BAOULÉ RELATED PARALLEL CORPORA FOR MACHINE TRANSLATION TASKS: Mtbci-1.0

Kouassi Konan Jean-Claude

Faculty of Computer Science via distance learning, Bircham International University, Madrid, Spain

ABSTRACT

According to the Ethnologue platform, we have 7,164 known living languages in the World, and not all of them have data available over the internet to facilitate Artificial Intelligence (AI) tasks such as Machine Translation (MT). Consequently, there is a need for thorough Data Engineering tasks for most of these languages. Especially, the Baoulé living language normalized as ISO 639-3 (bci) is not yet supported on popular worldwide free translation platform such as Microsoft Translator, nor on the Official Wikipedias. In this paper, we have proposed the "Baoulé Related Parallel Corpora for Machine Translation tasks: mtBCI-1.0" to make parallel Baoulé-related datasets available to the scientific community for AI tasks implying Machine Translation. We have shown that, after a brief presentation of the Baoulé language in the proposed approach, we will focus on the Data Engineering Process itself before providing a baseline proving that the collected data is of scientific interest.

KEYWORDS

Artificial Intelligence, Machine Learning, Machine Translation, Data Engineering, Dataset, Parallel Corpora, Baoulé language (bci)

1. INTRODUCTION

In their book "*Artificial Intelligence: A Modern Approach*", professors Stuart J. Russell and Peter Norvig revealed an important aspect of AI (thus including MT): Data Engineering (obtaining, cleaning, and representing data) is at least as important as Algorithm Engineering (cf. [1], introduction of section 18.11 and section 18.11.2), i.e., despite their equal value, in some situations Data Engineering importance could surpass those of Algorithm Engineering.

Indeed, in practical applications of AI, the data set is usually large, multidimensional, and messy. So, the expert (i.e., the Data Engineer) needs to perform Data Engineering tasks that consists first in acquiring the right data by deciding what type of data are necessary to accomplish a specific task that solves a given AI problem. Then, once the right data is found, the Data Engineer, via a data processing concept called Extract-Transform-Load (ETL), cleans the data (data reorganization or filtering) and proposes a representation by deciding the best structure of data (or features) to use during the processing. Therefore, the data design strongly depends on the specific tasks (Machine Translation, Natural Language Inference, Question-Answering, Object detection, etc.). The focus is on producing only relevant data features, dealing with missing data and irrelevant insights (from the same data) that could lead to underfitting (i.e., for a Data Engineer, too little features for covering the whole domain) or overfitting (i.e., for a Data Engineer, not enough data for representing and training each feature of the domain). Next, the so collected data by the Data Engineer is made available to a Data Scientist or a Machine Learning Engineer for Data Preparation tasks. For the Machine Learning Engineer, a failure in the Data

Engineering process could also lead to underfitting (where training and validation curves are very close, but at lower accuracies) or overfitting (where there is a gap between training and validation curves at higher accuracies). If the desired data is not available, it could be necessary to set up a social networking site to encourage people to share and correct data (cf. [1], page 770, para. 2). So, the quality of collected data strongly depends on the Data Engineer expert curation.

In the context of the Baoulé living language (bci), it is not yet supported (in September 2024) on popular worldwide free translation platform such as Microsoft's Bing Translator, Apple's Translate app, Facebook's AI translator, Amazon Translate, Yandex Translate, DeepL Translator, etc., nor yet supported among the Official Wikipedias, nor on popular Language identifier such as fastText and pycld2 (that are used in automatic Data Curator tools such as Bitextor or NeMo Curator). However, recently on 27th June 2024, an implementation of the bci language had been released on Google translate (cf. google-translate-new-languages-2024 on https://blog.google). That is a very good initiative that will serve to future data collection tasks, despite its failure to translate some statements such as 'ba nonman' or 'asan' that are respectively translated as 'nobody' and 'simple' instead of 'baby' and 'caterpillar'. Consequently, we still need human reviewers for the provided rough translation, but with the great advantage of less efforts. In conclusion, there is still clearly the need for making available, via Data Engineering, various Baoulé related datasets to the scientific community for AI tasks.

The Baoulé language is an alphabetic writing system based on Latin or French scripts. So, in our approach we proposed a first version of various Baoulé Related Parallel Corpora for Machine Translation tasks, combining Latin and/or French scripts. The data is collected from various sources (parallel and monolingual), including our own contribution via the Wikimedia Incubator and Translatewiki, but we also set up a social networking site for bringing more diversity in our corpora. And we finished our work by proposing a baseline model for language evaluation on our datasets, proving that these engineered parallel corpora are good enough for working on Machine Translation tasks implying the Baoulé language, and should yield significant contribution of interested people for curation of related datasets. [We provided between squared brackets […] very useful references to support the readers].

2. THE BAOULÉ LANGUAGE IN CÔTE D'IVOIRE

In this section, we provide an overview about the Baoulé language that will be the main workpiece of our intended Data Engineering tasks necessary for Machine Translation Systems.

2.1. A Brief Presentation of the Baoulé Language

The Baoulé language, also known as Baoule, Baule, Baule-Ando, Bawule or Wawle, is spoken in central and southern Ivory Coast (cf. Wikipedia, Baoulé language). According to Ethnologue (cf. Ethnologue, bci), there are 4,650,000 Baoulé speakers in Côte d'Ivoire alone, and total users in all countries are 4,654,060 (Leclerc 2017c). This means that about *99.91% of Baoulé speakers in the world are located in Côte d'Ivoire*. The Baoulé language includes today more than twenty (20) subgroups (tribes or dialects) [2] and the necessity to use it, and globally all national languages in a Country that has French as Official language, is a solved debate in Côte d'Ivoire [3]. Original bci-related scientific works had used only the Latin script [3,4], but new initiatives (even scientific) are mostly written with the French script (education of children at school, web or mobile applications, new versions of the Baoulé Bible, etc. are all written with the French script). Moreover, as local initiatives, some Primary Schools teach the bci language, and there exist some related shows on national TV [5,6]. This is a way for avoiding language death (unfortunately we have 3 extincted and 14 endangered languages in Côte d'Ivoire according to Ethnologue) [7] and

solving a more general problem for African people: the identity crisis (the strict rejection of mother tongues or dialects at least as a second language, in favor of the country official language only) [8].

2.2. Mathematical Modeling of the Creation Process of Baoulé Tribes

We already noticed in [2] that in Côte d'Ivoire there are more than twenty (20) subgroups or tribes for the Baoulé language only. But, according to Ethnologue, all claim to understand the standard variety, i.e., each Baoulé tribe equally and perfectly understand all the other more than 20 tribe variants. In this section, we show that Richard A. Blythe and William Croft in their paper "How individuals change language" [9], provided a plausible mathematical explanation of how tribes are created in terms of language variation. This could also explain the creation process of Baoulé tribes. Indeed, the authors produced a formal mathematical definition of the process by which *Language change at the population level*, as shown below.

$$P_{O,i}(T_{O,i}) = \omega_i e^{-\omega_i T_{o,j}}$$

Also using the **Wright-Fisher model** they produced a mathematical equation of ***Language change at the individual level***, as shown below.

$$T_M \dot{P}(x,t) = -s[x(1-x)P(x,t)]' + \frac{1}{2N_e}[x(1-x)P(x,t)]''$$

As a summary of the authors findings:

- The population size has a weak effect on the rate of grammatical change.
- There are three (03) types of changes from the Poisson baseline: Child-based models of language change (language change that arises through the process of childhood language acquisition), Usage-based models of language change (change in language usage during the lifespan), and Social network effects (related to interactions with the community). All these changes are weakly related to (or are not dependent on) the population size.
- They also demonstrated that when there are more opportunities for individual behaviour to change per unit time (innovations, political independence, agriculture, etc.), changes propagate through large speech communities *more quickly*.

From this analysis of Richard A. Blythe and William Croft, we can confidently make some conjectures about how the more than 20 tribes of the Baoulé language have been created over time. Indeed, looking at the historical moves of parts of the Baoulé people towards some regions of the country for economical, political or social convenance reasons, we can indisputably guess how changes happened so rapidly to reach more than 20 dialects inside a unique Baoulé living language; it is mostly because of more opportunities for individual behavior.

