KINSHIP OF SIBOLGA COASTAL MALAY AND MANDAILING LANGUAGE: STUDY OF LEXICOSTATISTICS

Tanty Aidullia Nasution[1*], Dwi Widayati[2], Tasnim Lubis[3]
[1] Postgraduate of Linguistics, Faculty of Cultural Sciences, Universitas Sumatera Utara, Medan, Indonesia
[2] Postgraduate of Linguistics, Faculty of Cultural Sciences, Universitas Sumatera Utara, Medan, Indonesia
[3] Postgraduate of Linguistics, Faculty of Cultural Sciences, Universitas Sumatera Utara, Medan, Indonesia
Pos-el:[1*] tantynasution2000@gmail.com, [2]dwiwidayati@usu.ac.id, [3]tasnimlubis@usu.ac.id

ABSTRACT

This study investigates the linguistic kinship between the Sibolga Coastal Malay Language (SCML) and the Mandailing Language (ML) using a lexicostatistical approach. Both languages belong to the Austronesian family and are spoken in adjacent regions of North Sumatra, which raises significant interest in their historical and phonological relationship. Utilizing a 200-word Swadesh list, the research identifies 75 cognate lexical items, categorized based on degrees of similarity such as identical forms, phonemic correspondence, phonetic similarity, and one-phoneme difference. The results show a lexical similarity of 38%, indicating that SCML and ML belong to the same language family. Using lexicostatistical formulas, the estimated time of separation is approximately 2,230 years ago, with a confidence range between 382 and 32 BCE. These findings provide insight into the gradual divergence of the two languages from a common proto-language due to geographical and sociocultural factors. This study contributes to the field of comparative lexicography by offering empirically grounded data, refining phonological correspondences, and demonstrating how lexicostatistics can illuminate historical language change. The results also emphasize the importance of documenting regional languages to support linguistic preservation and historical reconstruction.

Keywords: Language; Sibolga Coastal Malay; Mandailing; Kinship; lexicostatistical.

ABSTRAK

Penelitian ini mengkaji hubungan kekerabatan linguistik antara Bahasa Melayu Pesisir Sibolga (BMPS) dan Bahasa Mandailing (BM) dengan menggunakan pendekatan leksikostatistik. Kedua bahasa ini termasuk dalam rumpun bahasa Austronesia dan dituturkan di wilayah yang berdekatan di Sumatra Utara, sehingga *menimbulkan ketertarikan terhadap hubungan historis dan fonologisnya. Dengan menggunakan daftar Swadesh sebanyak 200 kosakata dasar, penelitian ini mengidentifikasi 75 pasangan leksikal kognat yang diklasifikasikan ke dalam kategori bentuk identik, korespondensi fonemis, kemiripan fonetik, dan perbedaan satu fonem. Hasilnya menunjukkan tingkat kesamaan leksikal sebesar 38%, yang mengindikasikan bahwa BMPS dan BM termasuk dalam satu keluarga bahasa. Berdasarkan rumus leksikostatistik, waktu pemisahan keduanya diperkirakan terjadi sekitar 2.230 tahun yang lalu, dengan rentang kepercayaan antara tahun 382 hingga 32 SM. Temuan ini menunjukkan bahwa kedua bahasa tersebut mengalami divergensi bertahap dari bahasa proto yang sama, dipengaruhi oleh faktor geografis dan sosial budaya. Studi ini memberikan kontribusi terhadap kajian leksikografi komparatif melalui data empiris, identifikasi korespondensi fonologis, dan penerapan leksikostatistik dalam menelusuri perubahan bahasa secara historis. Hasilnya juga menegaskan pentingnya dokumentasi bahasa daerah dalam mendukung pelestarian bahasa dan rekonstruksi sejarah linguistik.*

***Kata kunci**: Bahasa; Melayu Pesisir Sibolga; Mandailing; Kekerabatan; Leksikostatistik.*

A. INTRODUCTION

Language is a major element in daily life used by community groups in communicating (Meylani, 2024; Mz, 2019; F. Nasution & Tambunan, 2022; Puumala & Shindo, 2021; Rabiah, 2018). In this case, each community group has its own language that is only understood by people in this community group (Mz, 2019; Puumala & Shindo, 2021; Rabiah, 2018). This is also in line with the nature of humans as social creatures who certainly need help from other people (De Stefani & De Marco, 2019; Kinzler, 2021; Mz, 2019; Pickering & Garrod, 2021). Humans without language cannot be said to be social beings (Nababan, 1993; Hines & Stern, 2019; Kinzler, 2021; Meylani, 2024; Verma, 2024). In addition, humans also need language to be able to convey feelings, ideas, and desires to others using human speech organs (F. Nasution & Tambunan, 2022).

