Challenges and Limitations in Gathering Resources for Low-Resource Languages: The Case of Medumba

Tatiana Moteu Ngoli[1], MBUH Christabel[1], NJEUNGA YOPA[1]
[1]Metchoup
contact@metchoup.org

Abstract

Low-resource languages face significant challenges in natural language processing due to the scarcity of annotated data, linguistic resources, and the lack of language standardization, which leads to variations in grammar, vocabulary, and writing systems. This issue is particularly observed in many African languages, which significantly reduces their usability. To bridge this barrier, this paper investigates the challenges and limitations of collecting datasets for the Medumba language, a Grassfields Bantu language spoken in Cameroon, in the context of extremely low-resource natural language processing. We mainly focus on the specificity of this language, including its grammatical and lexical structure. Our findings highlight key barriers, including (1) the challenges in typing and encoding Latin scripts, (2) the absence of standardized translations for technical and scientific terms, and (3) the challenge of limited digital resources and financial constraints, highlighting the need to improve data strategies and collaboration to advance computational research on African languages. We hope that our study informs the development of better tools and policies to make knowledge platforms more accessible to extremely low-resource language speakers. We further discuss the representation of the language, data collection, parallel corpus development.

1 Introduction

The field of natural language processing (NLP) has made tremendous progress in improving low-resource languages in recent years. However, many languages remain underrepresented in computational linguistics. This is the case of Medumba, a Cameroonian language spoken by approximately 200.000 people in the western part of the country. Studies have been conducted on this particular language but these studies date back to the 90s, and focus primarily on its grammatical, structural, and

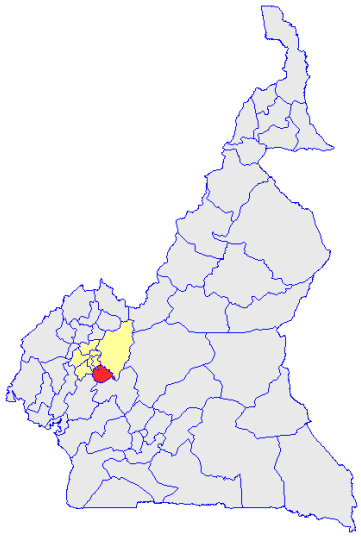

Figure 1: Representation of Medumba language

phonological aspects(Nganmou, 1991, Tchiegang, 1978, Kachin, 1990) . In addition, NLP researchers have developed benchmark datasets and parallel corpus covering specific language families, such as MasakhaNER (Adelani et al., 2021) the Sawa corpus (De Pauw et al., 2009), MasakhaNEWS (Adelani et al., 2023), WebCrawl African (Vegi et al., 2022) but, without including some extremely low-resource languages such as Medumba.

A language is considered as low-resource language by its limited linguistic resources and data, posing challenges in NLP in learning robust language patterns (Magueresse et al., 2020). On the other hand, Joshi et al. (2021) categorizes languages in six classes based on the availability of labeled and unlabeled data: *(The Left-Behinds (0)*, *(The Scraping-Bys (1)*, *The Hopefuls (2)*, *The Rising Stars (3)*, *The Underdogs (4)*, and *The Winners (5)*. In a simplified form, class 0 languages have neither labeled nor unlabeled data; class 1-4 languages have unlabeled data, but their labeled data quantity varies from virtually non-existent to high and, class 5 languages have both high volumes of labeled and

Proceedings of the Sixth Workshop on African Natural Language Processing (AfricaNLP 2025), pages 136–142

unlabeled data. However, the Medumba language might belong to either class 0 or 1 as it is very hard to find available resources, thus highlighting the need of more investigations into this particular language.

This study explores methods for building NLP resources for the Medumba language, contributing to the broader goal of enhancing language technology for African languages. We designed our analysis to mainly answer the research question: *What are the challenges and limitations of gathering and annotating an extremely low-resource language?*. To answer this question, we created a parallel French-Medumba corpus consisting of 2050 sentences translated by a professional linguist.

"Our study reveals a significant gap in categorization between the source language (French) and the target language (Medumba), making it difficult to find adequate equivalents due to the language's complexity. We summarize the main contributions of this paper as follows:

- We collected French sentences from open-source repositories related to African contexts from the web and asked a professional linguist to translate them
- We present the language background and the methodologies used to translate the sentences
- We present some baseline model results and discuss their performance
- We highlight the challenges and limitations encountered during data collection and propose solutions to overcome them

2 Related works

In this section, we provide an overview of related studies on extremely low-resource languages, specifically Medumba.

