Online version available in : http://arbitrer.fib.unand.ac.id

JURNAL ARBITRER

| 2339-1162 (Print) | 2550-1011 (Online) |

Review

Developing the Corpus of Minangkabau Language: Insights, Challenges, and Future Directions

Handoko Handoko

Faculty of Humanities, Universitas Andalas, Padang, Indonesia

Submission Track

Received: May 24, 2024
Final Revision: September 13, 2024
Accepted: September 15, 2024
Available Online: September 25, 2024

Keywords

Minangkabau corpus, language documentation, language preservation, corpus methodology, digital resources

Correspondence

E-mail: handoko@hum.unand.ac.id

A B S T R A C T

This paper discusses the design for developing the Minangkabau language corpus, especially regarding the opportunities and challenges. The corpus development of Minangkabau is a crucial project to document, preserve, and revive the treasure trove of culture within the language. The availability of a Minangkabau language corpus can open opportunities for more intensive research on the Minangkabau language with a more modern and data-based approach. It can also encourage the development of Minangkabau corpus-based teaching materials. The corpus is manually assembled using various sources' comprehensive data collection, annotation, and curation pipelines. These may be manuscripts, books, newspapers, or other written texts and spontaneous conversations, such as interviews or public speeches. Multimedia resources, such as television and radio broadcasts, audio-video recordings, and social media content, also add to the diversity of data gathered. The availability of accessible digital sources, such as online videos, online radio programs, and ebooks, can make data collection easier. However, several challenges may appear in developing the Minangkabau language corpus, such as limited technology accessibility, dialect variations, and the involvement of highly skilled human resources. This paper explains some opportunities for developing the Minangkabau language corpus and increasing the role of the corpus in revitalizing and documenting the Minangkabau language. Furthermore, the availability of the Minangkabau language corpus can also be a starting point for developing linguistic technology, such as voice recognition, text-to-speech, and natural language processing.

I. INTRODUCTION

The Minangkabau language is one of the Malay language families spoken by most of the people of West Sumatra, Indonesia. In addition, the Minangkabau language is also spoken in other regions, namely by people in Jambi, Riau, and Negeri Sembilan (Aman et al., 2019; Anwar, 1980). Although Minangkabau is closely related to Malay and the other languages grouped with it in the "Malayic" subgroup of Austronesian, distinct phonological, lexical, and grammatical features set it somewhat apart from the others (McGinn, 2009).

Over the past few decades, the Minangkabau language has been undergoing changes that threaten its vitality. This is becoming more common among the young as they are losing their competence in using the Minangkabau language since Indonesian has been replaced with the official national language that is used in school, mass media, and daily conversation (Suryani, 2018). Conversely, language is not only a means of communication but also part of identity and a reflection of the culture. Therefore, maintaining the Minangkabau language is key to preserving Minangkabu culture. Among the ways that can be done to preserve the Minangkabau language is to develop a well-designed corpus. This corpus can be a valuable data source for linguistic analysis, developing teaching material, as well as using linguistic data-based technology, such as speech recognition and speech generation (McEnery & Xiao, 2011; O'Keeffe & McCarthy, 2010).

DOI: https://doi.org/10.25077/ar.11.3.413-429.2024

Unlike the English Language or other major languages, the research on corpus linguistics in regional languages such as Minangkabau is still limited (Koto & Koto, 2020), Although some studies have engaged in constructing and analyzing corpora, general language documentation of Indonesian local languages including collaborative efforts for creating useful corpora are yet to be developed (Hamamah 2023). Advanced corpus tools, such as Sketch Engine, are available but their usage in Indonesian research is limited by subscription costs and the adoption of corpus linguistics by the academic community (Isti'anah et al., 2023).

The paper aims to investigate the development of the Minangkabau Language corpus and how available resources can be used as data for corpus development. This paper also explores the challenges in developing the corpus and research opportunities that can be carried out using the Minangkabau language corpus. Later, this paper will also discuss the methodology and some strategies in different places that have been successful in developing a regional language corpus so it can be applied to Minangkabau.

II. DEVELOPING THE MINANGKABAU LANGUAGE CORPUS

Over the decades, corpus linguistics has had many more uses and provided a great resource for language studies in general. Major languages, such as English, French, and Spanish, have large annotated corpora that consist of billions of words and are annotated with various linguistics annotations, such as phonology, morphology, lexical, syntax, semantics, pragmatics, sociolinguistics, and discourse analysis (Zufferey, 2020). Such corpora allow further research of various linguistic phenomena and serve applied linguistics, language teaching, cultural studies, and discourse analysis (Al-Hamzi et al., 2020). British National Corpus (BNC), for instance, demonstrates how large and complex annotated corpus can serve as an environment for testing linguistics theories, including phonology, semantics, syntax, and discourse analysis (Love et al., 2022; Mustafa & Yusuf, 2021; Nurmukhamedov & Sharakhimov, 2021; Peksoy, 2017)

Corpora also play a critical role in the evolution of linguistic theories, and language models. It has contributed significantly to modern corpus-based linguistic studies by providing valuable and reliable contextualized information (Xiao, 2008). Corpora provided a more reliable way to test linguistic hypotheses when considering the frequency and distribution of language in larger samples (Stubbs & Halbe, 2012). The role lies in verifying and modifying existing theories as well as developing new linguistic theories (Xiao, 2008). Working with corpora has given insights into adequate and accurate statistical analysis of language use, including concordances, collocations, and frequency studies (Anthony, 2013; Cushing, 2017; Meyer, 2002; Stubbs & Halbe, 2012). While the precise relationship between corpus linguistics and theoretical linguistics remains debatable, corpus-based research has come to be seen as an important complement to the traditional approach (Barlow, 2011).

The development of corpora in regional languages holds the key to conserving cultural history, education reform, and social inclusion. These corpora are valuable to researchers, learners, and teachers as they offer authentic texts for language and cultural studies (Juško-Štekele & Kļavinska, 2024). They facilitate the digitization and preservation of ethnocultural units, helping to maintain and renew national culture across generations (Barmenqulova, 2024; Felde, 2022; Rao et al., 2020). Bilingual corpora, in particular, serve as modern libraries of language heritage and reflect advanced contrastive analysis (Dimitrova & Garabík, 2012; Lane et al., 2022). Collaboration among corpus compilers is necessary for addressing common challenges and applying shared methods despite different needs and goals (Kretzschmar et al., 2006).

In language education, corpora can serve as a tool to increase access to multilingual quality education by offering authentic resources in the local languages of the different linguistic communities, which are indispensable for contributing towards promoting quality. In this way, community use of heritage languages within formal educational systems can reinforce cultural identity and resilience (Migge & Léglise, 2010; Oriyama, 2010; Poku, 2024; Tembe & Norton, 2008). Additionally, corpus-based methods work in synchrony with hands-on learning and inquiry-based approaches to language use — as opposed to the simplified examples common in structured teaching (McEnery & Xiao, 2011; O'Keeffe et al., 2007; Johns, 2002).

This enables effective language learning and allows for the creation of materials by educators that are reflective of the linguistic needs of certain communities, reinforcing cultural understanding and more successful learning outcomes (Boulton & Landure, 2016; Bennett, 2010; Breyer, 2009).

Overview of Existing Linguistic Corpora

Over the years, the field of corpus linguistics has developed considerably thanks to new methodological and technological advancements. In the beginning, corpora were mostly made manually with early corpus such as the Brown Corpus (1964) showing an organization of texts for linguistic analysis (Kytö, 2011)). Some of the well-known language corpora include the Corpus of Contemporary American English (COCA), which contains over 1 billion words from diverse genres; the British National Corpus (BNC), representing a wide range of British English; and the International Corpus of English (ICE), which allows for comparative studies of English varieties globally. Additional noteworthy corpora include the Michigan Corpus of Academic English (MICASE), focusing on academic discourse; the CHILDES Database for child language acquisition; and the Corpus of Global Web-Based English (GloWbE), which provides insights into online communication.