Language is also referred to as a symbol due to a series of sounds produced by the human vocal tract that are driven by a certain meaning, something and absorbed by humans (Keraf, 1997; Adelman et al., 2018; Johansson et al., 2020). Each language in a certain area is different, so it indirectly confirms that a language can also be a symbol of identity markers and behavior to recognizing a person's character (López-Narváez, 2023; Niwanda et al., 2024; Wijaya, 2024). This statement then becomes a supporter of the statement that states that in addition to being a means of communication, language is also a source of sound system that has an arbitrary nature and is used by group members in doing work, communicating, and identifying themselves (Kridalaksana, 1983; Dzhukaeva et al., 2024; Fabbro et al., 2022; Hines & Stern, 2019, 2019; Nasution & Tambunan, 2022).

The purpose and usefulness of the language then make language an object of study that is widely used in the field of scientific disciplines (Charity Hudley et al., 2023; Gajo & Berthoud, 2020). One of the disciplines that uses language as an object of study is the field of linguistics (Agha, 2007; Charity Hudley et al., 2023; Tanwete & Kombinda, 2020; Thoms et al., 2021). Linguistics is basically a field of scientific study that discusses language as a language itself (Tanwete & Kombinda, 2020; Thoms et al., 2021). This approach has a concept where language is a speech sound, is unique, as a system, and can change over time (Chaer, 2014).

As a main discipline, linguistics certainly has sub-disciplines consisting of diachronic linguistics and synchronic linguistics (Anderson, 2016; Radulović, 2023; Ye & Zhang, 2024). Diachronic linguistics is a sub-discipline of linguistics that studies language in an unlimited period of time, which can be from the birth of a language to the extinction of the language (Anderson, 2016; Radulović, 2023; Ye & Zhang, 2024). Diachronic linguistics is historical and comparative in nature, so it is also known as comparative historical linguistics (Anderson, 2016; Radulović, 2023). This linguistics is a branch of linguistics that discusses language in the field of time and changes in language elements that occur in that time span (Keraf, 1996). Comparative historical linguistics or also called CHL focuses on data from more than one language in at least two periods and is then carefully compared so that the rules that occur in the language change are found (Aisyah & Widayati, 2022; Harahap & Ritonga, 2024; Rahmawati, 2022).

In this field, there is a study method used to group languages that tend to prioritize static observation of words and then group the languages based on the percentage of language kinship (Keraf, 1996). The purpose of this language grouping is useful for showing the two languages compared to have kinship. Where language kinship is the relationship between two or more languages that are derived from the same language, ancient language (Kirdaklaksana, 2008). This language kinship is related to the comparative relationship between two languages due to similarities in terms of phonology, morphology, and syntax (Aisyah & Widayati, 2022; T. A. Nasution, 2023).

Previous research on Austronesian languages in North Sumatra, such as Aisyah & Widayati (2022), Batubara & Widayati (2022), and Nasution (2023), has demonstrated lexical kinship among closely related regional languages using the lexicostatistical approach. However, there remains a lack of focused investigation into the historical divergence between Sibolga Coastal Malay and Mandailing, particularly regarding the estimation of their separation time and phonemic correspondence. While these languages are often treated as distinct, their geographical proximity and social interactions suggest the potential for a shared linguistic ancestry. Therefore, this study fills the research gap by offering a systematic comparative analysis to quantify their lexical kinship, estimate divergence time, and interpret phonological developments using a standardized Swadesh list.

In this study, two languages will be compared, Sibolga Coastal Malay and Mandailing. Where both languages are in the same family, the Austronesian language family, so that the level of kinship can be known by grouping data using lexicostatistics. This study also aims to determine the level of kinship, separation time, and prediction of language age based on the level of kinship which is based on phonemic elements and lexicon.

B. THEORETICAL FRAMEWORK

The field of study known as comparative historical linguistics, or CHL, examines language throughout a particular period of time and the linguistic changes that take place in that language. In order to determine the time of divergence, this theory examines and compares data from at least two languages. Four fundamental presumptions are utilized as a standard in the pursuit of answers concerning the age of a language, often known as distinction between two or more languages (Keraf, 1996). The basic assumption referred to is:

1. When compared to other languages, some terminology from one language is hard to change. In fact, the basic vocabulary technique of language classification already makes this assumption. The lexicostatistic approach is where the idea of vocabulary originates. Basic vocabulary, also known as basic vocabulary, is made up of terms that are extremely important to a language's existence and that decide whether a language will survive or not.

   To correctly use this strategy, a restricted amount of language is taken, and it is then evaluated and tested. The objective is to compile a list of words with universal properties, which denote that the words are thought to be present in all languages from the very beginning of their evolution (Keraf, 1996). Among the words are; 1) Pronouns; 2) Number words; 3) Body part words (and their attributes or functions); 4) Nature and the environment, including the air, sky, water, mountains, and so forth, as well as their attributes or functions; 5) Daily essentials that have been around since the beginning of time, such as sticks, knives, houses, and so forth.
2. Basic vocabulary memory (resilience) remains consistent across time. According to the second fundamental premise, a language's basic vocabulary, or a specific percentage of it, will last for a millennium. If this presumption is true, then 200 words from a language's fundamental lexicon will indirectly survive at a specific percentage after 1,000 years, and the remaining words will also survive at the same rate.
3. All languages have seen the same shift in their foundational vocabulary. Thirteen languages have tested this third assumption, and some of those languages have produced written manuscripts. According to the research, a language's basic vocabulary survives between 86.4 and 74.4% of the time, or an average of 80.5%, every 1,000 years. However, while all of the languages utilized in the experiment (except from two) belong to the Indo-European family, the results do not directly state that all languages will survive with that average proportion.
4. The time of the two languages' separation can be computed if the percentage of the two cognate languages is known. The second and third basic assumptions logically lead to this fourth one. This presumption is valid provided that there are no obstacles or causes that hasten the separation (ceteris paribus). According to the explanation of the second, third, and fourth fundamental assumptions, if we know the percentage of cognate words in both languages, we can determine the separation age or separation time between languages A and B. Since the core vocabulary of both linked languages would decline at the same rate every 1,000 years, the period spent apart must be cut in half.