Research on Cameroonian languages has recently seen an evolution in the field of NLP. Echu (2004) investigate into the multilingualism and language policy since the colonial period of Cameroon while Olson and Meynadier (2015) assess the articulation and phonology of bilabial trills and vowels in Medumba. Moreover, a syntax of A-dependencies in Bamileke Medumba have been study (Keupdjio, 2020), and more recently, Zimmermann and Kouankem (2024) discuss the structural realization of contrastive focus in the Grassfields Bantu language Bamileke Medumba, and

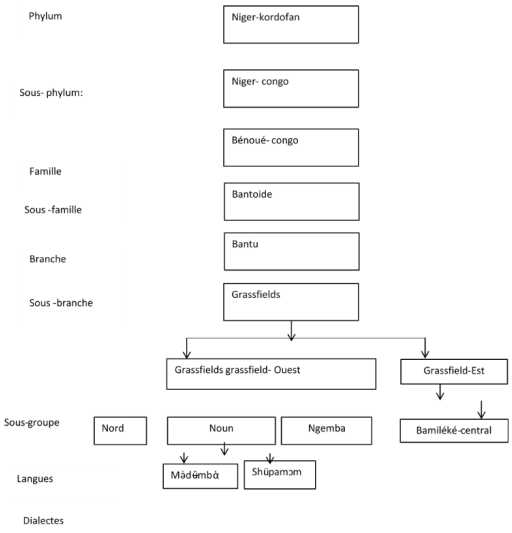

Figure 2: Family tree of the Medumba language.

Kouankem (2022) analyses the interaction between the syntactic structure and the semantic outcome of serial verb constructions in Medumba. Althought these studies investigates the Medumba language, the are more focused on the structural syntax and semantical aspect of the language, without highlighting the challenges of translating text into Medumba. In this study, we investigate the challenges of gathering resources in the medumba by highlighting the methodology, the challenges and some techniques used to translate sentences from a source language to the Medumba language.

3 Medumba language

3.1 Background

The Medumba (mèdʉ̂mbà) language is a Bamileke language primarily spoken in Cameroon in the Ndé department, West region, with the main settlements being *Bangangté, Bangoulap, Bakong, Bahouoc, Bagnoun, Bawouok, Tonga, Bamaha, Bagnoun*. It is also spoken in the North-West by the *Bahouoc* in the Bali district (futher details can be found in Figure 2). According to the Ethnology [1], this language belongs to Niger-Congo language family, the Eastern Grassfields group, and the Central Bamileke subgroup with over 210.000 speakers (htt, 2018). Medumba belongs to zone 9 of the Southern Grassfield languages, with Alcam code [997] (BIKOI, 2018). The Medumba language has a dialectal variant called nsî ntuñ spoken in *Tonga, Bandounga, Bassamba* and in part of *Bazou*. The standard reference variant is known as *bangangte*.

[1] https://www.ethnologue.com/language/byv/

Medumba language is governed by a set of rules. In terms of morphology, Medumba is monosyllabic, i.e the morphemes of this language are initially formed of one syllable. We can have examples like t´α/father, m´α/mother, nk$\ddot{u}\grave{n}$/the news, f´α/work, nvə̀n/the chief, etc. We also find disyllable and trisyllable words such as: ngə̀láη/paternal uncles or aunts, mϵntu$\grave{n}$/someone, ngàzìfə/the learner, etc. The morphemes of the Medumba language always begin with consonants and the tones are essentially marked on the vowels and on the consonant η. Vowels, on the other hand, always occupy the medial and final position in a word. The grammatical classes of this language are nouns, prepositions, adverbs, adjectives, conjunctions, verbs and pronouns. There are 5 noun classes, including 3 singular classes (classes 1, 3 and 5) and two plural classes (classes 4 and 6). The formation of the plural is done according to the noun class concerned. In general, the word (ba) is used as a plural marker. Compound nouns are written as a single word.
Syntactically, the sentence in Medumba generally follows the SVO (Subject-Verb-Object) structure. A set of orthographic principles governs this language. The following principles serve as examples:

- Do not write the same consonant twice in a word. This would simply mean that if at the time of pronunciation, we perceive a sound twice, we replace the first one with a sound that is close to the first, unless the first sound is separated from the second by the glottal stop. Example: bϵètə will be written bϵdtə , sà´

- The vowel /ɔ/ is never placed before /g/ and /η/ even if it is heard when pronouncing a word. Example: lɔη will be written loη in this word; the grapheme o is not read /ou/ as its alphabet requires, but it is read as /ɔ/

The phonology of Medumba is made up of 32 letters including ten 10 vowels, 22 consonants and five (5 tones. The different vowels of the Medumba language are / a, e, ∂, ϵ, i, u, 0, α, o, O /. Depending on the points of articulation, Table 1 and Table 2 summarize the classification of its different vowels.

The vowels of the Medumba language can be closed, half-closed, half-open or open. Among these vowels, we have two pairs of vowels that are represented differently in spelling, but are read the same way. These are (i/e and u and o). The concept of aspiration is crucial in distinguishing

	Anterior	**Central**	**Posterior**
Closed	i	~~ʉ~~	u
Half-closed	e	-	o
Half-open	ϵ	∂	ɔ
Open	-	α	a

Table 1: Medumba vowels

writing from reading. The consonants are b, d, c, k, f, s, g, j, h, sh, gh, l, m, n, v, z, y, η, ny, ‘, w and ts. Moreover, Kouankem (2012) summarizes these letters according to their place of articulation as follows, the punctual tones found in the Medumba language are the high tone, the low tone and the mid tone. The modulated tones are: the falling tone and the rising tone. In the writing of this language, the high tone and the low tone are not marked.

3.2 Data collection

We mainly worked on the translation of 2050 sentences from French to Medumba collected on the web. The sentences come from various categories and are based on African contexts (e.g. *Un seul projet est réalisé au Cameroun ou dans le cadre de la CEMAC un vaste programme de production d'engrais à la mesure des besoins de notre agriculture*); More examples can be found in Figure 3. This study made it possible to identify the specific obstacles linked to the absence of lexical equivalents and the differences in linguistic categorization between French and Medumba. To overcome these challenges, we adopted a methodical approach including:

- Consulting native speakers and existing documents on the Medumba lexicon

- Using translation techniques such as explanation and adaptation

- Lexical creation or neologism while respecting the grammatical principles of the target language

- Validation of translations with the Medumba language development committee

4 Methodology

We conducted a qualitative study based on the analysis of discussions from online forums and African content creators. We applied analysis to identify recurring problems and concerns encountered by

	Bilabial	Labio-dental	Alveolaire dental	Palatal	Velaire	Glottal
Plosives	b	-	t d	-	k	-
Nasals	m	-	n	-	η	-
Fricatives	-	f v	s sh z ts	-	gh	h
Glides	-	-	-	y	w	-
Laterals	-	-	l	-	-	-

Table 2: Medumba consonants

French	Medumba
Un seul réalisée au Cameroun ou dans le cadre de la CEMAC un vaste programme de production d'engrais à la mesure des besoins de notre agriculture.	Tà' nsâg nkǎzìn nə̀nə̀b mîcá' bò ghə̂ Kàmə̀rún kə̀ ntə̂m CEMAC mbɑ̌ à kú'nì nə̂m nzî zə̀bə̀ ngə̂mnə̀ ló
D'autres suivront avec l'aménagement du cours de la Sanaga.	Tsə̀mò' à' sə̂' bô nə̀nə̀b ntsə̂ Sanaga
Notre pays ne fait pas exception en Afrique.	Zə̀bə̀ lá' kə̂' tàg Àflìkà
La lutte contre le VIH/SIDA est une préoccupation importante pour les Synergies Africaines.	Zwə̂' VIH/SIDA bə́ ɑ̌ tà' nù tânjòŋ á cwə̌d ngʉ̀dnì tàmtə̀ ghǎfà' Aflìkà
M Samuel MVONDO AYOLO Directeur du Cabinet Civil de la Présidence de la République avec rang et prérogatives de Ministre.	Tɑ̌ Samuel MVONDO AYOLO, ngàcàgtə̀ Cabinet Civil ndə̂ŋɔ́ bô bìn nkə̂mŋɔ́
Pour d'autres il n'est pas toujours aisé de réunir toutes les pièces que l'administration exige.	À bə̂ ntâ tsə̀mə̂' bə̀nntə̀n nə̀ kʉ́mtə́ njôŋ fə̂ ŋwà'nì ngàcàgtə̀ cwə̌d mbə̂dtə́ lá
Le jeune garçon est le fils d'une cousine à elle et n'avait alors que 10 ans à cette époque.	Mə̂n mándûm lî bə́ nshûm bə̌mɑ́ ĩ. À ná' ngû ɑ̌ ngʉ̀' ghám ngə̀lâŋ bò ná' ndô' ĩ lá
Des campagnes de sensibilisation sont aussi organisées sur les dangers de la drogue.	Bó cwə̌d nɑ̀b nkǎzìn nə̀túm nzí'tə́ bə̀ntə̀n nə̂m cɔ̀kə̀bwɔ̀ fùkə̀bwɔ̀
Absente du domicile conjugal depuis vendredi dernier, la veuve du défunt n'est revenue que ce mardi dans la soirée.	Mbɑ̌ mfôg mènnzwî lî ná' túm ndɑ̌ ndú ĩ mfə̂nntɑ̂nndʉ́b, ntû' bə̂nnjàm ɑ̌ mvə̂dnjʉ́ mfə̂nntɑ̂nnkə̄'ə̄
Je crois que ce monument est une belle réalisation.	Mə́ kwâ mbə̀ sə̀n lôŋlá' lî bə́ ɑ̌ bwɔ̂ fà'|