3. MACHINE TRANSLATION (MT) AND DATA ENGINEERING

The general realm of Artificial Intelligence (AI) is made of Machines that Think and Act like Humans [10]. Therefore, in AI, an immense significance is attached to natural language and an intelligent system's ability to use it [11]. This is the reason why in this section we will deal with how machines handle humans' natural languages for communication, especially for Machine Translation (MT) tasks.

3.1. The Place of MT in the AI World

From [1], section 23.4, we know that Machine Translation (MT) is the automatic translation of text from one natural language (the source) to another (the target). Translation is difficult because, in the fully general case, it requires in-depth understanding of the text. Historically, there have been three (03) main applications of MT: Rough translation (*Free online translation* services such as Google Translate), Pre-edited translation (also known as *Document level-translation* generally used by companies), and Restricted-source translation (also known as *Domain-specific translation*, focused on highly stereotypical language such as a weather report, and is generally of excellent performance compared to the two previous groups).

In their paper [12], Aishwarya R. Verma and Dr. R. R. Sedamkar provided a very representative picture about existing approaches to Machine Translation, as shown below.

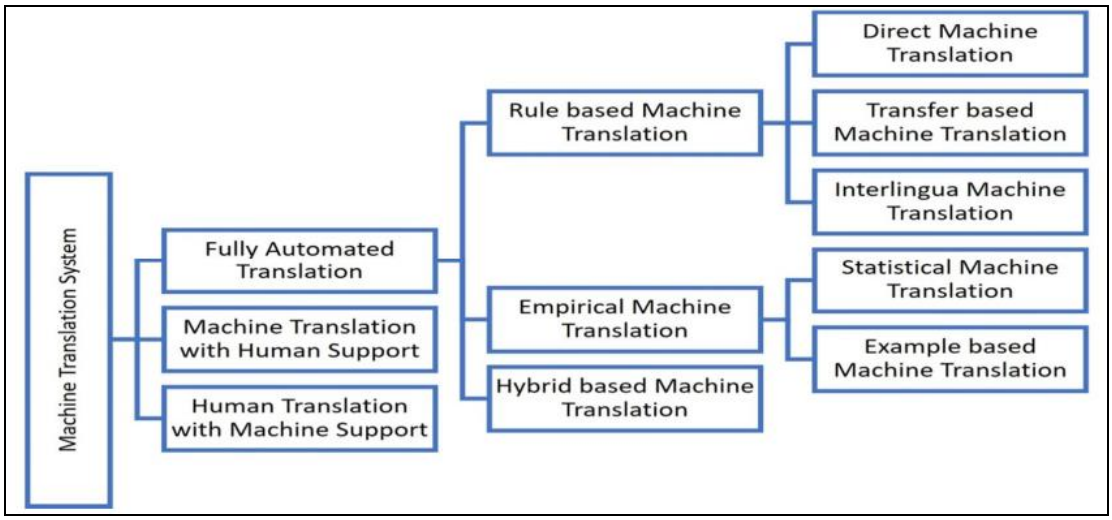

Figure 1. Approaches to Machine Translation [12]

In this schema, the three (03) main applications of MT presented above (Rough translation, Pre-edited translation and Restricted-source translation) should be found on the Fully Automated Translation branch. Then, the various MT techniques used under the fully automated branch could be either *Rule based Machine Translation (based on predefined logical rules that perform the translation)*, *Empirical Machine Translation (based on experimentations generally powered by dedicated datasets and machine learning models)* or *Hybrid based Machine Translation (a combination of previous methods)*. Just as an example, free online translation platforms perform their rough translations based on Hybrid MT techniques. Outside this path, there also exists *Business MT platforms* such as Omniscien Technologies that combine Hybrid and internal methods to provide higher performance than any existing rough model.

3.2. The importance of Data Engineering in Machine Translation (MT)

We know from the Introduction of our work that the Data Engineering task (obtaining, cleaning, and representing data) is at least as important as Algorithm Engineering. Especially, for MT tasks, in order to produce high quality translations, it is important to use reliable Language Pairs produced by an Expert in Language Pair Translation [13]. A Language Pair is a dataset of sentences from two (02) distinct languages (bilingual lexical data), where each sentence is reciprocally translated from one language (the source) *towards* the other (the reference), the best related *prediction*, *candidate* or *automatic translation* being the target (i.e., the predicted text into

the Target Language that we expect to be of high quality or very close to the reference). In case of several references per source, the target still needs to be, not any of the references, but the targeted prediction that is the closest possible to all these references, i.e., that better matches to all of them at once. In theory, a MT system that predicts the so defined target could make similar or even better propositions than the human expert references. Therefore, for the Reference-based Machine Translation technique that is the preferred approach in many MT-related works for obvious superior performance reasons, there is the need for producing High quality bilingual lexical data (Language Pairs or Source–Reference pairs, *cf. [14], Table 1*). This process of producing Language Pairs is called Language Pairing; it belongs to the Example based Machine Translation approach on the *Empirical MT* branch of Figure 1, that requires a good corpus for working well. But things are not as easy as we can think. As said M. Dion Wiggins, CTO and Founder of Omniscien Technologies: "High-Quality Machine Translation is Complex, any vendor that tells you that MT is easy is either trying to fool you or is incompetent. Time, Effort, Skill and Investment are Mandatory". Indeed, Grabbing High-Quality MT data is also a hard and expensive task that heavily depends on expert curation, even for the suboptimal reference-free technique, as mentioned in [15].

4. THE DATA COLLECTION PROCESS

In this section, we will show that a successful dataset creation process is not obtained hazardously, but through a thorough analysis and an excellent data collection strategy. Making decisions that correspond to your use case is essential, the copy paste strategy will not work here because the techniques and tools available for High Resource Languages for example will not compulsorily work with an Extremely Low-Resource Language. It is up to the Data Engineer to make a crafty use of his data engineering skills via various techniques, to find the best paths and produce the required data for Data Scientists and Machine Learning (ML) Engineers. The interest of all these is highlighted in some of Collibra's marketing emails as "Poor data quality leads to unreliable insights", and "Trust your data, trust your AI".

4.1. Data Categories and Complexity

As mentioned in [16], *section 6.2.2. Biases in Data*, there exist various categories of data depending on the degree of treatment applied on. We may so distinguish between Raw Data (unstructured original data), Dark Data (collected and stored classified data in a form that is not immediately usable), Dirty Data (well formatted data that contain bad or undesirable patterns or entries), and Clean Data (well formatted data that contain only good entries). In a domain-specific expertise (e.g., MT or Question Answering), Dirty Data and Clean Data are both generally collected via a data processing concept called Extract-Transform-Load (ETL) and then put into the same features; only the amount and quality of entries are different for both data classes. But for MT, we make additional classifications depending on the amount of useful data available or collected for a given language. Indeed, the translation quality strongly depend on the amount of parallel data (language pairs) and the baselines for high quality translation is reached above 10 million language pairs. However, we need billions of language pairs for highest translation quality (cf. [17], 30^{th} min). Therefore, based on various references we propose the following table for classifying languages according to their resource availability.

Table 1. Classification of languages based on their resource availability for Machine Translation tasks.

N°	Language Resource Category	Data range (number of language pairs)
1	**High Resource Languages or Resource-Rich Languages**	**>>10 m** (cf. [17], Table 1 of [18], Table 2 of [19])
2	**Medium Resource Languages**	**1m~10m** (cf. Table 5&8 of [14], [17], Table 1 of [18], Table 2 of [20])
3	**Low Resource Languages or Resource-Poor Languages**	**100k~1m** (cf. Table 1 of [18], [21], Table 4 of [22], [23])
4	**Very Low or Extremely Low-Resource Languages or Under-resourced Languages**	**<100k** (cf. Table 1 of [18], [21], Table 4 of [22], Abstract & Table 1 of [24], Table 11 of [25])

In this classification, Extremely Low-Resource Languages are Languages with a **digitally** versatile speaker community, but very limited support in terms of language technology (cf. [21], Abstract). Their main characteristics are:

1- Severe lack of language data (very limited documentation about the language, such as books, dictionaries, publications, etc., written into the language, *cf. [21], section 1.2*)
2- Severe lack of accessible language data that can be easily processed (existing data unavailable or very poorly available in digital form via the internet or open-source software).
3- Severe lack of language technological support such as NLP tools for various data curation tasks like data annotation tools including the intended language.
4- Very limited interoperability of available data and tools (available corpora are messy and not designed to work together).