Regional and minority language corpora emerged as a corollary to this development in response to the necessity of documenting languages that are under threat of extinction. Corpus linguistics has also been used in projects such as the Digital Indigenous Language Archive (DILA) and the Endangered Languages Documentation Programme (ELDP) to help preserve linguistic diversity (Nathan & Austin, 2004; Grenoble & Whaley, 2006). Today, there are many regional languages available, such as Corpus of Contemporary Arabic (CCA), The Indian Languages Corpora Initiative (ILCI), Bangla National Corpus, Uzbek National Corpus, Korpus Dewan Bahasa dan Pustaka (Malay), TITML Corpus (Japanese), Finnish National Corpus, Russian National Corpus (RNC), Thai National Corpus, Kurdish Corpus, and many more. Some minority languages also have collection of data in the form of a corpus, such as the Basque Corpus (spoken in the Basque region of Spain and France), the Sami Corpus (spoken by the Indigenous Sami people in Northern Scandinavia), Ainu Language Corpus (corpus developed for Ainu, an indigenous language of Japan), Nahuatl Language Corpus (language spoken in Mexico), Buryat Language Corpus (spoken in parts of Russia, Mongolia, and China), and many more. The creation of corpora for minority languages is fundamental for the documentation and preservation of linguistic diversity, offers resources to support language revitalization efforts, and provides data for academic research and language technology (Arkhangelskiy, 2019; Lane et al., 2022; Trosterud, 2002).

Previous Work on Minangkabau Language Documentation

Prior records of the Minangkabau language are mostly in descriptive linguistic research and dictionary compilation (Dipatuan, Rusmali et al., 1985; Saydam, 2004). Adelaar's (1992) comprehensive grammar of Minangkabau remains foundational, offering detailed descriptions of its phonological, morphological, and syntactic structures. Furthermore, research that examined the lexical language of Minagkabau was also carried out, such as lexical variations from several regions in West Sumatra (Heni & Suryadi, 2022; Nesti, 2016; Novita et al., 2021; Razin & Subiyanto, 2024; Reniwati & Khanizar, 2022), and comparison of lexicons with languages in Minangkabau outside West Sumatra (Reniwati et al., 2017).

Despite the foundational studies on Minangkabau, significant gaps remain in the comprehensive documentation and analysis of the language. One major gap is the lack of a large, annotated corpus that captures the linguistic diversity and usage patterns of Minangkabau across various social contexts (Koto & Koto, 2020). While some of the resources provide a wide range of free resources focused on supporting a particular dialect or genre, most do not give an elaborate view (Himmelmann, 2006). However, the applicability of most language processing tools has been limited to major languages, since they are predominantly designed for the description of major languages and have excluded minority languages like Minangkabau (Bird & Simons, 2003). These are critical gaps that need to be addressed in order for documentation and preservation work on Minangkabau to move forward.

Furthermore, community involvement in language documents has been insufficient as well. Good language documentation also requires the involvement of speakers and communities, not

just as assistants in data collection (Chelliah & de Reuse 2011). By soliciting active participation early on, the project can ensure that documentation efforts reflect the cultural values and needs of the language community.

Importance of Developing a Minangkabau Corpus

Corpus linguistics approaches have greatly facilitated language research in many fields. The methods behind it serve as the empirical foundation providing more insights into language use and structure through the analysis of massive textual data (McEnery & Hardie, 2011; Philip, 2018). The application of corpus-based methods has pushed forward several fields of linguistic research, including lexicology, discourse studies, and literary stylistics (Bhatia et al., 2008; Philip, 2018). Using corpus-based methods, researcher have addressed a deeper analysis of topics across discourse analysis, grammar in use, and meaning, both integrating sociocultural ideology and the cognitive evaluation of text (Conrad 2002; Deignan 2005; Gerbig & Mason 2008; Vessey, 2015). Corpus linguistics has been successful in two ways, promoting interdisciplinary research and joining linguistic disciplines to contact the humanities and social sciences (Ancarno, 2018). More recent studies on language variation and change have started to engage with more diverse data sources, such as social media platforms (Coats & Laippala, 2024). The field is still growing, with theoretical and methodological questions being addressed while empirical approaches to language study increase (Aijmer & Altenberg, 2004). Using the Minangkabau language corpus, linguistic research can be more comprehensive with real and varied data (data-driven language analysis). Moreover, a quantitative approach can also be used to describe linguistic phenomena in the Minangkabau language.

Moreover, corpora are commonly used for the development of NLP applications like speech recognition systems and machine translation tools to fit language into the modern technological environment (Musgrave, 2014). Recent work has provided Minangkabau corpora for sentiment analysis and machine translation to combat the lack of annotated resources in this language (Koto & Koto, 2020). Such digital resources can facilitate computational linguistics research and applications such as the Android-based Minangkabau language learning application (Oswari et al., 2020). These corpora repositories are critical resources for a variety of purposes, such as natural language processing tasks or linguistic applications for interactive language learning tools.

In education, the corpus can be used to develop teaching/learning materials based on Minangkabau linguistic features to give more natural language knowledge to purchasers. Corpus linguistics has contributed vastly to language education with practical implications in teaching and learning (Flowerdew, 2011). They can be effectively employed for syllabus development, materials creation, and data-driven learning to supplement both general and specialized language education (Römer, 2011). The integration of corpus linguistics in teacher education programs can improve teachers' research skills and language awareness, enabling them to create more effective classroom tasks (O'Keeffe & Farr, 2003). The empirical data integrated with the intuitions of potential speakers give corpus-based approaches an advantage over their non-empirical counterparts in terms of objectivity and representativeness (McEnery & Xiao, 2011).

Despite some criticism, corpora have become widely used in linguistics and language pedagogy, contributing to lexicography, grammar, language variation, translation studies, and discourse analysis (McEnery & Xiao, 2011). The combination of teaching and language corpora is increasingly becoming a convergence proving to be valuable in the field of language education (McEnery & Xiao, 2011). Additionally, it serves as a cultural preservation tool, capturing the oral traditions and linguistic practices of the Minangkabau people for future generations (Handoko et al., 2024; Nelisa et al., 2021; Sakti & Nakamura, 2013).

III. METHODOLOGY IN DEVELOPING CORPUS MINANGKABAU

Data and Source of Data

The development of a comprehensive Minangkabau corpus requires a systematic approach to data collection to ensure it accurately reflects the full diversity of the language. To capture the rich linguistic landscape of Minangkabau, data is sourced from a wide array of spoken and written forms. The most important are recordings of spoken language, such as natural conversations, interviews, oral traditions, and public speeches. The recordings

provide unique insights into everyday language practices, local dialects, and cultural expressions. Moreover, written sources are also important, including traditional and modern materials such as books, manuscripts, and digital content. Social media, blogs, and mailing lists are quoted to give you insight into the current usage of language in informal, online communication. By collecting data from these varied sources, the corpus portrays the living and changing character of the Minangkabau language in different social contexts and media and creates a resource for linguistic research, language preservation, and computational applications.