Glottochronological and lexicostatistical methods are employed in comparative historical linguistics (CHL). A technique for comparing two or more languages to ascertain when they diverged is called lexicostatistics. According to Keraf (1996), lexicostatistics is a method for grouping language data that focuses more on statistically grouping lexicons and then categorizing these languages according to the percentage of relatedness and non-relatedness. This method counts related words, gathers fundamental vocabulary, and then calculates the time of separation to determine whether or not words are linked. Words that are cognate include those that have the same phoneme, those that have phonemic correspondence, those that have phonetic resemblance, and those that differ by one phoneme.

This computation makes use of three formulas: the error range formula, the separation time formula, and the percentage calculation formula.

1. Formula for percentage calculation

$$c = \frac{k}{n}\, x\ 100\%$$

2. Formula for the separation time

$$W = \frac{\log C}{2 \log r}$$

3. Formula for the error range

$$s = \sqrt{\frac{C\,(1 - C)}{n}}$$

This computation also calls for logarithms derived from table 1.

Table. 1 Logarithms table

N	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,1	-2,303	-2,207	-2,120	-2,040	-1,966	-1,897	-1,833	-1,772	-1,715	-1,661
0,2	-1,609	-1,561	-1,514	-1,470	-1,427	-1,368	-1,347	-1,309	-1,273	-1,238
0,3	-1,204	-1,171	-1,139	-1,109	-1,079	-1,050	-1,022	-0,994	-0,968	-0,942
0,4	-0,916	-0,892	-0,868	-0,844	-0,821	-0,799	-0,777	-0,755	-0,734	-0,713
0,5	-0,693	-0,673	-0,654	-0,635	-0,616	-0,598	-0,580	-0,562	-0,545	-0,528
0,6	-0,511	-0,494	-0,478	-0,462	-0,446	-0,432	-0,416	-0,400	-0386	-0,371
0,7	-0,357	-0,342	-0,329	-0,315	-0,301	-0,288	-0,274	-0,261	-0,248	-0,236
0,8	-0,223	-0,211	-0,198	-0,186	-0,174	-0,163	-0,151	-0,139	-0,128	-0,117
0,9	-0,105	-0,094	-0,083	-0,073	-0,062	-0,051	-0,041	-0,030	-0,020	-0,010

C. METHOD

The method used in this study is the lexicostatistic method, which is a method of grouping languages to obtain the percentage of related word sets (Mahsun, 1995) (Aisyah & Widayati, 2022; Batubara & Widayati, 2022; T. A. Nasution, 2023; Nurmala & Widayati, 2022; Utama et al., 2023). In data collection, the benchmark used is the use of the Swadesh vocabulary, which is then searched for in each regional language that is the object of research. The steps in using this method are collecting data, determining related pairs, calculating age and separation time from the results of calculating the number of related words, and calculating the error term. The data found in both languages will also be presented in two written versions, namely using the alphabet and also phonemic transcription.

The data used in this study were collected through fieldwork conducted in Sibolga and Mandailing Natal regencies in North Sumatra. The primary data consist of 200 core vocabulary items based on the Swadesh list, elicited from native speakers of each language using direct interviews and structured elicitation techniques. The main instrument was a standardized wordlist that included glosses and prompts to ensure consistent lexical retrieval. The responses were then phonetically transcribed and classified based on lexicostatistical categories: identical forms, phonemic correspondence, phonetic similarity, and minimal phonemic difference. Data analysis followed the lexicostatistical calculation method as outlined by Keraf (1996), with additional computation of divergence time and standard error to determine the estimated separation period between the two languages.

In this study, the comparison between the Sibolga Coastal Malay Language and the Mandailing Language is represented using standard abbreviations to maintain consistency in data presentation. The abbreviation SCML refers to the Sibolga Coastal Malay Language, which is a local variety of Malay spoken along the eastern coastal area of Sibolga. Meanwhile, ML stands for the Mandailing Language, which is predominantly spoken in the Mandailing Natal Regency, both located in North Sumatra.

