Machine Translation for Nko: Tools, Corpora and Baseline Results

Moussa Koulako Bala Doumbouya *S,F **Baba Mamadi Diané** N **Solo Farabado Cissé** N
Djibrila Diané N **Abdoulaye Sow** F **Séré Moussa Doumbouya** F
Daouda Bangoura F **Fodé Moriba Bayo** F **Ibrahima Sory 2. Condé**K
Kalo Mory Diané N **Chris Piech** S **Christopher Manning** S

S Computer Science Department, Stanford University. 450 Jane Stanford Way, Stanford, CA 94305
N Nko USA Inc. 365 E 169th St. Bronx, NY, US 10456
F Friasoft. 9C5M+33, Fria, Guinea.
K Kofi Annan University. J986+7P Conakry, Guinea

Abstract

Currently, there is no usable machine translation system for Nko [1], a language spoken by tens of millions of people across multiple West African countries, which holds significant cultural and educational value. To address this issue, we present a set of tools, resources, and baseline results aimed towards the development of usable machine translation systems for Nko and other languages that do not currently have sufficiently large parallel text corpora available. (1) Fria||el: A novel collaborative parallel text curation software that incorporates quality control through copyedit-based workflows. (2) Expansion of the FLoRes-200 and NLLB-Seed corpora with 2,009 and 6,193 high-quality Nko translations in parallel with 204 and 40 other languages. (3) nicolingua-0005: A collection of trilingual and bilingual corpora with 130,850 parallel segments and monolingual corpora containing over 3 million Nko words. (4) Baseline bilingual and multilingual neural machine translation results with the best model scoring 30.83 English-Nko chrF++ on FLoRes-devtest.

1 Introduction

The Manding languages, including Bambara, Maninka, Mandinka, Dyula, and several others, are generally mutually intelligible and spoken by over 40 million people across West African countries including Mali, Guinea, Ivory Coast, Gambia, Burkina Faso, Sierra Leone, Senegal, Liberia, and Guinea-Bissau. Nko, which means 'I say' in all Manding languages, was developed as both the Manding literary standard language and a writing system by Soulemana Kanté in 1949 for the purpose of sustaining the strong oral tradition of Manding languages (Niane, 1974; Conde, 2017; Eberhard et al., 2023).[2] Nko thus serves a role for the Manding languages somewhat akin to Modern Standard Arabic for Arabic languages. It adequately transcribes their essential features such as vowel length, nasalization, and tone (Oyler, 2002; Conde, 2017; Donaldson, 2017) and enables the development of a shared literature.

Since its invention, the use of Nko has been growing. It is taught by literacy promotion associations, and used in newspapers, social media, and electronic communication (RFI, 2016; Rosenberg, 2011; Donaldson, 2019; Diane, 2022). Given that students learn best in their native language (Soh et al., 2021), Nko is particularly valuable for elementary native language education. Unfortunately, Nko and more generally West African languages remain marginalized in West African academic institutions (Kotey, 1975; Bryant, 2020). As a result, and despite the efforts of its courageous community, few academic resources are available in Nko.

Amongst numerous other benefits, computer-assisted translation could be used to facilitate the translation of academic content between Nko and other languages and facilitate projects such as Nko Wikipedia, which currently contains less than two thousand articles, in contrast with French and English Wikipedia with over 2 and 6 million articles respectively (Wikimedia, 2023). Unfortunately, to date, there isn't any usable machine translation (MT) system for Nko, in part due to the unavailability of large text corpora required by state-of-the-art neural machine translation (NMT) algorithms.

Nko is a representative case study of the broader issues that interfere with the goal of universal machine translation. Thousands of languages still don't have available or usable MT systems, mainly due to the unavailability of high-quality parallel text corpora. Recent corpora curation efforts have also resulted in sub-standard data quality for some languages. Some issues reported by (NLLB Team

*moussa@cs.stanford.edu

[1] Also spelled N'Ko, but speakers prefer the name Nko.

[2] ISO-639 code: nqo; ISO-15924 code: Nkoo.

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 312–343
December 6–7, 2023. ©2023 Association for Computational Linguistics

et al., 2022) and others that we address in this work (see Section 3.3, and 3.6) could have been avoided with the use of an adequate parallel text corpus curation system, which did not previously exist.

This work aims to bootstrap the development of MT systems for Nko and, in the process, to contribute open-sourced resources and tools applicable to other languages. Our main contributions include:

Novel Parallel Text Curation Software. Our first contribution is Fria||el (pronounced Friallel), a cloud-based collaborative parallel text curation software that helps human translators orchestrate copyediting processes resulting in high-quality corpora. Fria||el is presented in Section 2.

Extension of FLoRes-200 and NLLB-SEED. Our second contribution is the extension of FLoRes-200 and a multilingually aligned version of the NLLB-SEED (NLLB Team et al., 2022) corpora with high-quality Nko translations performed by Nko native speaker experts. Both FLoRes-200 and NLLB-SEED match our educational objective fairly well. Both are built over sentences drawn from Wikipedia, with NLLB-SEED, in particular, covering various fields of human knowledge and activity. They are therefore more diverse than other common parallel texts, such as religious texts.

Language Resource from the Nko Community. Our third contribution is the *nicolingua-0005* corpus, a collection of mono-, bi-, and trilingual corpora curated from data files donated by Baba Mamadi Diané, Solo Farabado Cissé, Djibrila Diané, Nafadji Sory Condé, and Kalo Mory Diané.

Baseline Machine Translation Results. Our fourth and last contribution consists of baseline NMT experiments from English, French, and Bambara transcribed in Latin script to Nko and vice versa. We present bilingual and multilingual transformer-based NMT systems (Vaswani et al., 2017) built using the fairseq toolkit (Ott et al., 2019). At present, results remain quite modest, with the best *eng_Latn* → *nqo_Nkoo* system scoring 30.83 chrF++ on FLoRes-devtest.

All presented software and tools have been publicly released to facilitate further progress on machine translation for Nko and other languages.[3]

2 Fria||el: Collaborative Parallel Corpus Curation System

Recent efforts on collecting multilingual parallel corpora involved sets of data file exchange between various translation teams (Federmann et al., 2022). This process is error-prone as it doesn't allow the systematic tracking of individual corpus entries through a curation quality process. Other recent similar efforts such as NLLB-SEED (NLLB Team et al., 2022), unnecessarily resulted in bi-text data rather than the intended multi-text because the reference files ended up being modified and re-ordered by various translation teams (see Section 3.3). Adequate software could have helped avoid such issues.

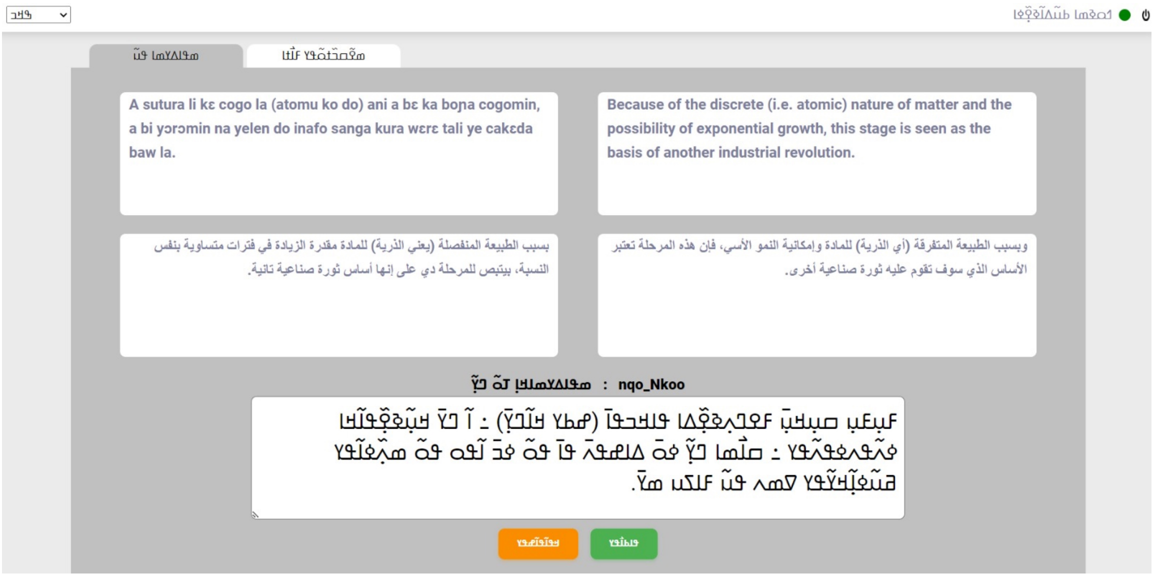
We propose Fria||el, a collaborative system designed to help distributed translation teams produce large multilingually aligned high-quality parallel text corpora. The system design particularly emphasizes suitability for use in various contexts, supporting web and mobile device usage and use in an offline mode. Its design goals include: itemized curation, automatic work organization, collaborative copyediting, and localization to translators' preferred user-interface language and preferred source languages to translate from (Figure 1).

2.1 Previous Tools and Multilingual Parallel Corpora Creation Processes

Masakhane Similarly to (Nekoto et al., 2020), this work is an effort towards African language technology development. Our work is participatory in the sense that we are a diverse team of computer scientists, linguists, and native speakers of Nko and other West African languages. We expect that our approach, and the parallel text curation software we release with this paper, Fria||el, will be valuable for MT technology development for other languages.

ParaText ParaText (SIL International & United Bible Societies, 2023) is specialized software for Bible translation projects. Its features include team management, task assignments, notes, collaborative document editing, multilingual dictionaries, and various biblical resources. It also allows a side-by-side comparison of biblical passages from various sources or in various languages. Paratext is not suited for general-purpose parallel corpus curation for MT. There is no indication that ParaText or any such software was used in the curation process of recent multilingual parallel corpora such as NLLB-SEED, FLoRes-200, and NTREX-128.

[3] Corpora and software on https://github.com/:
common-parallel-corpora/friallel
common-parallel-corpora/common-parallel-corpora
mdoumbouya/nicolingua-0005-nqo-nmt-resources
mdoumbouya/nko-nmt-wmt-2023



A sutura li kɛ cogo la (atomu ko do) ani a bɛ ka boɲa cogomin, a bi yɔrɔmin na yelen do inafo sanga kura wɛrɛ tali ye cakɛda baw la.

Because of the discrete (i.e. atomic) nature of matter and the possibility of exponential growth, this stage is seen as the basis of another industrial revolution.

بسبب الطبيعة المنفصلة (يعني الذرية) للمادة مقدرة الزيادة في فترات متساوية بنفس النسبة، بيتبص للمرحلة دي على إنها أساس ثورة صناعية تانية.

وبسبب الطبيعة المتفرقة (أي الذرية) للمادة وإمكانية النمو الأسي، فإن هذه المرحلة تعتبر الأساس الذي سوف تقوم عليه ثورة صناعية أخرى.

Figure 1: Fria||el's user interface for a Nko translator simultaneously inspecting multiple parallel variants of the same segment from the Multitext-NLLB-SEED corpus. All labels are localized to Nko. The source language fields are also localized to their own language's writing direction: LTR for Bambara in Latin script and English; and RTL for Moroccan Arabic and Egyptian Arabic. The translated text is localized to Nko's writing direction (RTL).

NLLB-SEED and FLoRes-200 The curation process of FLoRes-200 involved teams of translators and reviewers who underwent a vetting process. The QA team reviewed a 20% subset of data files with 3000 entries produced by translation teams. Data files falling below the 90% quality threshold were returned for rework. NLLB-SEED underwent a less rigorous quality control process. The curation process was English-centric. Translators were required to be proficient in English. Translation to the majority of languages was also done from English, with the following exceptions: In NLLB-SEED, Ligurian, was translated from Italian, In FLoRes-200, some Arabic languages were translated from Modern Standard Arabic. As noted by the authors, there are qualified translators who may not speak English, and several languages may be easier to translate from non-English sources.

NTREX-128 NTREX-128 (Federmann et al., 2022) was curated as follows. The English reference file was sent to a translation provider that produced translations. Source-based direct assessment was performed on the translated files by a different provider using the Appraise platform (Federmann, 2018) to generate segment-level quality scores. Segments with a score below a specified threshold were returned for correction. The translation process and quality control method of the translation provider were not specified.

Fria||el is a collaborative parallel text curation software system that tracks individual segments through a translation and copyedit workflow. Each segment is translated by one translator, and subsequently sequentially copyedited by other translators. Fria||el allows translators to simultaneously inspect variants of the source segment in multiple languages. This results in segments translated and copyedited in the context of different subsets of source languages. In addition to the final parallel corpus, Fria||el also yields copyedit logs, which could be valuable in various modeling scenarios.

2.2 Design Goals

Fria||el was designed with the following goals:

Itemized Curation Each corpus segment is individually tracked through the curation process in which it is translated to the target language and subsequently reviewed and copyedited several times.

Automatic Task Assignments Translation and copyediting tasks are automatically assigned to translators with fixed lease periods. Uncompleted tasks are automatically reassigned upon expiration.

Collaborative Copyediting Each segment is translated once and copyedited two or three times, following the workflow in Figure 2. Segments for which the first or second verification results in edits are copyedited a third time. A given translator can only perform a task on a given segment once.

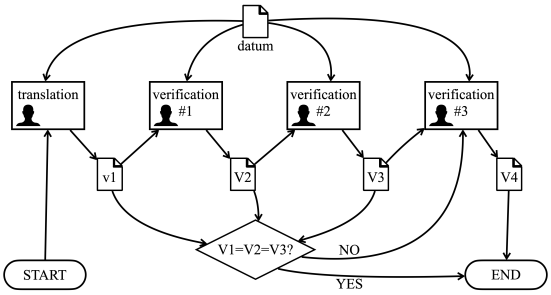


Figure 2: Translation workflow for a multilingual segment (datum). The initial translation (*v1*) is approved or copyedited by two other translators (*v2*) and (*v3*). If any copyediting occurs, a third copyediting task is assigned to a fourth translator who either approves the current translation or performs a final copyedit (*v4*).

Multilingual Sources While performing translation and copyediting tasks, translators can simultaneously inspect segments in several languages configured according to their preferences.

Machine-Generated Sources Datasets can be augmented with additional machine-generated variants of segments such as machine translations, transliterations, and detransliterations.

Responsive Web Design Fria||el is a web application that automatically adapts the layout of its component to the user's screen size. This makes it usable on desktop and laptop computers as well as on mobile phones and tablets.