In addition to data availability, another big challenge we meet during the Data Engineering process for MT is Data Complexity. Indeed, we face various complications due to the text nature of the dataset (cf. [1], section 23.3.5). These complications are: Time and tense understanding, Quantification, Pragmatics, Long-distance dependencies, Lexical ambiguity, Syntactic ambiguity, Semantic ambiguity, Figures of speech (*Metonymy, Metaphor*), etc. As contextual problems, they are often source of confusion and even strong ambiguities that the final MT system should overcome. Therefore, to deal with these various complications, we use the Disambiguation technique that is the process of recovering the most probable intended meaning of an utterance (we could remember here the similarity with the target in MT as the best automatic translation to reach). Disambiguation is properly done only when we combine the following four (04) models: the World Model (the likelihood that a proposition occurs in the world), the Mental Model (the likelihood that the speaker forms the intention of communicating a certain fact to the hearer), the Language Model (the likelihood that a certain string of words will be chosen, given the Mental Model of the speaker) and the Acoustic Model (the likelihood that a particular sequence of sounds will be generated, given the Language Model of the speaker) for spoken communication.

All the above struggles (data availability and complexity) should be considered and overcome during the Data Engineering process in order to produce a diverse and consistent training set for a High-Quality Machine Translation System.

4.2. The mtBCI-1.0 Corpus

In this section, based on the above discussion about data engineering techniques for MT tasks, the data categories and complexity, the evolution of the Baoulé language to yield more than 20 tribes that all claim to understand the standard variety, etc., we propose a first approach of a specific

corpus for covering the described AI problem. Indeed, the bci language has very limited documentation compared to the Resource-Rich Languages, is not yet supported (in September 2024) on famous worldwide free translation platform such as Microsoft's Bing Translator and Apple's Translate app, is not yet supported in any NLP tool for automatic language pair creation (sentence and document alignment), is not available among the official Wikipedias, and presents very limited interoperability between available data formats (Latin-based or French-based scripts) and existing tools (toolkits or toolboxes). Therefore, the Baoulé language could be classified as an Extremely Low-Resource Language (cf. Table 1 & comments, section 4.1).

4.2.1. Details about the Corpus

We know that "The writing system with which we learn to read may influence the way in which we process speech." [26]. Therefore, we make the hypothesis that the techniques used by the Data Engineer (*and so a Data Scientist or a Machine Learning Engineer*) to represent and transform the language data should also influence the AI model performance, mostly for the Neural-based models which are closer to the human brain structure (a_1). Additionally, [27] shows that the writing systems could influence the performance of language models (a_2). Subsequently, for creating the mtBCI-1.0 Corpus, we consider that we are in an Extremely Low-Resource Language context, and that the bci language is an alphabetic writing system containing 21 consonants and 12 vowels (cf. [3], pp.15-19), so 33 letters at all. We also found two (02) approaches for the Baoulé writing system: originally the Latin-based script, but also the French-based script that is mostly used in nowadays publications written in Baoulé. Furthermore, the Baoulé language, outside those who use or speak it very well as a second language, consists of more than 20 known tribes that all claim to understand the standard variety (i.e., *Agba, Aïtou, Oualèbo, Faafwè, N'zikpli, Nanafouè, N'gban, Saafwè, Ahaly, Akouè, Anôh, Elomoué, Dô'n, Fâly, Gbloh, Gôly, Kôdè, Satiklan, Sondo, Souhamlin, Yaourè*). Consequently, with the objective to creating a corpus that considers language complications (*cf. section 4.1*), the above insights (a_1 *&* a_2) and the limited availability of data resources (*cf. Table 1 & comments, section 4.1*), we split our mtBCI-1.0 corpus into several datasets or corpora:

- **Latin-based script Datasets** (en-bci, fr-bci);
- **French-based script Datasets** (en-bci, fr-bci);
- **Mixed Datasets** (en-bci, fr-bci), that is a mixture of Latin-based script and French-based script Datasets;
- **Synthesized Datasets** (en-bci, fr-bci), that is a French-based synthesis of expressive and much more readable words, based on justified evidences.

For empirical reasons, the subset that provides the best performance will be known only after experimentation.

4.2.1.1. Data Extraction technique

As mentioned in the introduction, the quality of this delicate, complex, and hard data creation stage will highly impact the performance of related Machine Translation models. Indeed, internet pages and raw documents cannot directly be applied to Algorithms (for AI tasks) without a watchful Data Engineering stage because they are just raw or dark data.

Therefore, now that we have much more information about the structure of the mtBCI-1.0 Corpus, we focus in this section on the sources of available Parallel Corpora including the Baoulé language, rather than their format. Our approach for finding parallel corpora sources is somehow similar to the one adopted by the EnKhCorp1.0 (cf. [28], section 2), but we bring more emphasis. Currently (September 2024), there exists no automatic language pair creation tool (auto bitext

alignment) that includes the Baoulé language. This generally starts with the creation of a language identifier algorithm that could require around 500k pairs for being performant enough. Indeed, there exist many popular free and open-source tools for web crawling and sentence alignment such as Bitextor, Bicleaner-AI, Interactive Clue Aligner (ICA), Alpaco, Microsoft Bilingual Sentence Aligner, LF Aligner, etc., where deprecated ones are still usable under the last release conditions. But none of them natively supports the Baoulé language that belongs to an Extremely Low-Resource Language context.

Indeed, in a context of an Under-Resourced Language that is not yet supported in any NLP tool for parallel sentence alignment (auto-aligned bitext creation), not yet available among the official Wikipedias, etc., we initiated since 2019 the creation of additional sources including Wikimedia Incubator, Translatewiki (cf. APPENDICES 11.1), and a Data Crowdsourcing website. To elaborate, looking at Figures 3,4,5 of [29], Figure 1 of [30], Figure 2 of [31], Figure 1 of [32] and Figure 1 of [33], we propose below a chart flow corresponding to our specific Data Collection Strategy shown in Figure 2 below, in which:

- The automatic data collection process is currently not yet available. We should move towards a Low-Resource Language context at least for allowing first good implementations.
- The Semi-Manual method combine manual operations and existing non-supported tools if possible.
- The Manual method is a pure manual data collection and alignment operation, by hand.

Once the Baoulé-related raw data is collected, the Data Transformation process consists in creating Baoulé-related High-Quality parallel sentences via various coding tasks and tools. The whole process stays updatable if the Raw Data Sources are updated.

Indeed, for aligning collected datasets, in addition to open-source tools such as LF Aligner, we also used new tools such as Google AI Studio since the Gemini 1.5 preview announcement on March 6, 2024 (for data alignment and page count, with between 40% and 60% of success; we provided ourselves the curated dataset to the platform, it is not generated). We also used other platforms such as ScrapeGraphAI (for various data extraction from Globse), Word Counter (for the number of words and unique words), spaCy larger model (for the number of tokens). The transformation of Wikimedia Incubator and Translatewiki data has been performed by code only. For scanned pages, we first tried our own implementation with tesseract that extracted very well the content of French words, but failed on some Latin characters such epsilon (ɛ) or ɔ. We so used Nanonets that provided better results (despite some additional manual annotation tasks). All these has so been done as explained in Figure 2 below; and there is still the need for huge manual work to obtain a somehow high-quality dataset. Therefore, we are still collecting and updating our data from various sources upon availability of new data.

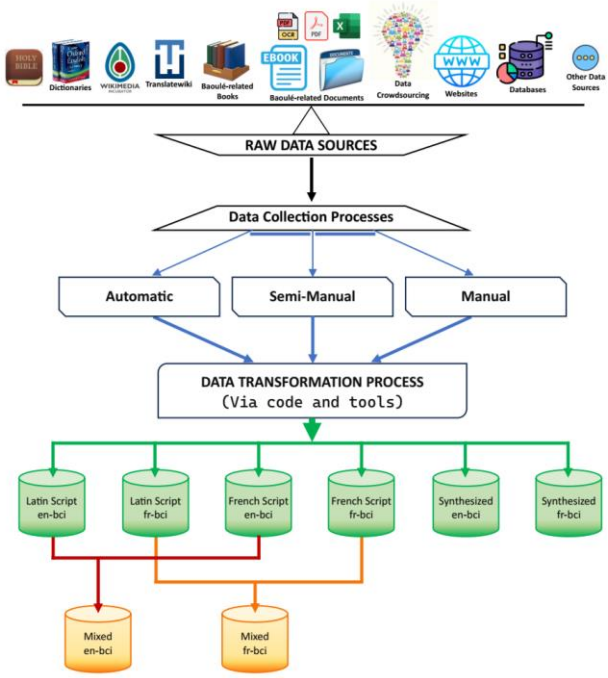

Figure 2. The Data Collection Strategy of the mtBCI-1.0 Corpus

4.2.1.2. Corpus Analysis and Statistics

Based on Table 4 of [34] and Table 6 of [35], we propose suitable tables about statistics of collected data (training and validation sets only, without the sets used for inference evaluation).

