nacional, [2]University of Gondar, [3]Northwest University Kano, [4]Bayero University Kano, [5]University of Huddersfield, [6]iCog Labs, [7]Pennsylvania State University, [8]University of Pretoria, [9]Masakhane, [10]University of Hamburg, [11]Imperial College London, [12]Wollo University

Abstract

Natural Language Processing (NLP) is undergoing constant transformation, as Large Language Models (LLMs) are driving daily breakthroughs in research and practice. In this regard, tracking the progress of NLP research and automatically analyzing the contributions of research papers provides key insights into the nature of the field and the researchers. This study explores the progress of African NLP (**AfricaNLP**) by asking (and answering) research questions about the progress of AfricaNLP (publications, NLP topics, and NLP tasks), contributions (data, method, and task), and contributors (authors, affiliated institutions, and funding bodies). We quantitatively examine two decades (2005–2025) of contributions to AfricaNLP research, using a dataset of 1.9K NLP papers, 4.9K contributing authors, and 7.8K human-annotated contribution sentences (`AfricaNLPContributions`), along with benchmark results. Our dataset and AfricaNLP research explorer tool will provide a powerful lens for tracing AfricaNLP research trends and hold potential for generating data-driven research approaches.

1 Introduction

Natural Language Processing (NLP) research continues to grow at a rapid rate and has undergone significant advancements. Assessing progress, identifying challenges, and suggesting future directions can advance research in languages by showing *where we are* and *where we can go* (Adebara and Abdul-Mageed, 2022). This need is urgent for African languages, as Africa is the most linguistically diverse continent, with over 2,000 languages, where each of them are spoken by hundreds of millions of people (Muhammad et al., 2023).

Some survey work has been performed at various levels of NLP for African languages, such as domain-specific (Imam et al., 2025; Hussen et al., 2025), region/language-specific (Tonja et al., 2023; Amol et al., 2024; Inuwa-Dutse, 2025), and continent-level to NLP language resources (Alabi et al., 2025). While valuable, existing surveys are limited in their coverage of regional/country, language, modalities, and domain, and are mainly non-reproducible. We provide a comprehensive 20-year African NLP survey without any disparities with an open-source tool that tracks the progress of African NLP.

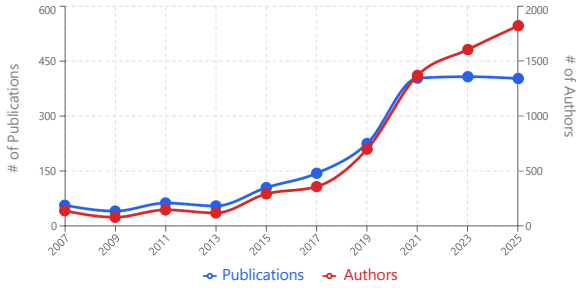

Figure 1: AfricaNLP publication history and author contributors (*2005 - 2025.July*). Papers are keyword-based searched from Semantic Scholar and ACL events - a total of 1.9K human-verified AfricaNLP papers with more than 4.9K authors.

In NLP, contributions may take the form of new knowledge, models, datasets, or methods that have significantly improved the field (Khurana et al., 2023). The rate of scientific literature continues to grow rapidly, and researchers face increasing challenges in keeping up with advancements in their respective fields. Thus, accurately identifying prior research and its key contributions is crucial for researchers to grasp which issues have been addressed and to effectively identify existing gaps in the literature (Snyder, 2019; Pramanick et al., 2025b).

In this study, we conduct a comprehensive analysis of AfricaNLP research papers published over a twenty-year span, specifically from 2005 to July 2025. We seek to answer the following descrip-

tive[1] research questions (RQ) on AfricaNLP works: 1) How has AfricaNLP evolved over the past two decades in terms of research focus and publication trends? 2) What types of contributions (knowledge, methods, datasets) are most prevalent in AfricaNLP publications? 3) Who are the main contributors (authors, institutions, and supporting organizations) to AfricaNLP? 4) Where are the AfricaNLP papers affiliated, inside or outside the continent? 5) How does the number of language speakers correlate with the availability of language resources/publications?

The main contributions of this study are summarized as follows:

- We present the first comprehensive, two-decade survey of AfricaNLP research (2005–2025), analyzing publications, authors, affiliations, supporters, NLP topics, and tasks (§4).
- We introduce `AfricaNLPContributions`, a large-scale dataset consisting of 1.9K papers, 4.9K authors, and 7.8K human-annotated contribution sentences (§5).
- We provide a detailed analysis of contributions, contributors, affiliations, and funding patterns in AfricaNLP (§5, §6).

2 Related Works

2.1 African NLP Surveys

Most previous surveys on AfricaNLP address domain-specific topics and studies focused on specific countries or languages. An overview of these efforts, including the languages, countries, and focus areas covered, is provided in Table 1.

Domain-specific Several surveys and progress reports focus on specific domains and NLP tasks, including automatic speech recognition (ASR) for African languages (Imam et al., 2025), LLMs for African languages (Hussen et al., 2025), NLP for public health in Africa (Hu et al., 2025) and NLP tools for African languages (Mussandi and Wichert, 2024).

Country/Language-specific Community-led initiatives have also conducted surveys documenting the progress of NLP research in specific countries and languages such as EthioNLP (Tonja et al., 2023) for the progress of Ethiopian languages, KenyaNLP (Amol et al., 2024) in Kenyan languages, NaijaNLP (Inuwa-Dutse, 2025) for Nigerian languages, GhanaNLP (Azunre et al., 2021) for Ghanaian languages and HausaNLP (Muhammad et al., 2025) for Hausa language.

Africa-wide At AfricaNLP, Adebara and Abdul-Mageed (2022) identified key linguistic and sociopolitical challenges to developing NLP technologies for African languages and outlined potential research directions. More recently and the only continent-level effort to assess NLP progress in Africa, the work by Alabi et al. (2025) analyzed 734 research papers on NLP for African languages published between 2019 and 2024. While valuable, their study was limited by a narrow set of search keywords (50 language names), restricted temporal coverage (5 years), and a focus on selected core NLP tasks and available datasets. This gap motivates our work.

2.2 NLP Contributions

The work by D'Souza and Auer (2020) proposed an NLP contribution annotation scheme at the sentence and phrase level. D'Souza et al. (2021) organized the first SemEval shared task on NLP contribution graph. Pramanick et al. (2025a) created a sentence-based NLP Contributions dataset by sampling five papers per year from ACL events between 1974 and February 2024. However, this work is limited to a specific database for the global-level contributions; where continent-level contributions, such as those from Africa, remain unexplored. In this work, we present the first comprehensive survey of AfricaNLP's progress over the past two decades, highlighting key contributions, contributors (authors, affiliated institutions, funding supporters, and computing resource supports), long-term trends, and the broader impact of AfricaNLP research on the global NLP community.

3 Survey Methodology

We follow the new PRISMA framework (Page et al., 2021) to identify and filter papers published between 2005 and 2025. Our pipeline combines keyword-based retrieval, a multi-source approach, and LLM-assisted annotation methods. Figure 2 summarizes the overall paper retrieval and annota-

[1]Descriptive questions: which are the most basic type of quantitative research question and seeks to explain the when, where, why or how something occurred.

Survey paper	# of papers	Years covered	Languages/Countries	Focus area(s)
Azunre et al. (2021)	< 20	— 2021	16 Ghanaian languages	About NLPGhana organization and the state of NLP in Ghana
Adebara and Abdul-Mageed (2022)	< 50	— 2022	512 African languages	African-centric approach to African languages
Tonja et al. (2023)	< 50	— 2023	4 Ethiopian languages	NLP tools, datasets and benchmarking it
Amol et al. (2024)	< 50	— 2024	15 Kenyan languages	NLP tools, datasets, and gaps
Mussandi and Wichert (2024)	< 50	— 2024	30 African languages	Identify languages that have NLP toolkit
Inuwa-Dutse (2025)	329	— 2025	3 Nigerian languages	Linguistic resources and key challenges
Muhammad et al. (2025)	< 50	— 2025	1 Nigerian languages	State of Hausa NLP (resources, LLMs, gaps)
Imam et al. (2025)	< 20	— 2025	African languages	Challenges of speech recognition for African languages
Hussen et al. (2025)	< 50	— 2025	42 African languages	Identifies LLMs that support African languages
Hu et al. (2025)	54	— 2025	African languages	NLP technologies for public health in Africa
Alabi et al. (2025)	732	5 (2019-2024)	53 African Languages	Progresses of NLP tasks, language resources
Ours	1902	21 (2005-2025)	517 African Languages	African NLP contributions, Contributors, progresses, community impact

Table 1: Summary of related works. The number of papers that the survey reviewed, in the starting year (—), is unknown or includes all available works until the cut-off year mentioned, the Languages/countries the survey targeted, and the main focus area of the survey, with in ascending order of publication year.

tion process.

Paper Sources We collected papers from two sources: Semantic Scholar and the ACL Anthology. Semantic Scholar[2] (Lo et al., 2020) aggregates publications from more than 50 partners (including ACL, ACM, IEEE, MT Press, arXiv, dblp, and others) and provides access through its public API. We also retrieve papers from the Association for Computational Linguistics (ACL)[3] anthology bibliographic file, which contains prominent conference proceedings (such as EMNLP, NAACL, EACL, and ACL), journals (TACL, CL), and workshops (such as AfricaNLP).

Searching Keywords: To identify relevant AfricaNLP papers, we compiled a comprehensive keyword list comprising: (1) 54 African country names, (2) their adjectival forms such as African, Ethiopian, Nigerian, (3) 517 language names that are listed from the AfroLID (Adebara et al., 2022) work, and (4) Grass-root African organization/community names that focus on the development of African AI/NLP, such as Masakhane, HausaNLP, EthioNLP, GhanaNLP, Lelapa, LesanAI, Lanfrica and others. The keywords co-occurrence network is presented in the Appendix, Figure 15.

Searching Criteria and Filtering: We retrieve paper metadata using keyword search from the title and abstract with the criteria: 1) title and abstract not empty, 2) based on the Semantic Scholar topic classifier, in the field of study, there should be either `Computer Science` OR `Linguistic`, and 3) at least one keyword in the title or abstract. We retrieved a total of 21,821 papers from the Semantic Scholar database over the past 20 years. Additionally, we explicitly retrieved 14,674 papers from ACL anthology `aclanthology.bib` file using the same set of keywords. This results in a total of 36,494 unique papers after deduplication. We retrieved the metadata of the papers, which includes the title, abstract, authors, URL, publication venue, year, citation, field of study, authors' details (name, affiliation, paper count, and h-index), and publication type (journal, conference).

3.1 Paper Annotation

LLMs are increasingly applied to research support tasks, such as LLMs as a judge (Gu et al., 2025), human-in-the-loop dataset annotation (Calderon and Reichart, 2025), and human-LLM collaborative annotation approach (Wang et al., 2024).

As shown in Figure 2, we used two LLMs (`GPT-4.1-mini` and `Gemini-2.5-flash`) to assist in paper annotation, specifically for identifying AfricaNLP papers, extracting contribution sentences, labeling NLP topics and tasks, and verifying contributors. A human acted as a third annotator to adjudicate the final labels. The annotation was conducted across four levels of granularity: 1) **AfricaNLP papers**: determining whether a paper is African AI/NLP-related (yes/no); papers not classified as AfricaNLP are discarded. 2) **Contribution sentences**: identifying sentences that describe contributions (knowledge and artifacts, with sub-categories described in §5.2). 3) **NLP topics and tasks**: labeling the relevant NLP topics and downstream tasks. 4) **Contributors**: extracting and verifying authors, affiliated institutions, and supporting organizations/funders from the paper. We used PyPDF2[4] to automatically extract affiliations and acknowledged bodies, with manual verifica-

[2] https://www.semanticscholar.org/product/api
[3] https://aclanthology.org/
[4] https://pypi.org/project/PyPDF2/

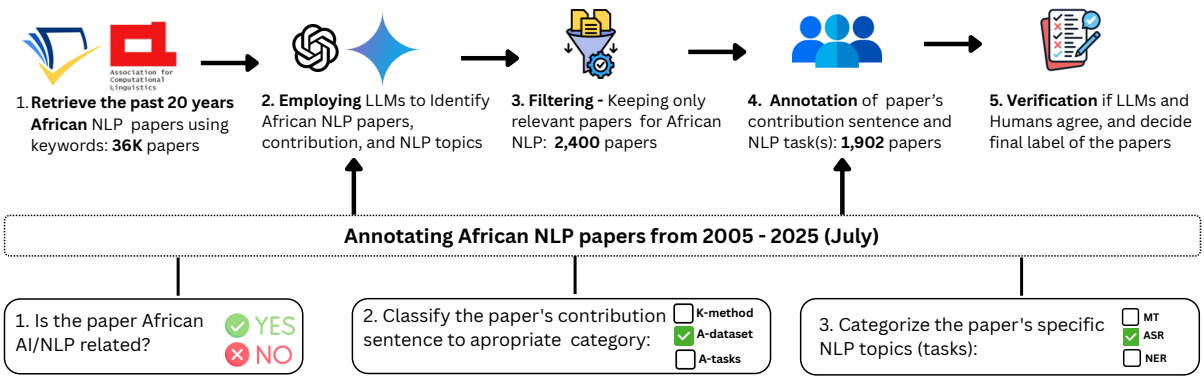

Figure 2: An illustration of the five-stage procedure for AfricaNLP paper extraction and annotation. Annotation tasks are 1) identifying AfricaNLP papers, 2) annotation of contribution sentences, and 3) categorizing the papers into the specific NLP topics (tasks). All tasks are first annotated by a pair of LLMs (GPT-4.1 and Gemini-2.5), taking into account their agreement, and then verified/annotated by humans. The overall paper annotations of task (2) and task (3) are in a **multi-label** selection options - a paper might have more than one contribution type, labeled under more than one NLP topic and task. The detailed annotation prompts of the LLMs are in Appendix E.

tion.