Resilience to Connectivity Disruptions Translators who temporarily lose their internet connectivity can seamlessly keep working offline on their currently assigned translation and verification tasks. Their work is automatically synchronized with the central database when their connectivity is restored.

Internationalization and Localization Fria||el is internationalized (i18n) in that all user-facing strings are externalized into a translatable resource file, and the writing direction and text alignment of translation source and target languages are configurable. As a result, the user interface is localized (L10n) to the translator's preferred user-interface language, and to each source language (Figure 1).

2.3 Software Components

This section provides details on Fria||el's software components that collectively realize the design goals specified in Section 2.2.

2.3.1 Workflow Manager

Both the Workflow and Task Managers are implemented as Firebase cloud functions that are triggered at fixed time intervals. A workflow entity is inserted for each parallel segment with an initial *active* state. The Workflow Manager periodically inspects workflow entities and (1) creates the next task if needed, and per the workflow management rules, (2) moves the workflow to the *completed* status if all related tasks have been completed and there is no need to create additional tasks or (3) nothing, if the workflow has an uncompleted task.

2.3.2 Task Manager

When triggered, the Task Manager revokes all expired task assignments and assigns unassigned translation and copyedit tasks to users according to their roles. The maximum number of tasks assigned to each user is fixed. A given user is never assigned a task related to a segment on which they have previously completed a task. The Task Manager also ensures that a copyedit task is only assigned to a user with the appropriate verification skill level (*L1*, *L2*, or *L3*) for the first, second, and third copyedit rounds. Each translator account is configured with specific verification skill levels.

2.3.3 Data Model and Storage

Google Firestore, a document-oriented NoSQL database, is used for data storage. The central application database is accessed by data import/export scripts, the WorkflowManager, the Task Manager, and the user interface. It contains the following collections of documents:

datasets: One collection per imported dataset. Each document represents a multilingual segment and contains all available translations of the segment, each annotated with its language and writing system. See Figure 10.

workflows: Each document represents a prioritized workflow entity. The WorkflowManager (Section 2.3.1) periodically inspects workflow entities by priority order and creates task entities as per the workflow management logic.

annotation-tasks: Each document is a task of a specific type (translation or copyedit) related to a specific multilingual segment. Each task has a status (unassigned, assigned, completed). Tasks are assigned to translators by the Task Manager.

users: Each document represents a translator and specifies whether they can be assigned translation (*isActiveTranslator*) and copyedit (*isActiveV-*

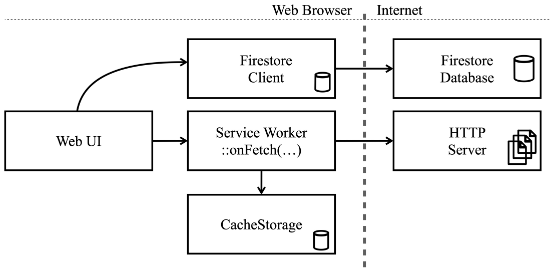


Figure 3: The software uses Firestore's client library's offline mechanism and *cached-aside* HTTP resources to be resilient to intermittent internet disruptions.

erifier) tasks. Translator documents also store the source languages the translator prefers to translate from, subject to availability in the source corpus. User documents also specify a *verifierLevel*, which indicates the maximum copyediting round the translator can participate in for a specific segment.

config: Contains language writing direction configuration. Languages are assumed to be left-to-right unless explicitly marked right-to-left.

2.3.4 User Interface

The user interface is a responsive web application that is usable on a variety of devices, including mobile phones, tablets, desktops, and laptops (Figure 1). It directs authenticated translators to their workspace where they can perform translation (first tab) and copyediting (second tab) tasks that are assigned to them. The task assignment process is transparent to translators. One task is displayed at a time. The prioritized list of tasks assigned to the current translator is kept in a cache for resilience to intermittent internet disruptions. The connection status is indicated by the green circle (top-right).

When performing translation tasks, translators simultaneously inspect the source segment in several languages (top four text fields) and write a translation in the target language text field (bottom). When the 'submit' button (green) is selected, the translation is recorded and the next task is displayed. Translators can also skip the current task by selecting the 'skip' button (orange). When performing copyediting tasks, the bottom text field is initialized with the latest version of the translated segment (Figure 2). The translator may leave the translation intact or copyedit it before submitting.

2.3.5 Offline Mode

The software is a web application designed to be resilient to intermittent internet disruptions. This is achieved with Google Firebase's client library (Google, 2023), which supports offline read and write operations by leveraging a client-side *eventually consistent* (Burckhardt et al., 2014) LRU cache, and *cached-aside* (Pamula et al., 2014) HTTP resources, implemented with two web APIs supported by the majority of web browsers: CacheStorage and ServiceWorker (w3.org, 2022; Mozilla, 2023b,a). After the initial loading of the web application in a web browser, a ServiceWorker is registered to intercept HTTP fetch events. If the remote web server is reachable, the ServiceWorker fetches remote HTTP resources (e.g., HTML, CSS, javascript, image files) and stores them in a CacheStorage before returning them to the caller; otherwise, cached resources are served. The entire process is transparent to the user. See Fig 3.

2.3.6 Translator Copyedit Logs

In addition to the final version of the translated segments, the data Fria||el also outputs their initial translation (*v1*), and the versions of the same entries after the first, second, and third copyediting rounds (*v2*, *v3*, *v4*) – see the workflow in Figure 2. Copyediting logs can be valuable in developing language and machine translation models.

2.3.7 Data Import and Export

Fria||el includes the following administrative Python scripts for importing and exporting parallel corpora and other reports. `load_dataset.py` imports a new parallel corpus from its original data files. Pre-processing may be required to adapt to various original dataset formats. `create_translation_workflows.py` creates active translation workflows for an imported dataset. `system_report.py` displays the number of workflows and tasks by status by dataset. `export_dataset.py` exports translations and translator edits for a curated dataset in a csv file. Post-processing may be required to adapt to a desired format. `accounting_statements.py` generate completed tasks by user by dataset by month. This data can be imported into an accounting system to generate payroll for translators.

2.4 Qualitative User Study

Nko translators used Fria||el to translate FLoRes-200 (dev, devtest) and Multitext-NLLB-SEED to nqo_Nkoo, and to copyedit each segment two or three times. The following sections present an analysis of their responses to a survey questionnaire

(Figure 6). Quantitative measures on their copyediting logs are also discussed in Section 2.4.6.

2.4.1 Usability

Nko translators praised the simplicity of the user interface. They appreciated the automatically organized itemized copyediting-based data curation process. They highlighted the localization features, particularly, the fact that the user-interface is available in Nko and that the presentation was adequate for both right-to-left and left-to-right source languages and the target language. They valued the offline functionality that allowed them to temporarily continue working without an internet connection. Furthermore, they found the task counters displayed on the user-interface helpful. They noted two usability-related limitations: First, it was not possible to directly go back to a task after submitting it. Second, although the software allowed them to continue working offline, it did not allow them to perform the initial authentication while offline.

2.4.2 Translation Process

Nko translators found the fact that source segments were visible in multiple languages beneficial. They said that the ability to inspect the same segment in multiple languages facilitated its translation to Nko. They also mentioned that the itemized translation tasks, which presented one segment at a time, decreased the likelihood of translation mistakes.

An improvement they requested is the addition of a translation memory including dictionary entries and previously translated expressions.

2.4.3 Copyediting Process

Nko translators found Fria||el's multi-pass copyediting process effective for finding and correcting translation mistakes. They mentioned that the fact that segments were consecutively assigned to different translators for copyediting led to higher-quality translations as it is easy to overlook one's own mistakes. Because each translator had a different translation source language configuration, Nko segments were translated from and copyedited against their versions in different sets of languages, which Nko translators found enriching.

2.4.4 Mistranslations

Types of mistranslations Nko translators noted during the copyediting process included typos, omitted words, grammatical errors, incorrect word sense translations, incorrect translations of named entities, and punctuation errors. They noted that word sense was sometimes hard to disambiguate without the full context of segments. For instance, the English word *state* maps to different Nko words based on the sense of the word (political community vs. a particular condition of a person, place, or thing). They also noted punctuation errors, particularly the use of the Arabic comma (U+060C) instead of the Nko comma (U+07F8), and spacing around that punctuation. Finally, they reported that translators using different source languages would sometimes disagree on named entity translations.

2.4.5 Disagreements

Nko translators reported few disagreements on language standards. They also reported using existing English-Nko and French-Nko dictionaries for consistency. During the translation of FLoRes-200, NLLB-SEED to Nko, translators participated in weekly team meetings and routinely consulted each other over video conferences and phone calls. They deferred the few cases of disagreement and perplexing questions to the most senior translator.

2.4.6 Copyediting Metrics

Table 1 summarizes the size of the translated corpora in segments and Nko words, as well as the percentage of segments that were edited in each verification round, and the related edit magnitudes, computed as edit distances. The number of edited segments and related edit magnitudes generally decreased as copy-editing rounds progressed.

3 Nko Corpora for Machine Translation

This section discusses the extension of FLoRes-200 and NLLB-SEED to Nko, which included the multilingual alignment of NLLB-SEED, and the use of Fria||el to translate those corpora to Nko. This section also introduces *nicolingua-0005*, a collection of monolingual corpora and bi- and trilingual parallel corpora donated by Nko community members.

3.1 Translation of FLoRes-200 and NLLB-Seed to Nko

Nko native speaker experts Baba Mamadi Diané, Solo Farabado Cissé, and Djibrila Diané, used Fria||el to translate Multitext-NLLB-SEED, FLoRes-dev, and FLoRes-devtest to Nko. They worked from Cairo (Egypt), Banankoro (Guinea), and New York (USA), and with the rest of the team, participated in weekly video conference meetings.

corpus	seg- ments	words	$v1 \rightarrow v2$ edited	 edit distance	$v2 \rightarrow v3$ edited	 edit distance	$v3 \rightarrow v4$ edited	 edit distance
FLoRes-dev	997	27,361	83%	38.75 ± 1.55	67%	50.48 ± 2.10	71%	11.74 ± 0.65
FLoRes-devtest	1,012	29,503	87%	61.74 ± 1.81	93%	9.69 ± 0.64	24%	2.79 ± 0.15
NLLB-SEED	6,193	184,138	48%	45.97 ± 1.11	35%	38.94 ± 1.16	35%	11.96 ± 0.48

Table 1: Percentage of edited Nko segments, and related mean± standard error of edit magnitudes (edit distance) resulting from the translation of FLoRes-dev, FLoRes-devtest, NLLB-SEED to Nko

3.2 Translation Process

The initial translations of FLoRes-dev, which our translators performed using spreadsheets, were imported into Fria||el after our software engineers completed its development. The copyediting tasks for FLoRes-dev, and the translation and copyediting tasks for FLoRes-devtest and Multitext-NLLB-SEED were entirely performed using Fria||el. The system was designed not to allow translators to copyedit their own translations or previous copyedits. This constraint made the proposed translation workflow impossible given the size of our team of translators. As a workaround, an additional user account was created for the two most experienced translators to allow third copyediting rounds.

Each segment was translated once and copyedited two or three times. The resulting curated Nko data files are summarized in Table 2. The multilingually aligned NLLB-SEED dataset (Multitext-NLLB-SEED), FLoRes-dev, and FLoRes-devtest, all extended with Nko translations along with copyedit logs, collectively make up common-parallel-corpora ver. 2023-06-19 summarized in Table 3.

Translations		
lines	words	file
6193	184138	SEED/nqo_Nkoo
997	27361	FLoRes/nqo_Nkoo.dev
1012	29503	FLoRes/nqo_Nkoo.devtest

Translator Edits		
lines	words	file
6193	170555	SEED/nqo_Nkoo.v1
6193	177703	SEED/nqo_Nkoo.v2
6193	182843	SEED/nqo_Nkoo.v3
6193	184138	SEED/nqo_Nkoo.v4
997	24455	FLoRes/nqo_Nkoo.dev.v1
997	25656	FLoRes/nqo_Nkoo.dev.v2
997	26541	FLoRes/nqo_Nkoo.dev.v3
997	27361	FLoRes/nqo_Nkoo.dev.v4
1012	25924	FLoRes/nqo_Nkoo.devtest.v1
1012	27771	FLoRes/nqo_Nkoo.devtest.v2
1012	29521	FLoRes/nqo_Nkoo.devtest.v3
1012	29503	FLoRes/nqo_Nkoo.devtest.v4

Table 2: Extensions of FLoRes-200 (dev, devtest) and Multitext-NLLB-SEED to Nko. The *nqo_Nkoo* data files are parallel with 40 other languages in NLLB-SEED, and 204 other languages in FLoRes-200. FLoRes-test, which is not publicly available, was not translated.

3.3 Multilingual Alignment of NLLB-SEED

The original NLLB-SEED dataset consists of pairwise parallel corpora between English and each other language but suffers from the complication that many of the source English sides are slightly different from each other, variously due to minor copyediting, and reordered and added entries.

Multitext-NLLB-SEED is a multilingually aligned version of NLLB-SEED that fixes this limitation. It was created as follows: A consensus *eng_Latn* reference file was manually edited by human comparison of all existing reference *eng_Latn* files. The lines of each *eng_Latn* file were matched (binary assignment matrix $M_{i,j}$) to the lines of the consensus *eng_Latn* file by minimizing the sum of the edit distances $E_{i,j}$ between matched lines (Equation 1). The optimal

CPC subset	lines	langs.	tr. edits langs.
Multitext-NLLB-SEED	6193	41	1
FLoRes-dev	997	205	1
FLoRes-devtest	1012	205	1

Table 3: Summary of common-parallel-corpora version 2023-06-19. All entries are parallel across all languages. Translator edits are only available for nqo_Nkoo.

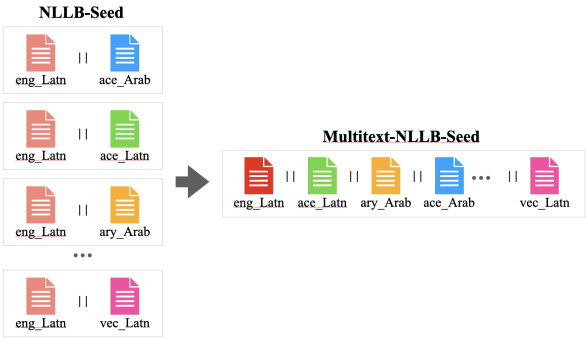


Figure 4: Multitext-NLLB-SEED is a multilingually aligned version of the original NLLB-SEED dataset.