The lexical data collected were organized using a gloss, which serves as a standardized semantic label or meaning reference for each vocabulary item. A gloss is a non-language-specific representation of a word's meaning, used to compare equivalent terms across different languages or dialects. For instance, the gloss "fire" would correspond to "api" in both SCML and ML. The use of glosses ensures uniform comparison across languages regardless of orthographic or phonological differences. These glosses were derived from the Swadesh 200-word list, which contains universally recognized core vocabulary items that are resistant to borrowing and relatively stable across time. For each gloss, corresponding lexical items were elicited from native speakers of SCML and ML and recorded for analysis. The data were then classified into categories such as identical forms, phonemic correspondences, phonetic similarities, and single-phoneme differences, which serve as the basis for the lexicostatistical calculations.

D. RESULT AND DISCUSSION

1. RESULT

From the results of data collection that has been carried out using 200 Swadesh vocabulary, several data were found that have been categorized according to the classification of related data analysis. In table 2 is the data of word pairs from both languages that are included in identical pairs. Identical word pairs are pairs of words from both languages that have identical similarities in terms of morphemes, phonemes, or phonology.

Table 2. identical word pairs

No	Gloss	SCML	ML	Phonemic Transcription (SCML)	Phonemic Transcription (ML)
1	child	anak	Anak	/anak/	/anak/
2	dog	anjing	Anjing	/and͡ʒiŋ/	/and͡ʒiŋ/
3	fire	api	Api	/api/	/api/
4	new	baru	Baru	/baru/	/baru/
5	rock	batu	Batu	/batu/	/batu/
6	star	bintang	Bintang	/bintaŋ/	/bintaŋ/
7	animal	binatang	Binatang	/binataŋ/	/binataŋ/
8	fruit	buah	Buah	/buah/	/buah/
9	hair	bulu	Bulu	/bulu/	/bulu/
10	worm	cacing	Cacing	/t͡ʃaciŋ/	/t͡ʃaciŋ/
11	rub	gosok	Gosok	/gosok/	/gosok/
12	count	etong	Etong	/etoŋ/	/etoŋ/
13	mother	umak	Umak	/umaʔ/	/umaʔ/
14	another	lain	Lain	/lain/	/lain/
15	shoot	manembak	Manembak	/manəmbak/	/manəmbak/
16	knock	manokɔk	manokɔk	/manokɔk/	/manokɔk/
17	dig	maɳorek	maɳorek	/maɳorɛk/	/maɳorɛk/
18	drink	minum	Minum	/minum/	/minum/
19	vomit	muta	Muta	/muta/	/muta/
20	short	pendek	Pendek	/pɛndɛk/	/pɛndɛk/
21	tree	batang	Batang	/bataŋ/	/bataŋ/
22	broom	sapu	Sapu	/sapu/	/sapu/
23	rope	tali	Tali	/tali/	/tali/
24	hand	tangan	Tangan	/taŋan/	/taŋan/
25	stab	tikam	Tikam	/tikam/	/tikam/
26	no	Inda	Inda	/inda/	/inda/
27	reject	Tulak	Tulak	/tulak/	/tulak/

The words pair that have phonemic correspondence is a pair of words from two different languages that are not identical, but have a systematic relationship between the sounds (phonemes) in the same position. In table 3 are word pairs that have phonemic correspondence.

Table 3. Pairs Have Phonemic Correspondence

No	Gloss	SCML	ML	Phonemic Transcription SCML	Phonemic Transcription ML
1	I	ambo	au	/ambo/	/aw/
2	water	ai	aek	/ai/	/aek/
3	heavy	barek	borat	/barek/	/borat/

4	blowing	mahambus	marhombus	/mahambus/	/marhombus/
5	near	dakek	donok	/dakek/	/donok/
6	sit	duduk	juguk	/duduk/	/d͡ʒuguk/
7	this	iko	on	/iko/	/ɔn/
8	that	itu	i	/itu/	/i/
9	evil	jahek	jahat	/d͡ʒahek/	/d͡ʒahat/
10	sewing	(man)jaik	jait	/mand͡ʒaik/	/d͡ʒait/
11	small	ketek	menek	/ketek/	/mənək/
12	dry	karing	goring	/kariŋ/	/goriŋ/
13	eat	makkan	mangan	/mak.an/ or /mak:an/	/maŋan/
14	squeeze	mamara	mamoro	/mamara/	/mamoro/
15	sucking	mangiso	mangitcop	/maŋiso/	/maŋit͡ʃop/
16	singing	balagu	marlagu	/balagu/	/marlagu/
17	blowing	mahambus	mangombus	/mahambus/	/maŋombus/
18	mosquito	raɲi	roɲit	/raɲi/	/roɲit/
19	long	panjaɳ	ginjaɳ	/pand͡ʒaɳ/	/gind͡ʒaɳ/
20	sand	kasik	horsik	/kasik/	/horsik/
21	think	pikki	pikir	/pik:i/	/pikir/
22	narrow	sampik	soppit	/sampik/	/sop:it/
23	land	tanah	tano	/tanah/	/tano/
24	snake	ula	ulok	/ula/	/ulok/
25	father	bapak	ayak	/bapak/	/ajak/ or /ajak/
26	road	bajalan	dalan	/bad͡ʒalan/	/dalan/

The next data is data between the two languages that have phonetic similarities. Words with phonetic similarity are words from two distinct languages that have an overall similarity in sound, but do not display a continuous phonemic change pattern or differ randomly yet still sound similar. Next in table 4 is a table of word pairs that have phonetic similarities.