Audio-visual documents serve as valuable resources for developing a comprehensive Minangkabau language corpus. TV and radio programs can be useful resources for developing the spoken corpus of the Minangkabau language. In West Sumatra, two prominent television stations, TVRI Sumatra Barat and Padang TV, offer rich data sources (see Figure 1 and Figure 2). The variety of programs in the Minangkabau language broadcast by these stations includes interviews, news reports, talk shows, scripted television dramas, and other spoken environments for which a set of authentic reference works is available. In addition, the use of audio-visual materials ensures the same linguistic features of language and regional dialectal variations, intonation and body language as accompanying cultural and social dynamics can be captured by researchers. These broadcasts provide formal and informal speech, including specialized language use in media settings. This will expand the corpus in both size and diversity regarding audio-visual content, thus rendering it a more extensive resource for analyzing Minangkabau language-in-use.

Although audio recordings transmitted via radio have a limited reach, they have an important role in disseminating information to the local area (Onyenankeya & Salawu, 2022). There are a lot of linguistic data that could be extracted from the recordings coming from these various communities in Minangkabau. Most cities and regencies in

Figure 1. Interview in Minangkabau Language on Padang TV (source: Padang TV, 2023)

Figure 2. Interview in Minangkabau Language on TVRI Sumatera Barat (source: TVRI Sumatera Barat, 2022)

Figure 3. Online Radio program in Minangkabau language (source: https://radio-online.id/rri-pro4-padang)

West Sumatra have local radio stations. The radio programs may consist of talk shows, interviews, local news, cultural discussions, and traditional storytelling, all authentic examples of the spoken language in various dialects and sociolects of the Minangkabau language.

The current digital age has made way for hundreds of online, and eventually digital, radio stations, which have further broadened the reach of these resources. Digital platforms provide wider and more convenient access to radio content by tearing down geographical barriers to collect and curate datasets from around the region (see Figure 3). Moreover, some program from various radio stations can now be accessed online with live streams or archived content often available on demand at the click of a button which can be essential sources for sophosticated corpus.

Digital radio and the broadcast of such information also allow for language data to be collected from various fieldwork contexts, including work with rural dialects, urban variations, and even cross-cultural influences (Adeyeye et al., 2021). Thus, radio is an extraordinarily important resource for constructing a Minangkabau language corpus since it enables us to capture written and spoken language on a formal-informal speech continuum directly from the media that use it and makes regional variations easily documentable. Adding radio broadcasts to the corpus data will also lead us to a greater understanding of how it is used in both high and low contexts in the larger, more general sense from local cultural expressions

Besides that, the Minangkabau also has a rich oral tradition (Amir et al, 2006). Various forms of oral tradition have been recorded on various media, such as cassette and VCD (Suryadi, 2010). This is certainly a useful resource for the making of Minangkabau language corpus based on the recordings of oral tradition. Suryadi (2010) has documented various Minangkabau oral traditions that have been recorded in the form of audio videos, such as *rabab Pariaman, indang, rabab Pesisir Selatan (or rabab Pasisia), dendang Pauah, sijobang, saluang (or bagurau), salawat dulang, randai, and traditional speech and pasambahan* (see Figure 4). More interestingly, these audio-visual recordings of Minangkabau oral traditions also indicate lexical variation because each form of oral tradition is closely related to the regional dialects in Minangkabau (Suryadi, 2010).

In terms of written resources, written language data is gathered from a variety of genres like newspapers, magazines, novels, religious texts, and so on. After the arrival of Islam, Minangkabau as an oral language had a written form in the 18th century and started to be written with the Jawi script, then later using Latin alphabet (Harun et al., 2018; Pramono et al., 2018). This script was used by officials in the 19th century, and as a symbol for identity and defiance against colonialism (Sulistiyo et al., 2023). The Minangkabau manuscripts on Islam studies topics indefinite with the Malay texts are hadith, shariah, history, and sufism (Taufiqurrahman et al., 2021). These manuscripts provide opportunities for research in philology, linguistics, and Islamic studies. Besides religious

Figure 4. Audiovisual recordings of Minangkabau oral tradition (source: Suryadi, 2010)

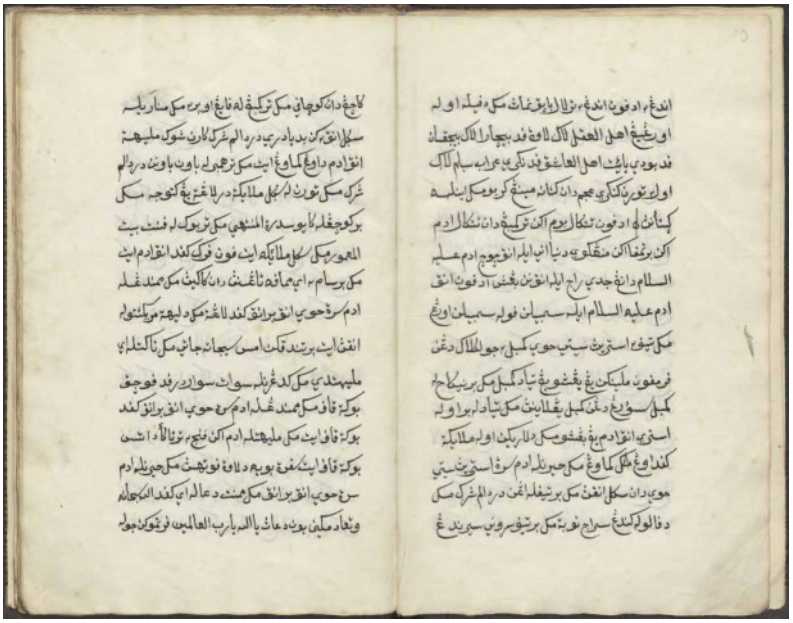


Figure 5. Kaba manuscript in Jawi script

texts, many Minangkabau manuscript talks about Minangkabau oral literature, known as *Tambo* and *Kaba* (Figure 5). While some contemporary works attempt to revitalize these texts, they may also weaken the image of traditional cultural values (Nurizzati & Nasution, 2021).

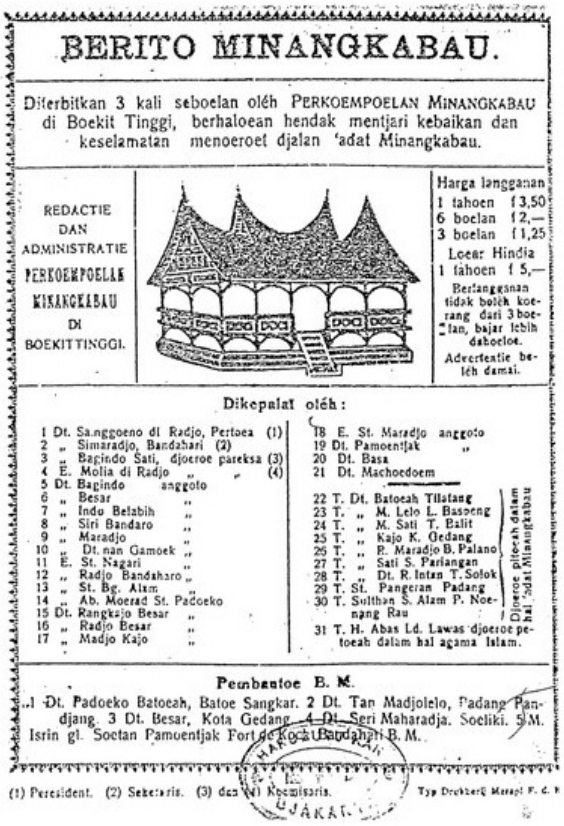
Written documents provide insights to those who study both formal and informal Minangkabau writing styles. Some sources even can be valuable resource for the classical Minangkabau language, which can be extracted from classic Minangkabau manuscripts, including *Tambo* (historical chronicles), *Kaba* (epic tales), and other traditional manuscripts (Abdullah, 1970; Jamaris, 2002). These texts provide rich historical and cultural context and are crucial for capturing the

traditional forms and uses of the language. Thus, the presence of these classical manuscripts ensures the corpus captures the richness and history of the Minangkabau language (Almos & Ladyanna, 2019; Hanif et al., 2022).