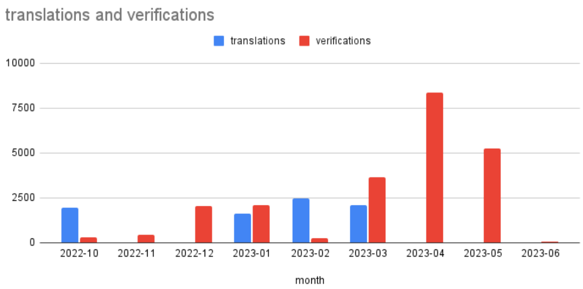


Figure 5: From October 2022 to June 2023, 8,202 translations and 22,426 verifications/edits were performed to produce high-quality translations of FLoRes-200 and Multitext-NLLB-SEED to Nko.

line matching M^*, obtained using the scipy package (Virtanen et al., 2020), was used to re-order each non-English language file to match the order of the consensus *eng_Latn* file. Two unmatched lines from (*eng_Latn*, *kas_Deva*) and one from (*eng_Latn*, *lij_Latn*) were discarded.

$$M^* = \arg\min_M \sum_{i,j} M_{i,j} E_{i,j} \qquad (1)$$

The resulting re-ordered non-English language files and the consensus eng_Latn file constitute the Multitext-NLLB-SEED corpus, containing 40 parallel language files; see Figure 4. Multitext-NLLB-SEED was loaded in Fria||el in lieu of the original NLLB-SEED corpus, enabling translators to inspect each segment in multiple languages, and resulting in an expanded multilingually aligned corpus.

3.4 Translation Source Languages

Source languages were configured in Fria||el according to the preferences of each translator. Collectively, they translated from *fra_Latn*, *eng_Latn*, *ary_Arab*, *arz_Arab*, and *bam_Latn*. Note that *fra_Latn* is not available in NLLB-SEED. *bam_Nkoo* was detransliterated from *bam_Latn* using a neural detransliterator (Doumbouya, 2022); however, translators did not find this source useful and preferred not to enable it in their configuration.

3.5 Further Notes on Manding languages

Nko was developed as a standardized form of the Manding languages. The aim was a standardized language and writing system, which could serve a similar role to Modern Standard Arabic with respect to various regional Arabic languages. Manding languages, which include Mandinka and Bambara, are a subgroup of the Mande language family and are generally mutually intelligible to speakers. Bambara, written in a Latin script, is currently the best-supported Manding language, available in Google Translate and in NLLB-SEED. Our Nko translators are also fluent in Bambara.

3.6 Quality of bam_Latn in NLLB-SEED

Our Nko translators noted the following quality issues with NLLB-SEED's *bam_Latn* data: (1) The data contains too much French vocabulary not enough Manding vocabulary. (2) Some entries do not match their English counterpart at all. (3) Some entries are entirely in French; examples are shown in Figure 11. (4) The *bam_Latn* data completely lacks tonal marks, which are important in Manding languages (e.g., many nouns are indistinguishable without tonal marks, such as *bird*, *belly*, *inside*; the definite and indefinite inflections of nouns cannot be distinguished without tonal marks (*I saw a person* vs. *I did not see any person*); and nouns that can be used as a verb and their verb form cannot be distinguished (*get out!* vs. *to get out*). *bam_Nkoo*, detransliterated from *bam_Latn* was included in the corpus; however, some Nko translators did not find it useful and preferred to not enable it as a source.

3.7 *nicolingua-0005* Corpus

nicolingua-0005 is curated from files donated by Nko community members for the purpose of developing machine translation for Nko. It is comprised of 3.9 million Nko words with 25K (Nko, English, French) parallel segments, 59K (Nko, English) parallel segments, 45K (Nko, French) parallel segments, and a monolingual corpus of 3.3 million Nko words. Included datasets were curated from files provided by Baba Mamadi Diané, Solo Farabado Cissé, Djibrila Diané, Nafadji Sory

type	languages	segments	nqo words
trilingual	*nqo_Nkoo, eng_Latn, fra_Latn*	25 848	256 934
bilingual	*nqo_Nkoo, eng_Latn*	59 442	283 279
bilingual	*nqo_Nkoo, fra_Latn*	45 560	129 789
monolingual	*nqo_Nkoo*	N/A	3 291 371
total		130 850	3 961 373

Table 4: Summary of *nicolingua-0005*

Condé, and Kalo Mory Diané. See Table 4 and Appendix D for more details on the constitution of the corpus. A datasheet questionnaire based on (Costa-jussà et al., 2020) is presented in Appendix E.

4 Baseline Machine Translation Experiments

This section describes Transformer (Vaswani et al., 2017) based encoder-decoder neural machine translation models built using the fairseq toolkit (Ott et al., 2019). Both bilingual and multilingual translation models are explored. At present, results remain quite modest, with the best model achieving a 30.83 *eng_Latn* → *nqo_Nkoo* chrF++ score on the CPC/FLoRes-devtest corpus.

Eight models were trained: The bilingual unidirectional models 200.11 and 200.16, the multilingual model 201.16, and its variant that is trained to also autoencode Nko segments 202.16, and Models 206.19, 207.19, 208.19 and 209.19, which explore three different ways of specifying language tokens.

4.1 Datasets

common-parallel-corpora (CPC) and *nicolingua-0005*, described in Section 3 are used to build baseline NMT models for the following translation directions: *nqo_Nkoo* ⇄ *eng_Latn*, *nqo_Nkoo* ⇄ *fra_Latn*, and *nqo_Nkoo* ⇄ *bam_Latn*. The subsets of those corpora used to train, validate, and test the models are specified in Tables 11 and 12.

4.2 Tokenization

Byte-pair encoding (BPE) (Sennrich et al., 2016) is employed to perform sub-word tokenization. In each training experiment, the BPE model is trained on a token corpus constructed by concatenating all data files containing the languages of interest in the training set. In all cases, the BPE model is trained to produce 15K sub-word units.

4.3 Models

Eight models were trained. The first two, 200.11 and 200.16, are unidirectional bilingual *nqo_Nkoo* ⇄ *eng_Latn* models. The last six, 201.16, 202.16, 206.19, 207.19, 208.19, and 208.19 are multilingual *nqo_Nkoo* ⇄ *eng_Latn*, *nqo_Nkoo* ⇄ *fra_Latn*, and *nqo_Nkoo* ⇄ *bam_Latn* models.

4.3.1 Bilingual Models

200.11 is the baseline bilingual *nqo* ← *eng* model. 200.16 differs from 200.11 in terms of model architecture and hyper-parameters. 200.16 and the multilingual models 201.16 and 202.16 have identical architectures and training hyper-parameters.

Model 200.11 is a Transformer-based (Vaswani et al., 2017) encoder-decoder sequence-to-sequence model consisting of 5 encoder and 5 decoder layers, each with a 512-dimensional token embeddings and 2048-dimensional feed-forward networks, 2 attention heads per layer, and a layer normalization module before each layer. Its architecture and training hyper-parameters are identical to the baseline system of the AmericasNLP 2021 Shared Task on Open Machine Translation (Mager et al., 2021), except for the following differences: (1) encoder and decoder embeddings are not shared, (2) Subword Regularization (Kudo, 2018) and BPE-dropout (Provilkov et al., 2020) are not employed in BPE tokenizer training, (3) larger batches are employed during training, (4) gradient clipping is applied during training.

Model 200.16 This model is only different from 200.11 in that it is deeper (6 encoder layers and 6 decoder layers), and that it is trained with a higher token dropout probability (0.6 instead of 0.4).

4.3.2 Multilingual Models

Our multilingual models are trained on parallel corpora obtained by concatenating all available (*nqo* ⇄ *eng*, *nqo* ⇄ *fra*, *nqo* ⇄ *bam*) bitext and prefixing the source segments with language tokens as introduced by (Johnson et al., 2017). Similarly to (Wicks and Duh, 2022), models 206.19, 207.19, 208.19, and 209.19 compare the effect of various approaches to constructing source-side prefixes.

Model 201.16 is the baseline multilingual model. It has the same architecture and training hyperparameters as the bilingual model 200.16, but it is trained on multilingual data and it employs target language token prefixes (Table 5).

Model 202.16 employs target language tokens just like 201.16, but its training set also contains *nqo* → *nqo* pairs where each side is the same sentence from monolingual Nko corpora in *nicolingua-*

model	prefix
200.11	(none)
200.16	
201.16	<to_tgt_Lang>
202.16	
206.19	
207.19	<from_src_Lang> <to_tgt_Lang>
208.19	<from> <src_Lang> <to> <tgt_Lang>
209.19	<from_src_Lang_to_tgt_Lang>

Table 5: Specification of source sequence language token prefixes used in our multilingual translation models.

0005. Consequently, 202.16 performs simultaneous multilingual translation and monolingual sequence autoencoding. Positive results from such a strategy were found in (Luong et al., 2016).

Models 20x.19 also perform simultaneous translation and monolingual sequence auto-encoding. However, their architecture is different from 202.16, and they explore different language token prefixing strategies. Compared to 202.16 models, in 20x.19 models, the encoder and decoder layers use 8 attention heads instead of 2. Also, the encoder's input embeddings and the decoder's input and output embeddings are all shared. Finally, the source and target token dictionaries are also shared.

Models 20x.19 explore four approaches of source-side prefix specification (Table 5). As an example, a source segment to be translated from English to Nko is prefixed as follows per model:
206.19: “<to_nqo_Nkoo> "
207.19: “<from_eng_Latn> <to_nqo_Nkoo> "
208.19: “<from> <eng_Latn> <to> <nqo_Nkoo> "
209.19: “<from><eng_Latn><to><nqo_Nkoo> "

4.4 Training

During training, dropout is used with the following probabilities: input token embedding dropout 0.4 (xxx.11) or 0.6 (xxx.16, xxx.19), attention dropout 0.2, ReLU dropout 0.2. The label-smoothed cross-entropy loss function is used with a smoothing rate of 0.2. Optimization is performed using Adam with a weight decay of 0.0001. The inverse squared root learning rate scheduler is used with an initial rate of 1e-7 and 4000 warm-up updates. Gradient clipping is employed with a norm threshold of 1. Effective batches of up to 65,536 tokens are used to train all models. Gradients are accumulated for 1 batch of up 65,536 tokens on A100 GPUs and 4 batches of up to 16384 on Titan XP GPUs before each update.

4.5 Model Selection and Stopping Criteria

Trainings are stopped when BLEU scores on the validation step do not improve after 20K gradient updates. Checkpoints with the highest BLEU scores on the validation set are selected. The average BLEU score across all supported translation directions is used for multilingual model selection.

4.6 Evaluation

CPC/FLoRes-dev and CPC/FLoRes-devtest are respectively used as validation and test sets. For each model, their subsets with languages of interest are considered (see Tables 12 and 11). The chrF++ score, which has been shown to align well with human assessments, especially for morphologically rich languages (Popović, 2017), is used as the main evaluation metric. The Sacre BLEU library (Post, 2018) is used to compute BLEU and chrF++ scores.

4.7 Results

Table 6 shows the test and validation BLEU and chrF++ scores for each model and supported translation direction. The best performing model 208.19 scores 26.00 mean chrF++ on the test set.

Layer Count and Regularization: Compared to 200.11, 200.16 with one extra encoder and decoder layer, and a higher token embedding dropout rate, scored $+0.34$ *nqo* $\leftarrow$ *eng* chrF++.

Multilinguality: Compared to 200.16, the multilingual model 201.16, which has the same architecture and training hyperparameters, scored -0.92 *nqo* $\leftarrow$ *eng* chrF++.

Monolingual Autoencoding: Compared to 201.16, 202.16, which performs simultaneous multilingual translation and monolingual autoencoding, scored $+0.14$ *nqo* $\leftarrow$ *eng* chrF++

Attention Heads and Shared Embeddings: Compared to 202.16, 206.19 which uses 8 attention heads in the encoder and decoder layers, and which shares all input and output embeddings and dictionaries scores $+1.59$ mean chrF++.

Language Token Prefixing: Compared to 206.19, which only specifies target language tokens in the source sequence, 207.19, which specifies the source and target languages as two separate tokens, scored $+0.08$ mean chrF++. 209.19, which specifies the source and target languages as a single token, scored +0.06 mean chrF++. 208.19, which specifies the source and target languages in a four-token clause scored +0.15 mean chrF++.

5 Discussions

5.1 Fria||el

Improving Usability: As noted by Nko translators, the usability of Fria||el could be improved by: (1) Allowing translators to review their recently submitted tasks before the Workflow Manager proceeds to the next stage of the curation process. (2) Implementing an offline authentication mechanism.

Adding a Translation Memory: Adding a translation memory could increase the productivity, accuracy, and consistency of translators. However, the effect of such a tool on the general quality of translations, including the diversity of synonyms and expression styles should not be overlooked.

Extensibility: Alternate copyediting workflows can be implemented in Fria||el by extending the Workflow Manager and Task Manager. The task presentation user interface can also be adapted to other text curation tasks, such as syntax annotation.

5.2 Parallel Corpora

Handling Short Sequences: The segments in *nicolingua-0005* are, on average, significantly shorter than those in FLoRes and NLLB-SEED. Despite being short, sequences such as ones from the Nko-Français dictionary and Unicode CLDR files, are too valuable to discard. To prevent biasing models towards shorter sequence lengths, we repeated the *(nqo_Nkoo, eng_Latn)* data from CPC/NLLB-SEED five times in the training set. A more principled approach should be considered.

Punctuations, Case and Diacritical Marks: Our models showed sensitivity to minor changes in Latin case, and punctuation as well as Nko diacritical marks (see Appendix G). Including augmented data with lowered case and stripped punctuation and diacritical marks in source sequences in the training corpora may help address this issue.

