

# Adja-French Parallel Corpus: A New Resource for Machine Translation of a West African Under-Resourced Language

Josue Godeme, Rolando Coto-Solano

Dartmouth College, Dartmouth College,  
josue.f.godeme.26@dartmouth.edu,rolando.a.coto.solano@dartmouth.edu,

## Abstract

We present the first parallel text corpus for Adja machine translation, an under-resourced Gbe language spoken by approximately 1,000,000 people in Benin and Togo. The corpus contains 10,000 French-Adja sentence pairs, providing a foundation for machine translation research. We establish baseline results using fine-tuned NLLB and ByT5 models, achieving a chrF++ of 28 in the French→Adja direction, and up to a chrF++ of 34 in the Adja→French direction. This work represents the first public machine translation resource for Adja. It provides benchmarks for future studies on this under-resourced West African language. The dataset is available at <https://huggingface.co/datasets/JosueG/french-adja-parallel-corpus>.

**Keywords:** Adja language, parallel corpus, machine translation, low-resource NLP, West African languages, Gbe languages

## 1. Introduction

Africa is home to over 2,000 languages (Eberhard et al., 2025b), representing one of the world’s richest areas in linguistic diversity, with around one third of the world’s languages. Among these is Adja, a Gbe language spoken by approximately 1,000,000 people primarily in southern Benin and southeastern Togo (Eberhard et al., 2025a). Despite its significant speaker population, Adja remains severely under-resourced in terms of natural language processing tools and datasets. To the best of our knowledge, there exist no publicly available parallel corpora, machine translation systems, or labeled computational resources for Adja.

The lack of language technology for Adja contributes to the digital marginalization of its speakers and limits access to information, education, and services in their language. This digital divide affects not only individual speakers but also the preservation and vitality of the language itself in an increasingly connected world (Kornai, 2013). Developing resources for under-resourced languages is crucial for linguistic diversity, digital inclusion, and equitable access to technology. (Kornai, 2013)

Machine translation (MT) systems, in particular, have the potential to bridge language barriers. However, the development of such systems requires parallel text corpora, which have been absent for Adja until now.

This paper makes the following contributions:

- We present a French-Adja parallel corpus containing 10,000 sentence pairs, establishing a foundation for computational work on Adja
- We provide baseline machine translation results using four approaches, spanning both

statistical and neural MT. Moses SMT (Koehn et al., 2007): a phrase-based statistical MT system using Moses as a non-neural baseline. We report median BLEU and chrF++ scores. (NLLB-200 (Team et al., 2022), mBART-50 (Liu et al., 2020), and ByT5 (Xue et al., 2022)), demonstrating both the utility of the corpus and the challenges inherent in extremely low-resource MT

- We establish evaluation benchmarks for future Adja machine translation research
- We release the corpus publicly under a non-commercial license to enable further academic research

## 2. Related Work

### 2.1. African Language Technology Initiatives

The movement to advance natural language processing for African languages has gained momentum in recent years through grassroots organizations and research communities. Masakhane (Orife et al., 2020), a pan-African research collective dedicated to strengthening NLP for African languages, has coordinated efforts across the continent to build MT systems for previously under-resourced languages (Wang et al., 2024; Adelani et al., 2023, 2022). Regional initiatives such as Ghana NLP (Ghana Natural Language Processing (NLP), 2020) and Lelapa AI (Lelapa AI) have focused on specific language groups, producing datasets and models for languages including Twi, Yoruba, and isiZulu.

Despite this progress, significant gaps remain. Of Africa’s 2,000 languages (Eberhard et al.,

2025b), only a small fraction have computational resources. Many languages, including Adja, remain completely absent from these initiatives.

## 2.2. Low-Resource Machine Translation

Recent advances in neural machine translation have demonstrated the viability of building systems for extremely low-resource languages. Multilingual pre-trained models such as NLLB-200 (Team et al., 2022) and mBART (Liu et al., 2020) leverage cross-lingual transfer learning to improve performance on languages with limited parallel data. These models have shown promising results on African languages including Fon (Emezue and Dossou, 2020), and Nko (Doubouya et al., 2023), even with training corpora of fewer than 10,000 sentence pairs.