For the task of identifying AfricaNLP papers, GPT-4.1 classified 1,630 papers[5], and Gemini classified 2,281 out of 36,494 papers, with 1,569 overlapping (69% agreement). We manually reviewed 2,281 titles and abstracts, including Gemini-only classifications and verified 1,902 as AfricaNLP papers for further annotation. Most of the disagreements were papers in Arabic (e.g., from Morocco and Egypt) involving tasks such as sign language or optical character recognition (OCR). Examples of the paper annotation disagreement between LLMs are shown in the Appendix, Table 8.

3.2 Annotation Agreement

As shown in Table 2, the Inter-annotator Agreement (IAA) score indicates that LLMs can achieve comparable agreement scores to humans for these specific tasks. Some of the limitations of LLM-based annotations we noticed are predicting many contribution labels for a sentence and no contribution prediction is mainly shown in Gemini. This might be because we employ the lightweight and cost-effective model versions. Notably, 82% of the papers annotated by LLMs are indeed relevant. Overall, LLMs can be used as helpful annotation assistance for NLP paper contributions and topic annotation, with a human-in-the-loop verification.

4 AfricaNLP Progresses

How has AfricaNLP research evolved over the past two decades? The progress of AfricaNLP

[5]The overall analysis is based on paper titles and abstracts

Agreement	AfricaNLP?	Contribution	NLP topic
Annotators (Anno.)	0.86	0.72	0.62
Anno. Vs GPT	0.82	0.71	0.60
Anno. Vs Gemini	0.82	0.67	0.61
GPT Vs Gemini	0.85	0.69	0.69

Table 2: Inter-annotator Agreement (IAA) among three annotators and LLMs using 500 sampled papers (25 papers per/year).

over the last 20 years is shown in Figure 3. AfricaNLP has experienced steady growth since 2005, with a sharp acceleration after 2019. This surge coincides with the emergence of grassroots communities that are working on African languages. These initiatives addressed the longstanding challenge of resource scarcity through participatory, community-driven research (RQ1). In addition, the rise of AfricaNLP is due to the emergence of the transformer architecture (Vaswani et al., 2017a) and its subsequent updates, such as T5 (Text-to-Text Transfer Transformer) (Raffel et al., 2020).

In addition, the advent of the transformer architecture (Vaswani et al., 2017b) and the subsequent rise of pre-trained language models (PLMs), including multilingual models such as mBERT (Pires et al., 2019), and XLM-R (Conneau et al., 2020), have further accelerated AfricaNLP progress. Their ability to handle multilingual and low-resource settings, combined with transfer learning and continual fine-tuning approaches, has enabled AfricaNLP researchers to achieve better performance in key tasks such as machine translation, named entity

Type	Description	Example
k-dataset	Describes new knowledge about datasets, such as their new properties or characteristics.	"Our evaluation reveals a significant performance gap between high-resource languages (such as English and French) and low-resource African languages." – Adelani et al. (2025)
k-language	Presents new knowledge about language, such as a new property or characteristic of language.	"When one homophone character is substituted by another, there will be a meaning change and it is against the Amharic writing regulation." – Belay, Tadesse Destaw and Ayele, Abinew Ali and Gelaye, Getie and Yimam, Seid Muhie and Biemann, Chris (2021)
k-method	Describes new knowledge or insights about NLP models or methods.	"This study investigates the effectiveness of Language-Adaptive Fine-Tuning (LAFT) to improve SA performance in Hausa." – Sani et al. (2025)
k-people	Presents new knowledge about people, humankind, society, or communities.	"One such group, Creole languages, have long been marginalized in academic study, though their speakers could benefit from machine translation (MT)." – Robinson et al. (2024)
k-task	Knowledge/insights about existing tasks or problem domains.	"These results emphasize the importance of selecting appropriate pre-trained models based on linguistic considerations and task requirements." – Aali et al. (2024)
a-dataset	Introduces a new NLP dataset (i.e., textual resources such as corpora or lexicon).	"Last, we release two Setswana LLM-translated benchmarks, MMLU-tsn and GSM8K-tsn, to measure Setswana knowledge and reasoning capabilities." – Brown and Marivate (2025)
a-method	Introduces or proposes a new or novel NLP methodological approach or model to solve NLP task(s).	"We design and implement a new MMT model framework suitable for our new generated dataset." – Xiao et al. (2025)
a-task	Introduces or proposes a new or novel NLP task formulation (i.e., well-defined NLP problem).	"In this work, we introduce the challenge of Meroitic decipherment as a computational task" – Otten and Anastasopoulos (2025)

Table 3: Description of the taxonomy for NLP research Knowledge (**k**) and artifacts (**a**) contributions with examples from the `AfricaNLPContributions` dataset.

recognition, and sentiment analysis with limited data (Howard and Ruder, 2018). This is evidenced by the fact that *efficient/low-resource methods for NLP* is the leading NLP topic for African languages and Keywords such as 'pre-training' (201 occurrences), 'transformer' (185), 'fine-tuning' (146), 'large language models' (145), and 'LLM' (141) frequently appear in paper titles and abstracts.

What are the main focus areas of NLP research in African languages? We categorize the 1.9K AfricaNLP paper under the predefined NLP topics, which encompass methods (including efficient methods for low-resource NLP), tasks (such as machine translation and speech recognition), and themes (including NLP applications, ethics, bias, and fairness). We adopt ACL Rolling Review (ARR) NLP research topics[6]. Figure 4 shows the distribution of these NLP topics. The "*efficient/low-resource methods for NLP*" NLP topic categories are the most frequently applied methods in a multilingual and linguistically diverse setup for low-resource African languages.

How is the progress of each downstream NLP task over the years? Classical tasks such as machine translation (MT), automatic speech recognition (ASR), language identification (LID), sentiment analysis, and parsing (including part-of-speech tagging and named entity recognition) are the most prevalent and widely explored tasks. Table 7 in Appendix I presents the distribution of the top downstream NLP tasks for African languages.

[6] `https://aclrollingreview.org/areas`

Which NLP task(s) are getting less attention over time? According to publication statistics across years for specific NLP tasks, lexicon-level NLP tasks have generally receiving limited attention. For example, Lexical Semantics had 5 papers in 2006, 15 in 2020, and 4 in 2024; Part-of-Speech Tagging had 10 papers in 2014, 25 in 2020, and 8 in 2024; Word Sense Disambiguation had 5 papers in 2022 and 3 in 2024; and Named Entity Recognition had 18 in 2023, and 5 in 2024. The reason might be that with the emergence of LLMs, they are taking over the traditional NLP tasks. With the emergence of LLMs, advanced NLP research topics have gained increasing attention. For instance, Text Generation had 14 papers in 2022, 17 in 2023, and 40 in 2024, while Language Modeling had 5 papers in 2018, 25 in 2020, and 36 in 2024. On the other hand, topics such as Reasoning (6 papers) and Vision-Language tasks (4 papers) remain relatively underexplored yet.

5 AfricaNLP Contributions

In this section, we discuss the analysis of `AfricaNLPContributions` dataset (§5.1) and contribution classification task experiments (§5.2).

5.1 AfricaNLPContributions Dataset

Contribution sentences are automatically annotated from the abstracts of 1902 papers with the help of LLMs and human approvals, as discussed in Section 3.1. The sentence-level contribution annotations comprised a single or a few sentences about the article's contribution inside the abstract (D'Souza et al., 2021; Pramanick et al., 2025a). We adopted the 8 classes NLPContributions annotation taxonomy (5 knowledge type and 3 arti-

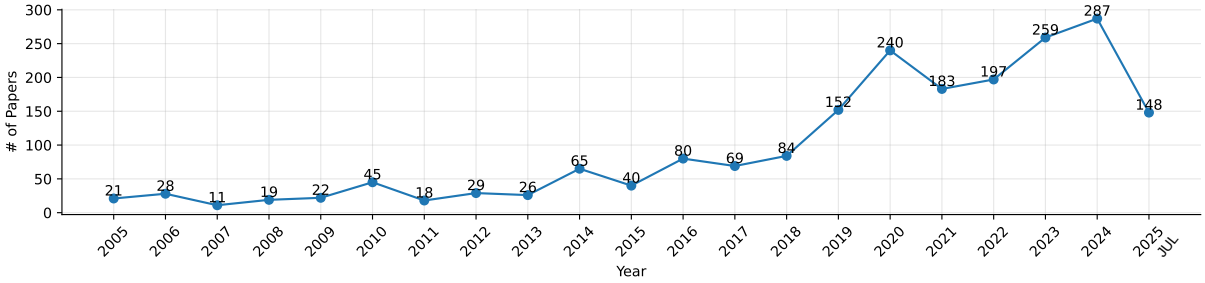

Figure 3: **The progress of AfricaNLP research over the past 20 years, 2005 – 2025.7**. The number of publications shows a steady increase starting from 2019, with the data for the cutoff year 2025 covering to the end of July.

facts type) from previous work (Pramanick et al., 2025a), The class names are knowledge about method (*k-method*), knowledge about dataset (*k-dataset*), knowledge about task (*k-task*), knowledge about language (*k-language*), and artifacts of new methods (*a-methods*), new datasets (*a-datasets*), and new tasks (*a-tasks*). The contribution classes are described with examples in Table 3.

What kinds of contributions are made in AfricaNLP papers? From the final 1,902 AfricaNLP papers, 7,796 unique sentences are annotated as knowledge and artifact contributions. The distributions are presented in Table 9. Among these, 27% of the sentences are knowledge of methods (k-method), 26% of the sentences are artifacts of methods (a-method), 12.9% of the sentences are artifacts of the dataset (a-dataset), and 10.7% of the sentences are knowledge of language (k-language). The least contributions are artifacts of new tasks (a-task) and knowledge about people (k-people), 5.3% and 4.3% respectively. The contributions *knowledge-method* and *artifact-method* are the most identified collection contribution sentences in the abstract. The detailed contribution statistics across 8 labels are shown in Figure 5.

How AfricaNLP contributions align with global NLP research? We compare the contributions of AfricaNLP (`AfricaNLPContributions`) with works done at global NLP (`NLPContributions`) by Pramanick et al. (2025b). In the knowledge category, the dominant `NLPContributions` class is k-task, which involves creating new downstream NLP tasks. In `AfricaNLPContributions`, the focus is on k-method - developing, adopting, or evaluating NLP methods. From the artifact category, a-dataset is the dominant one in both global and AfricaNLP contributions. In artifacts, `AfricaNLPContributions` aligns with the global NLP research contribution - `NLPContributions` (Pramanick et al., 2025a).

5.2 Contributions Classification Task

Task Definition. The task is to automatically classify a sentence into a predefined two broad contribution classes: **artifacts** - which encompass the development of new or novel resources (methods, tasks, or datasets) and **knowledge** - contributions that enrich the field with new insights or knowledge of language, methods, tasks, people, or datasets. Given an abstract A in plaintext format, the goal was to extract a set of contribution sentences $C_{sent} = \{C_{sent_1}, ..., C_{sent_N}\}$. We set up a multi-class classifier for the `AfricaNLPContributions` dataset to model and automatically classify a sentence into the 8 classes (5 knowledge and 3 artifact contribution classes), the details of the classes are shown in Table 3.