Table 2. Statistics about the English and French datasets related to the mtBCI-1.0 Corpus.

		English			**French**		
STATS	Number of pairs	Number of Sentences	Number of Words	Number of Tokens	Number of Sentences	Number of Words	Number of Tokens
TOTAL	**7191**	**8 326**	**44 563**	**62 815**	**7 986**	**46 866**	**66 593**
Number of unique Words (Vocabulary)		**8 766**			**11 364**		

Table 3. Statistics about the Baoulé datasets related to the mtBCI-1.0 Corpus.

		BCI - Latin Script			**BCI - French Script**		
STATS	Number of pairs	Number of Sentences	Number of Words	Number of Tokens	Number of Sentences	Number of Words	Number of Tokens
TOTAL	**7191**	**7 916**	**60 837**	**83 969**	**8 049**	**59 416**	**80 193**
Number of unique Words (Vocabulary)		**8100**			**8559**		

4.2.2. Coverage of our Contribution

All our process for creating the mtBCI-1.0 Corpus follows the scientific method (cf. [1], section 1.3.8, *AI adopts the scientific method (1987-present)* and page 770, para. 2). We proposed a first approach of a specific corpus for covering a well-described AI problem: the unavailability of at least a diverse toy bci dataset to the scientific community for performing NLP related experimentations (that matches the specificity of the bci language), and the lack of related NLP tools. Our strategy for overcoming this struggle is intended to lead to diversity, mostly via thorough data curation and engineering from various sources that cover different aspects of society such as religious material (including the Bible, for scientific interest here), literature, daily usage, and common statements (predominantly via a data crowdsourcing website), etc.

Let us note that we are in a unique or scarce case for a language including more than twenty (20) tribes that all claim to understand the standard variety. Further research work could perhaps try to disentangle them all, but we followed the standard variety path from our side. Consequently, we will find different spellings (as several vocabulary words) for some unique words, but we are sure that any Baoulé speaker will be able to perfectly understand them. For example, for the word "healthy" we found three (03) variants depending on the provider: juejue, dyuédyué, and djuédjué. And for the word "take" we found "le" or "de" as variants. Moreover, our methods are adaptable and reproducible to other very low-resource and low-resource languages. In Côte d'Ivoire alone, we have currently seventy-nine (79) very low-resource languages including three (03) extincted (cf. ethnologue.com, Côte d'Ivoire). And most of them have no existing dictionary or are not yet created among the Wikimedia incubator languages (the Baoulé language itself has been verified as eligible by Wikimedia incubator relatively recently on January 26th, 2017). Looking at our strategy, we are confident that the crafted corpus will, therefore, be useful to the NLP domain scientific community. Even the provided corpus could be useful for creating datasets for other NLP subdomains like Automated Speech Recognition (ASR), Question-Answering (QA), etc.

5. BASELINE SYSTEM

In this section, after the collection of a representative set for the mtBCI-1.0 Corpus via a thorough Data Engineering Process, we will provide a baseline model proving that the collected data is of scientific interest (Table 1 of [36] and Table 3 of [37] are excellent examples of this approach and its possibilities).

5.1. Experimental Setup

In order to provide a suitable experimentation environment for training the curated mtBCI-1.0 Corpus, we first planned to select a framework among a list of popular Open-source NMT toolkits provided in Table 5 of [38]. Especially, a tool like the OpenNMT framework that offers LLM-based fine-tuning is a good choice, because it has been proven recently that LLMs provide effectiveness towards solving the Disambiguation problem we mentioned in section 4.1 of our work (cf. [39]). But finally, our analysis led us to Predibase that provides latest open-source LLMs and suitable environment for fast prototyping and experimentation. So, after some experimentations, we fixed the same hyperparameters for all training datasets in order to mitigate with overfitting and provide a fair comparison of the results. We trained our baseline models on one A100 GPU with the following hyperparameters:

Table 4. Main training setup for the baseline models related to the mtBCI-1.0 Corpus.

Base LLM	Fine-Tuning Task	Adapter Rank	Target Modules	Epochs	Learning Rate	Early Stopping
llama-3-1-8b-instruct	Instruction Tuning	16	q_proj, v_proj, gate_proj, up_proj, down_proj	5	0.0001	Enabled

For the evaluation metrics, looking at their computation template the ROUGE metric fits better to our labels (one reference) than SacreBLEU and TER that work better on multiple references (at least two references). However, we kept the TER metric for getting an idea of the effort to make on each translation to match its reference. The SacreBLEU metric will give at least an idea about the number of Longest Common Subsequences between translations and references, given that the format of our labels could not guarantee evidence of this matching for a baseline model. Indeed, the ROUGE metric computes the similarity between the machine-generated text and the human reference text. The TER metric estimates the post-editing effort required on the translation to match the reference, and the SacreBLEU scores are based on the Longest Common Subsequence and Skip-Bigram Statistics.

Table 5. Chosen metrics.

Chosen METRICS	**Low Quality range**	**Medium Quality range**	**High Quality range**
ROUGE Score	[0.0 - 0.3]	[0.3 - 0.6]	[0.6 - 1.0]
TER Score	[0.7 - 1.0]	[0.4 - 0.7]	[0.0 - 0.4]
Score	[0 - 20]	[20 - 40]	[40 - 60]

During our experimentation, the mtBCI-1.0 Corpus was split as shown in Table 6 below.

Table 6. Statistics about the splits of the mtBCI-1.0 Corpus.

Data Sets	**Latin-based script**		**French-based script**		**Mixed Datasets**		**Synthesized Datasets**
	en-bci	**fr-bci**	**en-bci**	**fr-bci**	**en-bci**	**fr-bci**	**en-bci**
Training set	6471	6471	6471	6471	12943	12943	N/A
Validation set	720	720	720	720	1439	1439	N/A
Initial test set	7	7	7	7	7	7	
Simpler test set	7	7	7	7	7	7	N/A
All	**7205**	**7205**	**7205**	**7205**	**14396**	**14396**	**N/A**

5.2. Results

In this section, we will present the results of our experimentations in similar ways as Table 1 of [40], provided that our dataset splits are already available in Table 6 above.

We will look first at the behavior of our learning curves, and then provide an evaluation of each baseline model on our chosen metrics (cf. Table 5 of this paper).

The learning curves presented below show that for each of the six (06) datasets of the mtBCI-1.0 Corpus (cf. Table 6 above), the training and evaluation scores gradually decrease from 8 towards 1 and lower values.

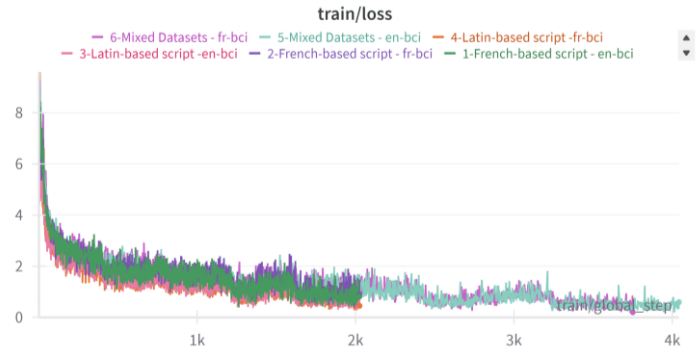

Figure 3. The mtBCI-1.0 Corpus baseline models' training curves

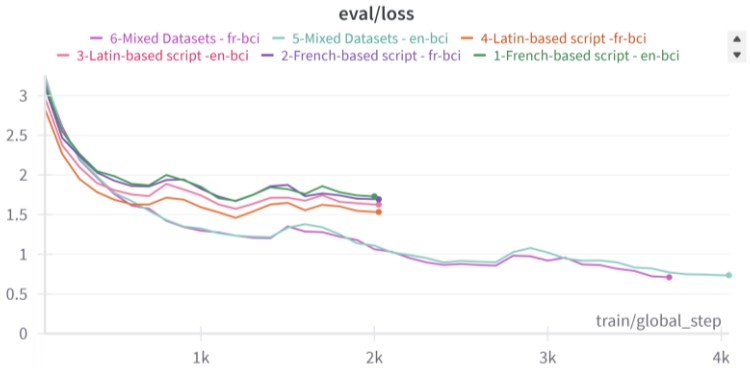

Figure 4. The mtBCI-1.0 Corpus baseline models' validation curves

We then reported below the performance of our baseline models based on our chosen metrics.