Figure 3: Samples translated sentences.

contributors. To improve the translation, we inquired whenever we were faced with a complex term whose translation was not immediately apparent. We thus verified the non-existence of the term itself before moving on to adopting a specific translation technique. For some terms, we drew inspiration from their explanations in French to translate them. In addition, we drew inspiration from the principle of forming the grammatical category to be translated in the target language to create a new word designating the term in the source language.

4.1 Medumba Dataset

The Medumba dataset is a translated version of French sentences collected from open-source repositories such as GitHub [2], covering multiple topics. After preprocessing, we use 31,679 tokens to train our baseline models. The dataset statistics are shown in Table 3. Furthermore, we split the dataset into train and test to train our baselines models as showed in 4.

4.2 Baselines Performance

To conduct our experiments, we chose to fine-tune custom pre-trained machine translation models, as our parallel corpus includes Medumba, a language not supported by most existing models. This approach enables the model to learn translation patterns specific to Medumba. For instance, we fine-tuned models such as opus-mt-fr-en[3], mbart50[4], and t5-small[5]. The results are reported in Table 5.

As metrics, we use:

- **BLEU** (Bilingual Evaluation Understudy): A metric that calculates n-gram precision for various n-gram lengths (typically 1 to 4) and combines these scores using a geometric mean. It also incorporates a brevity penalty to address the issue of overly short translations.

- **COMET** (Cross-lingual Optimized Metric for Evaluation of Translation): A metric that employs machine learning models to evaluate translations. Unlike traditional metrics, it does not rely solely on surface-level text comparisons. It assesses translations based on fluency, adequacy, and the preservation of meaning.

- **TER** (Translation Edit Rate): A metric that calculates the minimum number of edits required to transform a machine translation into one of the reference translations. The score is normalized by the total number of words in the reference translation.

The results reveal that only the T5-small model achieves a high BLEU score, while the other two models exhibit higher COMET scores. Since COMET is effective in scenarios requiring a deeper

[2] https://github.com/

[3] https://huggingface.co/Helsinki-NLP/opus-mt-fr-en

[4] https://huggingface.co/sarubi/mbart-50

[5] https://huggingface.co/google/flan-t5-small

	Tokens	Nbr documents	Vocab size
fr	33304	2052	6786
byv	31679	2052	4542

Table 3: Datasets tokens count

Train	Test
1846	206

Table 4: Datasets split

understanding of translation quality, it is particularly useful for evaluating translations where contextual and semantic accuracy are more important than literal word-for-word correspondence—an evaluation criterion well aligned with the characteristics of our Medumba dataset. The other results were expected, given the limited size of our dataset.

5 Challenges and Limitations

The translation of the 2.050 sentences from French to Medumba was mainly hampered by the lack of adequate equivalent terms in the target language and differences in categorization between the two languages.