Methods. We select the following pre-trained language models (PLMs) for benchmarking from different categories: BERT (Devlin et al., 2019) from general-purpose PLMs, SciBERT (Beltagy et al., 2019) from scientific text-based PLMs, AfroXLMR (Adelani et al., 2024) from African-language centric, and GPT-4.1-mini (OpenAI, 2025)) from state-of-the-art LLMs. For prompting LLM, we apply a zero-shot setting. The `AfricaNLPContributions` dataset is split into 5,847 training set, 1,170 test set, and 779 validation set, with a 75:15:10 splitting ratio, respectively. Training details and hyperparameters are presented in Appendix K.

Results and Discussion. Table 4 shows the results. We observe that SciBERT and AfroXLMR outperform BERT and GPT, likely due to the SciBERT pre-training on a collection of scientific documents, and AfroXLMR's familiarity with African coun-

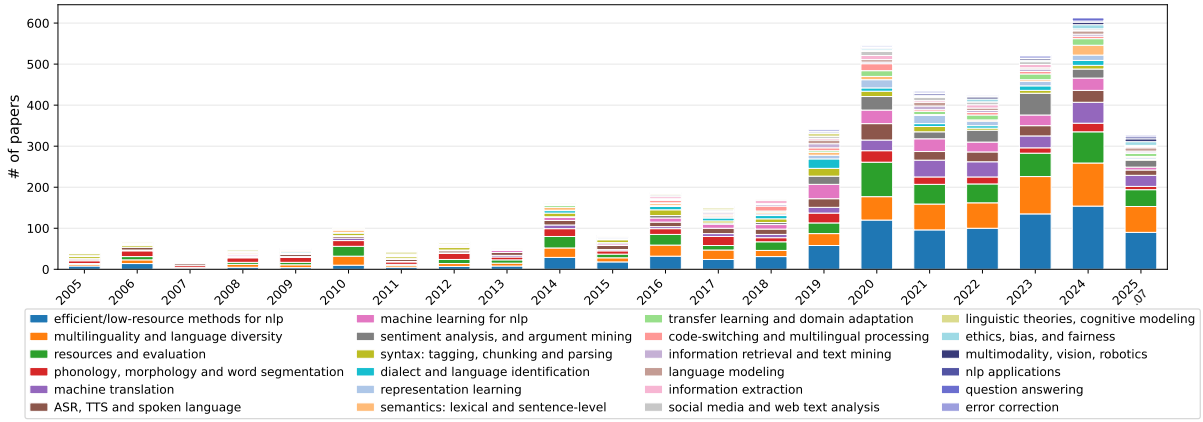

Figure 4: **AfricaNLP research increases in topic diversity**, (a paper may appear in multiple categories). The last year 2025 is cut off at July. The NLP topics are sorted in descending order based on the overall total frequency of papers. Through the years, the number of topics relating to African languages and their diversity has increased from 2019 to 2025.

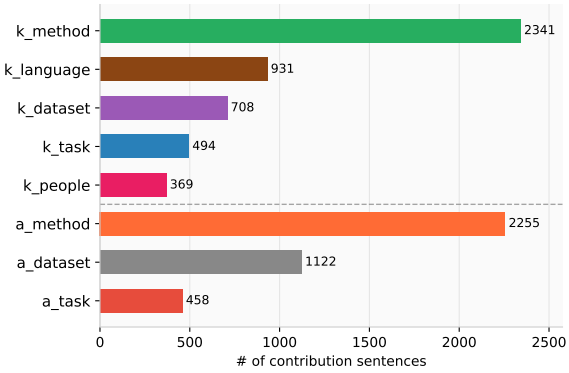

Figure 5: Contribution sentences occurrence statistics of different contribution types in the paper abstracts (a paper may have multiple contribution sentences).

Contrib.	BERT	SciBERT	AfroXLMR	GPT*
k-language	**0.69**	0.68	0.65	0.49
k-method	0.79	**0.80**	0.78	0.63
k-people	0.42	0.49	**0.54**	0.49
k-task	**0.41**	0.40	0.36	0.30
k-dataset	0.54	0.51	**0.62**	0.55
a-dataset	0.80	0.81	**0.83**	0.80
a-method	0.84	**0.85**	0.84	0.76
a-task	0.42	**0.48**	0.43	0.47
Overall	0.61	**0.63**	**0.63**	0.50

Table 4: Performance of different models (macro F1) for contribution statement classification in a multi-label setting. GPT* is `gpt-4.1-mini`. Bold results are the best performance for the specific contribution class.

tries and language names.

6 AfricaNLP Contributors

In this section, we provide analyses and insights to the following threefold contributors on AfricaNLP. **1) Authors**: We analyze the sets of authors (both first and co-authors) of the papers. **2) Affiliations**: We identify affiliated institutions, organizations, and independent researchers from the authors' affiliation section of the paper. **3) Supporters**: We identify funding, computing resource (GPU), and model API supports from the acknowledgment section of the NLP paper.

Who is supporting the development of AfricaNLP research? The National Science Foundation (NSF)[7] - independent United States Federal Government Agency, Google[8], National Research Foundation of South Africa (ZA)[9], Lacuna Fund [10] - an initiative co-founded by The Rockefeller Foundation, Google.org, and Canada's International Development Research Center and others are the main supporters of the development of AfricaNLP. As noted throughout the papers, we also acknowledge that the mention of organization names here are not reflect the official policies or positions of the funding or supporting organizations. The top contributors are presented in Figure 6; the complete list of supporter organizations is included in Appendix F.

Which institutions are affiliated with the development of AfricaNLP? We have identified a total

[7] https://www.nsf.gov/

[8] https://www.google.org/

[9] https://www.nrf.ac.za/

[10] https://lacunafund.org/

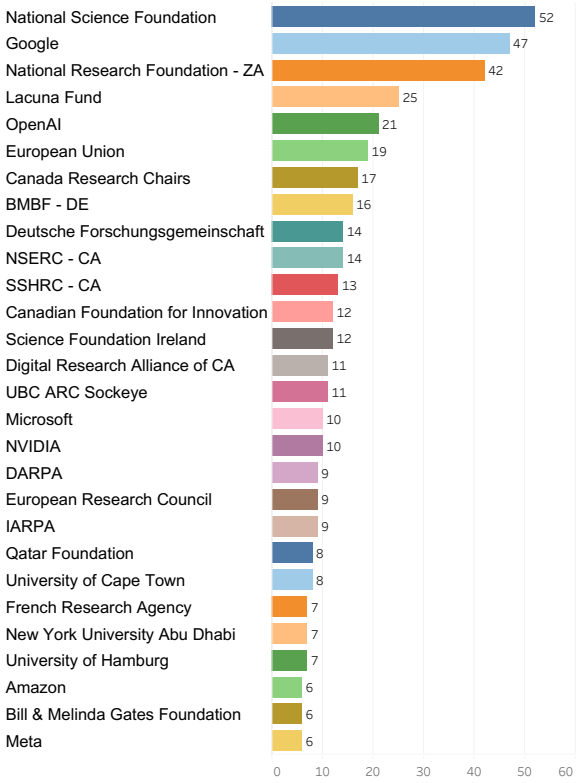

Figure 6: Top organizations with the number of publications supported, extracted from 1.9K AfricaNLP papers.

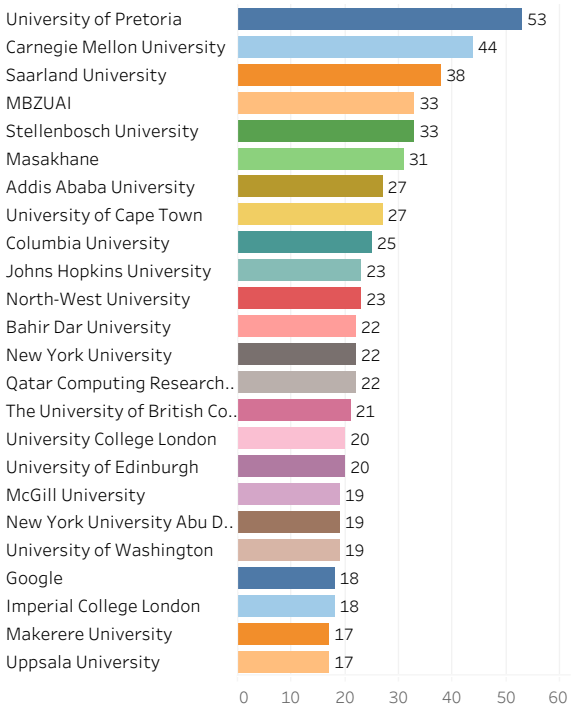

Figure 7: Top affiliated institutions/organizations/individuals of the authors (multiple institutions may be affiliated in a paper) with the number of publications, extracted from 1.9K AfricaNLP papers.

of around 200 unique organizations, and Figure 7 presents the top affiliated organization. Generally, the majority of the top contributing institutions to AfricaNLP are based outside Africa, while the *University of Pretoria* and *Stellenbosch University* are the leaders among Africa-based affiliated institutions. The entire list of affiliated organization names is given in Appendix Table G and their networking in Figure 7.

7 Applications and Future Work

Advancing NLP for African languages necessitates a multifaceted approach that addresses data scarcity, linguistic diversity, and cultural context. The following are some of the future directions for AfricaNLP: 1) Explore Advanced NLP research topics for AfricaNLP, such as Text generation, reasoning, and vision-language tasks, which are nearly underexplored for African languages. 2) Using `AfricaNLPContributions` dataset, explore better modeling of automatic contribution classification tasks for NLP research papers. 3) Using the **AfricaNLP** paper collections, explore the papers in various ways, such as how venues influence the nature of AfricaNLP research over time. The opportunities we identified for AfricaNLP are presented in Appendix D.

8 Conclusion

In this work, we presented a systematic survey of AfricaNLP spanning the past two decades, examining contributions in terms of research outputs, authorship, institutional affiliations, and supporting organizations. Despite persistent challenges, interest in African NLP has expanded significantly, fueled by community initiatives and the development of new tools. Over this period, publications on African languages have more than quadrupled, reflecting the growing momentum of the field. Through a set of guiding research questions, we quantitatively captured this progress, highlighting community-driven collaborations in producing datasets and benchmarks. Our surveying pipeline is fully reproducible, extendable, and adaptable - enabling continuous tracking of AfricaNLP progress tools and offering a framework that can be applied to other languages and regions.

Limitations

The analysis of the AfricaNLP papers is done with the help of a series of automated preprocessing tools and LLMs for reproducibility. The processes

may not be perfect, so some noise and errors may occur, starting from the scraping of available NLP papers. While Semantic Scholar accesses various research paper databases, some papers might be missed if the paper is not accessed by Semantic Scholar. This analysis aims to provide a higher-level view of what is happening in African NLP.

Scope Limitations of Paper Searching: Our paper searching scope encompasses the last 20 years, up to July 2025, based on the latest updates from Semantic Scholar and ACL databases. Additionally, we limited the scope of the study as follows.

- We survey papers that include at least one target keyword in their title or abstract; there might be interesting works that do not explicitly include an African country or language name in their title or abstract.
- Papers may not be included in the Semantic Scholar database, and papers might be published that are not accessed by Semantic Scholar and ACL anthology.
- Papers written only in English (the title and abstract) are included in the analysis.
- We exclude papers with very short abstracts (1/2 sentences), books, only linguistic papers without any computation, as our field of study is computer science OR linguistics, with/without other fields of study, based on the Semantic Scholar topic classifier models.

References

Yasamin Aali, Sardar Hamidian, and Parsa Farinneya. 2024. YSP at SemEval-2024 Task 1: Enhancing Sentence Relatedness Assessment using Siamese Networks. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 959–963, Mexico City, Mexico. Association for Computational Linguistics.

Ife Adebara and Muhammad Abdul-Mageed. 2022. Towards Afrocentric NLP for African Languages: Where We Are and Where We Can Go. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3814–3841, Dublin, Ireland. Association for Computational Linguistics.

Ife Adebara, AbdelRahim Elmadany, Muhammad Abdul-Mageed, and Alcides Inciarte. 2022. AfroLID: A Neural Language Identification Tool for African Languages. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1958–1981, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

David Ifeoluwa Adelani, Hannah Liu, Xiaoyu Shen, Nikita Vassilyev, Jesujoba O. Alabi, Yanke Mao, Haonan Gao, and En-Shiun Annie Lee. 2024. SIB-200: A Simple, Inclusive, and Big Evaluation Dataset for Topic Classification in 200+ Languages and Dialects. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 226–245, St. Julian's, Malta. Association for Computational Linguistics.