In addition to manuscripts in Jawi script, many Minangkabau books have also been written in Latin script. The use of Jawi script was widespread in the 19th century, serving as an official script and symbol of identity (Sulistiyo et al., 2023). However, Dutch colonization introduced the Latin script, resulting in the emergence of numerous Minangkabau literary works (Noranda, 2023; Nurizzati & Nasution, 2021). This transition included transcribing oral literature and writing contemporary works that transformed traditional stories and values (Nurizzati & Nasution, 2021).

The development of written languages in Minangkabau resulting various kinds of literature, including oral and written forms (Maryelliwati et al., 2018). The works were even then republished as classic literature by those in the printing industry who played an important part in the history of these books being popularized (Herbowo & Sulastri, 2020). This effort aimed to preserve Minangkabau folk literature, which contains valuable cultural wisdom and life philosophies (Herbowo & Sulastri, 2020). The existence of the publisher even provided the space for Minangkabau books to be published by local publishers, and a number were also published by national publisher Balai Pustaka Indonesia (Herbowo et al., 2021). These publications have contributed significantly to the development of the Minangkabau language corpus and literary tradition.

Books and manuscripts are among other valuable written sources for developing the Minangkabau language corpus together with magazines and newspapers. Although there are limited publications are solely in the Minangkabau language, newspapers and magazines such as *Padang Ekspress* and *Singgalang* have sections written in Minangkabau. These sections could contain articles, columns, or cultural stories from newspapers grounded in the ordinary practice of language among large groups of speakers and so provide useful tools for capturing current spoken norms. It is also feasible to access older newspapers and magazines, like *Barito Minangkabau* (Figure 6), which can help in building the classical corpus.



BERITO MINANGKABAU.

Diterbitkan 3 kali seboelan oléh PERKOEMPOELAN MINANGKABAU di Boekit Tinggi, berhaloean hendak mentjari kebaikan dan keselamatan menoeroet djalan 'adat Minangkabau.

REDACTIE DAN ADMINISTRATIE PERKOEMPOELAN MINANGKABAU DI BOEKITTINGGI.

Harga langganan
1 tahoen f 3,50
6 boelan f 2,—
3 boelan f 1,25
Loear Hindia
1 tahoen f 5,—
Berlangganan tidak boléh koerang dari 3 boelan, bajar lebih daboeloe.
Advertentie beléh damai.

Dikepalai oléh :

1 Dt. Sanggoeno di Radjo, Pertoea (1)
2 „ Simaradjo, Bandahari (2)
3 „ Bagindo Sati, djoeroe pareksa (3)
4 E. Molia di Radjo „ „ (4)
5 Dt. Bagindo anggoto
6 „ Besar „
7 „ Indo Belabih „
8 „ Siri Bandaro „
9 „ Maradjo „
10 „ Dt. nan Gamoek „
11 E. St. Nagari „
12 „ Radjo Bandaharo „
13 „ St. Bg. Alam „
14 „ Ab. Moerad St. Padoeko
15 Dt. Rangkajo Besar „
16 „ Radjo Besar „
17 „ Madjo Kajo „
18 E. St. Maradjo anggoto
19 Dt. Pamoentjak „
20 Dt. Basa
21 Dt. Machoedoem
22 T. Dt. Batoeah Tilatang
23 T. „ M. Lelo L. Basoeng
24 T. „ M. Sati T. Balit
25 T. „ Kajo K. Gedang
26 T. „ R. Maradjo B. Palano
27 T. „ Sati S. Parlangan
28 T. „ Dt. R. Intan T. Sofok
29 T. St. Pangeran Padang
30 T. Sulthan S. Alam P. Noenang Rau
31 T. H. Abas Ld. Lawas djoeroe peroeah dalam hal agama Islam.

Djoeroe pitoeah dalam hal 'adat Minangkabau

Pembantoe B. M.

1 Dt. Padoeko Batoeah, Batoe Sangkar. 2 Dt. Tan Madjolelo, Padang Pandjang. 3 Dt. Besar, Kota Gedang. 4 Dt. Seri Maharadja. Soeliki. 5 M. Isrin gl. Soetan Pamoentjak Fort de Kock. Bandahari B. M.

(1) Peresident. (2) Sekretaris. (3) dan (4) Koemisaris.

Typ Drukkerij Mesjid

Figure 6. Majalah *Barito Minangkabau* (source: https://commons.wikimedia.org/)

Data Preparation

In preparing corpus data, digitization and transcription of linguistic data become a very important process. Recorded audio data needs to be correctly transcribed in a written format for analysis (O'Keeffe & McCarthy, 2010). This involves a lot of time, manpower, and monetary input. Transcription is performed by individuals who are proficient in Minangkabaunese language, as well as technology-savvy to ensure conformity and productivity (Litosseliti, 2018). It is essential that these transcribers possess the necessary expertise to manage the technical aspects of transcription, ensuring that the resulting text is a reliable representation of the original spoken data. Wray and Bloomer (2013) provide a very comprehensive reference to standard orthographic transcription devices, including marking speaker identity, pauses, overlaps, backchannels, laughter as well as simple phonetic variability (see Figure 7). It can also be expanded to phonemic transcription, a system for representing pronunciation further down from basic textual form, and finally, phonetic transcription, which uses the International Phonetic Alphabet (IPA) to map out exact sounds. Transcriptions in multi-modal corpora can also

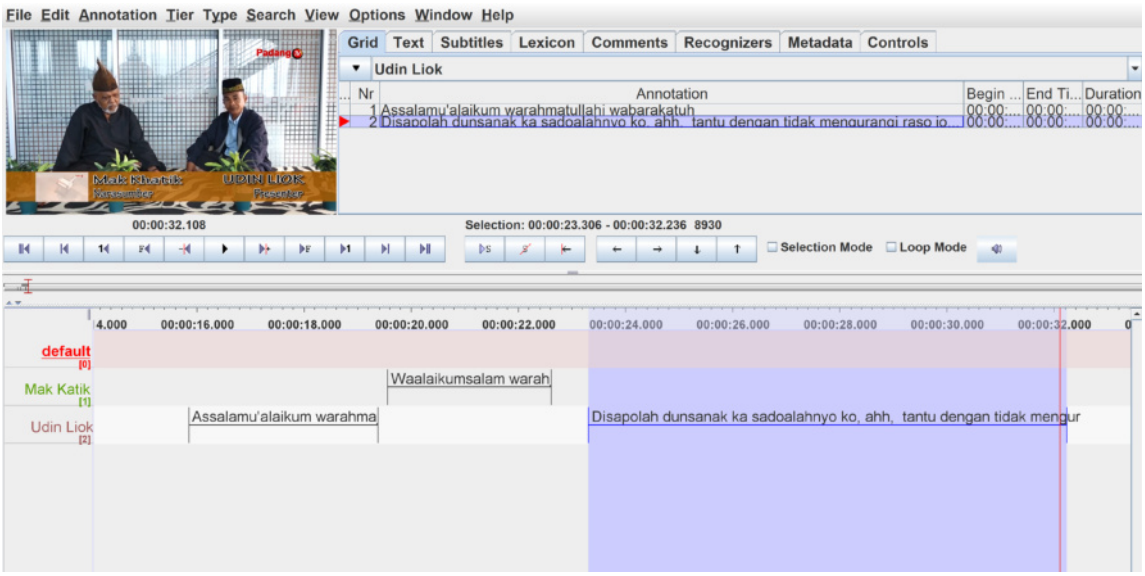


Figure 7. Transcription process with ELAN

Figure 8. Digitalizing Minangkabau manuscript (source: Pramono)

combine the unfolding of gestures with that of speech (O'Keeffe & McCarthy, 2010).