5.1 Challenges related to platform interfaces and language support

The Medumba language uses the Latin alphabet, which requires complex diacritical characters, making typing cumbersome. Platform updates sometimes disrupt existing input methods, causing frustration among contributors. In addition, we have faced some challenges in translating scientific and technological terms due to lack of consensus on local language equivalents. For example, terms like *spammer robots* or *word processing* had to be translated using periphrases in Medumba, while others, such as *JavaScript* and *thermal power station* remain untranslatable due to a lack of corresponding concepts. It was also impossible to translate scientific concepts from physics, such as *thermal power station* and *hydroelectric dam*, because there are no equivalents or realities that could provide inspiration for a satisfactory adaptation of these words. Some legal terms or expressions, such as *decree*, *democracy*, *order*, *Commander of the National Order of Value* and *State of the General Staff*, etc., have no equivalents in the Medumba language and have been maintained as borrowings in the target language. All in all, the absence of direct equivalents in the Medumba language has led to the use of periphrases and borrowings. On the other hand, the lack of spelling uniformity complicates access to information. Medumba has a great deal of variability in the writing of words and many homophones, which hinders the performance of search engines and automatic correctors. Furthermore, the difference in categorization between the Medumba language and the French language has also hampered the translation of certain specific concepts such as *ambassador* and *charge of mission* in two very different contexts, but the Medumba language classifies both under the generic term *ngàntùm/envoye*.

Models	BLEU	COMET	TER
opus-mt-fr-en	15.82	**0.80**	82.51
mbart50	20.36	**0.80**	77.15
T5-small	**83.20**	0.42	94.97

Table 5: Baselines results. Values in bold represent high scores.

5.2 Financial and material barriers

The lack of access to reliable internet, digital libraries and reference materials has greatly hampered work and generated significant costs. Furthermore, there is a shortage of online media, there are many African platforms [6] [7] [8] created, but very few promote Medumba. The Medumba language has a radio called Radio Medumba, however it is only accessible in the Ndé department. This media serves as a channel for broadcasting Medumba language learning programs through games, stories and the popularization of new words created by the Medumba language development committee mainly in the Medumba area. This implies that accessibility to this radio is limited. Given this reality, we therefore rely heavily on our own internal research work.

6 Conclusion

In this study, we investigate the challenges and limitations of gathering resources for an extremely low-resource language: Medumba. We present the language's background, the methodology used to translate sentences from French to Medumba, and particularly highlight the challenges encountered

[6] https://www.languagesafrica.com
[7] https://github.com/masakhane-io/lafand-mt
[8] https://github.com/masakhane-io/masakhane-mt

during the translation process. Our findings reveal that discrepancies in categorization between the source and target languages contribute to translation complexity. To address these limitations and advance the state of the art in low-resource languages, future research should explore additional techniques for resource gathering and enhance translation capabilities for extremely low-resource languages.

References

2018. Issue information. *IPPR Progressive Review*, 25(2):107–107.

David Ifeoluwa Adelani, Jade Abbott, Graham Neubig, Daniel D'souza, Julia Kreutzer, Constantine Lignos, Chester Palen-Michel, Happy Buzaaba, Shruti Rijhwani, Sebastian Ruder, Stephen Mayhew, Israel Abebe Azime, Shamsuddeen Muhammad, Chris Chinenye Emezue, Joyce Nakatumba-Nabende, Perez Ogayo, Anuoluwapo Aremu, Catherine Gitau, Derguene Mbaye, Jesujoba Alabi, Seid Muhie Yimam, Tajuddeen Gwadabe, Ignatius Ezeani, Rubungo Andre Niyongabo, Jonathan Mukiibi, Verrah Otiende, Iroro Orife, Davis David, Samba Ngom, Tosin Adewumi, Paul Rayson, Mofetoluwa Adeyemi, Gerald Muriuki, Emmanuel Anebi, Chiamaka Chukwuneke, Nkiruka Odu, Eric Peter Wairagala, Samuel Oyerinde, Clemencia Siro, Tobius Saul Bateesa, Temilola Oloyede, Yvonne Wambui, Victor Akinode, Deborah Nabagereka, Maurice Katusiime, Ayodele Awokoya, Mouhamadane MBOUP, Dibora Gebreyohannes, Henok Tilaye, Kelechi Nwaike, Degaga Wolde, Abdoulaye Faye, Blessing Sibanda, Orevaoghene Ahia, Bonaventure F. P. Dossou, Kelechi Ogueji, Thierno Ibrahima DIOP, Abdoulaye Diallo, Adewale Akinfaderin, Tendai Marengereke, and Salomey Osei. 2021. Masakhaner: Named entity recognition for african languages. *Preprint*, arXiv:2103.11811.