David Ifeoluwa Adelani, Jessica Ojo, Israel Abebe Azime, Jian Yun Zhuang, Jesujoba Oluwadara Alabi, Xuanli He, Millicent Ochieng, Sara Hooker, Andiswa Bukula, En-Shiun Annie Lee, Chiamaka Ijeoma Chukwuneke, Happy Buzaaba, Blessing Kudzaishe Sibanda, Godson Koffi Kalipe, Jonathan Mukiibi, Salomon Kabongo Kabenamualu, Foutse Yuehgoh, Mmasibidi Setaka, Lolwethu Ndolela, and 8 others. 2025. IrokoBench: A New Benchmark for African Languages in the Age of Large Language Models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2732–2757, Albuquerque, NM, USA. Association for Computational Linguistics.

Azzedine Aftiss, Salima Lamsiyah, Christoph Schommer, and Said Ouatik El Alaoui. 2025. Empirical Evaluation of Pre-trained Language Models for Summarizing Moroccan Darija News Articles. In *Proceedings of the 4th Workshop on Arabic Corpus Linguistics (WACL-4)*, pages 77–85, Abu Dhabi, UAE. Association for Computational Linguistics.

Khadija Ait ElFqih and Johanna Monti. 2024. Large Language Models as Legal Translators of Arabic Legislation: Do ChatGPT and Gemini Care for Context and Terminology? In *Proceedings of the Second Arabic Natural Language Processing Conference*, pages 111–122, Bangkok, Thailand. Association for Computational Linguistics.

Jesujoba O. Alabi, Michael A. Hedderich, David Ifeoluwa Adelani, and Dietrich Klakow. 2025. Charting the Landscape of African NLP: Mapping Progress and Shaping the Road Ahead.

Fakhraddin Alwajih, Gagan Bhatia, and Muhammad Abdul-Mageed. 2024. Dallah: A Dialect-Aware Multimodal Large Language Model for Arabic. In *Proceedings of the Second Arabic Natural Language Processing Conference*, pages 320–336, Bangkok, Thailand. Association for Computational Linguistics.

Cynthia Jayne Amol, Everlyn Asiko Chimoto, Rose Delilah Gesicho, Antony M. Gitau, Naome A. Etori, Caringtone Kinyanjui, Steven Ndung'u, Lawrence Moruye, Samson Otieno Ooko, Kavengi Kitonga, Brian Muhia, Catherine Gitau, Antony

Ndolo, Lilian D. A. Wanzare, Albert Njoroge Kahira, and Ronald Tombe. 2024. State of NLP in Kenya: A Survey. *Preprint*, arXiv:2410.09948.

Paul Azunre, Salomey Osei, Salomey Addo, Lawrence Asamoah Adu-Gyamfi, Stephen Moore, Bernard Adabankah, Bernard Opoku, Clara Asare-Nyarko, Samuel Nyarko, Cynthia Amoaba, Esther Dansoa Appiah, Felix Akwerh, Richard Nii Lante Lawson, Joel Budu, Emmanuel Debrah, Nana Boateng, Wisdom Ofori, Edwin Buabeng-Munkoh, Franklin Adjei, and 8 others. 2021. NLP for Ghanaian Languages. *Preprint*, arXiv:2103.15475.

David Bamutura, Peter Ljunglöf, and Peter Nebende. 2020. Towards computational resource grammars for Runyankore and rukiga. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2846–2854, Marseille, France. European Language Resources Association.

Belay, Tadesse Destaw and Ayele, Abinew Ali and Gelaye, Getie and Yimam, Seid Muhie and Biemann, Chris. 2021. Impacts of homophone normalization on semantic models for amharic. In *2021 International Conference on Information and Communication Technology for Development for Africa (ICT4DA)*, pages 101–106.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Nathan Brown and Vukosi Marivate. 2025. Pula: Training Large Language Models for Setswana. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6634–6656, Albuquerque, NM, USA. Association for Computational Linguistics.

Nitay Calderon and Roi Reichart. 2025. On Behalf of the Stakeholders: Trends in NLP Model Interpretability in the Era of LLMs. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 656–693, Albuquerque, NM, USA. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of ACL*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, MN, USA. Association for Computational Linguistics.

Jennifer D'Souza and Sören Auer. 2020. Nlpcontributions: An annotation scheme for machine reading of scholarly contributions in natural language processing literature. In *Proceedings of the 1st Workshop on Extraction and Evaluation of Knowledge Entities from Scientific Documents co-located with the ACM/IEEE Joint Conference on Digital Libraries in 2020 (JCDL 2020)*, Virtual Event, China.

Jennifer D'Souza, Sören Auer, and Ted Pedersen. 2021. SemEval-2021 task 11: NLPContributionGraph - structuring scholarly NLP contributions for a research knowledge graph. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 364–376, Online. Association for Computational Linguistics.

Tekabe Legesse Feleke. 2017. 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. Association for Computational Linguistics.

Sisay Fissaha Adafre. 2005. Part of speech tagging for Amharic using conditional random fields. In *Proceedings of the ACL Workshop on Computational Approaches to Semitic Languages*, pages 47–54, Ann Arbor, Michigan. Association for Computational Linguistics.

Google. 2025. Gemini 2.5 flash. Last accessed 17 Augest 2025.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. 2025. A Survey on LLM-as-a-Judge. *Preprint*, arXiv:2411.15594.

Najet Hadj Mohamed, Malak Rassem, Lifeng Han, and Goran Nenadic. 2023. AlphaMWE-Arabic: Arabic Edition of Multilingual Parallel Corpora with Multiword Expression Annotations. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 448–457, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Mariam Hassib, Nancy Hossam, Jolie Sameh, and Marwan Torki. 2022. AraDepSu: Detecting depression and suicidal ideation in Arabic tweets using transformers. In *Proceedings of the The Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 302–311, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Shehenaz Hossain, Fouad Shammary, Bahaulddin Shammary, and Haithem Afli. 2025. Enhancing Dialectal Arabic Intent Detection through Cross-Dialect Multilingual Input Augmentation. In *Proceedings of the 4th Workshop on Arabic Corpus Linguistics (WACL-4)*, pages 44–49, Abu Dhabi, UAE. Association for Computational Linguistics.

Jeremy Howard and Sebastian Ruder. 2018. Universal Language Model Fine-tuning for Text Classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339, Melbourne, Australia. Association for Computational Linguistics.

Songbo Hu, Abigail Oppong, Ebele Mogo, Charlotte Collins, Giulia Occhini, Anna Barford, and Anna Korhonen. 2025. Natural language processing technologies for public health in Africa: scoping review. *Journal of Medical Internet Research*, 27:1–25.

Yu Hu, Irina Matveeva, John Goldsmith, and Colin Sprague. 2005. Refining the SED heuristic for morpheme discovery: Another look at Swahili. In *Proceedings of the Workshop on Psychocomputational Models of Human Language Acquisition*, pages 28–35, Ann Arbor, Michigan. Association for Computational Linguistics.

Kedir Yassin Hussen, Walelign Tewabe Sewunetie, Abinew Ali Ayele, Sukairaj Hafiz Imam, Shamsuddeen Hassan Muhammad, and Seid Muhie Yimam. 2025. The State of Large Language Models for African Languages: Progress and Challenges. *Preprint*, arXiv:2506.02280.

Sukairaj Hafiz Imam, Babangida Sani, Dawit Ketema Gete, Bedru Yimam Ahmed, Ibrahim Said Ahmad, Idris Abdulmumin, Seid Muhie Yimam, Muhammad Yahuza Bello, and Shamsuddeen Hassan Muhammad. 2025. Automatic Speech Recognition for African Low-Resource Languages: Challenges and Future Directions. In *Proceedings of the Sixth Workshop on African Natural Language Processing (AfricaNLP 2025)*, pages 89–94, Vienna, Austria. Association for Computational Linguistics.

Isa Inuwa-Dutse. 2025. NaijaNLP: A Survey of Nigerian Low-Resource Languages. *Preprint*, arXiv:2502.19784.

Abdullah Khered, Youcef Benkhedda, and Riza Batista-Navarro. 2025. Dial2MSA-Verified: A Multi-Dialect Arabic Social Media Dataset for Neural Machine Translation to Modern Standard Arabic. In *Proceedings of the 4th Workshop on Arabic Corpus Linguistics (WACL-4)*, pages 50–62, Abu Dhabi, UAE. Association for Computational Linguistics.

Diksha Khurana, Aditya Koli, Kiran Khatter, and Sukhdev Singh. 2023. Natural language processing: state of the art, current trends and challenges. *Multimedia tools and applications*, 82(3):3713–3744.

Eliese-Sophia Lincke, Kirill Bulert, and Marco Büchler. 2019. Optical Character Recognition for Coptic fonts: A multi-source approach for scholarly editions. In *Proceedings of the 3rd International Conference on Digital Access to Textual Cultural Heritage*, DATeCH2019, page 87–91, New York, NY, USA. Association for Computing Machinery.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. 2020. S2ORC: The semantic scholar open research corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4969–4983, Online. Association for Computational Linguistics.

Sanad Malaysha, Mustafa Jarrar, and Mohammed Khalilia. 2024. NLU-STR at SemEval-2024 Task 1: Generative-based Augmentation and Encoder-based Scoring for Semantic Textual Relatedness. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 894–901, Mexico City, Mexico. Association for Computational Linguistics.

Million Meshesha and C. Jawahar. 2007. Optical Character Recognition of Amharic Documents. *African J. of Inf. & Commun. Technology*, 3.

So Miyagawa. 2025. RAG-enhanced neural machine translation of Ancient Egyptian text: A case study of THOTH AI. In *Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities*, pages 33–40, Albuquerque, NM, USA. Association for Computational Linguistics.

Hamdy Mubarak, Ammar Rashed, Kareem Darwish, Younes Samih, and Ahmed Abdelali. 2021. Arabic offensive language on Twitter: Analysis and experiments. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 126–135, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Shamsuddeen Hassan Muhammad, Idris Abdulmumin, Abinew Ali Ayele, Nedjma Ousidhoum, David Ifeoluwa Adelani, Seid Muhie Yimam, Ibrahim Sa'id Ahmad, Meriem Beloucif, Saif M. Mohammad, Sebastian Ruder, Oumaima Hourrane, Pavel Brazdil, Alipio Jorge, Felermino Dário Mário António Ali, Davis David, Salomey Osei, Bello Shehu Bello, Falalu Ibrahim, Tajuddeen Gwadabe, and 8 others. 2023. AfriSenti: A Twitter Sentiment Analysis Benchmark for African Languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13968–13981, Singapore. Association for Computational Linguistics.

Shamsuddeen Hassan Muhammad, Ibrahim Said Ahmad, Idris Abdulmumin, Falalu Ibrahim Lawan, Babangida Sani, Sukairaj Hafiz Imam, Yusuf Aliyu, Sani Abdullahi Sani, Ali Usman Umar, Tajuddeen Gwadabe, Kenneth Church, and Vukosi Marivate. 2025. HausaNLP: Current Status, Challenges and Future Directions for Hausa Natural Language Processing. *Preprint*, arXiv:2505.14311.

Joaquim Mussandi and Andreas Wichert. 2024. NLP Tools for African Languages: Overview. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 2*, pages 73–82, Santiago de Compostela, Galicia/Spain. Association for Computational Linguistics.

Jessica Ojo, Odunayo Ogundepo, Akintunde Oladipo, Kelechi Ogueji, Jimmy Lin, Pontus Stenetorp, and David Ifeoluwa Adelani. 2025. AfroBench: How good are large language models on African languages? In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 19048–19095, Vienna, Austria. Association for Computational Linguistics.

OpenAI. 2025. Introducing gpt-4.1. Last accessed 11 Augest 2025.

Joshua N. Otten and Antonios Anastasopoulos. 2025. Towards Ancient Meroitic Decipherment: A Computational Approach. In *Proceedings of the Second Workshop on Ancient Language Processing*, pages 87–98, The Albuquerque Convention Center, Laguna. Association for Computational Linguistics.

Matthew J Page, Joanne E McKenzie, Patrick M Bossuyt, Isabelle Boutron, Tammy C Hoffmann, Cynthia D Mulrow, Larissa Shamseer, Jennifer M Tetzlaff, Elie A Akl, Sue E Brennan, and 1 others. 2021. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *bmj*, 372.

Jiaxin Pei, Aparna Ananthasubramaniam, Xingyao Wang, Naitian Zhou, Apostolos Dedeloudis, Jackson Sargent, and David Jurgens. 2022. POTATO: The portable text annotation tool. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 327–337, Abu Dhabi, UAE. Association for Computational Linguistics.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual bert? *arXiv preprint arXiv:1906.01502*.

Aniket Pramanick, Yufang Hou, Saif M. Mohammad, and Iryna Gurevych. 2025a. The Nature of NLP: Analyzing Contributions in NLP Papers. *arXiv preprint*. ArXiv:2409.19505 [cs].