Table 7. Baseline neural model evaluation on the mtBCI-1.0 validation set.

Evaluation Metrics	**ON VALIDATION SET**						
	Latin-based script		**French-based script**		**Mixed Datasets**		**Synthesized Datasets**
	en-bci	**fr-bci**	**en-bci**	**fr-bci**	**en-bci**	**fr-bci**	**en-bci**
ROUGE Score	0.21719	0.234109	0.19526	0.19748	0.34624	**0.36254**	**N/A**
TER Score	1.47730	1.14986	1.65026	1.44619	0.88309	**0.86230**	**N/A**
SacreBLEU Score	0.07142	**0.07424**	0.06229	0.06388	0.05631	0.07368	**N/A**

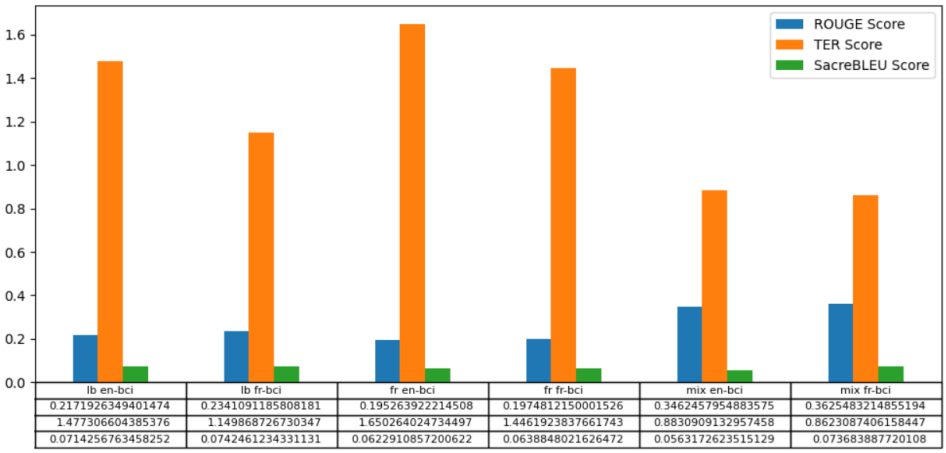

Figure 5. Chart of the evaluation on the mtBCI-1.0 validation set

As we mitigated overfitting during the training process, we provided just a few sets for evaluating the inference performance in order to get an idea of the generalization scheme.

Table 8. Baseline neural model evaluation on the mtBCI-1.0 initial unseen test set.

Evaluation Metrics	ON INITIAL UNSEEN TEST SET						
	Latin-based script		French-based script		Mixed Datasets		Synthesized Datasets
	en-bci	fr-bci	en-bci	fr-bci	en-bci	fr-bci	en-bci
ROUGE Score	0.25686	**0.34026**	0.18897	0.21389	0.10613	0.11790	**N/A**
TER Score	1.21428	1.11156	1.13333	1.20952	**0.96103**	**0.96103**	**N/A**
SacreBLEU Score	0.0	0.0	**0.03752**	0.0	0.0	0.0	**N/A**

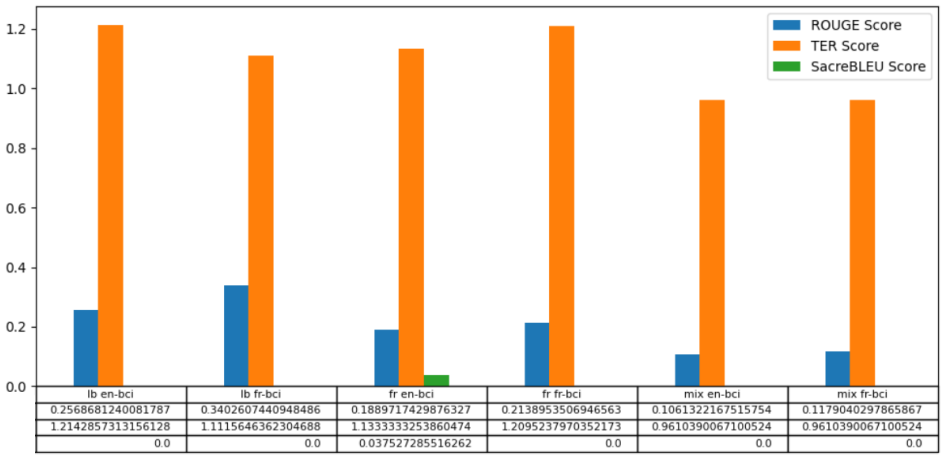

Figure 6. Chart of the evaluation on the mtBCI-1.0 initial unseen test set

Looking at the Rouge Scores of Figure 6 above, we made the hypothesis that a baseline model that turns around the bottom of the Medium Quality would certainly work better on simpler text, i.e., very short statements (about 5 tokens or words/subwords at max). And we provided another simpler unseen test set for this evaluation, as presented below.

Table 9. Baseline neural model evaluation on the mtBCI-1.0 simpler unseen test set.

Evaluation Metrics	**ON A SIMPLER UNSEEN TEST SET**						
	Latin-based script		**French-based script**		**Mixed Datasets**		**Synthesized Datasets**
	en-bci	**fr-bci**	**en-bci**	**fr-bci**	**en-bci**	**fr-bci**	**en-bci**
ROUGE Score	0.56771	0.43846	**0.58517**	0.57702	0.34386	0.28279	**N/A**
TER Score	0.62380	0.95476	**0.59523**	0.71904	0.79901	0.85116	**N/A**
SacreBLEU Score	0.0	**0.14285**	0.09553	**0.14285**	0.0	0.0	**N/A**

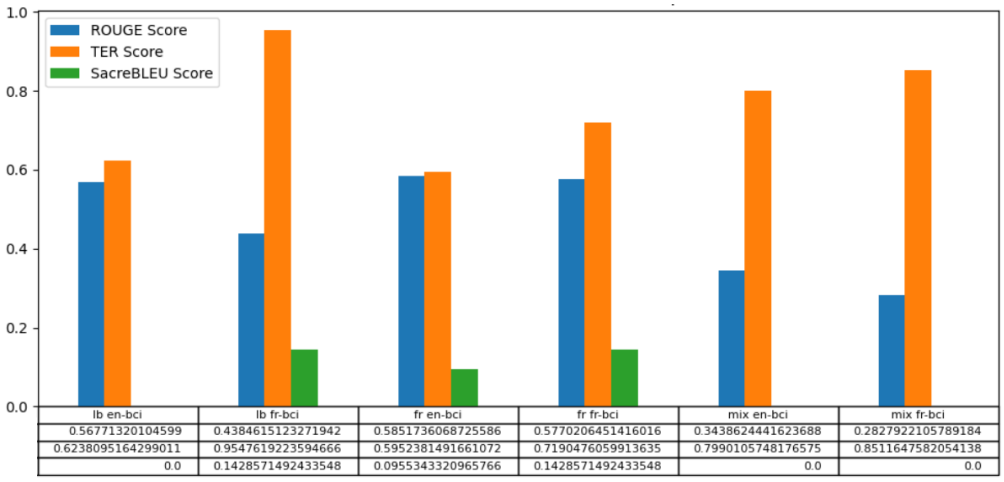

Figure 7. Chart of the evaluation on the mtBCI-1.0 simpler unseen test set

In our current work we did not create a baseline model for the synthesized datasets, but it will be done in another context of Benchmarking.

5.3. Analysis

In this section, we provide an analysis of the results of our experimentations on the mtBCI-1.0 Corpus. Indeed, from the above related figures and chart we have the following insights.