David Ifeoluwa Adelani, Marek Masiak, Israel Abebe Azime, Jesujoba Alabi, Atnafu Lambebo Tonja, Christine Mwase, Odunayo Ogundepo, Bonaventure F. P. Dossou, Akintunde Oladipo, Doreen Nixdorf, Chris Chinenye Emezue, sana al azzawi, Blessing Sibanda, Davis David, Lolwethu Ndolela, Jonathan Mukiibi, Tunde Ajayi, Tatiana Moteu, Brian Odhiambo, Abraham Owodunni, Nnaemeka Obiefuna, Muhidin Mohamed, Shamsuddeen Hassan Muhammad, Teshome Mulugeta Ababu, Saheed Abdullahi Salahudeen, Mesay Gemeda Yigezu, Tajuddeen Gwadabe, Idris Abdulmumin, Mahlet Taye, Oluwabusayo Awoyomi, Iyanuoluwa Shode, Tolulope Adelani, Habiba Abdulganiyu, Abdul-Hakeem Omotayo, Adetola Adeeko, Abeeb Afolabi, Anuoluwapo Aremu, Olanrewaju Samuel, Clemencia Siro, Wangari Kimotho, Onyekachi Ogbu, Chinedu Mbonu, Chiamaka Chukwuneke, Samuel Fanijo, Jessica Ojo, Oyinkansola Awosan, Tadesse Kebede, Toadoum Sari Sakayo, Pamela Nyatsine, Freedmore Sidume, Oreen Yousuf, Mardiyyah Oduwole, Tshinu Tshinu, Ussen Kimanuka, Thina Diko, Siyanda Nxakama, Sinodos Nigusse, Abdulmejid Johar, Shafie Mohamed, Fuad Mire Hassan, Moges Ahmed Mehamed, Evrard Ngabire, Jules Jules, Ivan Ssenkungu, and Pontus Stenetorp. 2023. Masakhanews: News topic classification for african languages. *Preprint*, arXiv:2304.09972.

Charles BIKOI. 2018. *BINAM BIKOI Ch.- (dir.)- Atlas linguistique du Cameroun, 2012, Yaoundé, éd. Cerdotola, 399 p.*

Guy De Pauw, Peter Waiganjo Wagacha, and Gilles-Maurice de Schryver. 2009. The SAWA corpus: A parallel corpus English - Swahili. In *Proceedings of the First Workshop on Language Technologies for African Languages*, pages 9–16, Athens, Greece. Association for Computational Linguistics.

George Echu. 2004. The language question in cameroon. *Linguistik online*, 18.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2021. The state and fate of linguistic diversity and inclusion in the nlp world. *Preprint*, arXiv:2004.09095.

Brigitte Kachin. 1990. *The phonological adaptation of English loan words in Medumba.*

Hermann Sidoine Keupdjio. 2020. *The syntax of A'-dependencies in Bamileke Medumba*. Ph.D. thesis, Universitu of British Columbia.

Constantine Kouankem. 2012. The syntax of the medumba determiner phrase. *Yaounde: University of Yaounde I dissertation*.

Constantine Kouankem. 2022. Issues on serial verb constructions in medumba. *Language in Africa*.

Alexandre Magueresse, Vincent Carles, and Evan Heetderks. 2020. Low-resource languages: A review of past work and future challenges. *Preprint*, arXiv:2006.07264.

Alise Nganmou. 1991. *Modalités verbales: Temps, Aspect et Mode en Mdūmba*. Ph.D. thesis, Université de Yaoundé I.

Kenneth Olson and Yohann Meynadier. 2015. On medumba bilabial trills and vowels.

Luc Tchiegang. 1978. Bangangté-deutsch-konversationsbuch. Thesis, Saarbrücken: Universität des Saarlandes.

Pavanpankaj Vegi, Sivabhavani J, Biswajit Paul, Abhinav Mishra, Prashant Banjare, Prasanna K R, and Chitra Viswanathan. 2022. WebCrawl African : A multilingual parallel corpora for African languages. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1076–1089, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Malte Zimmermann and Constantine Kouankem. 2024.
Focus fronting in a language with in situ marking:
The case of mdmbà. *Languages*, 9(4).