Aniket Pramanick, Yufang Hou, Saif M. Mohammad, and Iryna Gurevych. 2025b. The Nature of NLP: Analyzing Contributions in NLP Papers. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 25169–25191, Vienna, Austria. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1).

Nathaniel Robinson, Raj Dabre, Ammon Shurtz, Rasul Dent, Onenamiyi Onesi, Claire Monroc, Loïc Grobol, Hasan Muhammad, Ashi Garg, Naome Etori, Vijay Murari Tiyyala, Olanrewaju Samuel, Matthew Stutzman, Bismarck Odoom, Sanjeev Khudanpur, Stephen Richardson, and Kenton Murray. 2024. Kreyòl-MT: Building MT for Latin American, Caribbean and Colonial African Creole Languages. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3083–3110, Mexico City, Mexico. Association for Computational Linguistics.

Abdulwahab Sahyoun and Shady Shehata. 2023. AraDiaWER: An explainable metric for dialectical Arabic ASR. In *Proceedings of the Second Workshop on NLP Applications to Field Linguistics*, pages 64–73, Dubrovnik, Croatia. Association for Computational Linguistics.

Kyle Rui Sang, Tahseen Rabbani, and Tianyi Zhou. 2025. Federated Meta-Learning for Low-Resource Translation of Kirundi. In *Proceedings of the Third Workshop on Resources and Representations for Under-Resourced Languages and Domains (RESOURCEFUL-2025)*, pages 190–194, Tallinn, Estonia. University of Tartu Library, Estonia.

Sani Abdullahi Sani, Shamsuddeen Hassan Muhammad, and Devon Jarvis. 2025. Investigating the Impact of Language-Adaptive Fine-Tuning on Sentiment Analysis in Hausa Language Using AfriBERTa. In *Proceedings of the First Workshop on Language Models for Low-Resource Languages*, pages 101–111, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Hussien Seid and Björn Gambäck. 2005. A speaker independent continuous speech recognizer for Amharic. In *Interspeech 2005*, pages 3349–3352.

Kathleen Siminyu, Jade Abbott, Kola Tubosun, Aremu Anuoluwapo, Blessing K Sibanda, Kofi Yeboah, David Adelani, Masabata Mokgesi-Selinga, Frederick R Apina, Angela Thandizwe Mthembu, and 1 others. 2023. Consultative engagement of stakeholders toward a roadmap for African language technologies. *Patterns*, 4(8).

Hannah Snyder. 2019. Literature review as a research methodology: An overview and guidelines. *Journal of Business Research*, 104:333–339.

Ofra Tirosh-Becker, Michal Kessler, Oren Becker, and Yonatan Belinkov. 2022. Part-of-speech and morphological tagging of Algerian Judeo-Arabic. In *Northern European Journal of Language Technology, Volume 8*, Copenhagen, Denmark. Northern European Association of Language Technology.

Atnafu Lambebo Tonja, Tadesse Destaw Belay, Israel Abebe Azime, Abinew Ali Ayele, Moges Ahmed Mehamed, Olga Kolesnikova, and Seid Muhie

Yimam. 2023. 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. Association for Computational Linguistics.

Andreea Ioana Tudor and Tsegaye Misikir Tashu. 2025. Mapping Cross-Lingual Sentence Representations for Low-Resource Language Pairs Using Pre-trained Language Models. In *Proceedings of the First Workshop on Language Models for Low-Resource Languages*, pages 249–257, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017a. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017b. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Xinru Wang, Hannah Kim, Sajjadur Rahman, Kushan Mitra, and Zhengjie Miao. 2024. Human-llm collaborative annotation through effective verification of llm labels. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–21.

Bushi Xiao, Qian Shen, and Daisy Zhe Wang. 2025. From Text to Multi-Modal: Advancing Low-Resource-Language Translation through Synthetic Data Generation and Cross-Modal Alignments. In *Proceedings of the Eighth Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2025)*, pages 24–35, Albuquerque, New Mexico, U.S.A. Association for Computational Linguistics.

Michelle Yuan, Mozhi Zhang, Benjamin Van Durme, Leah Findlater, and Jordan Boyd-Graber. 2020. Interactive refinement of cross-lingual word embeddings. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5984–5996, Online. Association for Computational Linguistics.

Appendix

A Processing of Affiliated Institutions and Acknowledged Organizations

From 1,902 papers, the support contributors are identified in the acknowledgment section of 412 papers. The remaining papers are either do not have an acknowledgment section or acknowledge the support of research idea development, reviewers, readers, volunteers, or data annotators. During acknowledged organization annotation, similar organization names, such as Google DeepMind and Google Research are merged into a single organization (Google), together with support from different regional Google offices. The same approach is applied to acknowledged organizations: departments, faculties, and research labs or groups, which are grouped under the umbrella of the same organization names are recorded as the same affiliation. We extract the list of affiliated institutions from the actual papers. The affiliation institutions are identified from 1,367 papers; the other papers do not have an affiliation, either independent researchers, require a subscription to read/access the paper, or are unable ot access the PDF.

B Additional Analysis

To what extent does this survey cover AfricaNLP research papers? To ensure that as many AfricaNLP papers as possible are included in our study, we manually collect a random 50 sample research papers from various publication venues such as `Springer Nature` and contributions (model, dataset, and methods) that involve African languages, and we cross-checked to ensure that these papers are included in our collection. Among these samples, 94% of the papers are included in our collection. The AfricaNLP papers that do not have at least a single keyword in the title/abstract from the set of searching keywords might be excluded.

How effective are LLMs for paper contribution and NLP topic annotation? The papers are first annotated by LLMs and then verified/re-annotated by three human annotators, who are PhD students working in the NLP research area. We randomly sampled 500 papers (26% of the collection) and annotated them using both LLMs and the three human annotators. We then measured the inter-annotator agreement (IAA) among pairs of LLMs, human annotators, and between humans and LLMs for each annotation task. The agreement between human annotators, as measured by Fleiss' Kappa (among the three annotators), is 84%, while Cohen's kappa (pairwise agreement between human annotators, then averaged) is 86%.

Which NLP topic(s) and languages received attention earlier in AfricaNLP? From the set of NLP research topics, 1) Syntax: tagging, chunking, and parsing (Fissaha Adafre, 2005), 2) Phonology, morphology, and word segmentation (Hu et al., 2005), 3) ASR (Seid and Gambäck, 2005), 4) OCR (Meshesha and Jawahar, 2007) are the early researched downstream NLP tasks since 2005. Languages such as Amharic, Afrikaans, Kinyarwanda, Swahili, Setswana, and Yoruba are among the earliest researched African languages since 2005. More details about the progress of specific NLP tasks are presented in Appendix H.

Who are the authors contributing to AfricaNLP research? To examine author contributions and the distribution of papers, we conducted a statistical analysis of 1,902 publications. Parallel to the growth in publications, the number of contributing authors has increased exponentially over time. In 2006, there were 21 papers authored by 78 researchers. By 2024, this had grown to 287 papers with 1,103 authors. From 1,902 papers overall, there are 4,901 unique authors, with an average of 2.6 authors per paper. Among them, 52 authors have contributed more than 10 papers, 166 authors have contributed between 5 and 10 papers, and the remaining 1,683 authors have contributed between 1 and 5 papers. The authors and co-authorship networks are presented in Figure 13, and authors who co-author in 10 or more papers are in Figure 14.

C `AfricaNLPContributions` Dataset Distributions

Figure 8 shows the co-occurrence of contribution types. Table 9 shows the statistics of the 8 contribution types.

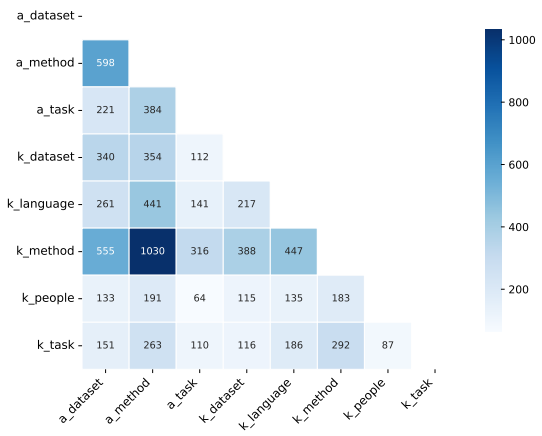

Figure 8: Contribution co-occurrence across the 8 contribution labels, 8,678 contribution sentences with 882 (11.3%) statements received multiple contribution labels

Type	Sub-typ.	Prop. %	# Sent
Knowledge	k-dataset	8.16	708
	k-language	10.73	931
	k-method	26.98	2341
	k-people	4.25	369
	k-task	5.69	494
Artifact	a-dataset	12.93	1122
	a-method	25.99	2255
	a-task	5.28	458

Figure 9: Distributions of `AfricaNLPContributions` classes: **Prop.** is proportion in percent and **Sent.** is number of sentences for the corresponding contribution class.

D Opportunities for AfricaNLP

To contribute more to AfricaNLP, we identified the following opportunities.

D.1 Community Engagement

We conducted a survey and compiled a list of communities and initiatives working on African AI/NLP. The following communities are frequently mentioned, such as Maskahene[11], GhanaNLP[12], HuasaNLP[13], EthioNLP[14], NaijaVoices[15], Lanfrica[16], and others are involving native speakers of African languages and Open for research collaborations including data collection, annotation, validation and model result error analysis ensures linguistic and cultural relevance in NLP systems. The details about each community can be found on their website.

D.2 Policy and Funding Support

Increased funding and supportive policies have the potential to significantly strengthen NLP research in Africa by fostering innovations that address local needs. Supporting organizations play a central role in advancing research on African languages, and their involvement has expanded in parallel with the growth of publications. Specifically, the number of supporting organizations over the years was 26 in 2016, 14 in 2017, 28 in 2018, 39 in 2019, 64 in 2020, 63 in 2021, 103 in 2022, 146 in 2023, and 165 in 2024, corresponding to a total of 1,902 papers. As shown in Figure 6, the leading supporting organizations are predominantly based outside Africa, underscoring both the reliance on external institutions and the opportunities for these organizations to contribute to the development of low-resource languages. At the same time, the implementation of flexible AI policies at governmental and institutional levels (Siminyu et al., 2023) is a critical pathway for advancing AfricaNLP.

D.3 Science Assessment Reform and Inclusive Recognition

Efforts to advance NLP for African languages must also engage with the structural enablers and barriers of scientific publishing. Many African institutions use rigid metrics that discourage conference participation, limiting visibility and uptake of local innovations in global NLP. Supporting reform of science assessment practices—to better align with the norms of fast-moving computational fields—is crucial. Advocacy is

[11] https://www.masakhane.io/
[12] https://ghananlp.github.io/
[13] https://hausanlp.github.io/
[14] https://ethionlp.github.io/
[15] https://naijavoices.com/
[16] https://lanfrica.com/en

needed to: 1) Encourage recognition of top-tier NLP conference proceedings as equivalent to journal publications in research assessments; 2) Promote co-indexing of AfricaNLP outputs in both conference and journal repositories; 3) Create pathways for researchers to submit extended versions of ACL papers to indexed journals, increasing reward alignment; and 4) Support travel and registration sponsorships for African authors to present at global venues, ensuring that local innovations are seen, cited, and built upon. Such actions would not only increase participation but also elevate the global impact of African research communities.

D.4 Limitations of Scientific Incentives and Dissemination Pathways

A key limitation not captured in current bibliometric or semantic analyses is the influence of science incentive systems in different African countries. In many contexts, institutional and national research evaluation frameworks heavily prioritize journal publications—particularly those indexed in Scopus or Web of Science—over conference proceedings, even for fast-moving fields like NLP and machine learning. This may discourage African researchers from submitting to premier venues such as ACL, EMNLP, NAACL, or COLING, which are often conference-based, despite their high global impact. As a result, valuable African NLP work may remain underrepresented in global discourse or be siloed in lower-visibility outlets. ACL and other top NLP conferences might not reach most Africans who are in countries or institutions that reward journals. This publication misalignment contributes to systemic undercounting in automated surveys like ours, especially for papers published outside of major conference ecosystems.

E LLM prompts for Annotation

E.1 LLM prompt to Identify AfricaNLP papers

LLM prompts for Identify Africa NLP papers

You are an assistant that determines whether a research paper focuses on Natural Language Processing (NLP) for African languages.
Based on the title and abstract below, determine if this paper is about NLP research involving African languages that develops NLP datasets, tools, models, techniques, or studies computational linguistics aspects of African languages.
Do not consider it if the title and abstract are not written in English.
Title: title
Abstract: abstract
Answer with only "yes" or "no" without any additional explanation.