Learning from Translator Edits: Translator edits, as recorded by Fria||el throughout the copy-edit process, could be useful for various modeling and quality estimation tasks. This data could also be used for an auxiliary copy-edit reconstruction task that may improve the accuracy of a multitask NMT

		Intl. BLEU		chrF++	
model	direction	valid	test	valid	test
200.11	*nqo ← eng*	5.40	5.11	28.80	29.73
200.16	*nqo ← eng*	5.85	5.25	29.06	30.07
201.16	*nqo → bam*	1.19	1.12	16.73	17.04
201.16	*nqo ← bam*	2.86	3.19	22.07	23.01
201.16	*nqo → eng*	3.65	3.78	26.31	26.99
201.16	*nqo ← eng*	6.11	5.71	28.64	29.15
201.16	*nqo → fra*	2.33	2.35	22.27	22.61
201.16	*nqo ← fra*	4.50	4.29	25.55	25.89
201.16	mean	3.44	3.41	23.60	24.12
202.16	*nqo → bam*	1.14	1.00	16.68	16.82
202.16	*nqo ← bam*	2.83	3.11	22.33	23.11
202.16	*nqo → eng*	4.27	4.26	26.86	27.61
202.16	*nqo ← eng*	6.18	5.80	28.63	29.44
202.16	*nqo → fra*	2.31	2.74	22.46	22.89
202.16	*nqo ← fra*	4.18	4.51	25.22	25.68
202.16	mean	3.49	3.57	23.70	24.26
206.19	*nqo → bam*	1.69	1.50	19.04	19.34
206.19	*nqo ← bam*	3.63	3.43	23.26	23.81
206.19	*nqo → eng*	5.22	5.15	28.70	28.85
206.19	*nqo ← eng*	6.79	6.50	29.97	30.66
206.19	*nqo → fra*	3.28	3.41	25.26	25.42
206.19	*nqo ← fra*	4.77	5.05	26.59	27.03
206.19	mean	4.23	4.17	25.47	25.85
207.19	*nqo → bam*	1.52	1.53	19.09	**19.43**
207.19	*nqo ← bam*	3.56	3.28	23.10	23.77
207.19	*nqo → eng*	5.10	5.05	28.61	28.69
207.19	*nqo ← eng*	6.96	6.21	29.92	30.45
207.19	*nqo → fra*	3.26	3.51	25.44	**25.98**
207.19	*nqo ← fra*	5.23	5.14	26.89	**27.25**
207.19	mean	4.27	4.12	25.51	25.93
208.19	*nqo → bam*	1.44	1.52	18.83	19.08
208.19	*nqo ← bam*	3.32	3.37	23.38	**24.00**
208.19	*nqo → eng*	4.78	5.05	28.64	**29.13**
208.19	*nqo ← eng*	6.99	6.44	30.05	**30.83**
208.19	*nqo → fra*	3.20	3.61	25.15	25.79
208.19	*nqo ← fra*	5.04	4.78	26.73	27.17
208.19	mean	4.13	4.13	25.46	**26.00**
209.19	*nqo → bam*	1.60	1.47	19.00	19.25
209.19	*nqo ← bam*	3.45	3.43	23.29	23.80
209.19	*nqo → eng*	5.07	4.79	28.67	28.82
209.19	*nqo ← eng*	6.96	6.58	30.10	30.78
209.19	*nqo → fra*	3.49	3.13	25.39	25.76
209.19	*nqo ← fra*	5.13	4.92	26.56	27.06
209.19	mean	4.28	4.05	25.50	25.91

Table 6: Our bilingual and multilingual models measured for accuracy on FLoRes-dev (valid) and FLoRes-devtest (test) using the Intl. BLEU (Sacre BLEU with Unicode-aware tokenization) and chrF++ metrics.

model. Finally, translator edit data can be used to train and align translators on consistency standards.

5.3 Neural Machine Translation

Tokenization: Subword regularization, as discussed in (Kudo, 2018) and the dropout-based approach presented by (Provilkov et al., 2020), may lead to increased translation performance for Nko.

Language Token Prefixes: The choice of source-side prefixing strategy had a marginal impact on translation accuracy. Our best model employs a four-token prefix, consisting of source and target language tokens joined with the '<from>' and '<to>' tokens. Our results and those of (Wicks and Duh, 2022), suggest that the specification of translation directions as source-side prefixes in multilingual NMT models merits further investigation.

Learning from Monolingual Data: The use of monolingual Nko data in 202.16 led to marginal improvements in most translation directions. Additional unsupervised tasks such as masked language modeling and denoising should also be explored.

Data Augmentation: Back-translation-based data augmentation, and the generalized data augmentation method in (Xia et al., 2019) could significantly increase NMT performance for Nko.

International BLEU Our BLEU scores are computed with sacreBLEU using international tokenization because sacreBLEU's current default tokenizer (v13a) is inappropriate for Nko; it doesn't properly interpret the Nko Unicode block, particularly its punctuations, to detect word boundaries.

BLEU vs chrF++ The BLEU scores of our models are rather low. This was surprising given the training data size and given Nko translators' feedback on generated translations. This observation is in line with (Popović, 2017)'s hypothesis that chrF++ correlates better with human judgment than BLEU for morphologically rich languages.

6 Conclusion

This work presented Fria||el, a collaborative parallel text curation system with copyediting-based quality workflows. Fria||el enabled the extension of existing multilingual corpora, FLoRes-200 and NLLB-SEED with high-quality Nko translations. Those, and a new corpus we introduced, *nicolingua-0005*, served to build baseline bilingual and multilingual NMT systems for Nko, with the best model achieving the accuracy of 30.84 *eng_Latn* → *nqo_Nkoo* chrF++. We have released Fria||el to facilitate the development and extension of multilingual parallel corpora to more languages. We have also released resources and tools to enable the reproducibility of our results, and further progress towards usable MT systems for Nko.

Acknowledgements

Sources of support that made this project possible include: A generous unrestricted research gift from Meta Platforms, Inc., the Stanford Graduate Fellowship (SGF), the Stanford Natural Language Processing (NLP) group, FriaSoft, and Nko USA Inc. We extend our deep gratitude to FriaSoft for selflessly donating numerous software engineering hours and to Nko USA Inc. for providing Nko language expertise and resources. The commitment of Nko USA Inc. and FriaSoft to advancing West African language technology was instrumental in realizing this project. We thanks to John Hewitt and Tolúlọpẹ́ Ògúnrẹ̀mí for giving invaluable feedback. We are grateful to Asmaou Diallo, Nantenin Camara, Koulako Camara, Moussa Thomas Doumbouya, and Ibrahima Doumbouya for creating a supportive environment conducive to the execution of portions of this work in Conakry, Fria, and Boké.

References

Kelly Duke Bryant. 2020. Education and Politics in Colonial French West Africa. In *Oxford Research Encyclopedia of African History*.

Sebastian Burckhardt et al. 2014. Principles of Eventual Consistency. *Foundations and Trends® in Programming Languages*, 1(1-2):1–150.

Laye Camara. 1953. *L'enfant Noir*. Éditions Plon.

Nafadji Sory Conde. 2017. *Introduction au N'ko: Une Alternative Linguistique pour l'Afrique*. Presses de l'Université Kofi Annan and Harmattan Guinée.

Marta R Costa-jussà, Roger Creus, Oriol Domingo, Albert Domínguez, Miquel Escobar, Cayetana López, Marina Garcia, and Margarita Geleta. 2020. Mt-Adapted Datasheets for Datasets: Template and Repository. *arXiv preprint arXiv:2005.13156*.

Baba Mamadi Diane. 2022. Kanjamadi – Kanjamadi for Nko. `https://web.archive.org/web/20231011145800/https://kanjamadi.org/baju/`. Accessed on 2023-10-11.

Diane, Baba Mamadi. 2021. Translation of the Meanings of the Noble Qur'an - N'ko

Translation. https://quranenc.com/en/browse/ankobambara_dayyan/1. Accessed on 2023-10-23.

Coleman Donaldson. 2017. *Clear Language: Script, Register and the N'ko Movement of Manding-Speaking West Africa*. Ph.D. thesis, University of Pennsylvania, Philadelphia, PA. Archived from the original on 2019-02-21. Retrieved 2019-02-21.

Coleman Donaldson. 2019. Linguistic and Civic Refinement in the N'ko Movement of Manding-Speaking West Africa. *Signs and Society*, 7:156 – 185.

Moussa Koulako Bala Doumbouya. 2022. Detransliterator. https://github.com/mdoumbouya/detransliterator.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig. 2023. Ethnologue: Languages of the world. Online version.

Christian Federmann. 2018. Appraise Evaluation Framework for Machine Translation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 86–88, Santa Fe, New Mexico. Association for Computational Linguistics.

Christian Federmann, Tom Kocmi, and Ying Xin. 2022. NTREX-128–news Test References for MT Evaluation of 128 Languages. In *Proceedings of the First Workshop on Scaling Up Multilingual Evaluation*, pages 21–24.

Firebase Documentation Google. 2023. Access data offline. https://firebase.google.com/docs/firestore/manage-data/enable-offline. Accessed on 2023-08-04.

International Center, Noor. 2018. Translation of the Meanings of the Noble Qur'an - French Translation. https://quranenc.com/en/browse/french_montada/1. Accessed on 2023-08-15.

International, Saheeh. 2022. Translation of the Meanings of the Noble Qur'an - English Translation. https://quranenc.com/en/browse/english_saheeh/1. Accessed on 2023-08-15.

Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al. 2017. Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Paul Amon Kotey. 1975. The Official Language Controversy: Indigenous versus Colonial.

Taku Kudo. 2018. Subword Regularization: Improving Neural Network Translation Models with Multiple Subword Candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75, Melbourne, Australia. Association for Computational Linguistics.

Minh-Thang Luong, Quoc V. Le, Ilya Sutskever, Oriol Vinyals, and Lukasz Kaiser. 2016. Multi-task Sequence to Sequence Learning. In *International Conference on Learning Representations*.

Manuel Mager, Arturo Oncevay, Abteen Ebrahimi, John Ortega, Annette Rios Gonzales, Angela Fan, Ximena Gutierrez-Vasques, Luis Chiruzzo, Gustavo Giménez-Lugo, Ricardo Ramos, et al. 2021. Findings of the Americasnlp 2021 Shared Task on Open Machine Translation for Indigenous Languages of the Americas. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 202–217.

MDN Contributors Mozilla. 2023a. CacheStorage - Web APIs | MDN. https://developer.mozilla.org/en-US/docs/Web/API/CacheStorage. Accessed on 2023-08-04.

MDN Contributors Mozilla. 2023b. Service Worker API - MDN Web Docs. https://developer.mozilla.org/en-US/docs/Web/API/Service_Worker_API. Accessed on 2023-08-04.

Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Taiwo Fagbohungbe, Solomon Oluwole Akinola, Shamsuddeen Muhammad, Salomon Kabongo Kabenamualu, Salomey Osei, Freshia Sackey, Rubungo Andre Niyongabo, Ricky Macharm, Perez Ogayo, Orevaoghene Ahia, Musie Meressa Berhe, Mofetoluwa Adeyemi, Masabata Mokgesi-Selinga, Lawrence Okegbemi, Laura Martinus, Kolawole Tajudeen, Kevin Degila, Kelechi Ogueji, Kathleen Siminyu, Julia Kreutzer, Jason Webster, Jamiil Toure Ali, Jade Abbott, Iroro Orife, Ignatius Ezeani, Idris Abdulkadir Dangana, Herman Kamper, Hady Elsahar, Goodness Duru, Ghollah Kioko, Murhabazi Espoir, Elan van Biljon, Daniel Whitenack, Christopher Onyefuluchi, Chris Chinenye Emezue, Bonaventure F. P. Dossou, Blessing Sibanda, Blessing Bassey, Ayodele Olabiyi, Arshath Ramkilowan, Alp Öktem, Adewale Akinfaderin, and Abdallah Bashir. 2020. Participatory Research for Low-Resourced Machine Translation: A Case Study in African Languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2144–2160, Online. Association for Computational Linguistics.

Djibril Tamsir Niane. 1974. Histoire et Tradition Historique du Manding. *Présence africaine*, (1):59–74.

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.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A Fast, Extensible Toolkit for Sequence Modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

Dianne Oyler. 2002. Re-Inventing Oral Tradition: The Modern Epic of Souleymane Kanté. *Research in African Literatures*, 33(1):75–93.

Narendra Babu Pamula, K Jairam, and B Rajesh. 2014. Cache-aside Approach for Cloud Design Pattern. *International Journal of Computer Science and Information Technologies*, 5(2):1423–1426.

Maja Popović. 2017. chrf++: Words Helping Character n-grams. In *Proceedings of the second conference on machine translation*, pages 612–618.

Matt Post. 2018. A Call for Clarity in Reporting BLEU Scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Ivan Provilkov, Dmitrii Emelianenko, and Elena Voita. 2020. BPE-dropout: Simple and Effective Subword Regularization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1882–1892, Online. Association for Computational Linguistics.

RFI. 2016. Mandenkan, la Vitalité d'une Langue. https://www.rfi.fr/fr/afrique/20161018-mandenkan-vitalite-une-langue. Accessed on 2023-10-11.

Tina Rosenberg. 2011. Everyone Speaks Text Message - The New York Times. https://www.nytimes.com/2011/12/11/magazine/everyone-speaks-text-message.html. Accessed on 2023-10-23].

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

SIL International & United Bible Societies. 2023. Paratext. https://paratext.org/. Accessed on 2023-10-17.

Yew Chong Soh, Ximena V. Del Carpio, and Liang Choon Wang. 2021. *The Impact of Language of Instruction in Schools on Student Achievement: Evidence from Malaysia using the Synthetic Control Method*. Policy Research Working Papers. The World Bank.

Unicode. 2023a. Unicode CLDR Project. https://cldr.unicode.org/. Accessed on 2023-06-15.

Unicode. 2023b. Unicode cldr project - acknowledgments. https://cldr.unicode.org/index/acknowledgments. Accessed on 2023-06-15.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. *Advances in neural information processing systems*, 30.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. 2020. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272.

Valentin Vydrin, Andrij Rovenchak, and Kirill Maslinsky. 2016. Maninka Reference Corpus: A Presentation. In *TALAf 2016: Traitement Automatique des Langues Africaines (Écrit et Parole). Atelier JEP-TALN-RECITAL 2016-Paris le.*

w3.org. 2022. Service Workers. https://www.w3.org/TR/service-workers/. Accessed on 2023-08-04.

Rachel Wicks and Kevin Duh. 2022. The Effects of Language Token Prefixing for Multilingual Machine Translation. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 148–153, Online only. Association for Computational Linguistics.