Table 4. Phonetically Similar Pairs

No	Gloss	SCML	ML	Phonemic Transcription SCML	Phonemic Transcription ML
1	binding	mangabe	mangkobet	/maɲabe/	/maɲkobɛt/
2	nose	hidung	igung	/hiduɳ/	/iguɳ/

Finally in table 5 are word pairs that only differ by one phoneme. A pair of words from two languages that have the same meaning and very similar phonological forms, but differ by only one phoneme — whether it be a vowel or consonant sound, in any position within the word (beginning, middle, or end).

Table 5. Different One Phoneme Pairs

No	Gloss	SCML	ML	Phonemic Transcription SCML	Phonemic Transcription ML
1	ash	habu	abu	/habu/	/abu/
2	big	gadang	godang	/ɡadaŋ/	/ɡodaŋ/
3	flower	bungo	bunga	/buŋo/	/buŋa/
4	in	di	i	/di/	/i/
5	two	duo	dua	/duo/	/dua/
6	tails	iku	ikur	/iku/	/ikur/
7	liver	ati	ate-ate	/ati/	/ateate/
8	rain	ujan	udan	/ud͡ʒan/	/udan/
9	fog	kabuk	kabut	/kabuk/	/kabut/
10	wood	kayu	hayu	/kaju/	/haju/
11	we (1pl)	kami	hami	/kami/	/hami/
12	skin	kuli	kulit	/kuli/	/kulit/
13	flea	kutu	utu	/kutu/	/utu/
14	sky	langi	langit	/laŋi/	/laŋit/
15	fat	lamak	lomak	/lamak/	/lomak/
16	five	limo	lima	/limo/	/lima/
17	eyes	mato	mata	/mato/	/mata/
18	dead	mati	mate	/mati/	/mate/
19	cooking	mamasak	marmasak	/mamasak/	/marmasak/
20	splitting	mambala	mambola	/mambala/	/mambola/
21	spider	lawa-lawa	laba-laba	/lawa-lawa/	/laba-laba/
22	washing	mambasuh	mamasuh	/mambasuh/	/mamasuh/
23	up	naek	naik	/naek/	/naik/
24	year	taun	taon	/taun/	/taon/
25	sharp	tajam	tajom	/tad͡ʒam/	/tad͡ʒom/
26	thin	tipis	nipis	/tipis/	/nipis/
27	hunting	barburu	marburu	/barburu/	/marburu/
28	walking	(ba)jalan	(mar)dalan	/bad͡ʒalan/	/dalan/ or /mardalan/
29	sun	matohari	matahari	/matohari/	/matahari/
30	dust	habu	abu	/habu/	/abu/

The number of related words between the Sibolga Coastal Malay language and the Mandailing language has a total of 75 word pairs. From the total number of word pairs, the percentage of kinship between the two languages can then be calculated. The calculation is as follows:

$$c = \frac{k}{n}\ x\ 100\%$$

$$= \frac{75}{200}\ x\ 100\%$$

= 0,375 x 100% (rounded up to 0,38)

= 0,38 x 100%

$= 38\%$

From the results of this kinship calculation, it can be seen that both languages are included in the category of language family groups. Through this percentage, the separation time between the two languages can then be calculated. The calculation of the separation time between the two languages is;

$$W = \frac{\log C}{2 \log r}$$
$$= \frac{\log 38}{2 \log 805}$$
$$= \frac{968}{2 \, x \, 217}$$
$$= \frac{968}{434}$$
$$= 2.230 \text{ years ago}$$

This finding 2,230 years ago became the beginning of a hypothesis, where these two languages are estimated to be a single language 2,230 years ago. This result then became a prediction that these two languages separated from the proto language in the 21st century BC. However, the discovery of this number is not a definite number at the time of separation. This is because the language will separate gradually, slowly at the same time. Through this result, the error period can then be calculated to avoid errors in interpreting the separation time. In calculating this separation time, there is a standard error period used, which is 70% of the estimated truth. So, it is necessary to calculate the error period.

$$s = \sqrt{\frac{C\,(1-C)}{n}}$$
$$= \sqrt{\frac{0{,}38\,(1-0{,}38)}{200}}$$
$$= \sqrt{\frac{0{,}38 \, x \, 0{,}62}{200}}$$
$$= \sqrt{\frac{0{,}2356}{200}}$$
$$= \sqrt{0{,}001178}$$
$$= 0{,}0343220 \text{ (rounded up to 0,03)}$$

The number found from the calculation of the error term is then added to the percentage of kinship with details of 0.38 + 0.03 = 0.41. The result of this addition is called the new cognate value and this value will then be used to calculate the new separation time with the same separation time calculation formula;

$$W = \frac{\log C}{2 \log r}$$
$$= \frac{\log 0{,}41}{2 \log 805}$$
$$= \frac{892}{2 \, x \, 217}$$
$$= \frac{892}{434}$$
$$= 2.055$$

This final result will then be reduced by the long separation time. This action aims to minimize errors in calculating the separation time with details of 2,230 - 2,055 = 175. This final result will then be added and subtracted by the long separation time to be able to find the age prediction between the two languages. From this result, it is then obtained by using the standard error term, which is 0.7 from the truth. So, the results of the age prediction between the two languages are obtained;

a. Sibolga Coastal Malay and Mandailing were a single language 2,230 ± 175 years ago.
b. Sibolga Coastal Malay and Mandailing were a single language around 2,405 – 2055 years ago.
c. Sibolga Coastal Malay and Mandailing separated from the proto language around 382 – 32 BC (calculated from 2023).