Character-level and byte-level models such as ByT5 (Xue et al., 2022) have proven particularly effective for morphologically rich and agglutinative languages, as they bypass the need for language-specific tokenization. This approach has strong results on languages with complex scripts and limited preprocessing resources (Clark et al., 2022).

## 2.3. Parallel Corpus Creation for Low-Resource Languages

Building parallel corpora for extremely low-resource languages presents significant challenges. Common approaches include mining web data, translating existing benchmark datasets like FLORES-200 (Team et al., 2022), and crowdsourcing through platforms such as Tatoeba (Tatoeba Community), and Mozilla Common Voice (Ardila et al., 2020). While web mining can yield large-scale data, it is often unavailable for languages with minimal online presence, like Adja, a primarily spoken language. Benchmark translation projects produce high-quality data but require significant coordination and funding. Tatoeba offers a middle ground, providing community-contributed translations that, while not domain-specific, establish a baseline for MT development.

## 2.4. The Gbe Language Family and Adja

Adja belongs to the Gbe language cluster, which includes Fon, Ewe, Gen, and other closely related languages spoken across Benin, Togo, and Ghana (Kluge, 2005; Tompkins and Kluge, 2002). Concurrent with our work, Justin et al. (2025) introduced Eyaa-Tom, a multi-language speech dataset for Togolese languages that includes a small amount of Adja audio data for ASR and language identification tasks. However, no parallel text corpus for Adja machine translation existed prior to our work. While some computational work has been initiated

for Fon (Emezue and Dossou, 2020), Adja has received no prior attention in the NLP literature. To the best of our knowledge, this work represents the first published MT computational resource for Adja.

# 3. Corpus Creation

## 3.1. Data Collection

**Source Selection.** We selected 10,000 French sentences from Tatoeba (Tatoeba Community)<sup>1</sup>, a collaborative platform for collecting translated sentences contributed by people worldwide. Sentences were selected through uniform random sampling from the full set of available French-language entries on the platform, without filtering by topic, length, or domain. While Tatoeba provides diverse sentence structures and vocabulary, the sentences are not domain-specific, there is no quality control on the submitted sentences, limitations we address in Section 5.

**Translation Process.** We assembled a team of five native Adja speakers in the Couffo region of Benin, a region where Adja is predominantly spoken. The team included two translators, a government-accredited Adja language instructor and an experienced fluent speaker, and three transcribers, all native Adja speakers.

The translation process proceeded as follows over a six-month period:

1. French sentences were read aloud and their meanings discussed to ensure accurate comprehension
2. Translators produced Adja translations orally, maintaining natural language use
3. Transcribers recorded the Adja translations in writing
4. Translations were typed and formatted for computational use

This collaborative, manual process ensured high translation quality through direct involvement of native speakers with strong literacy in both French and Adja. While labor-intensive, this approach was necessitated by limited internet connectivity and varying levels of digital literacy in the region, common challenges in low-resource language documentation. Future work could explore digital tools to streamline data collection, though infrastructure barriers remain significant.

**Data Processing and Quality Control.** Following digitization, sentence pairs were processed through a dedicated cleaning pipeline to standardize encoding, normalize spacing, and ensure punctuation consistency between the French source and

---

<sup>1</sup><https://tatoeba.org>

Adja target. The pipeline applied the following steps in order:

1. **Unicode normalization (NFKC).** All text was normalized to NFKC form, resolving encoding inconsistencies such as composed vs. decomposed diacritics that arise when typing Adja characters on different keyboards.
2. **Spacing normalization.** Double spaces were collapsed; spaces before sentence-final punctuation (. , ! ? ; :) were removed.
3. **Non-standard character normalization.** The Latin retroflex click (Unicode U+01C3), occasionally introduced by keyboards as a visually similar substitute, was replaced with standard ASCII !.
4. **Bidirectional punctuation consistency.** For each pair, sentence-final punctuation was checked in both directions: if the French sentence ends with ? or !, the Adja sentence is required to match; mismatches detectable from interrogative function words are corrected automatically, and ambiguous cases are flagged for human review.
5. **Quotation mark matching.** If the French sentence contains guillemets (« ») or standard double quotes, matching quotation marks are added to the Adja sentence.