For written data, digitization is essential, particularly when the source is a physical document like a book or manuscript. They need scanning with standard or applicable scanners to develop digital images (see Figure 8). This technology allows scanned images to be transformed into text via Optical Character Recognition (OCR). Yet manuscripts in Jawi script, especially in manuscript form and handwritten, are more complex personalities. That would require transliteration from the Jawi script to the Latin script (Ghani et al., 2009; Singh et al., 2023).

After the image data has been converted into text using OCR, the next crucial step is to verify the accuracy of the conversion. Since OCR still frequently fails, especially with complicated fonts, handwritten text, or non-Latin scripts, it leads to differences between the produced text and the original source. Hence, it is important that the content should be compared with their OCR results. The process of cleaning up typos, punctuation, and formatting is called post-editing. Text checks and corrections Text is only considered as text if listed in the Corpus and it becomes reliable for textual operations. As it ensures the corpora data is trustworthy and correct, which then are critical for any linguistic analysis afterwards.

Data Processing and Annotation

After being collected, data is processed and annotated extensively for linguistic analysis and different applications. In order to enable efficient analysis of data, we will need to encode three types of information in our corpus: metadata, markup elements extracted from the texts, and annotations for language (O'Keeffe & McCarthy, 2010).

Metadata provides information for the text that exists beyond a specific context. For written texts, metadata typically includes details such as

the author's name, the language of the text, the date of publication, and the genre or source. For spoken data, metadata might include information about the speakers, such as age, gender, dialect, and the context in which the conversation or speech took place (Bird & Simons, 2003). Textual markup is used to describe non-linguistic or paralinguistic facets of the text and capture the document structure and appearance. This may involve things like the placement of italics, line breaks, and image or diagram positioning in written documentation. The markup for text corpora could be transcribed pauses, fillers, or interruptions, which would give us a clearer idea of how the language can appear in context.

Linguistic annotation is the third layer of encoding and involves adding explicit linguistic information to the corpus. This may consist of POS tagging, lemmatization, parsing or chunking grammatical structure or named entities, thematic roles, etc. Tools like TreeTagger and UAM CorpusTool facilitate these annotations, ensuring a rich linguistic description (Schmid, 1994). These linguistic annotations expose more profound patterns in the data, such as syntactic dependencies or the association of specific constructs and offer a greater depth of linguistics analysis. Over this, the metadata, textual markup, and linguistic annotation layers form a broad model to be used when analyzing the corpus in diverse directions as they offer both qualitative and quantitative new opportunities for linguistics.

The functionality and accessibility of the Minangkabau corpus both depend on the quality of design and architecture. A sub-corpora organization is used, found in the commonality across spoken, written, or multimedia and various genres (e.g., narratives, conversations, monologues, formal). The hierarchical structure of these types of data enables users to navigate with ease proving the capacity for recovery (Leech, 1991). As the data is linked to the content in its original form, annotations, and transcriptions can also be produced per video or audio file, making it possible for researchers to cross-reference their data with transcripts. This linkage improves the corpus maintainability and accessibility (Wittenburg et al., 2006). The corpus is indexed using robust database management systems, to support efficient data retrieval and analysis. Advanced search functionalities, including keyword searches, pattern matching, and frequency analysis, are implemented to enhance user access to the corpus (Wittenburg et al., 2006).

IV. CHALLENGES AND FUTURE DIRECTIONS

Developing a comprehensive Minangkabau language corpus will face several challenges, including technical challenges such as mastering technology and challenges in data collection. Therefore, it is necessary to design solutions to face these challenges and build sustainable projects. The main challenge in developing a Minangkabau language corpus is the wide geographic distribution of Minangkabau language speakers. Despite the large area of West Sumatra, Minangkabau speakers are also widely spread in various regions and even abroad (Novita et al., 2021; Reniwati et al., 2017). This migration has dispersed Minangkabau speakers to numerous cities and regions, making it difficult to collect a representative corpus of

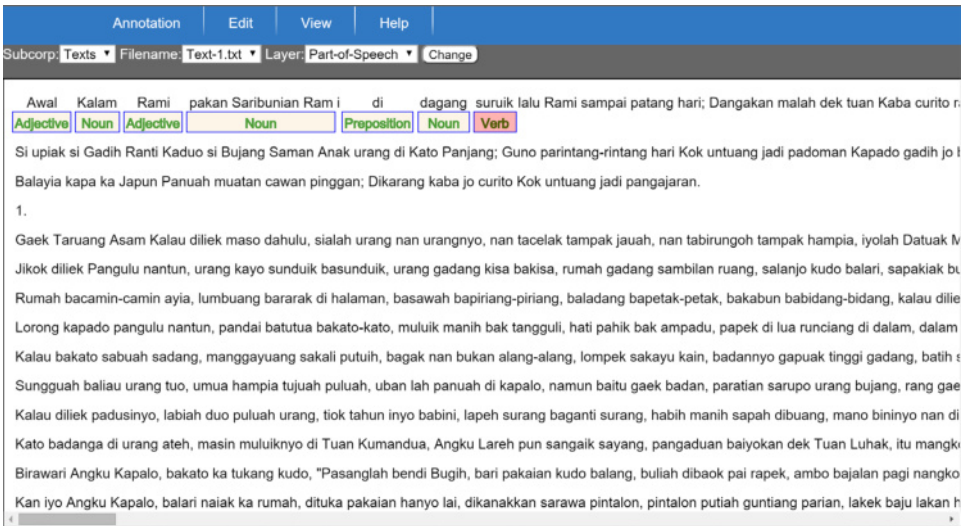


Figure 9. Data Annotation with UAM CorpusTool

spoken language data that reflects the full range of linguistic variation, including both dialectal and sociolectal differences. The scatter makes it challenging to collect data from multiple dialects, each of which may have developed differently across regions due to geographic isolation or the influence of neighboring languages and cultures (Honkola et al., 2018). To reflect the linguistic richness of Minangkabau, it is essential to collect data from urban areas, where modern versions of the language are probably spoken, and rural areas, where some level of traditional varieties (dialects) might be preserved. In addition to geographic barriers, local sociolinguistic factors may further complicate data collection. For example, speakers in urban areas may have adopted elements of other languages, such as Indonesian or English, resulting in code-switching and language mixing that differ from the purer forms spoken in more isolated regions (Schreier, 2013).

Another problem is related to technical challenges, especially dealing with digitization, annotation, and data management. The first main barrier is the need for advanced software or tools to handle linguistic data in a small language like Minangkabau. Most of the existing linguistic tools are designed for high-resourced languages and might not be directly applicable to low-resourced ones that have phonological, grammatical, or syntactic features different from those commonly considered in popular languages (Koto & Koto, 2020). This incompatibility can result in problems when both transcribing and annotating data, which makes it harder to manage data accuracy and consistency.

Annotating the data is another obstacle that requires linguistic skills and software capable of tagging phonemes, morphemes, syntax, and semantics in high detail (O'Keeffe & McCarthy 2010). Since there are not a set of existing general annotation guidelines specifically for Minangkabau, individual corpus developers have to come up with customised guidelines representing the specific characteristics of this language. This, in turn, demands an investment both in terms of technology and also deeper collaboration between linguists and software developers.

Another set of challenges arises from the linguistic diversity within Minangkabau. Minangkabau has various features for each of its dialects (Antoni et al., 2019; Novita et al., 2021; Pramono, 2018; Reniwati et al., 2017). This variation takes a significant investment in fieldwork and planning to ensure that all of the dialectal variation is properly captured. The diversity of both the extent of use and proficiency will present a variety of sociolinguistic challenges in the Minangkabau language. Most Minangkabau speakers are bilingual or multilingual, with proficiency in Indonesian only or in other regional languages (Marnita, 2017). As for the former, multilingualism can affect their use of Minangkabau, which in turn encourages code-switching and borrowing that complicates linguistic analysis.