Discussion

Through the data grouping that has been done which have been grouped based on the division and provisions of lexicostatistics (Aisyah & Widayati, 2022; Batubara & Widayati, 2022; T. A. Nasution, 2023; Nurmala & Widayati, 2022), it was obtained that from the two languages there are 27 identical word pairs, 26 word pairs that have phonemic correspondence, 2 word pairs with phonetic similarity, and 30 word pairs that only differ by one phoneme.

The findings of this study, which revealed a lexical similarity of 38% between the Sibolga Coastal Malay Language (SCML) and the Mandailing Language (ML), contribute significantly to the field of historical linguistics and lexicostatistics. This contribution is twofold: first, by reinforcing the utility of lexicostatistics in assessing linguistic kinship and estimating separation time; and second, by expanding the empirical database of Austronesian languages in North Sumatra, particularly underrepresented varieties like SCML.

Lexicostatistics, a method originally formalized by Morris Swadesh, operates on the assumption that basic vocabulary items are relatively stable over time and therefore suitable for tracing historical relationships between languages (Keraf, 1996). In this context, our study affirms that the 200-word Swadesh list is a reliable instrument for measuring lexical retention and divergence, as has also been shown in previous studies such as Aisyah & Widayati (2022), who applied similar techniques to compare coastal Malay dialects in Pasar, Kampung, and Sorkam. Like our findings, their results yielded cognate percentages that supported close linguistic kinship and shared ancestry.

Our study goes further by providing a detailed phonemic transcription and categorization of lexical data into four major relational types: identical pairs, phonemic correspondences, phonetic similarities, and one-phoneme differences. This nuanced classification is essential in lexicography and comparative linguistics because it reflects the various degrees of lexical relatedness and supports the reconstruction of phonological evolution over time. For instance, the identification of 27 identical word pairs and 26 pairs with regular phonemic correspondences points to systematic sound change, which is a cornerstone of the comparative method (Agha, 2007).

More specifically, the presence of systematic phonemic changes—such as the correspondence between SCML /k/ and ML /h/ in words like "*kayu*" vs. "*hayu*"—aligns with established principles in comparative phonology. These patterns not only support the kinship hypothesis but also offer concrete data for refining the reconstruction of proto-forms in regional Malayic languages. Similar comparative research in other Austronesian contexts (Batubara & Widayati, 2022; Rahmawati, 2022) has also identified such correspondences as key evidence of historical relatedness.

Furthermore, the calculation of separation time—estimated at 2,230 years ago, with a standard error range of ±175 years—demonstrates how quantitative approaches can complement qualitative linguistic analysis. While some scholars have critiqued glottochronology for its statistical assumptions (e.g., its reliance on constant lexical replacement rates), our study supports the view of Keraf (1996) and Mahsun (1995) that, with sufficient data and error control, lexicostatistics remains a valuable heuristic in diachronic linguistics.

Importantly, our results also highlight the role of geographical and sociocultural factors in linguistic divergence. The SCML and ML communities are located in adjacent yet ecologically and socially distinct areas of North Sumatra. Historical migration, trade, intermarriage, and sociopolitical separation likely contributed to the gradual divergence observed in their lexicons. This finding supports the sociolinguistic perspectives of Nababan (1993) and Puumala & Shindo (2021), who emphasize the interplay of social context and language change.

In relation to lexicography, this study provides empirically grounded data that can inform future dictionary compilation and comparative lexical databases for Austronesian languages. Our approach—documenting phonemic transcriptions and organizing words based on relational similarity—can serve as a model for constructing lexicons that are both diachronically informative and typologically sensitive. This aligns with the goals of modern lexicographic theory, which seeks to balance descriptive accuracy with historical depth (Charity Hudley et al., 2023).

Moreover, the classification of word pairs into "one-phoneme difference" and "phonetic similarity" categories offers valuable insight into the transitional stages of word change, which often precede total lexical replacement. Such categories are underutilized in many lexicostatistical studies but prove essential in understanding the gradience of lexical innovation and retention. As noted by Johansson et al. (2020), phonetic features often carry emotional and semantic salience that can influence lexical stability across generations.

Contrasted with related studies in the same field, this research adds novelty by combining field-based data collection, precise phonemic transcription, and error-calibrated time depth estimation. While earlier studies like Nurmala & Widayati (2022) compared European languages (e.g., English, German, Dutch), our study focuses on closely related Austronesian varieties, thus filling a gap in regional language documentation and typological comparison.