The use of two independent translators working from the same source material provided implicit quality validation, as both translators shared a common linguistic baseline and produced consistent translations. The cleaned corpus is referred to as the v2-normalized dataset and is used in all experiments reported in this paper.

### 3.2. Data Splits

We create two complementary splits from the same 10,000 sentence pairs.

**Random split.** Pairs are shuffled uniformly at random (seed 42) and sliced 80/10/10, yielding 8,000 train / 1,000 dev / 1,000 test pairs.

**Length-stratified split.** To ensure train, dev, and test sets share the same sentence-length distribution, we first bin all pairs by the length of the French source sentence: *short* ( $\leq 8$  words), *medium* (9–16 words), and *long* ( $> 16$  words). We then sample 80/10/10 proportionally within each bin and concatenate the results, yielding 7,999 train / 999 dev / 1,002 test pairs. Table 1 reports the resulting distributions.

Reporting results on both splits allows us to assess whether length distribution in the test set affects evaluation scores. The test sets are held out for final evaluation and will serve as standardized benchmarks for future Adja MT research.

| Split | Total | Short<br>( $\leq 8$ ) | Medium<br>(9–16) | Long<br>( $> 16$ ) |
|-------|-------|-----------------------|------------------|--------------------|
| Train | 7,999 | 6,456                 | 1,481            | 62                 |
| Dev   | 999   | 807                   | 185              | 7                  |
| Test  | 1,002 | 807                   | 186              | 9                  |

Table 1: Sentence count by length stratum for the stratified split. All three sets share a mean French source length of 6.4 words.

| Metric                         | French | Adja   |
|--------------------------------|--------|--------|
| <i>Corpus Statistics</i>       |        |        |
| Total sentence pairs           | 10,000 |        |
| Total tokens                   | 66,245 | 66,661 |
| Vocabulary size                | 11,385 | 12,560 |
| Type-Token Ratio               | 0.1719 | 0.1884 |
| HAPAX count                    | 7,196  | 8,318  |
| <i>Sentence Length (words)</i> |        |        |
| Mean                           | 6.62   | 6.67   |
| Median                         | 6.0    | 6.0    |
| Std. deviation                 | 2.92   | 3.18   |
| Min – Max                      | 1 – 63 | 1 – 68 |
| <i>Data Splits</i>             |        |        |
| Train sentences                | 8,000  |        |
| Dev sentences                  | 1,000  |        |
| Test sentences                 | 1,000  |        |

Table 2: Comprehensive statistics for the Adja-French parallel corpus.

### 3.3. Corpus Statistics

Table 2 presents comprehensive statistics for our parallel corpus of 10,000 Adja-French sentence pairs. The corpus exhibits several characteristics that are particularly favorable for neural machine translation in a low-resource setting. Table 3 shows representative sentence pairs from the corpus.

The statistical analysis reveals four key properties of the corpus:

**Balanced sentence length distribution.** Both languages demonstrate median sentence lengths of 6 words with moderate standard deviations (French:  $\sigma = 2.92$ , Adja:  $\sigma = 3.18$ ), indicating a well-balanced mix of sentence complexities. The corpus comprises approximately 39% short sentences ( $\leq 5$  words), 59% medium-length sentences (6–15 words), and 1.5% long sentences ( $> 15$  words). This distribution suggests exposure to a variety of syntactic structures, from simple clausal constructions in short sentences to more complex sentences with subordination and modification, while avoiding the data sparsity issues associated with very long sentences in low-resource contexts.

**Lexical richness and morphological complexity.** Adja demonstrates notably higher lexical diversity than French, with a Type-Token Ratio of 0.188 compared to French’s 0.172 (9.5% higher). This difference, combined with Adja’s larger vocabulary

size (12,560 vs 11,385 unique tokens despite comparable token counts), reflects the morphological richness characteristic of Gbe languages. The high HAPAX counts: 66% for Adja and 63% for French are typical of low-resource corpora and suggest opportunities for future expansion.