Despite the challenge in developing the corpus, the Minangkabau language has a great opportunity to be explored with the most robust technologies. This rich set of language resources is a serious candidate for building an extensive Minangkabau language corpus. Using these different sources of data, many forms of corpora can be formed, including spoken corpus and written corpus (O'Keeffe & McCarthy 2010), wherein each focuses on different functionalities of the language. Moreover, specific subcorpora or small corpora may be created to emphasize a particular theme or a particular source of data, such as the corpus of the classical Minangkabau language, journalistic corpus, Minangkabau dialects corpus, and corpus on the social media usage of the Minangkabau language. The subcorpora offers more detailed access to certain linguistic domains and allows for targeted studies of the particularities of Minangkabau in different environments.

Similarly, it would drastically enhance the state-of-the-art of corpus-based Minangkabau language research. Making the Minang dataset available to other researchers enables them to readily query and investigate their research with little or no struggle as per traditional constraints. The availability of features like this opens access to a wider range of scholars, democratizing linguistic research. Furthermore, having an extensive corpus allows different ways of conducting research: quantitative methods like frequency analysis or corpus-based statistical techniques can now be combined with the usual qualitative approaches. By combining methods, such as these, the incorporation of more robust and data-driven insights could ultimately push Minangkabau linguistic research forward. In short, it will be a valuable resource that facilitates research and avoids the extinction of the

Minangkabau Language as an academic Culture so they can overcome problems and endure greater sociocultural impacts.

V. CONCLUSION

The Minangkabau corpus development is an important project to document and save the Minangkabau language and culture. The corpus is a rich repository of data on which linguistic research can be carried out for educational purposes as well as technological advancements and has been curated through detailed data collection, annotation, and organization. Drawing from a wide range of spoken, written, and multimedia data, the corpus provides a representation of the rich complexity and versatility present in Minangkabau language use across contexts and types. Moreover, the availability of secondary data from television and radio broadcasts, audio-video recordings, books and manuscripts, and social media is a very valuable opportunity for developing a Minang language corpus. The corpus can cover a wide range of linguistic and nonlinguistic information, which will help researchers study different levels of structure, free variation, and developments in a language. Furthermore, the corpus functions as an indispensable catalyst for language preservation and revitalization within the Minangkabau community. Through the development of language teaching materials, curriculum, and cultural records, the corpus enables community members to participate in their heritage languages by forging intergenerational continuity.

ETHICS STATEMENT

The authors have read and followed the ethical requirements for publication in Jurnal Arbitrer. The current work does not involve human subjects, animal experiments, or any data collected from social media platforms.

ACKNOWLEDGEMENTS

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

DECLARATION OF COMPETING INTERESTS

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

REFERENCES

Abdullah, T. (1970). Some notes on the Kaba Tjindua Mato: An example of Minangkabau traditional literature. *Indonesia*, 9, 1. https://doi.org/10.2307/3350620

Adelaar, K. A. (1992). *Proto malayic: The reconstruction of its phonology and parts of its lexicon and morphology*.

Adeyeye, B., Amodu, L., Odiboh, O., Oyesomi, K., Adesina, E., & Yartey, D. (2021). Agricultural radio programmes in indigenous languages and agricultural productivity in North-Central Nigeria. *Sustainability*, *13*(7), 3929. https://doi.org/10.3390/su13073929

Aijmer, K., & Altenberg, B. (2004). *Advances in corpus Linguistics*.

Al-Hamzi, A. M. S., Gougui, A., Amalia, Y. S., & Suhardijanto, T. (2020). Corpus linguistics and corpus-based research and its implication in applied linguistics: A systematic review. *PAROLE Journal of Linguistics and Education*, *10*(2), 176–181. https://doi.org/10.14710/parole.v10i2.176-181

Almos, R., & Ladyanna, S. (2019). Lexicons classics of fishing in Minangkabau community. In *Sciendo eBooks* (pp. 230–235). https://doi.org/10.2478/9783110680027-033

Aman, I., Jaafar, M. F., & Awal, N. M. (2019). Language and identity: A reappraisal of Negeri Sembilan Malay language. *Kajian Malaysia*, *37*(1), 27–49. https://doi.org/10.21315/km2019.37.1.2

Amir, A., Zuriati, & Anwar, K. (2006). *Pemetaan sastra lisan Minangkabau*.

Ancarno, C. (2018). Interdisciplinary approaches in corpus linguistics and CADS. In *Routledge eBooks* (pp. 130–156). https://doi.org/10.4324/9781315179346-7

Anthony, N. L. (2013). A critical look at software tools in corpus linguistics. *Linguistic Research*, *30*(2), 141–161. https://doi.org/10.17250/khisli.30.2.201308.001

Antoni, C., Irham, I., & Ronsi, G. (2019). Language variation in Minang colloquial language spoken in

Kabun region: Sociolinguistic study on millennial citizens. *Jurnal Arbitrer*, *6*(2), 92–98. https://doi.org/10.25077/ar.6.2.92-98.2019

Anwar, K. (1980). Language use in Minangkabau society. *Indonesia Circle School of Oriental & African Studies Newsletter*, *8*(22), 55–63. https://doi.org/10.1080/03062848008723789

Arkhangelskiy, T. (2019). Corpora of social media in minority Uralic languages. In *Proceedings of the Fifth International Workshop on Computational Linguistics for Uralic Languages*. https://doi.org/10.18653/v1/w19-0311

Barlow, M. (2011). Corpus linguistics and theoretical linguistics. *International Journal of Corpus Linguistics*, *16*(1), 3–44. https://doi.org/10.1075/ijcl.16.1.02bar

Barmenqulova, A. (2024). Transfer of ethnocultural units stored in the regional lexicon to the national corpus. *Tiltanym*, *2*, 193–200. https://doi.org/10.55491/2411-6076-2024-2-193-200

Bennett, G. (2010). *Using Corpora in the language learning classroom*.

Bhatia, V. K., Flowerdew, J., & Jones, R. H. (2008). Advances in discourse studies.

Bird, S., & Simons, G. (2003). Seven dimensions of portability for language documentation and description. *Language*, *79*(3), 557–582. https://doi.org/10.1353/lan.2003.0149

Boulton, A., & Landure, C. (2016). Using Corpora in language teaching, learning and use. *Recherche Et Pratiques PéDagogiques En Langues De SpéCialité*, *Vol. 35 N° 2*. https://doi.org/10.4000/apliut.5433

Breyer, Y. (2009). Learning and teaching with corpora: reflections by student teachers. *Computer Assisted Language Learning*, *22*(2), 153–172. https://doi.org/10.1080/09588220902778328

Chelliah, S. L., & De Reuse, W. J. (2011). *Handbook of descriptive linguistic fieldwork*.

Coats, S., & Laippala, V. (2024). *Linguistics across disciplinary borders*.

Conrad, S. (2002). 4. Corpus linguistic approaches for discourse analysis. *Annual Review of Applied Linguistics*, *22*, 75–95. https://doi.org/10.1017/s0267190502000041

Cushing, S. T. (2017). Corpus linguistics in language testing research. *Language Testing*, *34*(4), 441–449. https://doi.org/10.1177/0265532217713044

Deignan, A. (2005). *Metaphor and corpus linguistics.*

Dimitrova, L., & Garabík, R. (2012). Bilingual corpus – digital repository for preservation of language heritage. *Digital Presentation and Preservation of Cultural and Scientific Heritage*, *2*, 132–141. https://doi.org/10.55630/dipp.2012.2.5

Felde, O. V. (2022). Electronic corpus of linguaculture of the Northern Angara Region: Foundations, structure, and application. *Bulletin of Kemerovo State University*, *23*(4), 1086–1095. https://doi.org/10.21603/2078-8975-2021-23-4-1086-1095

Flowerdew, L. (2011). *Corpora and language education*.