Lastly, these findings have implications for language planning and revitalization. Recognizing the historical ties between SCML and ML can foster interethnic understanding and may serve as a foundation for integrated cultural preservation initiatives. As regional languages face increasing pressure from national and global languages, studies like this reaffirm the deep-rooted interconnectedness of local linguistic heritage.

In conclusion, the present study not only confirms the kinship between SCML and ML but also enriches the methodological and empirical foundations of lexicostatistics and comparative lexicography. It demonstrates that quantitative lexical comparison, when grounded in detailed phonological analysis and sociocultural context, remains a powerful tool for uncovering the hidden histories of language families.

E. CONCLUSION

The results of this study demonstrate that the Sibolga Coastal Malay and Mandailing languages share a lexical similarity of 38%, placing them within the same language family. The lexicostatistical analysis estimates their divergence from a common proto-language occurred approximately 2,230 years ago, with a confidence range placing the separation between 382 and 32 BCE. This finding reinforces the hypothesis that these languages evolved from a shared linguistic origin and underwent gradual differentiation influenced by geography and sociocultural interaction. The implications of this research extend to historical linguistics, particularly for reconstructing proto-languages and understanding language diffusion in North Sumatra. It also supports the integration of quantitative methods in regional linguistic studies to better map language kinship and change. However, this study is limited by its reliance on a single list of lexical items and the relatively small number of informants. Future research should incorporate additional linguistic features, such as morphosyntactic and phonological rules, and expand the data set to include more dialectal variants for deeper comparative insight.

F. ACKNOWLEDGMENT

The University of Sumatra Utara and the course lecturers who helped with the journal writing are both acknowledged by the researcher. The resource people who were willing to supply research data are also thanked by the author. Every step of the study process is carried out entirely by the researcher alone, without any other funding. The authors declare that there is no conflict of interest related to the research, writing, and publication of this article. The authors have no financial affiliations, personal relationships, or professional interests that could affect the objectivity of the research results.

REFERENCES

Adelman, J. S., Estes, Z., & Cossu, M. (2018). Emotional sound symbolism: Languages rapidly signal valence via phonemes. *Cognition*, *175*, 122–130. https://doi.org/10.1016/j.cognition.2018.02.007

Agha, A. (2007). The object called "language" and the subject of linguistics. *Journal of English Linguistics*, *35*(3), 217–235. https://doi.org/10.1177/0075424207304240

Aisyah, S., & Widayati, D. (2022). Hubungan kekerabatan bahasa pesisir pasar, kampung, dan sorkam (kajian linguistik historis komparatif). *Aksara:*

Jurnal Ilmu Pendidikan Nonformal, *8*(3), 2367. https://doi.org/10.37905/aksara.8.3.2367-2376.2022

Anderson, S. R. (2016). Synchronic versus diachronic explanation and the nature of the language faculty. *Annual Review of Linguistics*, *2*(1), 11–31. https://doi.org/10.1146/annurev-linguistics-011415-040735

Batubara, N. A., & Widayati, D. (2022). Language kinship of English, German, and Dutch: A comparative historical linguistic study. *International Journal Of Humanities Education and Social Sciences (IJHESS)*, *1*(6). https://doi.org/10.55227/ijhess.v1i6.189

Charity Hudley, A. H., Clemons, A. M., & Villarreal, D. (2023). Language across the disciplines. *Annual Review of Linguistics*, *9*(1), 253–272. https://doi.org/10.1146/annurev-linguistics-022421-070340

Chaer, Abdul. (2014). *Linguistik umum.* Jakarta: PT. Rineka Cipta.

De Stefani, E., & De Marco, D. (2019). Language, gesture, and emotional communication: An embodied view of social interaction. *Frontiers in Psychology*, *10*, 2063. https://doi.org/10.3389/fpsyg.2019.02063

Dzhukaeva, M. A., Abueva, N. N., & Mamedova, G. B. (2024). Language as a tool of communication: How it reveals cultural reality and forms our spirit. *SHS Web of Conferences*, *195*, 02006. https://doi.org/10.1051/shsconf/202419502006

Fabbro, F., Fabbro, A., & Crescentini, C. (2022). The nature and function of languages. *Languages*, *7*(4), 303. https://doi.org/10.3390/languages7040303

Gajo, L., & Berthoud, A.-C. (2020). Issues of multilingualism for scientific knowledge: Practices for assessing research projects in terms of linguistic diversity. *European Journal of Higher Education*, *10*(3), 294–307. https://doi.org/10.1080/21568235.2020.1777451

Harahap, R. I. F., & Ritonga, H. J. (2024). Nilai-nilai "markobar" dalam pernikahan adat Mandailing dan keterkaitannya dengan bimbingan konseling islami. *Ganaya : Jurnal Ilmu Sosial dan Humaniora*, *7*(2), 224–236. https://doi.org/10.37329/ganaya.v7i2.3202

Hines, P. J., & Stern, P. (2019). More than a tool for communication. *Science*, *366*(6461), 48–49. https://doi.org/10.1126/science.aaz4133

Johansson, N. E., Anikin, A., Carling, G., & Holmer, A. (2020). The typology of sound symbolism: Defining macro-concepts via their semantic and phonetic features. *Linguistic Typology*, *24*(2), 253–310. https://doi.org/10.1515/lingty-2020-2034

Keraf, G. (1996). Linguistik bandingan historis. Jakarta: Gramedia Pustaka Utama.