The learning curves presented above (figures 3 and 4) show that the baseline models related to our mtBCI-1.0 Corpus effectively learn well based on the training and validation curves. During the experimental setup we chose the parameters in order to avoid or mitigate overfitting. Also, our data collection strategy was intended to mitigate the underfitting phenomenon. For example, for the learning curves of the mixed-fr-bci baseline model that reached the best ROUGE and TER scores on the validation set among all mtBCI-1.0 Corpus baseline models, we got the following chart.

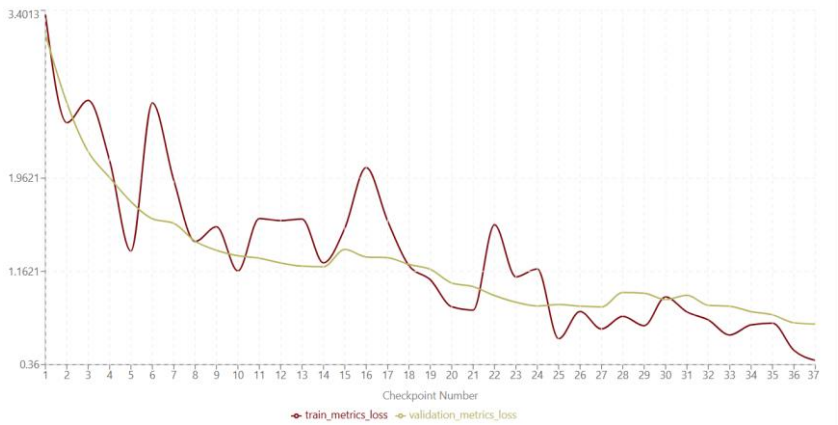

Figure 8. Learning curves of the mixed-fr-bci baseline model

Further improvement techniques towards the creation of a benchmark model should lead to more smooth learning curves. But that is not the scope of our current work.

For our hypothesis that a baseline model that turns around the bottom of the Medium Quality should certainly work better on simpler text (very short statements), we were right as shown by the Rouge Scores of Figures 6 and 7, now being very close to the bottom of High Quality for the French-based scripts (en-bci and fr-bci where bci is written with the French-based script). We think that the specificity of top scores with the French-based script are surely due to the vocabulary size of the bci French-based script that is of 8,559 unique words, against 8,100 unique words for the bci Latin-based script. But globally, all mtBCI-1.0 Corpus baseline models performed better on simpler text (with short statements) than on the initial test set; they generalized better on the unseen simpler test set. Just as an example, let's look at the following table (predictions are given as is, for both French-based and Latin-based scripts).

Table 10. Top 3 prediction samples from the mtBCI-1.0 Corpus baseline models.

Type of prompts	Unseen prompts in French or English	Best bci predictions from the mtBCI-1.0 Corpus' **baseline models** (Top 3)	Expected results from a human translator
Longer text (above 5 tokens or words/subwords)	Je m'appelle Jean-Claude, et toi comment t'appelles-tu ?	N flê mi Jean-Claude, ô fata kê ô flê i sô?	Bé flê mi Jean-Claude, yê ô li, bé flê wô sê ?
		Be flɛ mi Zan-Klɔ, ɔ flɛ wɔ nuan?	Be flɛ mi Jean-Claude, yɛ ɔ li, be flɛ wɔ sɛ ?
	My name is Jean-Claude, and you, what is your name?	bé flê mi Jean-Claud, ô, ô flê ô?	Bé flê mi Jean-Claude, yê ô li, bé flê wô sê ?
		Ɔ se kɛ: «Be flɛ min Jean-Claude, ɔ flɛ wɔ n?»	Be flɛ mi Jean-Claude, yɛ ɔ li, be flɛ wɔ sɛ ?
Simpler text (about 5 tokens or words/subwords at max)	Appelle le.	ɔ flɛ ɔ.	Flê i. / Flɛ i.
	Call him.	Fle i.	Flê i.
		Flɛ i.	Flɛ i.
	D'où viens-tu ?	A fin ni?	A fin ni? / A fi ni? Amoun fin ni?/Amoun fi ni?
	Where do you come from?	Afin ni? Amoun fi ni?	
	Je mange de la nourriture.	N di like.	N su di aliɛ. / N sou di aliê. N su di like. / N sou di liké.
	I eat some food.	ɔ di like.	

For a baseline model, we think that the above answers are effectively of medium quality, with some reaching the bottom of high quality, matching perfectly the above ROUGE scores' rating (cf. Figures 6 & 7). The next step should consist in providing a good model that performs very well on both, simpler and longer texts.

Another observation is that while the mixed dataset models performed better on the validation sets, the generalization has been better on more specific sets; either Latin-based initial test set or French-based simpler test sets (looking at the ROUGE Scores).

Also, the translation from French, that have the highest vocabulary, provided better performance than the Latin-based one on each case for the Validation set and the initial test set. But the opposite happened when we took the simpler test set; here the translation from English provided the best scores in each case (looking at the ROUGE scores of Figure 7).

As in this work we just provided a baseline implementation via a useful dataset for experimentations, further investigations and endeavors towards creating a Dataset and a Model Benchmarks should provide more insights and explanations about all these aspects.

6. LIMITATIONS AND FUTURE RESEARCH DIRECTIONS

In our work, we created the mtBCI-1.0 Corpus via a thorough Data Engineering strategy, and we provided an evaluation on related baseline models, proving that the collected data is of scientific interest. In this section, after highlighting that despite the good job, some big struggles remain on the path for the creation of High-Quality Resources for the Baoulé language, we will focus on Future Research Directions in order to overcome these challenges.

6.1. Limitations

A translation system has the advantages of not being affected by human bias (stereotypes), and it could perform the translation faster and with a rather objective stance. However, the machine (language model) has its own internal limitations such as data dependency (cf. p.4 of [41]) while translation professionals (human translators) can bring more judgment, consideration, flexibility and subjectivity in their translations (cf. section 4 of [42] and abstract of [43]). Therefore, despite their scientific value, the collected mtBCI-1.0 Corpus and the related baseline models could be still considered as a toy implementation (initial steps) towards a greater objective. Indeed, in the context of the extremely low resource Baoulé language, we still face to various limitations due to lack of resources, and very limited interoperability of available data and tools (available corpora are messy and not designed to work together). Let us elaborate.

First, despite our work provided useful insights (cf. section 5.3) including its medium performance on simpler text, the mtBCI-1.0 Corpus and related models are a baseline (no system instruction, no back-translation, and any other improvement techniques, etc.) and longer prompts lead to the left-in-the-middle phenomenon and there is the need for handling the problem of diversity in the collected dataset.

Second, for some words, publishers proposed different writings for the same word (visible even among various versions of Baoulé dictionaries). This must be solved certainly by linguists before adopting the right writing style or spelling. For example, the English word "thing" could be translated as "ninnge" (Latin script) or "ningué" (French script), and the English word "language" could be translated as "aniɛn" (Latin script) or "aniein" (French script) when we

consider words for which there is no variation in spelling among Baoulé tribes. Also, the English word “tongue” could be translated as “tafiman” (Latin script) or “tafliman” (French script), and the English word “take” could be translated as “le” (Latin script) or “dé” (French script), for words for which there exist variations in spelling among tribes.

Third, as a very low resource language, some French or English words such as “Artificial Intelligence”, “Mathematics”, etc., have no equivalent translation in the Baoulé language, because they don't exist in the language. They have never been created by ancient users because these speakers didn't know them (science words, medicine terms, etc.) and there is no official pronunciation in Baoulé for them. We think that here too, linguists must bring their contributions before adopting the right writing style (we will update upon their work).