Ghani, R. A., Zakaria, M. S., & Omar, K. (2009). Jawi-Malay transliteration. In *2009 International Conference on Electrical Engineering and Informatics*. https://doi.org/10.1109/iceei.2009.5254799

Grenoble, L. A., & Whaley, L. J. (2006). *Saving Languages: An Introduction to Language Revitalization*.

Hajič, J., Hajičová, E., Mírovský, J., & Panevová, J. (2016). Linguistically annotated corpus as an invaluable resource for advancements in linguistic research: a case study. T*he Prague Bulletin of Mathematical Linguistics*, *106*(1), 69–124. https://doi.org/10.1515/pralin-2016-0012

Handoko, H., Kaur, S., & Kia, L. S. (2024). Cultivating sustainability: A cultural linguistic study of Minangkabau environmental proverbs. *Jurnal Arbitrer*, *11*(1), 72–84. https://doi.org/10.25077/ar.11.1.72-84.2024

Hanif, A., Afrina, C., Putra, H., & Rudiamon, S. (2022). Investigating Minangkabau's scattered manuscript: Philological studies of religious manuscripts in West Sumatera. *Proceedings of the 6th Batusangkar International Conference, BIC 2021, 11 - 12 October, 2021, Batusangkar-West Sumatra, Indonesia*.

https://doi.org/10.4108/eai.11-10-2021.2319433

Harun, M. H., Aziz, M. K. N. A., Rahim, E. a. A., Shuhairimi, A., & Ahmad, Y. (2018). Jawi writing in Malay archipelago manuscript: A general overview. *MATEC Web of Conferences*, *150*, 05054. https://doi.org/10.1051/matecconf/201815005054

Heni, A. N., & Suryadi, M. (2022). Variasi leksikal bahasa Minangkabau di Kanagarian Kubang Putiah, Kabupaten Agam: Kajian sosiodialektologi. *Widyaparwa*, *50*(1), 151–161. https://doi.org/10.26499/wdprw.v50i1.911

Herbowo, N. a. S., & Sulastri, S. (2020). Reprinting of Kaba and Tambo books by Kristal Multimedia Publisher. *Wanastra Jurnal Bahasa Dan Sastra*, *12*(2), 223–228. https://doi.org/10.31294/w.v12i2.8744

Himmelmann, N. P. (2006). *Essentials of language documentation.*

Honkola, T., Ruokolainen, K., Syrjänen, K. J. J., Leino, U., Tammi, I., Wahlberg, N., & Vesakoski, O. (2018). Evolution within a language: environmental differences contribute to divergence of dialect groups. *BMC Evolutionary Biology*, *18*(1). https://doi.org/10.1186/s12862-018-1238-6

Jamaris, E. (2002). *Pengantar sastra rakyat Minangkabau*.

Juško-Štekele, A., & Kļavinska, A. (2024). Developing corpus literacy: A perspective of Latgalian language and cultural studies. *Journal of Multilingual and Multicultural Development*, 1–13. https://doi.org/10.1080/01434632.2024.2359020

Koto, F., & Koto, I. (2020). Towards computational linguistics in Minangkabau language: Studies on sentiment analysis and machine translation. In *Proceedings of the 34th Pacific Asia Conference on Language, Information and Computation* (pp. 138–148). Association for Computational Linguistics.

Kretzschmar, W. A., Anderson, J., Beal, J. C., Corrigan, K. P., Opas-Hänninen, L. L., & Plichta, B. (2006). Collaboration on corpora for regional and social analysis. *Journal of English Linguistics*, *34*(3), 172–205. https://doi.org/10.1177/0075424206293598

Kytö, M. (2011). Corpora and historical linguistics. *Revista Brasileira De Lingüística Aplicada*, *11*(2), 417–457. https://doi.org/10.1590/s1984-63982011000200007

Lane, P., Hagen, K., Nøklestad, A., & Priestley, J. (2022). Creating a corpus for Kven, a minority language in Norway. *Nordlyd*, *46*(1). https://doi.org/10.7557/12.6345

Leech, G. (1991). The state of the art in corpus linguistics. In K. Aijmer & B. Altenberg (Eds.), *English Corpus Linguistics* (pp. 8-29). Longman.

Litosseliti, L. (2018). *Research methods in linguistics*.

Love, R., Dembry, C., Hardie, A., Brezina, V., & McEnery, T. (2022). The spoken BNC2014. *International Journal of Corpus Linguistics*, 319–344. https://doi.org/10.1075/ijcl.22.3.02lov

McGinn, R. (2009). *Studies in Austronesian languages and cultures: Papers in honor of René van den Berg*.

Marnita, R. (2017). Pergeseran bahasa dan identitas sosial dalam masyarakat Minangkabau Kota: Studi kasus di kota padang. *Masyarakat Indonesia*, *37*(1), 139–163. https://doi.org/10.14203/jmi.v37i1.607

Maryelliwati, M., Rahmat, W., & Kemal, E. (2018). A reality of Minangkabau language and literature and its transformation to a creation of performance works. *Gramatika STKIP PGRI Sumatera Barat*, *4*(1). https://doi.org/10.22202/jg.2018.v4i1.2422

McEnery, T., & Xiao, R. (2011). *What Corpora can offer in language teaching and learning*.

Meyer, C. F. (2002). *English Corpus Linguistics*.

Migge, B., & Léglise, I. (2010). Integrating local languages and cultures into the education system of French Guiana. In *Creole language library* (pp. 107–132). https://doi.org/10.1075/cll.36.05mig

Musgrave, S. (2014). Language documentation and sociolinguistics: Capturing variation in discourse.

Language Documentation & Conservation, 8, 121–136.

Mustafa, F., & Yusuf, S. B. (2021). Transitivity of try and V construction in British and American English. *Langkawi Journal of the Association for Arabic and English*, *7*(2), 197. https://doi.org/10.31332/lkw.v7i2.3166

Nathan, D., & Austin, P. K. (2004). Reconceiving metadata: Language documentation through thick and thin. In P. K. Austin (Ed.), *Language documentation and description* (Vol. 2, pp. 179–187). SOAS.

Nelisa, M., Ardoni, N., & Rasyid, Y. (2021). Preservation of Minangkabau local wisdom as media for cultural literacy. *Advances in Social Science, Education and Humanities Research/Advances in Social Science, Education and Humanities Research*. https://doi.org/10.2991/assehr.k.211201.024

Nesti, M. R. (2016). Variasi leksikal bahasa minangkabau di Kabupaten Pesisir Selatan. *Jurnal Arbitrer*, *3*(1), 46–61. https://doi.org/10.25077/ar.3.1.46-61.2016

Noranda, A. (2023). Minangkabo dalam naskah kuno. *Jurnal Ceteris Paribus*, *2*(2), 37–66. https://doi.org/10.25077/jcp.v2i2.18

Novita, R., Firdaus, W., & Budiono, S. (2021). Minangkabau language mapping verification in West Sumatra Province. *Advances in Social Science, Education and Humanities Research/Advances in Social Science, Education and Humanities Research*. https://doi.org/10.2991/assehr.k.211226.056

Nurizzati, N., & Nasution, M. I. (2021). The profile of Kaba Si Tungga manuscript and the play script of Anggun Nan Tongga by Wisran Hadi: An overview of the transcription and transformation of Minangkabau oral literary texts. *Advances in Social Science, Education and Humanities Research/Advances in Social Science, Education and Humanities Research*. https://doi.org/10.2991/assehr.k.211201.031

Nurmukhamedov, U., & Sharakhimov, S. (2021). Corpus-based vocabulary analysis of English podcasts. *RELC Journal*, *54*(1), 7–21. https://doi.org/10.1177/0033688220979315

O'Keeffe, A., & Farr, F. (2003). Using language corpora in initial teacher education: Pedagogic issues and practical applications. *TESOL Quarterly*, *37*(3), 389. https://doi.org/10.2307/3588397

O'Keeffe, A., McCarthy, M., & Carter, R. (2007). *From corpus to classroom*.