Keraf, Gorys. (1997). *Komposisi.* Jakarta. Ikrar Media Mandiri

Kinzler, K. D. (2021). Language as a social cue. *Annual Review of Psychology*, *72*(1), 241–264. https://doi.org/10.1146/annurev-psych-010418-103034

Kridalaksaa, Harimurti. (1983). *kamus linguistik.* Jakarta. Gramedia

López-Narváez, J. (2023). Language and identity in the literature of fiction. The translation of idiolect and its effects in literary characterisation in Tess of the D'urbervilles' male characters. *TRANSFER*, *18*(1). https://doi.org/10.1344/transfer.2023.18.40325

Mahsun. (1995). Dialektologi diakronis. Yogyakarta: Gadjah Mada University Press.

Meylani, A. (2024). The role of Indonesian as a communication tool in learning. *Journal of Education, Linguistics, Literature, and Art*, *2*(2), 58–64. https://doi.org/10.62568/ella.v2i2.128

Mz, Z. (2019). The Role of language communication with the society and culture. *Vernacular: Linguistics, Literature, Communication and Culture Journal*, *1*(1), 1–9.

Nababan, P. W. J. (1993). Sosiolinguistik: Suatu pengantar. Jakarta: Gramedia.

Nasution, F., & Tambunan, E. E. (2022). *Language and communication*.

Nasution, T. A. (2023). Kekerabatan bahasa Melayu pesisir Sibolga, bahasa Toba, dan bahasa Mandailing: Kajian leksikostatistik. *Universitas Sumatera Utara*.

Niwanda, A., Harahap, M. A., & Rahmadani, P. (2024). Bahasa dan budaya sebagai cerminan kepribadian seseorang perspektif kasus budaya Jawa. *PUSTAKA: Jurnal Bahasa dan Pendidikan*, *4*(3), 184–192. https://doi.org/10.56910/pustaka.v4i3.1485

Nurmala, D., & Widayati, D. (2022). Lexicostatistics of English, German and Dutch. *International Journal of Research and Review*, *9*(5), 281–289. https://doi.org/10.52403/ijrr.20220536

Pickering, M. J., & Garrod, S. C. (2021). *Understanding dialogue: Language use and social interaction*. Cambridge University Press. https://doi.org/10.1017/9781108610728

Puumala, E., & Shindo, R. (2021). Exploring the links between language, everyday citizenship, and community. *Citizenship Studies*, *25*(6), 739–755. https://doi.org/10.1080/13621025.2021.1968696

Rabiah, S. (2018). *Language as a tool for communication and cultural reality discloser*. INA-Rxiv. https://doi.org/10.31227/osf.io/nw94m

Radulović, M. (2023). A review of methodologies and methods in linguistic research: Diachronic and synchronic approaches. *Facta Universitatis, Series: Linguistics and Literature*, 095. https://doi.org/10.22190/FULL230330008R

Rahmawati, R. (2022). Proto language relationship with Mandailing language. *Randwick International of Education and Linguistics Science Journal*, *3*(2), 362–367. https://doi.org/10.47175/rielsj.v3i2.482

Tanwete, C. S., & Kombinda, N. (2020). Object of study and linguistic subdisciplinary. *Macrolinguistics and Microlinguistics*, *1*(1), 23–36. https://doi.org/10.21744/mami.v1n1.3

Thoms, S. L., Bonviglio, T., & Suryasa, W. (2021). Linguists study language structure. *Linguistics and Culture Review*, *5*(1), i–iv. https://doi.org/10.21744/lingcure.v5n1.1844

Utama, S. S., Nuswantoro, A. W., Febrianto, A., & Mulyono, S. (2023). Hubungan kekerabatan bahasa Jawa dan bahasa Melayu (kajian linguistik historis komparatif). *Jurnal Pendidikan, Bahasa dan Budaya*, *2*(3), 60–76. https://doi.org/10.55606/jpbb.v2i3.1972

Verma, S. (2024). Language and literature: A reflection of social change. *The Creative Launcher*, *9*(6), 159–169. https://doi.org/10.53032/tcl.2024.9.6.18

Wijaya, B. S. (2024). The language of brand relationships: Symbolic, social, and political dimensions. *Review of Communication Research*, *12*, 18–32. https://doi.org/10.52152/RCR.V12.2

Ye, J., & Zhang, J. (2024). An analysis of metaphorical thinking and its flaws in the synchronic and diachronic linguistic sections of Saussure's institutes of linguistics. *IRA International Journal of Education and Multidisciplinary Studies*, *20*(2), 163. https://doi.org/10.21013/jems.v20.n2.p8