**Comparable corpus sizes.** The near-identical total token counts (French: 66,245 tokens; Adja: 66,661 tokens) ensure balanced bidirectional training, preventing the model bias that can arise from asymmetric parallel corpora.

**Adequate vocabulary coverage.** With over 11,000 unique tokens per language, the corpus captures substantial lexical variation suitable for initial neural MT experiments, though the high proportion of singleton words indicates potential benefits from additional parallel data.

| French                              | Adja                            |
|-------------------------------------|---------------------------------|
| [Je sue tous les jours.]            | [ŋ kɔ nɔ ade tɛgbɛ ɛ.]          |
| [Je pense que tu devrais voir ça.]  | [ŋ bumɔ wɔ a kpɛ alo wɔ a nya.] |
| [Tu te perds.]                      | [E búbu ɔ deki.]                |
| [Cesse de rêver et ouvre les yeux.] | [Mi edrɔ kukú ahùn ŋkuvɪ wo.]   |
| [Ce n'était pas moi. C'était Tom.]  | [Enyè yɔ go ! Tom yɔ.]          |

Table 3: Representative sentence pairs from the corpus.

## 4. Baseline Experiments

We establish baseline results using three state-of-the-art approaches and one classical approach:

**Moses Statistical MT** (Koehn et al., 2007): We trained a Moses model because older, statistical methods can still be the best way to train very small datasets. We trained for 5 iterations.

**NLLB-200** (Team et al., 2022): We fine-tune the `nllb-200-distilled-600M` model, initialized from the Ewe (`ewe_Latn`) language token as a proxy for Adja, leveraging cross-lingual transfer from the closest available Gbe language. We apply early stopping on chrF (patience 10 evaluations) to prevent overfitting.

**mBART-50** (Liu et al., 2020): We fine-tune `mbart-large-50-many-to-many-mmt`, initializing from the French (`fr_XX`) source-language embedding as a starting point for the Adja target. We register a new `aj_Latn` language token and resize the token embeddings accordingly.

**ByT5** (Xue et al., 2022): We fine-tune `byt5-base`, a byte-level model that requires no language-specific tokenization, making it well-suited for a language with limited preprocessing resources.

All three models are optimized with Adafactor (learning rate  $1 \times 10^{-4}$ , no relative step), batch

size 16, and early stopping on validation chrF with patience 10. We train for up to 50 epochs per model. We run each configuration with 5 random seeds (42, 123, 456, 789, 1024) and report mean  $\pm$  standard deviation. Results on both the random and length-stratified splits are reported. Evaluation uses BLEU (Papineni et al., 2002) and chrF++ (Popović, 2017).

Table 4 presents our baseline results for both translation directions. The stratified ByT5 training had the best results in the Adja-to-French direction. In the French-to-Adja direction, stratified ByT5 had the best ChrF++, but the older Moses had the best BLEU. Adja is not a morphologically rich language, so it makes sense that BLEU n-grams would give Moses an advantage over newer models.

## 5. Discussion and Conclusion

We have presented the first publicly available parallel text corpus and machine translation baselines for Adja, an under-resourced West African language. Our 10,000-sentence corpus, collected through a six-month collaborative effort with native Adja speakers in Benin, provides a foundation for future research on Adja language technologies.

### 5.1. Future Directions

Several avenues exist for extending and improving this work:

**Domain-Specific Corpora.** We plan to develop targeted corpora in high-priority domains such as healthcare, agriculture, education, and government services. These domain-specific datasets would better serve practical translation needs in Benin and Togo.

**Need for ASR development** Adja is primarily a spoken language, and the most impactful language technologies for its speakers would likely be speech-based. The oral translation process we used in which translators produced Adja translations aloud before transcription was designed in part to preserve natural spoken Adja. ASR development for Adja would complement the MT work presented here, enabling end-to-end spoken language translation pipelines that could serve the many Adja speakers who communicate primarily in speech rather than writing. Machine Translation was only the first step.

**Data Availability.** The dataset is available at <https://huggingface.co/datasets/JosueG/french-adja-parallel-corpus> under a non-commercial license to protect against exploitative use while enabling academic research.