Onyenankeya, K., & Salawu, A. (2022). Community radio acceptance in rural Africa: The nexus of language and cultural affinity. *Information Development*, *39*(3), 567–580. https://doi.org/10.1177/02666669211073458

Oriyama, K. (2010). Heritage language maintenance and Japanese identity formation: What role can schooling and ethnic community contact play? *Heritage Language Journal*, *7*(2), 237–272. https://doi.org/10.46538/hlj.7.2.5

Oswari, T., Hastuti, E., & Chandra, R. (2020). Minangkabau language learning based on android application. In proceedings of the 4th International Conference on Arts Language and Culture (ICALC 2019). https://doi.org/10.2991/assehr.k.200323.033

Padang TV. (2023, January 20). Duduak Baselo - hukum hukum adat di Minangkabau [Video]. YouTube. https://www.youtube.com/watch?v=skKJowyVGbA

Peksoy, E. (2017). Corpus based authenicity analysis of language teaching course books. *International Journal of Languages Education*, *1*(Volume 5 Issue 4), 287–307. https://doi.org/10.18298/ijlet.2324

Philip, G. (2018). Corpus linguistics.

Poku, F. A. (2024). Linguistics of Ghanaian language: A platform to embed formal education in culture. *International Journal of Research and Innovation in Social Science*, VIII(III), 1337–1346. https://doi.org/10.47772/ijriss.2024.803098

Pramono, P., Yusuf, M., & Hidayat, H. N. (2018). Bahasa melayu dan Minangkabau dalam khazanah naskah Minangkabau. *Jurnal Pustaka Budaya*, *5*(2), 24–35. https://doi.org/10.31849/pb.v5i2.1483

Rao, D. L., Pala, V. R., Herndon, N., & Gudivada, V. N. (2020). A deep learning architecture for corpus creation for Telugu language. In *Advances in intelligent systems and computing* (pp. 1–16). https://

doi.org/10.1007/978-981-15-4029-5_1

Razin, T., & Subiyanto, A. (2024). Pola perubahan fonologi antara bahasa Minangkabau umum dan subdialek Minangkabau Selayo. *Widyaparwa*, *52*(1), 206–220. https://doi.org/10.26499/wdprw.v52i1.1719

Reniwati, R., & Khanizar, K. (2022). Leksikon nama peralatan rumah tangga masyarakat Minangkabau: gambaran dinamika masyarakat. *Ranah Jurnal Kajian Bahasa*, *11*(1), 141. https://doi.org/10.26499/rnh.v11i1.4169

Reniwati, R., Midawati, M., & Noviatri, N. (2017). Lexical variations of Minangkabau language within West Sumatra and Peninsular Malaysia: A dialectological study. *Malaysian Journal of Society and Space*, *13*(3), 1–10. https://doi.org/10.17576/geo-2017-1303-01

Römer, U. (2011). Corpus research applications in second language teaching. *Annual Review of Applied Linguistics*, *31*, 205–225. https://doi.org/10.1017/s0267190511000055

Rusmali, M., Usman, A. H., Nikelas, S., Husin, N., & Busri, B. (1985). *Kamus Minangkabau-Indonesia.*

Sakti, S., & Nakamura, S. (2013). Towards language preservation: Design and collection of graphemically balanced and parallel speech corpora of Indonesian ethnic languages. In International Conference Oriental COCOSDA held jointly with 2013 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE) 2013. https://doi.org/10.1109/icsda.2013.6709907

Saydam, G. (2004). *Kamus lengkap bahasa Minang: Minang-Indonesia, Indonesia-Minang.*

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In Proceedings of International Conference on New Methods in Language Processing.

Schreier, D. (2013). Collecting ethnographic and sociolinguistic data. In Research Methods in Language Variation and Change (pp. 17–35). https://doi.org/10.1017/cbo9780511792519.004

Singh, S. R., Anand, A., & Chauhan, S. (2023). Handwritten documents conversion to digital documents. In 2023 9th International Conference on Smart Computing and Communications (ICSCC). https://doi.org/10.1109/icscc59169.2023.10335069

Sneddon, J. N. (2003). *The Indonesian language: Its history and role in modern society*.

Stubbs, M., & Halbe, D. (2012). Corpus linguistics: Overview. The Encyclopedia of Applied Linguistics. https://doi.org/10.1002/9781405198431.wbeal0033

Sulistiyo, R., Sani, A., & Rusli, R. (2023). Manuskrip Beraksara Jawi pada Khazanah Pustaka EAP British Library. *Ulumuddin Jurnal Ilmu-ilmu Keislaman*, *13*(1), 115–136. https://doi.org/10.47200/ulumuddin.v13i1.1625

Suryadi, S. (2010). The impact of the West Sumatran regional recording industry on Minangkabau oral literature. *Wacana Journal of the Humanities of Indonesia*, *12*(1), 35. https://doi.org/10.17510/wjhi.v12i1.45

Suryani, E. (2018). The survival of local languages in Indonesia: A case study of Minangkabau and Sundanese. *Asian Journal of Humanities and Social Studies*, 6(1), 1-10.

TVRI Sumatera Barat. (2022, August 1). Limbago adaik Minangkabau di maso kini - budaya alam minangkabau TVRI Sumbar (full) [Video]. YouTube. https://www.youtube.com/watch?v=xJQmYlqzI5E

Taufiqurrahman, T., Hidayat, A. T., Efrinaldi, Sudarman, & Lukmanulhakim. (2021). The Existence of the Manuscript in Minangkabau Indonesia and its field in Islamic studies. *Journal of Al-Tamaddun*, *16*(1), 125–138. https://doi.org/10.22452/jat.vol16no1.9

Tembe, J., & Norton, B. (2008). Promoting local languages in Ugandan Primary Schools: the community as stakeholder. *Canadian Modern Language Review/ La Revue Canadienne Des Langues Vivantes*, *65*(1), 33–60. https://doi.org/10.3138/cmlr.65.1.33

Trosterud, T. (2002). *Parallel corpora as tools for investigating and developing minority languages.*

Velini, R. S., & Suryadi, M. (2023). Usaha pemertahanan Bahasa Minangkabau melalui permainan dan tradisi budaya lokal di Kota Padang, Sumatera Barat. *Jurnal Sastra Indonesia*, *12*(1), 71–80. https://doi.org/10.15294/jsi.v12i1.59370

Vessey, R. (2015). Corpus approaches to language ideology. *Applied Linguistics*, *38*(3), 277–296. https://doi.org/10.1093/applin/amv023

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., & Sloetjes, H. (2006). ELAN: A Professional Framework for Multimodality Research. In Proceedings of LREC 2006.

Wray, A., & Bloomer, A. (2012). *Projects in linguistics and language studies*.

Xiao, R. (2009). 46. Theory-driven corpus research: Using corpora to inform aspect theory. *In Corpus Linguistics: An International Handbook*.

Zufferey, S. (2020). *Introduction to Corpus Linguistics*.