6.2. Future Research Directions

After the creation of the mtBCI-1.0 Corpus and related baseline models, we face to various challenges that offer opportunities for future research directions that we present below.
First, we must collect enough parallel data for transferring the bci language from very low-resource to at least low-resource or medium-resource setup. Indeed, under low-resource or medium-resource conditions for the bci language, we will be able to thoroughly integrate it in well-known automation platforms such as Bitextor, Bicleaner, etc., for which the language identifier classifier requires a clean parallel corpus of at least 500k pairs of sentences for providing best performance. Second, we must complete the related Wikimedia Incubator and translatewiki.net in order to bring the Baoulé language among Official Wikipedias. This includes but does not restrict to the completion of the translatewiki.net interface translation, the creation of a Main Page for the Baoulé language on the Official WikiPedia, the creation of several new pages, and the translation of existing pages from other languages into the Baoulé language. Third, it is necessary to provide a Data and Model Benchmarking for MT tasks implying the Baoulé language. One good example of this approach could be seen in [44] that created several benchmark datasets, benchmark models and a toolkit. Creating these Baoulé-related data and model benchmarks, we must ensure that not only the datasets contain linguistic phenomena, but also the accuracy of linguistic phenomena does not drop while training them, as mentioned in conclusions of [45]. Fourth, as a last possible research direction, further investigations could perhaps try to disentangle the more than twenty (20) Baoulé tribes in order to identify each tribe separately. As we followed the standard variety in our data collection strategy (because all Baoulé tribes claim to understand very well each one the other), the provided models cannot distinguish them. For example, our models cannot tell if they are generating Agba, Aïtou, or Oualèbo tribes’ languages, but they ensure to generate globally the Baoulé language so that any Baoulé speaker will be able to understand it.

We encourage professional and independent researchers, advanced students (master and Ph.D.), newbies, enthusiasts and hobbyists who are interested in contributing in the bci language expansion towards medium resource language and even more, to build upon the mtBCI-1.0 Corpus (make useful experimentations based on our initiative). They can work on related dataset improvement, on our proposed pairing platform or on their own project, etc., and share their results to the research community. We so shared some insightful files of our implementations (cf. APPENDICES 11.1) in order to allow to any volunteer to contribute. Let us note also that the new google translate could help in providing good bci related datasets with less reviewing efforts than it was necessary before.

7. CONCLUSION

In this paper, we provided useful insights about Data Engineering tasks and focused on the (AI) problem of unavailability of Baoulé-related Machine Translation resources (including good Data and Model Benchmarks) to the scientific community to build upon.

In order to mitigate this AI problem, we proposed "Baoulé Related Parallel Corpora for Machine Translation tasks: mtBCI-1.0" to make parallel Baoulé-related datasets and various related tools available to the scientific community, for AI tasks implying Machine Translation. In our approach, we started by a brief presentation of the Baoulé language that is specific for including up to more than twenty (20) tribes for which the creation process is plausibly scientifically explainable (cf. section 2.2). Then we focused on the corresponding Data Engineering Process itself that is of great importance (as mentioned in sections 3.2 and 4) before providing a baseline proving that the collected data is of scientific interest (cf. sections 4.2.2 and 5). Indeed, from a preliminary analysis, we provided a clear classification of languages based on their resource availability for Machine Translation tasks (cf. Table 1, section 4.1). Then, we found that the Baoulé language could be considered as an Extremely Low-Resource Language, and this allowed us to determine the right Data Collection strategy for our unique or scarce case for a language including more than twenty (20) tribes that all claim to understand the standard variety. From the collected data, we created the mtBCI-1.0 Corpus via a thorough Data Engineering Process and provided baseline models proving that the collected data is of scientific interest.

However, despite its scientific value, the scope of our work is about providing a baseline as a launchpad for building useful related sets. Therefore, the collected mtBCI-1.0 Corpus and the related baseline models could be still considered as a toy implementation towards a greater objective, due to various limitations discussed in section 6.1. Indeed, there exist the lack of resources and very limited interoperability of available data and tools, that should be overcome via thorough future research directions (cf. section 6.2). These scientific investigations include the insertion of the Baoulé language among Official Wikipedias, the creation of a Data and Model Benchmarking for MT tasks implying the Baoulé language, and the disentanglement of the more than twenty (20) Baoulé tribes in order to identify each tribe separately.

We have also shown that, regardless of the particular case, the proposed methods are adaptable and reproducible to other very low-resource and low-resource languages worldwide. We hope this work will illuminate anyone interested in Artificial Intelligence (including newbies, enthusiasts, hobbyists, etc.).

ACKNOWLEDGEMENT

We would like to thank Professor William Martin our Supervisor, Professor ABDO Miled Abou Jaoude and the BIU staff for the reviews of the reports of our Ph.D. program, and for the access to a partial scholarship leading to this highest level of specialization in Artificial Intelligence. Invaluable things we learned there are the supports for writing this paper.

We are also thankful to the Ivoirian Civil Service, as we got access to this Ph.D. program in 2018 as a Computer Scientist Civilian (Civil Servant or Official) from Côte d'Ivoire. Local initiatives encourage commitment to training and to the field of research for civilians.

DECLARATIONS

Funding	No funding was received to assist with the preparation of this manuscript.
Conflicts of interests	No conflicts of interest to the best of our knowledge.
Ethical Approval and Consent to Participate	No, the article does not require ethical approval and consent to participate with evidence.
Availability of Data and Material/ Data Access Statement	A toy dataset of the mtBCI-1.0 Corpus is available for scientific experimentations (cf. APPENDICES 11.1).
Code availability	Yes, (cf. APPENDICES 11.1).
Authors Contributions	One author article.

Credit: Declaration statement of DOI: 10.35940/ijese.C9803.12040324

REFERENCES

[1] Russell, Stuart J. & Norvig, Peter (2018) *Artificial Intelligence: A Modern Approach*, Pearson India Education Services Pvt. Ltd., India, Third Edition, twelfth Impression 2018.

[2] Abonoua Rachelle, YAO (2008) "Valeurs Culturelles Du Peuple Baoulé: Culture et Mariage.", *Memoire Online*, Université de Bouaké, https://www.memoireonline.com/12/09/2947/Valeurs-culturelles-du-peuple-Baoule-culture-et-mariage.html (accessed Feb. 02, 2022).

[3] Judith Tymian, Jérémie Kouadio N'Guessan & Jean-Noël Loucou (directeurs), (2003) *Dictionnaire baoulé-français*, Nouvelles Editions Ivoiriennes (NEI), 610 pages, Préface, p.5, para. 1.

[4] RICHARD R., DAY & ALBERT B., SARAKA (1968) *An Introduction to Spoken Baoule, Preliminary Text*.

[5] RTI Officiel. (2019). *Les Trésors Du Monde : Apprenons nos langues "La langue Agba (Baoulé de Dimbokro)* [Video]. YouTube. https://www.youtube.com/watch?v=DWAtiTCCH60.

[6] RTI Officiel. (2021, March 4). *Les Trésors du monde | Apprenons nos langues* [Video]. YouTube. https://www.youtube.com/watch?v=iYdMacPy1qg.

[7] Benjamin, Plackett, "Is Latin a dead language?", *Live Science*, 01 Jun. 2021, https://www.livescience.com/did-latin-die.html (accessed Jan. 16, 2024).

[8] Kim, Chakanetsa, "Africa's lost languages: How English can fuel an identity crisis.", *The Comb podcast*, BBC World Service, 16 May 2021, https://www.bbc.com/news/world-africa-57093347. (accessed Jan. 16, 2024).

[9] Blythe RA, Croft W. (2021, June 2). "How individuals change language," PLoS ONE 16(6): e0252582, https://doi.org/10.1371/journal.pone.0252582.

[10] Konan Jean-Claude, Kouassi, (2022) "A Comprehensive Overview of Artificial Intelligence.", *In Proceedings of CS & IT - CSCP*, Vol. 12, No. 23, pp. 173-194, Figure 1., doi:10.5121/csit.2022.122314.

[11] Mariusz, Flasiński (2016) *Introduction to Artificial Intelligence*, Springer Nature, Switzerland, P.7, para. 1.

[12] Aishwarya R. Verma1, Dr. R. R. Sedamkar, (2021) "Comparative Analysis of Language Translation and Detection System Using Machine Learning.", *International Journal for Research in Applied Science & Engineering Technology (IJRASET)*, Vol. 9, No. 8, pp. 1202, Fig. 1, doi:10.22214/ijraset.2021.37577.

[13] Ulatus. "The Importance of Language Pairs in Academic and Professional Translation." Ulatus, 14 January 2015, https://www.ulatus.com/translation-blog/the-importance-of-language-pairs-in-academic-and-professional-translation/ (accessed Jan. 17, 2024).

[14] Zhang, Biao; Williams, Philip; Titov, Ivan; Sennrich, Rico (2020) "Improving Massively Multilingual Neural Machine Translation and Zero-Shot Translation.", *In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online, 6 July 2020 - 10 July 2020. Association for Computational Linguistics, 1628-1639, doi:10.5167/uzh-188226.