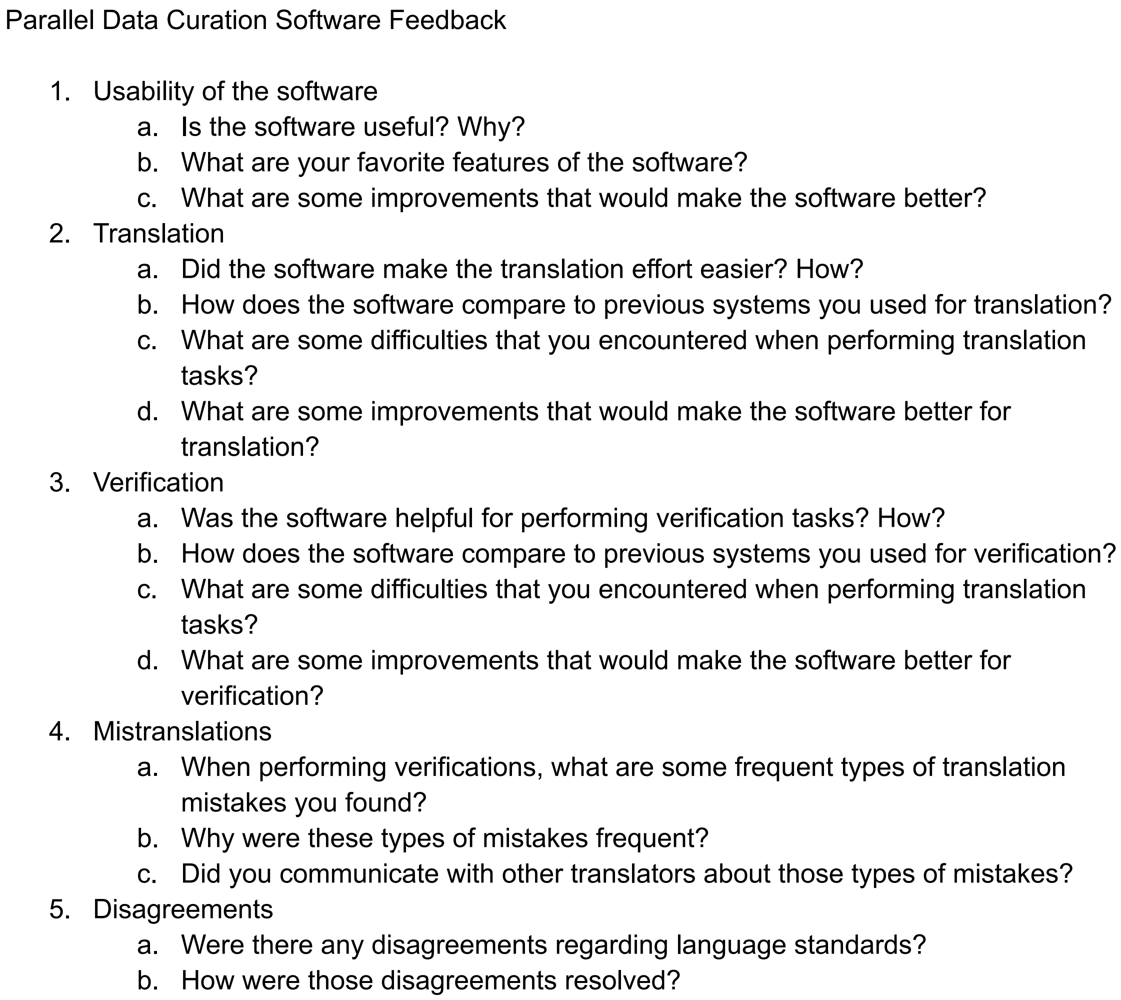
Wikimedia. 2023. List of Wikipedias. https://meta.wikimedia.org/wiki/List_of_Wikipedias. Accessed on 2023-10-14.

Mengzhou Xia, Xiang Kong, Antonios Anastasopoulos, and Graham Neubig. 2019. Generalized Data Augmentation for Low-Resource Translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, volume 57.

Appendices

A Fria||el User Study Feedback Questionnaire

The feedback questionnaire sent to N'Ko translators appears in Figure 6.



Parallel Data Curation Software Feedback

1. Usability of the software
    a. Is the software useful? Why?
    b. What are your favorite features of the software?
    c. What are some improvements that would make the software better?
2. Translation
    a. Did the software make the translation effort easier? How?
    b. How does the software compare to previous systems you used for translation?
    c. What are some difficulties that you encountered when performing translation tasks?
    d. What are some improvements that would make the software better for translation?
3. Verification
    a. Was the software helpful for performing verification tasks? How?
    b. How does the software compare to previous systems you used for verification?
    c. What are some difficulties that you encountered when performing translation tasks?
    d. What are some improvements that would make the software better for verification?
4. Mistranslations
    a. When performing verifications, what are some frequent types of translation mistakes you found?
    b. Why were these types of mistakes frequent?
    c. Did you communicate with other translators about those types of mistakes?
5. Disagreements
    a. Were there any disagreements regarding language standards?
    b. How were those disagreements resolved?

Figure 6: Survey questions sent to translators after they translated flores-200, nllb-seed, and ntrex-128 to N'Ko

B Fria||el Software Engineering Diagrams

On the next three pages appear:

- Workflow and task management sequence diagrams
- Workflow and Task State-Transition Diagrams
- Logical Data Model
- Physical Data Storage Model in Google Firestore

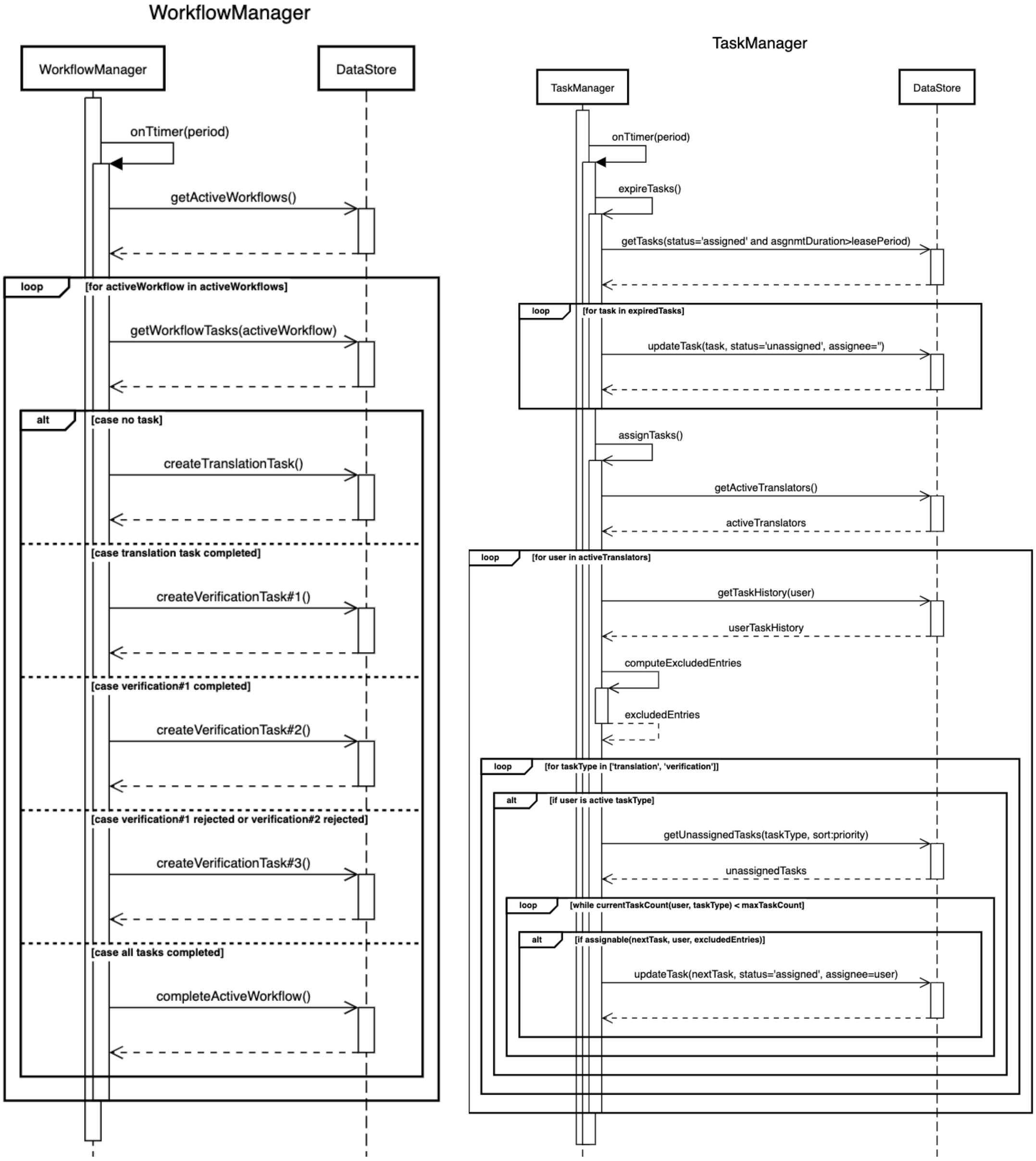

WorkflowManager
WorkflowManager
DataStore
onTtimer(period)
getActiveWorkflows()
loop
[for activeWorkflow in activeWorkflows]
getWorkflowTasks(activeWorkflow)
alt
[case no task]
createTranslationTask()
[case translation task completed]
createVerificationTask#1()
[case verification#1 completed]
createVerificationTask#2()
[case verification#1 rejected or verification#2 rejected]
createVerificationTask#3()
[case all tasks completed]
completeActiveWorkflow()
TaskManager
TaskManager
DataStore
onTtimer(period)
expireTasks()
getTasks(status='assigned' and asgnmtDuration>leasePeriod)
loop
[for task in expiredTasks]
updateTask(task, status='unassigned', assignee='')
assignTasks()
getActiveTranslators()
activeTranslators
loop
[for user in activeTranslators]
getTaskHistory(user)
userTaskHistory
computeExcludedEntries
excludedEntries
loop
[for taskType in ['translation', 'verification']]
alt
[if user is active taskType]
getUnassignedTasks(taskType, sort:priority)
unassignedTasks
loop
[while currentTaskCount(user, taskType) < maxTaskCount]
alt
[if assignable(nextTask, user, excludedEntries)]
updateTask(nextTask, status='assigned', assignee=user)

Figure 7: Sequence Diagram: Workflow Manager and Task Manager

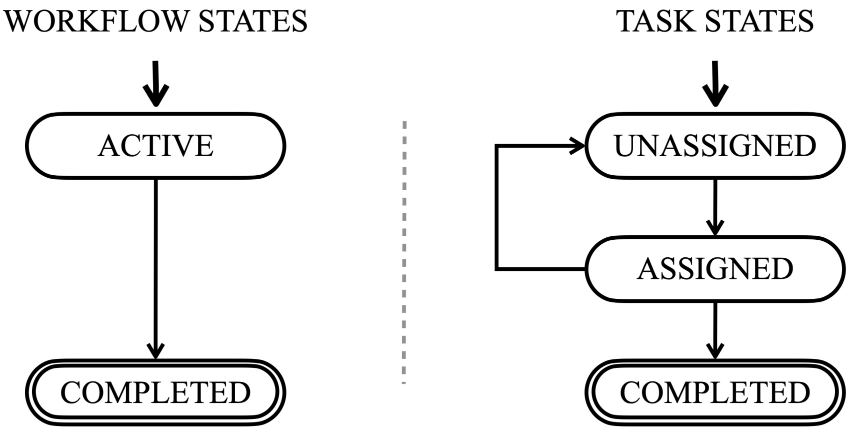
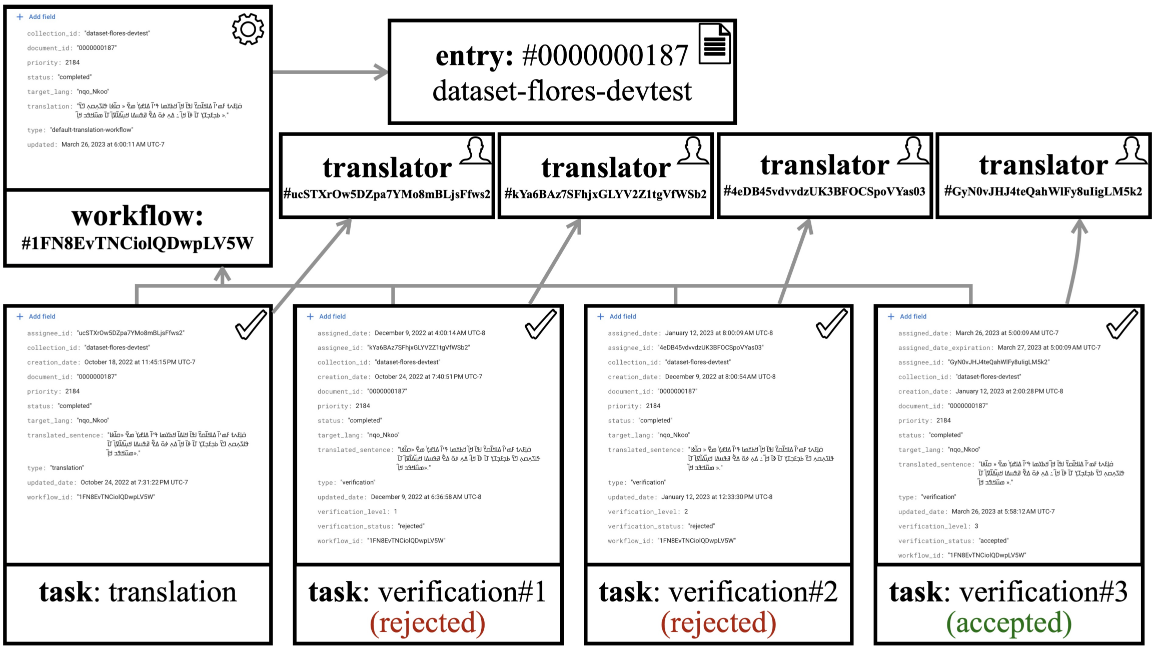


Figure 8: State Transition Diagrams for Tasks and Workflows. A translation workflow entity in the *active* state is created for each dataset entry. The workflow manager creates related *unassigned* tasks as needed, per the rules of the workflow. The Task Manager assigns tasks to users as appropriate. Uncompleted tasks are moved back to the *unassigned* status when not completed within the lease period. The workflow manager moves workflows to the *completed* status when all related tasks are *completed* and there is no need to create additional tasks.



```
collection_id: "dataset-flores-devtest"
document_id: "0000000187"
priority: 2184
status: "completed"
target_lang: "nqo_Nkoo"
translation: "[illegible]"
type: "default-translation-workflow"
updated: March 26, 2023 at 6:00:11 AM UTC-7
```