E.2 LLM prompt to Annotate NLP Topics

LLM prompt for annotating NLP topics

You are an expert at analyzing NLP research papers. Based on the title and abstract below, extract the following information:
NLP TOPICS
Identify the specific NLP topic (category) addressed from the following list:
- *# List of NLP TOPICS that shown in Table 7 goes here ...*
Output format:
{ "nlp-tasks": ["task1", "task2", ...] }
Title: {title}
Abstract: {abstract}"""
Title: [PAPER-TITLE]
Abstract: [PAPER-ABSTRACT]
Answer:

E.3 LLM prompt to Annotate NLP Tasks

LLM prompt for annotating NLP tasks

You are an expert at analyzing NLP research papers. Based on the title and abstract below, extract the following information:
NLP TASKS
Identify the specific NLP tasks addressed from the following list:
- *# List of NLP TASKS goes here ...*
Output format:
{ "nlp-tasks": ["task1", "task2", ...] }
Title: {title}
Abstract: {abstract}"""
Title: [PAPER-TITLE]
Abstract: [PAPER-ABSTRACT]
Answer:

E.4 LLM prompt for NLP Paper Contribution Annotations

LLM prompt for annotating abstracts for contribution and NLP tasks

You are an expert at analyzing NLP research papers. Based on the title and abstract below, extract the following information:

For each sentence in the abstract that describes a contribution, classify it into the appropriate category and extract the exact sentence. Group sentences by contribution type:
artifact-method: Introduces or proposes a new or novel NLP methodological approach or model to solve NLP tasks. [*example here*]
artifact-dataset: Introduces a new NLP dataset (i.e., textual resources such as corpora or lexicons). [*example here*]
artifact-task: Introduces or proposes a new or novel NLP task formulation (i.e., well-defined NLP problem). [*example here*]
knowledge-method: Describes new knowledge or insights about NLP models or methods. [*example here*]
knowledge-dataset: Describes new knowledge about datasets, such as their new properties or characteristics. [*example here*]
knowledge-task: Knowledge/insights about existing tasks or problem domains. [*example here*]
knowledge-language: Presents new knowledge about language, such as a new property or characteristic of language. [*example here*]
knowledge-people: Presents new knowledge about people, humankind, society, or communities. [*example here*]

OUTPUT FORMAT:

{{ "contribution": {{ "artifact-task": ["sentence 1", "sentence 2"],"artifact-method": ["sentence 1"], "artifact-dataset": [], "knowledge-method": ["sentence 1", "sentence 2"],"knowledge-dataset": ["sentence 1"], "knowledge-task": [], "knowledge-language": [], "knowledge-people": [] }},
"nlp-tasks": ["task1", "task2", ...]
}}
INSTRUCTIONS:
- A sentence might be assigned to more than one contribution category
- Multiple sentences can belong to the same category
Title: {title}
Abstract: {abstract}"""
Abstract: [PAPER-ABSTRACT]
Answer:}

F List of Support Contributor Organizations

Table 5 shows all the lists of organization names that support the papers.

Organization	# Paper(s)	Organization	# Paper(s)
National Science Foundation	52	European Regional Development Fund	2
Google	47	Stanford University	2
National Research Foundation of South Africa	42	Howard University	2
Lacuna Fund	25	Carnegie Corporation of New York	2
OpenAI	21	Masaryk University	2
European Union	19	King Fahd University of Petroleum and Minerals	2
Canada Research Chairs	17	Czech Science Foundation	2
German Federal Ministry of Education and Research	16	Ministry of Education Youth and Sports of the Czech Republic	2
Deutsche Forschungsgemeinschaft	14	Allen Institute for Artificial Intelligence	2
Natural Sciences and Engineering Research Council of Canada	14	Federal Government of Nigeria	2
Social Sciences and Humanities Research Council of Canada	13	U.S. Army Research Office	2
Canadian Foundation for Innovation	12	Qatar Computing Research Institute	2
Science Foundation Ireland	12	Clarkson Open Source Institute	2
Digital Research Alliance of Canada	11	Swiss National Science Foundation	2
UBC Advanced Research Computing-Sockeye	11	Makerere University	2
Microsoft	10	University of Cambridge	2
NVIDIA	10	University of Oslo	2
European Research Council	9	Samsung	2
Intelligence Advanced Research Projects Activity	9	Research Council of Norway	2
Defense Advanced Research Projects Activity	9	University of Malawi	2
University of Cape Town	8	Covenant University	2
Qatar Foundation	8	University of Texas	2
French Research Agency	7	Indiana University	2
University of Hamburg	7	University of Sheffield	2
New York University Abu Dhabi	7	Qatar National Research Fund	2
Meta	6	American University of Beirut	2
Amazon	6	Birzeit University	2
Bill & Melinda Gates Foundation	6	National Endowment for the Humanities	2
Oracle	5	Israeli Science Foundation	2
University of Bremen	5	German Ministry of Education and Science	2
Johns Hopkins University	5	National Key Research and Development Program of China	2
Masakhane	5	Princeton Laboratory for Artificial Intelligence	2
University of Gothenburg	5	Nigerian Artificial Intelligence Research Scheme	1
South African Centre for Digital Language Resources	5	Label Studio	1
Swedish Research Council	5	MUR FARE 2020	1
Advanced Micro Devices	5	Trelis Research	1
University of Edinburgh	4	UK Research and Innovation	1
African Institute for Mathematical Sciences	4	Luxembourg Ministry of Foreign and European Affairs	1
Foundation for Science and Technology, PT	4	CONAHCYT	1
South African Department of Arts and Culture	4	AI for Good	1
Intron Health	4	TUM School of Social Sciences and Technology	1
University of Porto	4	Hasso Plattner Institute for Digital Engineering	1
Addis Ababa University	4	CNPq,FAPEMIG,FAPESP, Brazil	1
German Academic Exchange Service	4	University of Information Technology,Vietnam	1
National Natural Science Foundation of China	4	Alberta Innovates	1
Stony Brook University	3	Ministry of Science Technology and Innovation, Brazil	1
University of Oregon	3	Stanford UIT	1
Toloka	3	ABSA Data Science	1
Huawei	3	KAIST-NAVER AI Center	1
Ghent University	3	Korea Government MSIT	1
TensorFlow	3	GENCI-IDRIS	1
Jimma University	3	ESPERANTO	1
George Mason University	3	Helmholtz Programme-oriented Funding	1
Ashesi University	3	GRAIN	1
Bahir Dar University	3	Science Foundation Ireland Digital Reality	1
Japan Society for the Promotion of Science	3	Canada CIFAR AI Chair	1
Waterloo AI Institute	3	German Research Center for Artificial Intelligence	1
Microsoft Research	3	University of Hull	1
South African National Research Foundation	3	TRADEF	1
U.S. Army Research Laboratory	3	Stony Brook Research Computing	1
Mila - Quebec AI Institute	3	University of Murcia	1
National Research Foundation of Korea	3		
Compute Canada	3		

Table 5: List of organizations with the number of papers mentioned by

G Partial List of Affiliation Organizations

Table 6 shows all the lists of institution names that support the papers.

Institutions	# Paper(s)	Organization	# Paper(s)
University of Pretoria	53	University of the Witwatersrand	10
Carnegie Mellon University	44	Nnamdi Azikiwe University	10
Saarland University	38	Dublin City University	10
MBZUAI	33	Mila Quebec AI Institute	10
Stellenbosch University	33	University of Waterloo	10
Masakhane	31	University of Stuttgart	10
University of Cape Town	27	Qatar University	10
Addis Ababa University	27	University of Maryland	9
Columbia University	25	University of Sheffield	9
Johns Hopkins University	23	Massachusetts Institute of Technology	9
North-West University	23	University of Illinois	9
Bahir Dar University	22	Georgetown University	9
New York University	22	Kaduna State University	9
Qatar Computing Research Institute	22	LMU Munich	9
The University of British Columbia	21	University of Ibadan	9
University College London	20	African Institute for Mathematical Sciences	9
University of Edinburgh	20	Stony Brook University	9
New York University Abu Dhabi	19	University of Porto	9
McGill University	19	University of Bergen	9
University of Washington	19	University of Zimbabwe	9
Imperial College London	18	Cairo University	9
Google	18	Bayero University Kano	8
Uppsala University	17	Cardiff University	8
Makerere University	17	Montclair State University	8
Technical University of Munich	16	University of British Columbia	8
Indiana University	16	American University of Beirut	8
University of Cambridge	16	Instituto Politécnico Nacional	7
University of Pennsylvania	16	University of Minnesota	7
University of Sfax	16	Amazon	7
Stanford University	15	Luleå University of Technology	7
Lelapa AI	14	Invertible AI	7
George Mason University	14	University of Oslo	7
Lancaster University	13	Idiap Research Institute	7
University of Hamburg	13	University of Southern California	7
University of Leeds	13	The George Washington University	7
Alexandria University	13	University of Nairobi	7
Hamad Bin Khalifa University	13	Arba Minch University	7
University of Gothenburg	13	University of the Western Cape	7
University of South Africa	13	Maseno University	6
Northeastern University	12	SADiLaR	6
Universität Hamburg	12	University of Amsterdam	6
University of California	12	Avignon University	6
Ahmadu Bello University	12	Meta	6
Bayero University	12	Inria	6
HausaNLP	11	Clarkson University	6
Wollo University	11	University of Melbourne	6
Rochester Institute of Technology	11	University of Helsinki	6
Microsoft	11	University of Limpopo	6
Ghent University	11	LIMSI-CNRS	6
Obafemi Awolowo University	11	TshwaneDJe HLT	6
University of Toronto	10	University of Birmingham	6

Table 6: List of institutions/organizations affiliated with AfricaNLP papers with the number of papers mentioned. There are other more than 100 affiliations mentioned less than 6 times.

H Full List of NLP Topics

Table 7 shows all the lists of NLP topics with the number of papers.

Task name	# Paper(s)	Task name	# Paper(s)
Low-resource language processing	896	Text generation	23
Multilinguality and language diversity	638	Educational NLP	22
Resources and evaluation	541	Conversational AI and chatbots	21
Phonology, morphology and word segmentation	291	Natural language understanding	20
Machine translation	276	Discourse and pragmatics	19
Speech recognition, text-to-speech and spoken language understanding	273	Few-shot and zero-shot learning	17
Machine learning for NLP	247	Sign language processing	16
Sentiment analysis, stylistic analysis, and argument mining	209	Benchmarking and testing	15
Syntax: tagging, chunking and parsing	145	Computational social science and cultural analytics	13
Dialect and language identification	108	Interpretability and analysis of models for NLP	12
Representation learning	102	Computational linguistics theory	11
Semantics: lexical and sentence-level	89	Human-centered NLP	10
Transfer learning and domain adaptation	86	Legal and biomedical NLP	9
Code-switching and multilingual processing	75	Text summarization	8
Information retrieval and text mining	67	Dialogue and interactive systems	8
Language modeling	60	Reasoning	6
Information extraction	59	Knowledge graph construction and reasoning	4
Social media and web text analysis	57	Privacy and security in NLP	4
Linguistic theories, cognitive modeling and psycholinguistics	53	Vision-language tasks	4
Efficient/low-resource methods for NLP	51	Document understanding and layout analysis	2
Ethics, bias, and fairness	36	Creative writing and storytelling	2
Multimodality and language grounding to vision, robotics and beyond	30	Text-to-code generation	1
NLP applications	28	Paraphrasing and text rewriting	1
Error correction	25	Personality and psychological profiling from text	1
Question answering	24	Temporal information processing	1

Table 7: NLP task distribution by number of publications/papers

I NLP tasks and progresses

Table 10 shows lists of NLP tasks progress across years.

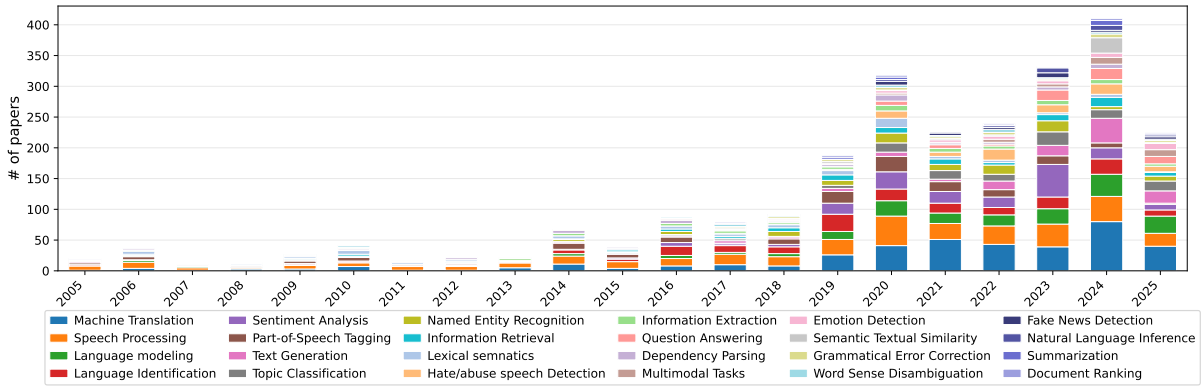

Figure 10: AfricaNLP Downstream NLP task progresses categories, (a paper may appear in multiple categories). The year 2025 is cut off at 07 (July month). The NLP topics are sorted in descending order based on the overall total frequency of papers.