[15] Liu, Danni et al., (2022) "Learning an Artificial Language for Knowledge-Sharing in Multilingual Translation", *arXiv:2211.01292v2*.

[16] Konan Jean-Claude, Kouassi, (2023) "Understanding the Worldwide Paths Towards the Creation of True Intelligence for Machines", *International Journal of Computer Science & Information Technology (IJCSIT)*, Vol. 15, No. 1, pp. 43-68, Academy and Industry Research Collaboration Center (AIRCC), doi:10.5121/ijcsit.2023.15104.

[17] Omniscien Technologies. (2022, July 14). Part 1: Advances in Artificial Intelligence and Machine Translation: 2022 and Beyond [Video]. YouTube. https://youtu.be/ADohX83vHMA.

[18] Lin, Zehui et al., (2021) "Pre-training Multilingual Neural Machine Translation by Leveraging Alignment Information", *arXiv:2010.03142v3*.

[19] Shaham, Uri et al., (2003) "Causes and Cures for Interference in Multilingual Translation", *arXiv:2212.07530v3*.

[20] Felix, Stahlberg et al., (2022) "Jam or Cream First? Modeling Ambiguity in Neural Machine Translation with SCONES", *arXiv:2205.00704v1*.

[21] Rosner, Michael et al., (2022) "Cross-Lingual Link Discovery for Under-Resourced Languages", *In Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022)*, pages 181–192. Marseille, 20-25 June 2022.

[22] Trieu, Hai-Long et al., (2020) "MT-WIKI: A Wikipedia-based Multilingual Parallel Corpus for Machine Translation on Low-Resource Languages".

[23] Thanh Duong, Long, (2017) "Natural Language Processing for Resource-Poor Languages", Ph.D. Thesis, University of Melbourne, Australia.

[24] Maali, TARS et al., (2022) "Cross-lingual Transfer from Large Multilingual Translation Models to Unseen Under-resourced Languages", *Baltic J. Modern Computing*, Vol. 10, No. 3, pp. 435–446, doi:10.22364/bjmc.2022.10.3.16.

[25] Samuel, Cahyawijaya et al., (2023) "NusaWrites: Constructing High-Quality Corpora for Underrepresented and Extremely Low-Resource Languages", *arXiv:2309.10661v2*.

[26] News, Neuroscience, (2022) "Literacy Influences Understanding of Speech", *Neuroscience News*, 18 Oct. 2022, https://neurosciencenews.com/writing-language-speech-21666/ (accessed Oct. 20, 2022).

[27] News, Neuroscience, (2023) "In Bilingual Readers, the Visual Cortex Processes Latin and Chinese Characters Differently", *Neuroscience News*, 7 Apr. 2023, https://neurosciencenews.com/bilingual-visual-processing-22965/ (accessed Apr. 13, 2023).

[28] Laskar, Sahinur Rahman et al., (2021) "EnKhCorp1.0: An English–Khasi Corpus", *In Proceedings of the 18th Biennial Machine Translation Summit*, pp. 89-95, 4th Workshop on Technologies for MT of Low Resource Languages, Virtual USA, August 16 - 20, 2021.

[29] Paul, Sujni, (2011) "Parallel and Distributed Data Mining", *New Fundamental Technologies in Data Mining*, Prof. Kimito Funatsu (Ed.), ISBN: 978-953-307-547-1, pp. 43–54.

[30] Nguyen, Van-Vinh et al., (2022) "KC4MT: A High-Quality Corpus for Multilingual Machine Translation", *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022)*, pages 5494–5502. Marseille, 20-25 June 2022.

[31] Thangkhanhau, Haulai, (2023) "Construction of Mizo – English Parallel Corpus for Machine Translation", *ACM*, https://doi.org/10.1145/3610404.

[32] Ali Yuksel, Kamer et al., (2022) "Efficient Machine Translation Corpus Generation", *In Proceedings of the 15th Biennial Conference of the Association for Machine Translation in the Americas*, pp. 11-17, Workshop 2: CoCo4MT, Orlando, USA, September 12-16, 2022.

[33] Ranganathan, Karthika et al., (2022) "Building and Analysis of Tamil Lyric Corpus with Semantic Representation", *In Proceedings of the 15th Biennial Conference of the Association for Machine Translation in the Americas*, pp. 18-27, Workshop 2: CoCo4MT, Orlando, USA, September 12-16, 2022.

[34] Knowles, Rebecca et al., (2022) "Translation Memories as Baselines for Low-Resource Machine Translation", *In Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022)*, pages 6759–6767, Marseille, 20-25 June 2022.

[35] Abdulmumin, Idris et al., (2022) "Separating Grains from the Chaff: Using Data Filtering to Improve Multilingual Translation for Low-Resourced African Languages", *arXiv:2210.10692v2*.

[36] Jon, Josef et al., (2023) "Negative Lexical Constraints in Neural Machine Translation", *arXiv:2308.03601v1*.

[37] Trieu, Hai Long et al., (2023) "VBD-MT Chinese↔Vietnamese Translation Systems for VLSP 2022.", *arXiv:2308.07601v1*.

[38] Tan, Zhixing et al., (2020) "Neural Machine Translation: A Review of Methods, Resources, and Tools", *arXiv:2012.15515v1*.
[39] Iyer, Vivek et al., (2023) "Towards Effective Disambiguation for Machine Translation with Large Language Models", *arXiv:2309.11668v1*.
[40] Brunda B N., Lavanya Santhosh., Brunda N C., Veena Potdar and Indu N (2023); COMPARATIVE STUDY OF MACHINE TRANSLATION TECHNIQUES *Int. J. of Adv. Res. 11 (Jun)*. 387-402, doi:10.21474/IJAR01/17083.
[41] Dong, Jun, (2023) "Transfer Learning-Based Neural Machine Translation for Low-Resource Languages", *ACM Transactions on Asian and Low-Resource Language Information Processing*, doi:10.1145/3618111.
[42] Wang, Lan, (2023) "The Impacts and Challenges of Artificial Intelligence Translation Tool on Translation Professionals", *In Proceedings of the 8th International Conference on Social Sciences and Economic Development (ICSSED 2023)*, SHS Web of Conferences, Vol. 163, No 02021, doi:10.1051/shsconf/202316302021.
[43] Sheng, Anfeng et al., (2023) "Neural machine translation and human translation, A political and ideological perspective", *Babel*, Vol. 69, No. 4, pp. 483 - 498, doi:10.1075/babel.00332.she.
[44] Haq, Ijazul et al., (2023) "NLPashto: NLP Toolkit for Low-resource Pashto Language", *(IJACSA) International Journal of Advanced Computer Science and Applications*, Vol. 14, No. 6, pp. 1344-1352.
[45] Stadler, Patrick et al., (2021) "Observing the Learning Curve of Neural Machine Translation with regard to Linguistic Phenomena", https://api.semanticscholar.org/CorpusID:260083896.

AUTHOR

Kouassi Konan Jean-Claude

- Senior Computer Scientist, Civilian (Public Sector) in Côte d'Ivoire since February 2013.
- Currently (September 2024) in service at the Ministry of Environment, Sustainable Development and Ecological Transition. Head of the Study, Development and Environmental Information System Service (inside the IT Department).
- Totalizing 12+ years of cumulated experiences in the field of Computer Science, in private enterprises as well as in the Ivorian Public Sector as a Civilian (Civil Servant).
- From Junior AI Expert (0-2 years of expe rience) to Middle-Level AI Expert (2-5 years of experience), I am now a Confirmed or Advanced AI Expert (5-10 years of experience) with Deep expertise in specific AI domains, strong research background, and ability to innovate new algorithms. But I am not yet fully what I am called to be; a Senior or Very Advanced AI Expert (10+ years of experience in AI) with Extensive knowledge across multiple AI domains, strategic thinking, and leadership abilities.

ORCID iD: https://orcid.org/0000-0002-4744-1335.

APPENDICES

Open-Ended Contribution to BCI-Related Data Creation

BCI Wikimedia Incubator: https://incubator.wikimedia.org/wiki/Wp/bci.
BCI Translatewiki: https://translatewiki.net/, the bci language is mentioned here as wawle (Thank you very much to Amir Aharoni for the translation rights.)
Open-ended Data Crowdsourcing website: https://pairing.excellence-integration.org.
Github: https://github.com/Kjeanclaude/mtBCI-1.0-Corpus.