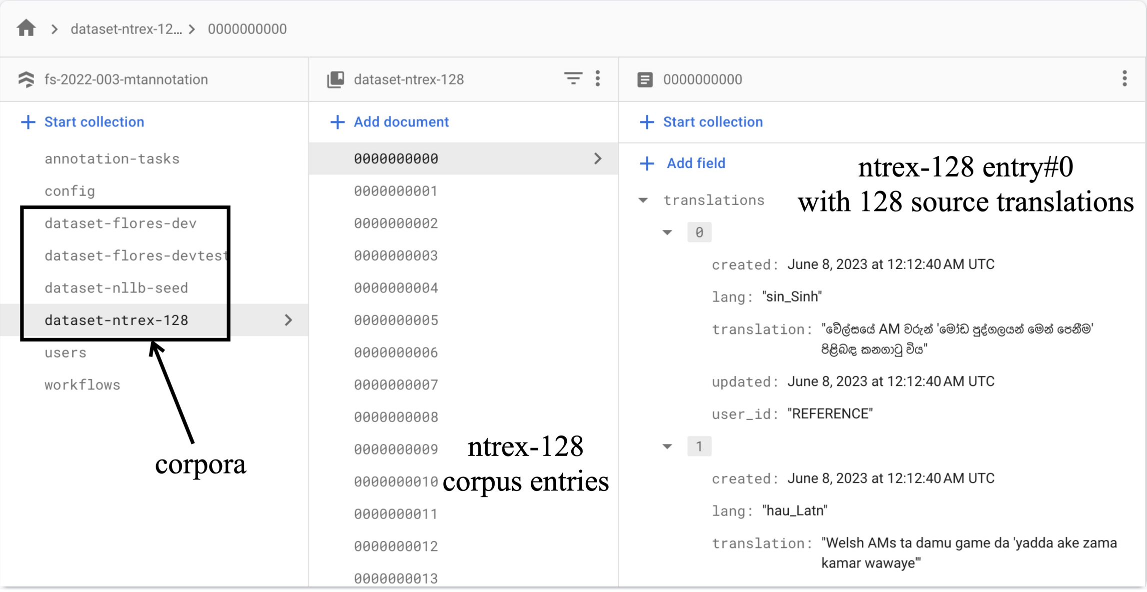
```
assignee_id: "ucSTXrOw5DZpa7YMo8mBLjsFfws2"
collection_id: "dataset-flores-devtest"
creation_date: October 18, 2022 at 11:45:15 PM UTC-7
document_id: "0000000187"
priority: 2184
status: "completed"
target_lang: "nqo_Nkoo"
translated_sentence: "[illegible]"
type: "translation"
updated_date: October 24, 2022 at 7:31:22 PM UTC-7
workflow_id: "1FN8EvTNCiolQDwpLV5W"
```

```
assigned_date: December 9, 2022 at 4:00:14 AM UTC-8
assignee_id: "kYa6BAz7SFhjxGLYV2Z1tgVfWSb2"
collection_id: "dataset-flores-devtest"
creation_date: October 24, 2022 at 7:40:51 PM UTC-7
document_id: "0000000187"
priority: 2184
status: "completed"
target_lang: "nqo_Nkoo"
translated_sentence: "[illegible]"
type: "verification"
updated_date: December 9, 2022 at 6:36:58 AM UTC-8
verification_level: 1
verification_status: "rejected"
workflow_id: "1FN8EvTNCiolQDwpLV5W"
```

```
assigned_date: January 12, 2023 at 8:00:09 AM UTC-8
assignee_id: "4eDB45vdvvdzUK3BFOCSpoVYas03"
collection_id: "dataset-flores-devtest"
creation_date: December 9, 2022 at 8:00:54 AM UTC-8
document_id: "0000000187"
priority: 2184
status: "completed"
target_lang: "nqo_Nkoo"
translated_sentence: "[illegible]"
type: "verification"
updated_date: January 12, 2023 at 12:33:30 PM UTC-8
verification_level: 2
verification_status: "rejected"
workflow_id: "1FN8EvTNCiolQDwpLV5W"
```

```
assigned_date: March 26, 2023 at 5:00:09 AM UTC-7
assigned_date_expiration: March 27, 2023 at 5:00:09 AM UTC-7
assignee_id: "GyN0vJHJ4teQahWlFy8uIigLM5k2"
collection_id: "dataset-flores-devtest"
creation_date: January 12, 2023 at 2:00:28 PM UTC-8
document_id: "0000000187"
priority: 2184
status: "completed"
target_lang: "nqo_Nkoo"
translated_sentence: "[illegible]"
type: "verification"
updated_date: March 26, 2023 at 5:58:12 AM UTC-7
verification_level: 3
verification_status: "accepted"
workflow_id: "1FN8EvTNCiolQDwpLV5W"
```

Figure 9: Logical model of entities involved in the curation process of entry#187 of the FLoRes-devtest dataset. Each entity is stored as a document in the Firestore database. The Workflow Manager created one translation task, and three verification tasks, each assigned to a different translator. The third verification task was created because at least one of the previous two resulted in translator edits. Arrows point from referencing to referenced documents.



annotation-tasks
config
dataset-flores-dev
dataset-flores-devtest
dataset-nllb-seed
dataset-ntrex-128
users
workflows

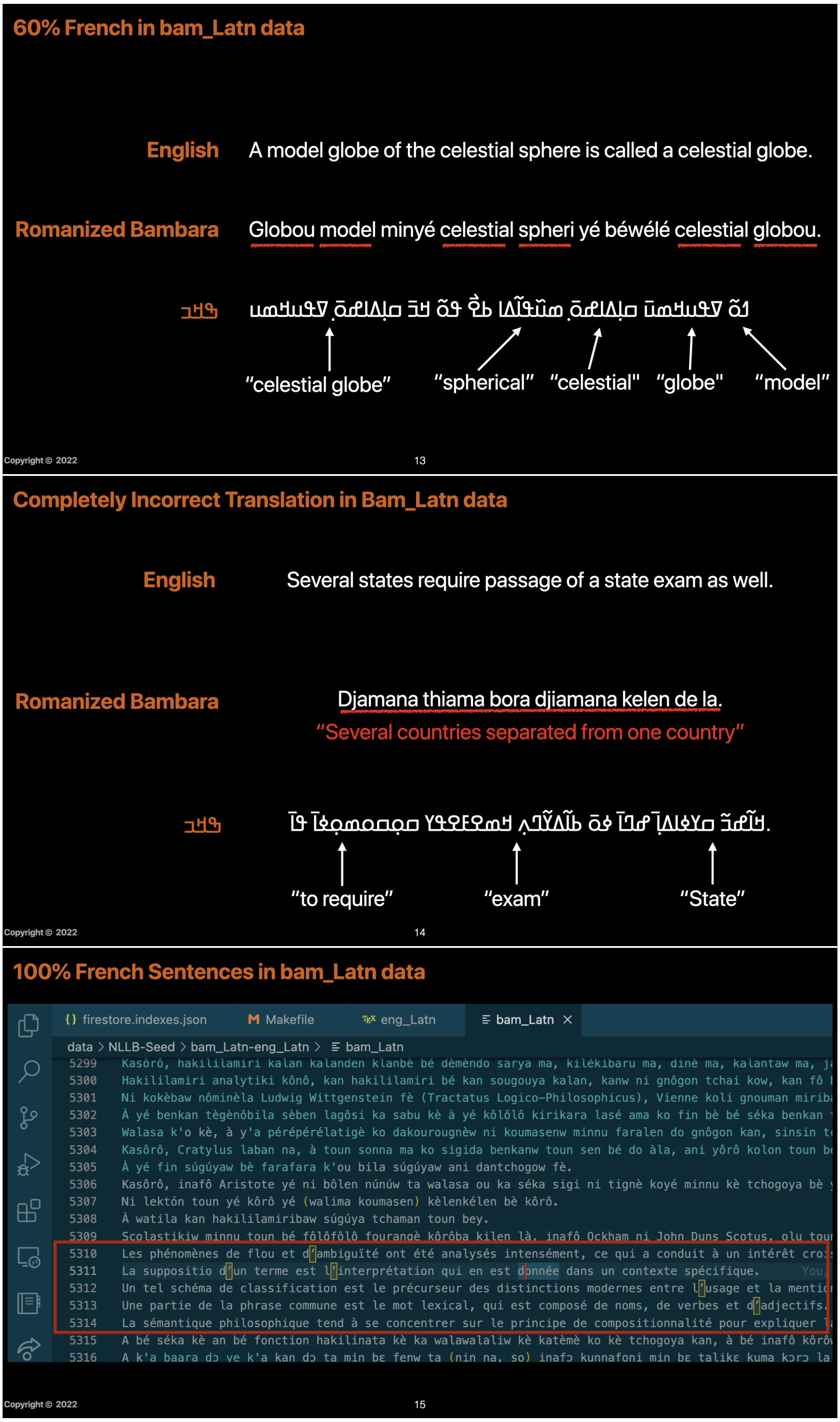
0000000000
0000000001
0000000002
0000000003
0000000004
0000000005
0000000006
0000000007
0000000008
0000000009
0000000010
0000000011
0000000012
0000000013

```
translations
  0
    created: June 8, 2023 at 12:12:40 AM UTC
    lang: "sin_Sinh"
    translation: "වේල්සයේ AM වරුන් 'මෝඩ පුද්ගලයන් මෙන් පෙනීම'
                 පිළිබඳ කනගාටු විය"
    updated: June 8, 2023 at 12:12:40 AM UTC
    user_id: "REFERENCE"
  1
    created: June 8, 2023 at 12:12:40 AM UTC
    lang: "hau_Latn"
    translation: "Welsh AMs ta damu game da 'yadda ake zama
                 kamar wawaye'"
```

Figure 10: Data Storage in Google Firestore. Each corpus is stored as a collection of documents (left), each of which is identified by its position in the original data files (middle). Each entry contains an array of source translations. Each translation is labeled with its language and script codes (ISO-639_ISO-15924) (right). The system also uses the *users*, *config*, *workflows* and *annotation-tasks* for user, configurations, and data curation workflow management.

C NLLB-SEED *bam_Latn* Quality Issues

Examples of quality issues in NLLB-SEED *bam_Latn* data file appear on the following page.



60% French in bam_Latn data

English A model globe of the celestial sphere is called a celestial globe.

Romanized Bambara Globou model minyé celestial spheri yé béwélé celestial globou.

Completely Incorrect Translation in Bam_Latn data

English Several states require passage of a state exam as well.

Romanized Bambara Djamana thiama bora djiamana kelen de la.

"Several countries separated from one country"

100% French Sentences in bam_Latn data

{} firestore.indexes.json
M Makefile
eng_Latn
bam_Latn ×
data > NLLB-Seed > bam_Latn-eng_Latn > bam_Latn

```
Kasôrô, hakililamiri kalan kalanden klanbè bé dèmèndo sarya ma, kilékibaru ma, dinè ma, kalantaw ma, ja
Hakililamiri analytiki kônô, kan hakililamiri bé kan sougouya kalan, kanw ni gnôgon tchai kow, kan fô l
Ni kokèbaw nôminèla Ludwig Wittgenstein fè (Tractatus Logico-Philosophicus), Vienne koli gnouman miriba
À yé benkan tègènôbila sèben lagôsi ka sabu kè à yé kôlôlô kirikara lasé ama ko fin bè bé séka benkan t
Walasa k'o kè, à y'a pérépérélatigè ko dakourougnèw ni koumasenw minnu faralen do gnôgon kan, sinsin to
Kasôrô, Cratylus laban na, à toun sonna ma ko sigida benkanw toun sen bé do àla, ani yôrô kolon toun be
À yé fin súgúyaw bè farafara k'ou bila súgúyaw ani dantchogow fè.
Kasôrô, inafô Aristote yé ni bôlen núnúw ta walasa ou ka séka sigi ni tignè koyé minnu kè tchogoya bè y
Ni lektón toun yé kôrô yé (walima koumasen) kèlenkélen bè kôrô.
À watila kan hakililamiribaw súgúya tchaman toun bey.
Scolastikiw minnu toun bé fôlôfôlô fourangè kôrôba kilen là. inafô Ockham ni John Duns Scotus. olu tour
Les phénomènes de flou et d'ambiguïté ont été analysés intensément, ce qui a conduit à un intérêt crois
La suppositio d'un terme est l'interprétation qui en est donnée dans un contexte spécifique.      You
Un tel schéma de classification est le précurseur des distinctions modernes entre l'usage et la mentio
Une partie de la phrase commune est le mot lexical, qui est composé de noms, de verbes et d'adjectifs.
La sémantique philosophique tend à se concentrer sur le principe de compositionnalité pour expliquer la
A bé séka kè an bé fonction hakilinata kè ka walawalaliw kè katèmè ko kè tchogoya kan, à bé inafô kôrô
A k'a baara dɔ ye k'a kan dɔ ta min bɛ fenw ta (nin na, so) inafɔ kunnafoni min bɛ talikɛ kuma kɔrɔ la
```

Figure 11: Examples of quality issues in NLLB-SEED *bam_Latn* data file. (1) Sentence with 60% borrowed French words. (2) Incorrect translation. (3) a block of sentences entirely in French. Notice that tonal marks are missing from *bam_Latn* text.

D nicolingua-0005 details

This section provides details on the monolingual, bilingual and trilingual parallel corpora, donated by Nko community members, collectively making up the nicolingua-0005 corpus.

D.0.1 Trilingual Corpora $(nqo\_Nkoo, eng\_Latn, fra\_Latn)$

baba_mamadi_diane_parallel_002 This corpus is composed of parallel Quran translations in Nko (Diane, Baba Mamadi, 2021), English (International, Saheeh, 2022), and French (International Center, Noor, 2018). The Quran's translation in Nko was originally performed by Baba Mamadi Diane for Islamic education purposes.

kalo_mory_diane_parallel_00{1,2,3} This corpus contains various short phrases collected and translated by Kalo Mory Diane for the purpose of machine translation system development.

solo_farabado_cisse_parallel_002 This corpus contains various short phrases collected and translated by Solo Farabado Cisse for the purpose of machine translation system development.

solo_farabado_cisse_parallel_001 Nko localization strings from the Unicode Common Locale Data Repository (CLDR) (Unicode, 2023a) to which Solo Farabado Cisse and Baba Mamadi Diane contributed (Unicode, 2023b). Corresponding CLDR strings in Nko, English, and French were compiled to make this trilingual parallel corpus.