This work represents the first step in a longer journey toward comprehensive language technology support for Adja. We hope this resource catalyzes further research and demonstrates the feasibility

| Model                 | Split      | FR→ADJ      |                   | ADJ→FR            |                   |
|-----------------------|------------|-------------|-------------------|-------------------|-------------------|
|                       |            | BLEU        | chrF++            | BLEU              | chrF++            |
| (lr)3-4(lr)5-6        |            |             |                   |                   |                   |
| <i>Statistical MT</i> |            |             |                   |                   |                   |
| Moses SMT             | random     | <b>6.43</b> | 14.07             | 6.77              | 14.11             |
|                       | stratified | 5.84        | 13.87             | 6.87              | 14.03             |
| <i>Neural MT</i>      |            |             |                   |                   |                   |
| NLLB-600M             | random     | 4.5 ± 0.1   | 26.1 ± 0.3        | 11.8 ± 0.6        | 30.3 ± 0.5        |
|                       | stratified | 4.7 ± 0.2   | 27.1 ± 0.3        | 13.6 ± 0.6        | 31.8 ± 0.3        |
| mBART-50              | random     | 3.4 ± 0.3   | 22.6 ± 0.5        | 8.7 ± 0.1         | 26.1 ± 0.3        |
|                       | stratified | 3.6 ± 0.2   | 23.8 ± 0.3        | 9.5 ± 0.7         | 26.9 ± 0.6        |
| ByT5-base             | random     | 4.3 ± 0.4   | 26.1 ± 0.4        | 11.5 ± 0.8        | 31.2 ± 0.8        |
|                       | stratified | 4.9 ± 0.2   | <b>27.5 ± 0.4</b> | <b>14.3 ± 0.7</b> | <b>33.7 ± 0.7</b> |

Table 4: Baseline MT results on the random and length-stratified test sets (mean ± std over 5 seeds). Best per direction in **bold**.

of building NLP tools for truly under-resourced languages.

## 5.2. Limitations and Challenges

While this corpus represents an important first step, several limitations must be acknowledged:

**Potential Data Overlap.** NLLB-200 was pre-trained on a large multilingual corpus that might include Tatoeba data for many language pairs. However, since no Adja-language data existed in any public corpus at the time of NLLB’s training, our Adja test references are not present in NLLB’s training data. The French source sentences may appear in NLLB pre-training data (in French–English or other pairs), but this does not constitute target-side contamination for the French→Adja evaluation direction. ByT5 training dataset similarly contains no Adja data. We therefore consider the risk of meaningful data contamination to be negligible.

**Source Data Quality.** Our French sentences originate from Tatoeba, a crowdsourced platform where any user can contribute translations. Tatoeba does not verify whether contributors are native speakers, nor does it enforce quality control measures beyond community reporting. Consequently, the source French sentences may contain grammatical errors, unnatural phrasing, or non-standard usage that could propagate into the Adja translations.

**Lack of Domain Specificity.** The Tatoeba sentences cover random topics without thematic coherence, jumping from one subject to another without domain focus. This randomness limits the corpus’s utility for domain-specific applications such as healthcare, education, or government services: contexts where machine translation would be most valuable to Adja speakers.

**Dialectal and Regional Variation.** The French sentences reflect global French usage rather than the specific variety of French spoken in Benin and Togo. Regional expressions, cultural references, and locally relevant vocabulary may be underpre-

sented. Similarly, while our translators were based in the Kouffo region, Adja itself may have dialectal variation across different communities that our corpus does not capture.

**Sentence-Level Scope.** Our corpus consists of isolated sentences rather than connected discourse. This limits its applicability for tasks requiring discourse-level understanding, such as document translation, conversational systems, or context-dependent interpretation.

**Corpus Size.** While 10,000 sentences represents a significant achievement for a previously undocumented language, it remains small compared to the millions of sentence pairs typically used to train high-quality MT systems. The low BLEU scores we observe (Section 4) reflect this data scarcity.

**Translation Direction.** Since our corpus was produced by translating French sources into Adja, future work collecting naturally produced Adja text and translating it into French would complement this resource and enable evaluation on both original-language directions.