J Human Annotation Interface

Figure 11 shows the human annotation interface for identifying AfricaNLP papers, contributions, and NLP topic annotations

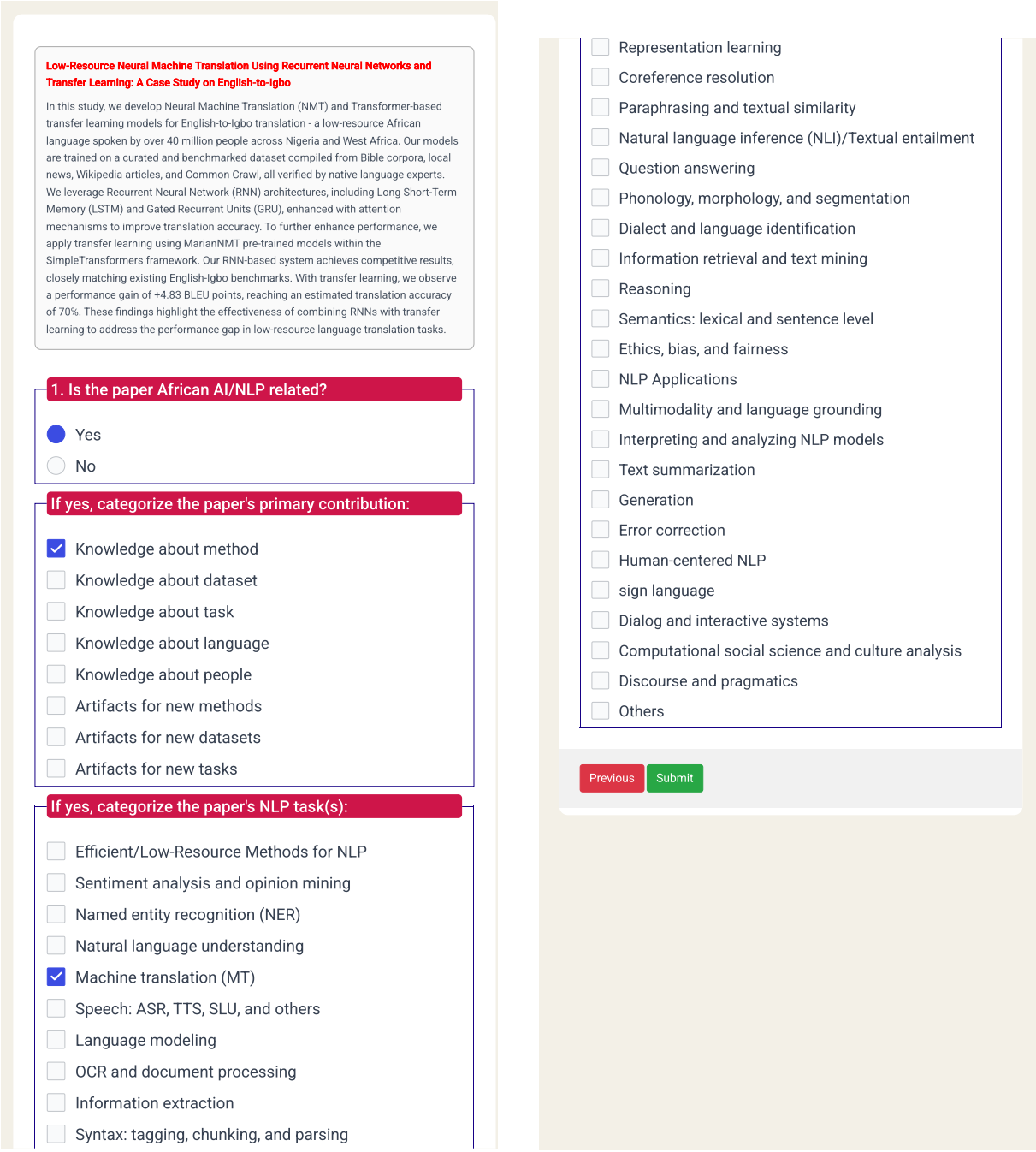

Figure 11: NLP paper annotation tool interface screenshot, customized version of POTATO annotation tool (Pei et al., 2022).

K Contribution Classifier Training Details

- The PLMs are loaded from the Transformer on Huggingface
- BERT = google-bert/bert-base-uncased
- SciBERT = allenai/scibert_scivocab_uncased

- AfroXLMR = Davlan/afro-xlmr-large-76L
- GPT[17] is Gpt-4.1-mini-2025-04-14 (OpenAI, 2025) and Gemini-2.5-flash (Google, 2025)
- Traing details for all PLMs are batch size = 16, epoch = 3, zero shot for LLMs

L Affiliation institutions networks of AfricaNLP

Figure 12 shows the networks of Affiliation institutions, organizations, and individual researchers from AfricaNLP papers. The visualization is done with the VOSviewer[18] bibliometric analysis tool.

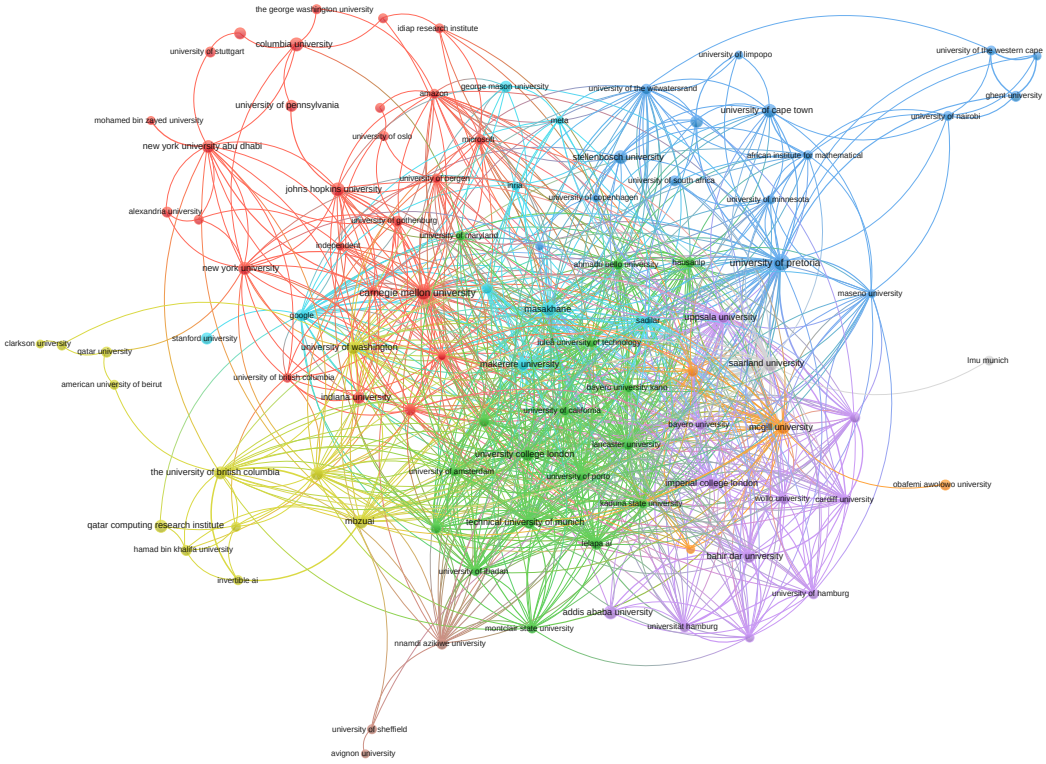

Figure 12: **Affiliation networks**: The networks are generated using VOSviewer with a minimum of five documents per author for better readability. The size of the nodes (circles) and edges (lines) corresponds to the frequency of affiliations and links, respectively. In the network visualization, each node represents an affiliation, and a cluster with the same color indicates affiliations within a close network or collaboration. Connected authors account for 123 out of more than 200 affiliations, grouped into 10 clusters. The figure can be enlarged for better readability.

[17] https://openai.com/index/gpt-4-1/
[18] https://www.vosviewer.com/

M Author Networks of AfricaNLP Papers

Figure 13 shows the networks of authors' co-authorship from AfricaNLP papers.

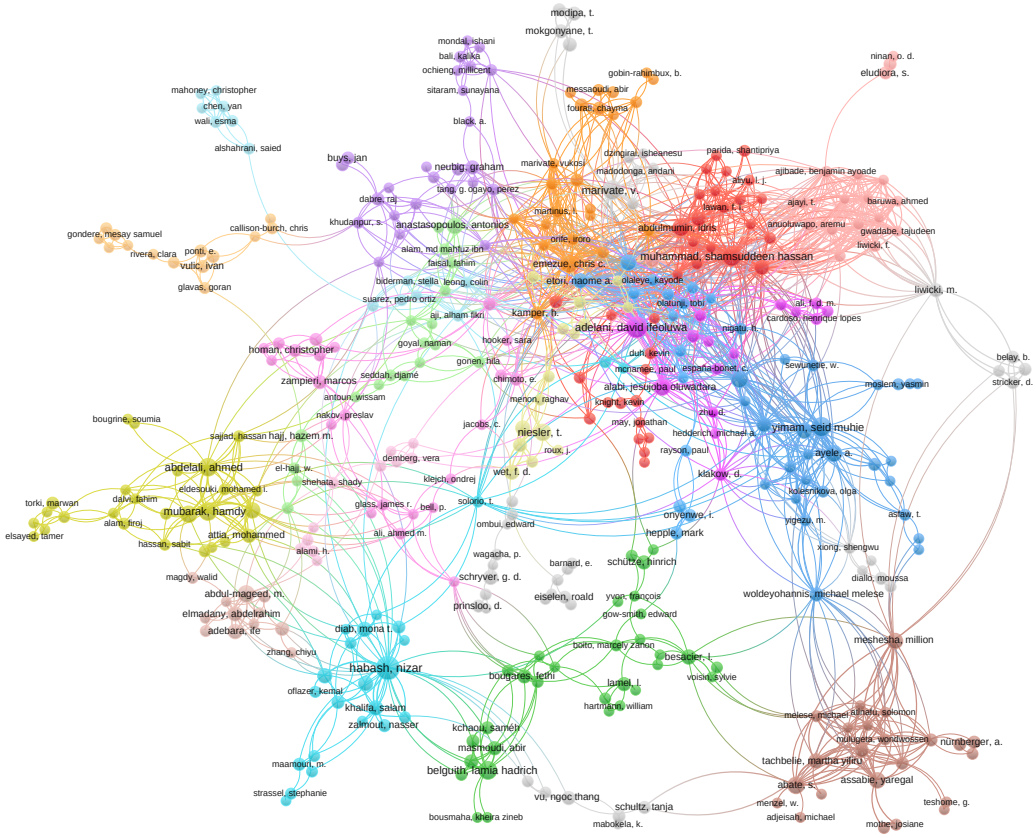

Figure 13: **Author co-authorship network analysis** generated using VOSviewer with a minimum number of papers of an author equal to five for better readability. Connected authors comprise 173 out of 4,901 authors, grouped into 6 clusters. In the network visualization, each node represents an author, and a cluster with the same color indicates authors with a close relationship. The size of the node shows the frequency of occurrence contributing to research in the past 20 years. The figure can be enlarged for better readability.

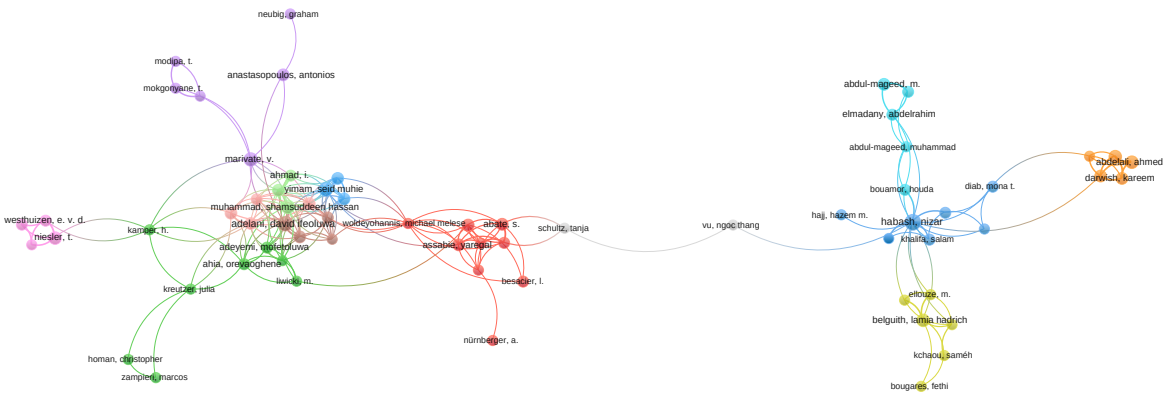

Figure 14: **Author co-authorship network analysis** with authors appearing in a minimum of 10 papers, 66 authors from 4901 authors, 12 clusters. The figure can be enlarged for better readability.