D.0.2 Bilingual Corpora (nqo_Nkoo, eng_Latn)

baba_mamadi_diane_parallel_003 This corpus contains segments manually chunked from the Quran and translated by Baba Mamadi Diane specifically for the purpose of creating a corpus usable for machine translation system development.

baba_mamadi_diane_parallel_004 This corpus contains the localization strings of a custom Android build translated by Baba Mamadi Diane.

djibrila_diane_parallel_003 This corpus contains short phrases collected and translated by Djibrila Diane. The phrases also include some basic scientific terminology. The corpus was originally created for education purposes only.

djibrila_diane_parallel_001 This corpus contains short phrases in various tenses collected and translated by Djibrila Diane to serve of MT system development.

djibrila_diane_parallel_002 This corpus contains various short phrases composed and translated by Djibrila Diane for the purpose of MT system development.

D.0.3 Bilingual Corpora (nqo_Nkoo, fra_Latn)

baba_mamadi_diane_parallel_001 Nko-French dictionary authored by Baba Mamadi Diane for education purposes. Dictionary entries in french with multiple forms (e.g. gender) were automatically expanded using regular expressions.

nafadji_sory_conde_parallel_001 This corpus contains various short phrases composed and translated by Nafadji Sory Conde for the purpose of machine translation system development.

nafadji_sory_conde_parallel_003 This corpus contains phrases from Camara Laye's 1953 novel "L'enfant Noir" (Camara, 1953). The translation was originally done by Nafadji Sory Conde for the purpose of expanding available literature in Nko.

nafadji_sory_conde_parallel_002 This corpus contains various phrases related to Guinean society and sociology. It was created by Nafadji Sory Conde for the purpose of MT system development.

nafadji_sory_conde_parallel_004 This corpus contains segments extracted from the Guinean constitution. It was originally translated by Nafaji Sory Conde for education purposes.

D.0.4 Monolingual Corpora (nqo_Nkoo)

nafadji_sory_conde_monolingual_001 This corpus, composed by Nafadji Sory Conde and his collaborators, contains extracts of books and newspapers in Nko. A substantial part of the corpus was harvested from Kanjamadi.com. This corpus may overlap with the Maninka Reference Corpus (Vydrin et al., 2016).

baba_mamadi_diane_monolingual_00{1,2} These corpora were extracted from various Nko books and articles in various domains including history, religion, philosophy, literature and Science. The corpora were originally composed by Baba M. Diane for the purpose of auto-completion algorithm development for Nko.

lines	words	file	originator	description
6236	175382	baba_mamadi_diane_parallel_002.nqo_Nkoo	Baba Mamadi Diane	Traductions of the Quran
6236	151323	baba_mamadi_diane_parallel_002.eng_Latn		
6236	171085	baba_mamadi_diane_parallel_002.fra_Latn		
7001	28626	kalo_mory_diane_parallel_001.nqo_Nkoo	Kalo Mory Diane	Short Phrases
7001	17558	kalo_mory_diane_parallel_001.eng_Latn		
7001	21593	kalo_mory_diane_parallel_001.fra_Latn		
4001	18864	kalo_mory_diane_parallel_003.nqo_Nkoo	Kalo Mory Diane	Short Phrases
4001	12891	kalo_mory_diane_parallel_003.eng_Latn		
4001	15050	kalo_mory_diane_parallel_003.fra_Latn		
3999	17903	kalo_mory_diane_parallel_002.nqo_Nkoo	Kalo Mory Diane	Short Phrases
3999	12237	kalo_mory_diane_parallel_002.eng_Latn		
3999	14495	kalo_mory_diane_parallel_002.fra_Latn		
3052	13420	solo_farabado_cisse_parallel_002.nqo_Nkoo	Solo Farabado Cisse	Short Phrases
3052	9615	solo_farabado_cisse_parallel_002.eng_Latn		
3052	11308	solo_farabado_cisse_parallel_002.fra_Latn		
1559	2739	solo_farabado_cisse_parallel_001.nqo_Nkoo	Solo Farabado Cisse	Unicode CLDR Strings
1559	2382	solo_farabado_cisse_parallel_001.eng_Latn		
1559	2338	solo_farabado_cisse_parallel_001.fra_Latn		

Table 7: nicolingua-0005's trilingual subsets in Nko (*nqo_Nkoo*), English (*eng_Latn*) and French (*fra_Latn*)

lines	words	file	originator	description
21590	154238	baba_mamadi_diane_parallel_003.nqo_Nkoo	Baba Mamadi Diane	Segments Chunked from the Quran
21590	133369	baba_mamadi_diane_parallel_003.eng_Latn		
36211	119536	baba_mamadi_diane_parallel_004.nqo_Nkoo	Baba Mamadi Diane	Localization Strings for a Custom Android Build
36211	72612	baba_mamadi_diane_parallel_004.eng_Latn		
492	4666	djibrila_diane_parallel_003.nqo_Nkoo	Djibrila Diane	Various Short Phrases and Basic Sci. Terms
492	4122	djibrila_diane_parallel_003.eng_Latn		
1001	3536	djibrila_diane_parallel_001.nqo_Nkoo	Djibrila Diane	Short Phrases in Various Tenses
1001	3487	djibrila_diane_parallel_001.eng_Latn		
148	1303	djibrila_diane_parallel_002.nqo_Nkoo	Djibrila Diane	Various Short Phrases
148	1361	djibrila_diane_parallel_002.eng_Latn		

Table 8: nicolingua-0005's bilingual subsets in Nko (*nqo_Nkoo*) and English (*eng_Latn*)

lines	words	file	originator	description
37894	40436	baba_mamadi_diane_parallel_001.nqo_Nkoo	Baba Mamadi Diane	Nko-Francais Dictionary
37894	41598	baba_mamadi_diane_parallel_001.fra_Latn		
3604	39020	nafadji_sory_conde_parallel_001.nqo_Nkoo	Nafadji Sory Conde	Various Short Phrases
3604	35037	nafadji_sory_conde_parallel_001.fra_Latn		
1141	22379	nafadji_sory_conde_parallel_003.nqo_Nkoo	Nafadji Sory Conde	Segment from "L'enfant Noir"
1141	21049	nafadji_sory_conde_parallel_003.fra_Latn		
2200	16091	nafadji_sory_conde_parallel_002.nqo_Nkoo	Nafadji Sory Conde	Phrases related to Guinean Society and Sociology
2200	15413	nafadji_sory_conde_parallel_002.fra_Latn		
721	11863	nafadji_sory_conde_parallel_004.nqo_Nkoo	Nafadji Sory Conde	Guinean Constitution
721	11345	nafadji_sory_conde_parallel_004.fra_Latn		

Table 9: nicolingua-0005's bilingual subsets in Nko (*nqo_Nkoo*) and French (*fra_Latn*)

lines	words	file	originator	description
134000	2017158	nafadji_sory_conde_monolingual_001.nqo_Nkoo	Nafadji Sory Conde	Various Books and News Papers
44604	853464	baba_mamadi_diane_monolingual_002.nqo_Nkoo	Baba Mamadi Diane	Various Books and Articles
10195	420749	baba_mamadi_diane_monolingual_001.nqo_Nkoo	Baba Mamadi Diane	Various Books and Articles

Table 10: nicolingua-0005's monolingual subsets in Nko (*nqo_Nkoo*)

E Datasheet Questionnaire for *nicolingua-0005*

E.1 Motivation

E.1.1 Who created the dataset(e.g., which team, research group) and on behalf of which entity (e.g. company, institution, organization)?

nicolingua-0005 was curated by Moussa Doumbouya (Stanford University). Its constituent corpora were provided by the following members of Nko USA Inc: Baba Mamadi Diane, Solo Farabado Cisse, Djibrila Diane, Nafadji Sory Conde, Kalo Mory Diane.

E.1.2 Did they fund it themselves? If there is an associated grant, please provide the name of the grantor and the grant name and number.

Nko community members voluntarily composed the included corpora.

E.1.3 For what purpose was the data set created? Was there a specific task in mind? If so, please specify the result type (e.g. unit) to be expected.

Some included corpora were composed specifically for the development of MT systems while others were originally created for educational purposes. See Appendix D for details.

E.1.4 Could any of these uses, or their results, interfere with human will or communicate a false reality?

Not to the best of our knowledge.

E.1.5 What is the antiquity of the file? Provide, please, the current date.

July 19 2023.

E.1.6 Has there been any monetary profit from the creation of this dataset?

No.

E.2 Composition

E.2.1 Is there any synthetic data in the dataset? If so, in what percentage?

The corpus doesn't contain any synthetic data.

E.2.2 Are there multiple types of instances or is there just one type? Please specify the type(s), e.g. Raw data, preprocessed, symbolic.

The corpus contains monolingual and parallel text corpora.

E.2.3 What do the instances (of each type, if appropriate) that comprise the data set represent? (e.g. documents, photos, people, countries).

The instances represent segments of text in Nko, English, and French.

E.2.4 How many instances (of each type, if appropriate) are there in total?

See Tables 4, 7, 8, 9 and 10

E.2.5 Does the dataset contain all possible instances or is it just a sample of a larger set? i.e. Is the dataset different than an original one due to the preprocessing process? In case this dataset is a subset of another one, is the original dataset available?

This dataset is a collection of corpora from various sources. Some sources were integrally sampled (e.g. quran), while other sources were composed by individual translators.

E.2.6 Is there a label or a target associated with each of the instances? If so, please provide a description.

The multilingual subsets of the corpora are matching segments of text in multiple languages.

E.2.7 What is the format of the data? e.g. .json, .xml, .csv .

The files are text files encoded in UTF-8 that have the following extensions matching the iso standard code of the language and writing system they contain: .nqo_Nkoo, .eng_Latn, .fra_Latn.

E.2.8 Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g. because it was unavailable). This does not include intentionally removed information, but might include, e.g. redacted text.

There is no missing information to report.

E.2.9 Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description. Do not include missing information here.

The sentences were benevolently translated by various individuals. A minimal quality control process was adopted during the curation phase. The data may contain some errors. The corpus baba_mamadi_diane_parallel_003 was created by sampling Quran phrases from baba_mamadi_diane_parallel_002. Some parallel Nko segments may be repeated in the monolingual Nko corpora.

E.2.10 Is there any verification that guarantees there is not institutionalization of unfair biases? Both regarding the dataset itself and the potential algorithms that could use it.

no.

E.2.11 Are there recommended data splits, e.g. training, development/validation, testing? If so, please provide a description of these splits explaining the rationale behind them.

The corpora are intended to be used to train natural language processing algorithms.

E.2.12 Is the dataset self-contained, or does it link to or otherwise rely on external resources? e.g., websites, tweets, other datasets. If it links to or relies on external resources, a) Are there any guarantees that they will exist, and remain constant over time? b) Are there official archival versions of the complete dataset? i.e. including the external resources as they existed at the time the dataset was created. c) Are there any restrictions (e.g. licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, if appropriate.

nicolingua-0005 is self-contained.

E.2.13 Does the dataset contain data that might be considered confidential? e.g. data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals non-public communications. If so, please provide a description.

Not to the best of our knowledge.

E.2.14 Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.

Not to the best of our knowledge. Notes: (1) *nicolingua-0005* contains religious text that some people may find offensive or threatening. (2) Some words contained in *nicolingua-0005*, such as the name of certain human body parts included in the Nko-Francais dictionary, may be considered vulgar or offensive.

E.2.15 Does the dataset relate to people? If so, please specify a) Whether the dataset identifies subpopulations or not. b) Whether the dataset identifies indivual people or not. c) Whether it contains information that could vulnerate any individuals or their rights. c) Any other verified information on the topic that can be provided.

The data includes news articles that may reference specific people and people groups. The data also includes literature relating to West African people and people groups and their history.

E.2.16 Does the dataset cover included languages equally?

No. The sizes of various parallel and monolingual subsets have been specified in Table 4.

E.2.17 Is there any evidence that the data may be somehow biased? i.e. towards gender, ethics, beliefs.

The data includes religious texts, articles, and books that may reflect various types of biases. The data may contain biases inherent in historical and current Manding culture such as work organization between men and women, young and old people. Nko doesn't have masculine vs. feminine noun classes. Therefore genders are not distinguished in

Nko nouns and pronouns, which may reduce the potential for gender-based bias.

E.2.18 Is the data made up of formal text, informal text or both equitably?

The data mostly contains formal text.

E.2.19 Does the data contain incorrect language expressions on purpose? Does it contain slang terms? If that's the case, please provide which instances of the data correspond to these.

Not to the best of our knowledge. The dataset may contain unintentional errors.

E.3 Collection Process

E.3.1 Where was the data collected at? Please include as much detail; i.e. country, city, community, entity and so on.

Most data was collected in Conakry, Guinea, and Banakoro, Guinea. Some contributors also worked in Bamako, Mali (Solo F Cisse, Baba M Diane), Egypt (Baba M Diane) and USA (Djibrila Diane) while collecting the datasets.

E.3.2 If the dataset is a sample from a larger set, what was the sampling strategy? i.e. deterministic, probabilistic with specific sampling probabilities.

N/A

E.3.3 Are there any guarantees that the acquisition of the data did not violate any law or anyone's rights?

Not to the best of our knowledge.

E.3.4 Are there any guarantees that prove the data is reliable?

No.

E.3.5 Did the collection process involve the participation of individual people? If so, please report any information available regarding the following questions: Was the data collected from people directly? Did all the involved parts give their explicit consent? Is there any mechanism available to revoke this consent in the future, if desired?

The dataset authors are authors of this paper. They gave their explicit consent.

E.3.6 Has an analysis of the potential impact of the dataset and its use on data subjects been conducted? i.e. a data protection impact analysis. If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.

No.

E.3.7 Were any ethical review processes conducted?

No.

E.3.8 Does the data come from a single source or is it the result of a combination of data coming from different sources? In any case, please provide references.

The data was curated from a combination of different sources.

E.3.9 If the same content was to be collected from a different source, would it be similar?

Not Applicable.

E.3.10 Please specify any other information regarding the collection process. i.e. Who collected the data, whether they were compensated or not, what mechanisms were used. Please, only include if verified.