Despite these limitations, this corpus establishes a critical baseline. The challenges we identify point toward directions for future improvement rather than undermining the value of the current resource.

## Acknowledgements

We are deeply grateful to the Adja-speaking community members in the Couffo region of Benin who dedicated their time and expertise to translating and transcribing this corpus over a six-month period. Their commitment to documenting and advancing their language made this work possible. We also thank the broader Adja-speaking community for their support of this project.

## Ethical Considerations

This work involves a language community. We have worked with native Adja speakers in the data

collection process. The corpus will be released under a non-commercial license to protect against exploitative use while enabling academic research.

## 6. Bibliographical References

- 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, Mardiyah 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. 2023. [MasakhaNEWS: News Topic Classification for African languages](#). In *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. Association for Computational Linguistics.
- David Ifeoluwa Adelani, Graham Neubig, Sebastian Ruder, Shruti Rijhwani, Michael Beukman, Chester Palen-Michel, Constantine Lignos, Jesujoba O. Alabi, Shamsuddeen H. Muhammad, Peter Nabende, Cheikh M. Bamba Dione, Andiswa Bukula, Rooweither Mabuya, Bonaventure F. P. Dossou, Blessing Sibanda, Happy Buzaaba, Jonathan Mukiibi, Godson Kalipe, Derguene Mbaye, Amelia Taylor, Fatoumata Kabore, Chris Chinenye Emezue, Anuoluwapo Aremu, Perez Ogayo, Catherine Gitau, Edwin Munkoh-Buabeng, Victoire M. Koagne, Allahsera Auguste Tapo, Tebogo Macucwa, Vukosi Marivate, Elvis Mboning, Tajuddeen Gwadabe, Tosin Adewumi, Orevaoghene Ahia, Joyce Nakatumba-Nabende, Neo L. Mokono, Ignatius Ezeani, Chiamaka Chukwuneke, Mofetoluwa Adeyemi, Gilles Q. Hacheme, Idris Abdulmumin, Odunayo Ogundepo, Oreen Yousuf, Tatiana Moteu Ngoli, and Dietrich Klakow. 2022. [MasakhaNER 2.0: Africa-centric Transfer Learning for Named Entity Recognition](#).
- Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. [Common Voice: A Massively-Multilingual Speech Corpus](#). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4218–4222, Marseille, France. European Language Resources Association.
- Jonathan H. Clark, Dan Garrette, Iulia Turc, and John Wieting. 2022. [Canine: Pre-training an Efficient Tokenization-Free Encoder for Language Representation](#). *Transactions of the Association for Computational Linguistics*, 10:73–91.
- Moussa Koulako Bala Doumbouya, Baba Mamadi Diané, Solo Farabado Cissé, Djibrila Diané, Abdoulaye Sow, Séré Moussa Doumbouya, Daouda Bangoura, Fodé Moriba Bayo, Ibrahima Sory 2. Condé, Kalo Mory Diané, Chris Piech, and Christopher Manning. 2023. [Machine Translation for Nko: Tools, Corpora and Baseline Results](#).
- David M. Eberhard, Gary F. Simons, and Charles D. Fennig. 2025a. Aja. <https://www.ethnologue.com/language/ajg>. In *Ethnologue: Languages of the World*, Twenty-eighth edition. Dallas, Texas: SIL International.
- David M. Eberhard, Gary F. Simons, and Charles D. Fennig, editors. 2025b. *Ethnologue: Languages of the World*, twenty-eighth edition. SIL International, Dallas, Texas. Online version: <https://www.ethnologue.com/insights/continents-most-indigenous-languages/>.
- Chris Chinenye Emezue and Femi Prance Bonaventure Dossou. 2020. [FFR v1.1: Fon-French neural machine translation](#). In *Proceedings of the Fourth Widening Natural Language Processing Workshop*, pages 83–87, Seattle, USA. Association for Computational Linguistics.
- Ghana Natural Language Processing (NLP). 2020. Ghana nlp: An open source initiative for natural language processing of ghanaiian languages. <https://ghananlp.github.io/>. Accessed: 2025-10-22.