N Search Keyword Networks

Figure 15 shows the networks of keyword co-occurrence from AfricaNLP papers. Out of the set of more than 600 country names (including adjectival forms like African), language names, and initiative organization community names, only 184 keywords have at least one paper.

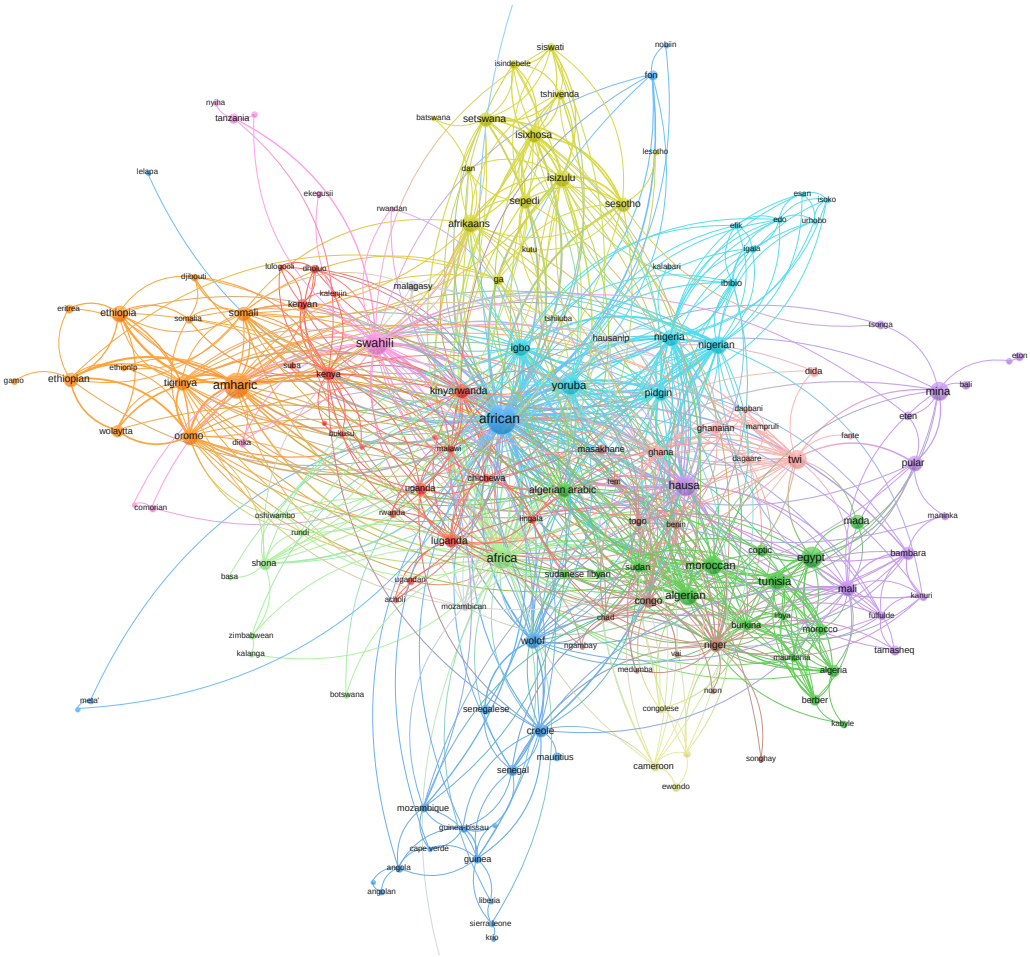

Figure 15: **Keyword co-occurrence network analysis** generated using VOSviewer with a minimum occurrence of five for better readability. Connected keywords comprise 121 out of 184 keywords, grouped into 11 clusters. In the network visualization, each node represents a keyword, and a cluster with the same color indicates keywords they are appear together in the paper. The size of the node indicates the frequency of occurrence in the AfricaNLP papers. The figure can be enlarged for better readability.

O Examples of Paper annotation disagreement between LLMs

Table 8 shows examples of papers that the LLMs disagree on when classifying as AfricaNLP NLP papers.

Paper	Title	GPT	Gemini
Hossain et al. (2025)	Enhancing Dialectal Arabic Intent Detection through Cross-Dialect Multilingual Input Augmentation	✗	✓
Khered et al. (2025)	Dial2MSA-Verified: A Multi-Dialect Arabic Social Media Dataset for Neural Machine Translation to Modern Standard Arabic	✗	✓
Miyagawa (2025)	RAG-Enhanced Neural Machine Translation of Ancient Egyptian Text: A Case Study of THOTH AI	✗	✓
Malaysha et al. (2024)	NLU-STR at SemEval-2024 Task 1: Generative-based Augmentation and Encoder-based Scoring for Semantic Textual Relatedness	✗	✓
Alwajih et al. (2024)	Dallah: A Dialect-Aware Multimodal Large Language Model for Arabic	✗	✓
(Ait ElFqih and Monti, 2024)	Large Language Models as Legal Translators of Arabic Legislation: Do ChatGPT and Gemini Care for Context and Terminology?	✗	✓
Hadj Mohamed et al. (2023)	AlphaMWE-Arabic: Arabic Edition of Multilingual Parallel Corpora with Multiword Expression Annotations	✗	✓
(Sahyoun and Shehata, 2023)	AraDiaWER: An Explainable Metric For Dialectical Arabic ASR	✗	✓
Tirosh-Becker et al. (2022)	Part-of-Speech and Morphological Tagging of Algerian Judeo-Arabic	✗	✓
(Hassib et al., 2022)	AraDepSu: Detecting Depression and Suicidal Ideation in Arabic Tweets Using Transformers	✗	✓
Mubarak et al. (2021)	Arabic Offensive Language on Twitter: Analysis and Experiments	✗	✓
Lincke et al. (2019)	Optical Character Recognition for Coptic fonts: A multi-source approach for scholarly editions	✗	✓
Feleke (2017)	The similarity and Mutual Intelligibility between Amharic and Tigrigna Varieties	✗	✓
Aftiss et al. (2025)	Empirical Evaluation of Pre-trained Language Models for Summarizing Moroccan Darija News Articles	✓	✗
Sang et al. (2025)	Federated Meta-Learning for Low-Resource Translation of Kirundi	✓	✗
Tudor and Tashu (2025)	Mapping Cross-Lingual Sentence Representations for Low-Resource Language Pairs Using Pre-trained Language Models	✓	✗
Bamutura et al. (2020)	Towards Computational Resource Grammars for Runyankore and Rukiga	✓	✗
Yuan et al. (2020)	Interactive Refinement of Cross-Lingual Word Embeddings	✓	✗

Table 8: LLMs answer for the question *Is this African AI/NLP paper?* using the title and abstract of the paper. Note that the prompt includes an abstract of the paper. For simplicity, we provide here the title of the paper only. ✗ is not AfricaNLP paper response and ✓ is Yes reprocess. These papers are part of our collection, which is handled by human evaluations after the LLM predictions.

P Number of publication across languages versus number of (both L1 and L2) speakers of the language

How does the number of language speakers correlate with the availability of language resources/publications?. The 17 top languages, along with their number of speakers and publications from 1902 papers, are presented in Table 9. The analysis reveals a weak correlation between the number of speakers and research attention across African languages. While Swahili and Nigerian Pidgin have the largest speaker bases (150M and 121M, respectively), they are comparatively underrepresented in research publications, particularly Nigerian Pidgin with only 25 papers. In contrast, Amharic, with 57M speakers, shows a disproportionately high research presence (204 papers), suggesting that academic focus and resource availability outweigh population size in influencing research activity. East African languages such as Amharic and Swahili receive more attention overall compared to West African and Southern African languages, despite the latter regions hosting several high-population languages like Hausa and Igbo. Interestingly, languages with relatively smaller populations, such as Twi and Afrikaans, also demonstrate notable research engagement, underscoring the role of institutional support, data accessibility, and linguistic policy rather than speaker population in driving research output. There is little to no direct correlation between the number of language speakers and the availability of language resources or publications, as research attention is more strongly influenced by institutional support, data availability, and linguistic policy than by speaker population size.

Language name	Region (of Africa)	# of Speakers	# of papers ↓
Amharic (amh)	East	57M	204
Swahili (swa)	East/Central	150M	140
Twi (twi)	West	9M	105
Yoruba (yor)	West	46M	84
Hausa (hau)	West	77M	73
Afrikaans (afr)	Southern	7M	47
Oromo (orm)	East	37M	42
Isizulu (zul)	Southern	27M	40
Igbo (ibo)	West	31M	39
Wolof (wol)	West	5M	32
Kinyarwanda (kin)	East	10M	29
Sesotho (sot)	Southern	6M	28
Isixhosa (xho)	Southern	19M	28
Nigerian Pidgin (pcm)	West	121M	25
Sepedi (nso)	Northeastern	6M	25
Tigrinya (tir)	East	9M	24
Setswana (tsn)	Southern	14M	24

Table 9: Languages number of L1 and L2 speakers vs number papers that has > 20 papers. Note that this frequency reflects the number of language keyword occurrences in the titles and abstracts of papers; some papers may involve a language without explicitly mentioning its name in the title or abstract. Number speakers are sourced from Ojo et al. (2025).

Q AfricaNLP Research Explorer Tool

Figure 16 shows the underconstruction AfricaNLP research explorer tool. It features paper searching capabilities, including the ability to search for paper titles and abstracts, as well as categories by language, NLP topic, and NLP task. The tool link will be provided during the camera-ready submission of the paper.

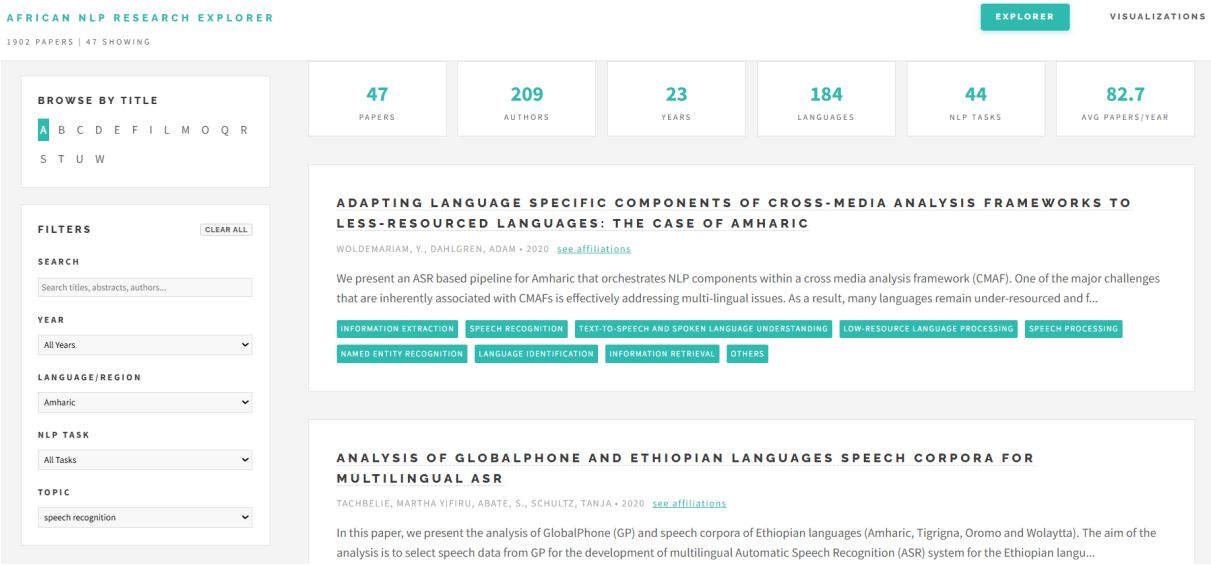

Figure 16: AfricaNLP research explorer tool screenshot (under improvement) and for Amharic language - speech recognition papers as a demonstration