E.4 Preprocessing/Cleaning/Labelling

E.4.1 Please specify any information regarding the preprocessing that you may know (e.g. the person who created the dataset has somehow explained it) or be able to find (e.g. there exists and informational site). Please, only include if verified. i.e. Was there any mechanism applied to obtain a neutral language? Were all instances preprocessed the same way?

The data was normalized with Unicode normalization form NFC: Canonical Decomposition followed by Canonical Composition. Non-Nko characters were stripped from monolingual Nko text. Extra punctuations were removed from some sources. Some entries in Baba Mamadi Diane's Nko-Francais dictionary were expanded using regular expressions so that separate forms of the same words (e.g. gendered, plural) were repeated as separate entries.

E.5 Uses

E.5.1 Has the dataset been used already? If so, please provide a description.

The data was used to build baseline neural machine translation algorithms for Nko. See Section 4.

E.5.2 Is there a repository that links to any or all papers or systems that use this dataset? If so, please provide a link or any other access point.

https://github.com/mdoumbouya/nicolingua-0005-nqo-nmt-resources https://github.com/mdoumbouya/nicolingua-0005-nqo-nmt-resources

E.5.3 What (other) tasks could the dataset be used for? Please include your own intentions, if any.

Any natural language processing tasks including language modeling and machine translation.

E.5.4 Are there tasks for which the dataset should not be used? If so, please provide a description.

Not to the best of our knowledge.

E.6 Distribution

E.6.1 Please specify the source where you got the dataset from.

The datasets came from the following individuals:

E.6.2 When was the dataset first released?

July 19 2023.

E.6.3 Are there any restrictions regarding the distribution and/or usage of this data in any particular geographic regions?

No.

E.6.4 Is the dataset distributed under a copyright or other intellectual property (IP) license? And/or under applicable terms of use (ToU)? Please cite a verified source.

The dataset is openly available under the Attribution-ShareAlike 4.0 International (CC BY-SA 4.0) license.

E.7 Maintenance

E.7.1 Is there any verified manner of contacting the creator of the dataset?

The authors of this paper can be contacted via email.

E.7.2 Specify any limitations there might be to contributing to the dataset. i.e. Can anyone contribute to it? Can someone do it at all?

The dataset is openly available under the Attribution-ShareAlike 4.0 International (CC BY-SA 4.0) license.

E.7.3 Has any erratum been notified?

No.

E.7.4 Is there any verified information on whether the dataset will be updated in any form in the future? Is someone in charge of checking if any of the data has become irrelevant throughout time? If so, will it be removed or labeled somehow?

The dataset will be maintained on GitHub. Any updates will be made available in the same GitHub repository.

E.7.5 Is there any available log about the changes performed previously in the dataset?

Any future modifications will be tracked in GitHub's version control.

E.7.6 Could changes to current legislation end the right-of-use of the dataset?

Not to the best of our knowledge.

E.7.7 Are there any lifelong learning updates, such as vocabulary enrichment, automatically developed?

No.

F Train, Valid and Test Subset Details

Details on the training, validation, and test subset composition for each model appear on the following page.

TRAIN

lines	words	file	200	201	202-9
6193	148442	common-parallel-corpora/multitext-nllb-seed/bam_Latn		✓	✓
6193	136157	cpc/multitext-nllb-seed/eng_Latn	✓	✓	✓
6193	184138	cpc/multitext-nllb-seed/nqo_Nkoo	✓	✓	✓
6236	151323	nicolingua-0005/baba_mamadi_diane_parallel_002.eng_Latn	✓	✓	✓
6236	171085	nicolingua-0005/baba_mamadi_diane_parallel_002.fra_Latn		✓	✓
6236	175382	nicolingua-0005/baba_mamadi_diane_parallel_002.nqo_Nkoo	✓	✓	✓
7001	17558	nicolingua-0005/kalo_mory_diane_parallel_001.eng_Latn	✓	✓	✓
7001	21593	nicolingua-0005/kalo_mory_diane_parallel_001.fra_Latn		✓	✓
7001	28626	nicolingua-0005/kalo_mory_diane_parallel_001.nqo_Nkoo	✓	✓	✓
4001	12891	nicolingua-0005/kalo_mory_diane_parallel_003.eng_Latn	✓	✓	✓
4001	15050	nicolingua-0005/kalo_mory_diane_parallel_003.fra_Latn		✓	✓
4001	18864	nicolingua-0005/kalo_mory_diane_parallel_003.nqo_Nkoo	✓	✓	✓
3999	12237	nicolingua-0005/kalo_mory_diane_parallel_002.eng_Latn	✓	✓	✓
3999	14495	nicolingua-0005/kalo_mory_diane_parallel_002.fra_Latn		✓	✓
3999	17903	nicolingua-0005/kalo_mory_diane_parallel_002.nqo_Nkoo	✓	✓	✓
3052	9615	nicolingua-0005/solo_farabado_cisse_parallel_002.eng_Latn	✓	✓	✓
3052	11308	nicolingua-0005/solo_farabado_cisse_parallel_002.fra_Latn		✓	✓
3052	13420	nicolingua-0005/solo_farabado_cisse_parallel_002.nqo_Nkoo	✓	✓	✓
1559	2382	nicolingua-0005/solo_farabado_cisse_parallel_001.eng_Latn	✓	✓	✓
1559	2338	nicolingua-0005/solo_farabado_cisse_parallel_001.fra_Latn		✓	✓
1559	2739	nicolingua-0005/solo_farabado_cisse_parallel_001.nqo_Nkoo	✓	✓	✓
21590	133369	nicolingua-0005/baba_mamadi_diane_parallel_003.eng_Latn	✓	✓	✓
21590	154238	nicolingua-0005/baba_mamadi_diane_parallel_003.nqo_Nkoo	✓	✓	✓
36211	72612	nicolingua-0005/baba_mamadi_diane_parallel_004.eng_Latn	✓	✓	✓
36211	119536	nicolingua-0005/baba_mamadi_diane_parallel_004.nqo_Nkoo	✓	✓	✓
1001	3487	nicolingua-0005/djibrila_diane_parallel_001.eng_Latn	✓	✓	✓
1001	3536	nicolingua-0005/djibrila_diane_parallel_001.nqo_Nkoo	✓	✓	✓
148	1361	nicolingua-0005/djibrila_diane_parallel_002.eng_Latn	✓	✓	✓
148	1303	nicolingua-0005/djibrila_diane_parallel_002.nqo_Nkoo	✓	✓	✓
492	4122	nicolingua-0005/djibrila_diane_parallel_003.eng_Latn	✓	✓	✓
492	4666	nicolingua-0005/djibrila_diane_parallel_003.nqo_Nkoo	✓	✓	✓
37894	41598	nicolingua-0005/baba_mamadi_diane_parallel_001.fra_Latn		✓	✓
37894	40436	nicolingua-0005/baba_mamadi_diane_parallel_001.nqo_Nkoo		✓	✓
3604	35037	nicolingua-0005/nafadji_sory_conde_parallel_001.fra_Latn		✓	✓
3604	39020	nicolingua-0005/nafadji_sory_conde_parallel_001.nqo_Nkoo		✓	✓
2200	15413	nicolingua-0005/nafadji_sory_conde_parallel_002.fra_Latn		✓	✓
2200	16091	nicolingua-0005/nafadji_sory_conde_parallel_002.nqo_Nkoo		✓	✓
1141	21049	nicolingua-0005/nafadji_sory_conde_parallel_003.fra_Latn		✓	✓
1141	22379	nicolingua-0005/nafadji_sory_conde_parallel_003.nqo_Nkoo		✓	✓
721	11345	nicolingua-0005/nafadji_sory_conde_parallel_004.fra_Latn		✓	✓
721	11863	nicolingua-0005/nafadji_sory_conde_parallel_004.nqo_Nkoo		✓	✓
134000	2017158	nicolingua-0005/nafadji_sory_conde_monolingual_001.nqo_Nkoo			✓
10195	420749	nicolingua-0005/baba_mamadi_diane_monolingual_001.nqo_Nkoo			✓
44604	853464	nicolingua-0005/baba_mamadi_diane_monolingual_002.nqo_Nkoo			✓

Table 11: Data files included in the training set of each model family

VALID

lines	words	file	200	201	202-9
997	21565	common-parallel-corpora/flores-200-dev/bam_Latn.dev		✓	✓
997	20954	common-parallel-corpora/flores-200-dev/eng_Latn.dev	✓	✓	✓
997	23957	common-parallel-corpora/flores-200-dev/fra_Latn.dev		✓	✓
997	27361	common-parallel-corpora/flores-200-dev/nqo_Nkoo.dev	✓	✓	✓

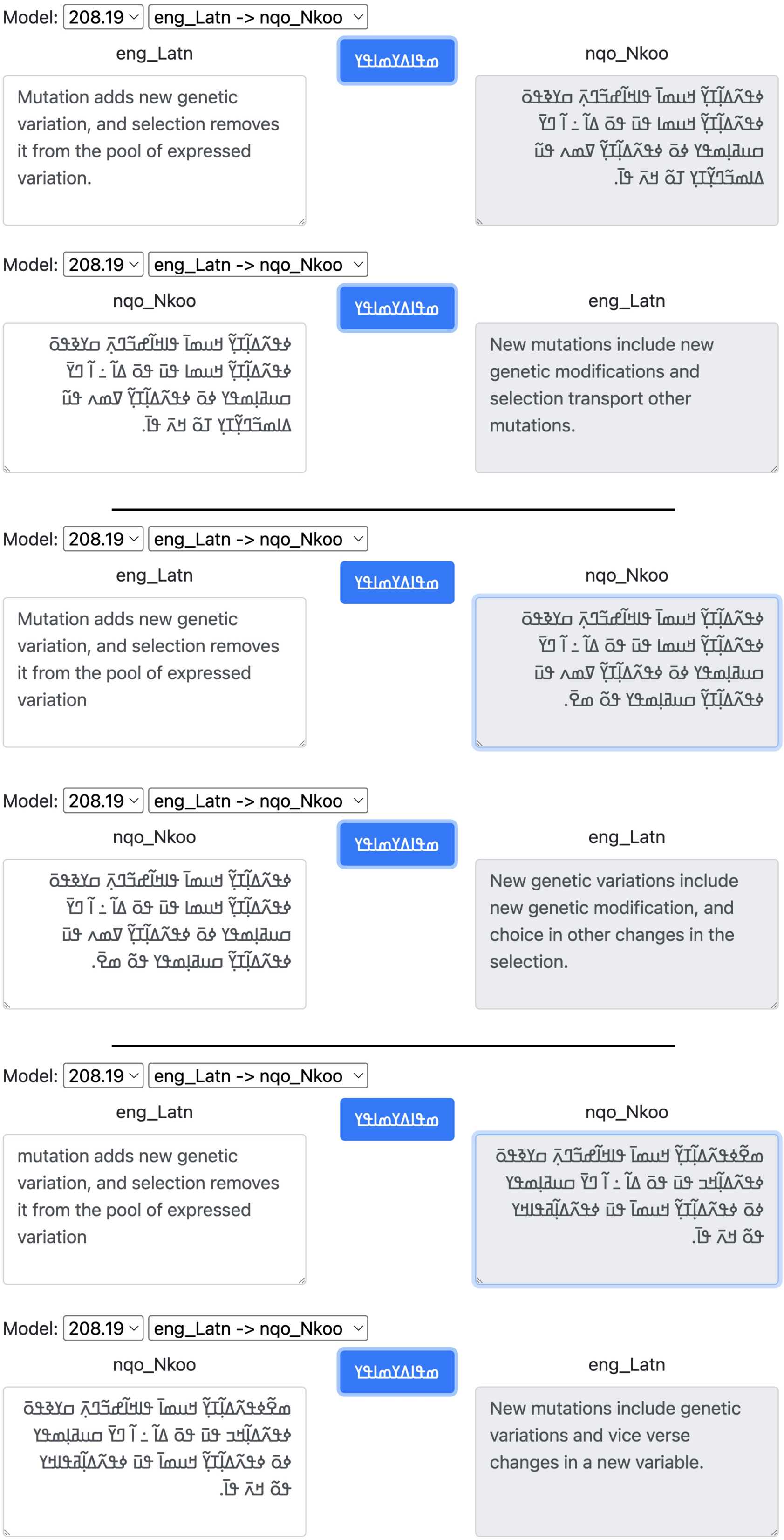
TEST

lines	words	file	200	201	202-9
1012	22565	common-parallel-corpora/flores-200-devtest/bam_Latn.devtest		✓	✓
1012	21901	common-parallel-corpora/flores-200-devtest/eng_Latn.devtest	✓	✓	✓
1012	25319	common-parallel-corpora/flores-200-devtest/fra_Latn.devtest		✓	✓
1012	29503	common-parallel-corpora/flores-200-devtest/nqo_Nkoo.devtest	✓	✓	✓

Table 12: Data files included in the validation and test sets of each model family

G Examples of Translations

Examples of generations highlighting the sensitivity our ouf baseline NMT system to punctuation and case appear on the following page.



Model: 208.19 eng_Latn -> nqo_Nkoo

eng_Latn

Mutation adds new genetic variation, and selection removes it from the pool of expressed variation.

nqo_Nkoo

[illegible]

Model: 208.19 eng_Latn -> nqo_Nkoo

nqo_Nkoo

[illegible]

eng_Latn

New mutations include new genetic modifications and selection transport other mutations.

Model: 208.19 eng_Latn -> nqo_Nkoo

eng_Latn

Mutation adds new genetic variation, and selection removes it from the pool of expressed variation

nqo_Nkoo

[illegible]

Model: 208.19 eng_Latn -> nqo_Nkoo

nqo_Nkoo

[illegible]

eng_Latn

New genetic variations include new genetic modification, and choice in other changes in the selection.

Model: 208.19 eng_Latn -> nqo_Nkoo

eng_Latn

mutation adds new genetic variation, and selection removes it from the pool of expressed variation

nqo_Nkoo

[illegible]

Model: 208.19 eng_Latn -> nqo_Nkoo

nqo_Nkoo

[illegible]

eng_Latn

New mutations include genetic variations and vice verse changes in a new variable.

Figure 12: A sentence from the FLoRes-200-devtest corpus translated from English to Nko and back-translated to English using model 208.19. The three examples highlight the sensitivity of our baseline system to punctuation and case. Top: original sentence; Middle: removed final period; Bottom: removed initial capitalization and final period.