- Bakoubo Essowe Justin, Kodjo François Xegbe, Catherine Nana Nyaah Essuman, and Kossi Mawouéna Samuel Afola. 2025. YodiV3: NLP for Togolese languages with Eyaa-Tom dataset and the Lom metric. In *Proceedings of the Sixth Workshop on African Natural Language Processing (AfricaNLP 2025)*, pages 143–149. Association for Computational Linguistics.
- Angela Kluge. 2005. *A Synchronic Lexical Study of Gbe Language Varieties: The Effects of Different Similarity Judgment Criteria*. *Linguistic Discovery*, 3(1). PDF available: 2477k.
- Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.
- András Kornai. 2013. *Digital Language Death*. *PLOS ONE*, 8(10):e77056.
- Lelapa AI. Lelapa AI: Building AI for Africa, by Africans. <https://lelapa.ai/home/>. Accessed: 2025-10-22.
- Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. *Multilingual Denoising Pre-training for Neural Machine Translation*.
- Iloro Orife, Julia Kreutzer, Blessing Sibanda, Daniel Whitenack, Kathleen Siminyu, Laura Martinus, Jamii Toure Ali, Jade Abbott, Vukosi Marivate, Salomon Kabongo, Musie Meressa, Espoir Murhabazi, Orevaoghene Ahia, Elan van Biljon, Arshath Ramkilowan, Adewale Akinfaderin, Alp Öktem, Wole Akin, Ghollah Kioko, Kevin Degila, Herman Kamper, Bonaventure Dossou, Chris Emezue, Kelechi Ogueji, and Abdallah Bashir. 2020. *Masakhane – Machine Translation For Africa*.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. *BLEU: a Method for Automatic Evaluation of Machine Translation*. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.
- Maja Popović. 2017. *chrF++: words helping character n-grams*. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.
- Tatoeba Community. Tatoeba: A Collection of Sentences and Translations. <https://tatoeba.org/en/>. Collaborative, open, and free database of multilingual sentences and translations. Accessed: 2025-10-22.
- NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. *No Language Left Behind: Scaling Human-Centered Machine Translation*.
- Barbara Tompkins and Angela Kluge. 2002. *Sociolinguistic Survey of the Aja Language Area*. Technical report, SIL International. SIL Electronic Survey Reports 2002-020.
- Jiayi Wang, David Ifeoluwa Adelani, Sweta Agrawal, Marek Masiak, Ricardo Rei, Eleftheria Briakou, Marine Carpuat, Xuanli He, Sofia Bourhim, Andiswa Bukula, Muhidin Mohamed, Temitayo Olatoye, Tosin Adewumi, Hamam Mokayed, Christine Mwase, Wangui Kimotho, Foutse Yuehgoh, Anuoluwapo Aremu, Jessica Ojo, Shamsuddeen Hassan Muhammad, Salomey Osei, Abdul-Hakeem Omotayo, Chiamaka Chukwuneke, Perez Ogayo, Oumaima Hourrane, Salma El Anigri, Lolwethu Ndolela, Thabiso Mangwana, Shafie Abdi Mohamed, Ayinde Hassan, Oluwabusayo Olufunke Awoyomi, Lama Alkhaled, Sana Al-Azzawi, Naome A. Etori, Millicent Ochieng, Clemencia Siro, Samuel Njoroge, Eric Muchiri, Wangari Kimotho, Lyse Naomi Wamba Momo, Daud Abolade, Simbiat Ajao, Iyanuoluwa Shode, Ricky Macharm, Ruqayya Nasir Iro, Saheed S. Abdulahi, Stephen E. Moore, Bernard Opoku, Zainab Akinjobi, Abeeb Afolabi, Nnaemeka Obiefuna, Onyekachi Raphael Ogbu, Sam Brian, Verah Akinyi Otiende, Chinedu Emmanuel Mbonu, Sakayo Toadom Sari, Yao Lu, and Pontus Stenetorp. 2024. *AfriMTE and AfriCOMET: En-*

hancing COMET to Embrace Under-resourced African Languages.

Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. 2022. [ByT5: Towards a token-free future with pre-trained byte-to-byte models.](#)