Ampersand 8 (2021) 100074

Contents lists available at ScienceDirect

Ampersand

journal homepage: www.elsevier.com/locate/amper

Ethiosemitic languages: Classifications and classification determinants

Tekabe Legesse Feleke

AcqVa Aurora Research Center, the Arctic University of Norway, Tromso, Norway

ARTICLE INFO

Keywords:
Areal classification
Determinants
Ethiosemitic
Linguistic distance
Language similarity

ABSTRACT

The present study addresses three concerns: (1) presents the areal classification of Ethiosemitic languages; (2) identifies major determinants of the distance among the languages; (3) challenges previous genealogical classifications of Ethiosemitic languages. To address these objectives, cluster analyses were performed on randomly selected 147 word lists. Multidimensional scaling was employed for the cluster validation. The cluster analyses performed on the phonetic and lexical distance matrices show that Ethiosemitic languages can be classified into six major groups: {Chaha, Gura, Gumer, Ezha, Mesqan, Muher}, {Amharic, Argobba}, {Endegagn, Inor, Gyeto}, {Wolane, Silt'e, Zay}, {Dobbi, Kistane}, and {Ge'ez, Tigrigna, Tigre}. Harari has an unstable position that swings based on the type of classification parameter used. The areal classifications obtained from the analyses fairly match the genealogical classifications previously proposed by historical linguists, resulting in a significant degree of overlap between the areal and genealogical classifications. The study further examined selected linguistic and non-linguistic variables that underpin the distance among Ethiosemitic languages, using Multiple Linear Regression. The results of the regression analyses show that lexical diffusion among Ethiosemitic languages, geographical distance and diffusion of phonetic features from Oromo to the Ethiosemitic languages are the major determinants of the distance among Ethiosemitic languages.

1. Introduction

Ethiosemitic languages are variants of the Semitic languages. As can be seen from Fig. 1, they are spoken in Ethiopia and present day of Eritrea. Amharic, Tigrigna and Tigre are among the major Ethiosemitic languages, in terms of the number of speakers. There are also other several varieties which are traditionally called 'Gurage languages'. The Gurage languages do not refer to a single genetically confirmed unit (Hetzron, 1972, p.119; Meyer, 2011, p.1221), and some of the speakers of these languages do not consider themselves as Gurage (Meyer, 2011, p.1223). Gafat is an extinct Ethiosemitic language, whereas Ge'ez is the oldest language of the early Abyssinian civilization, literature and philosophy, but currently functionally restricted to religious services in the Ethiopian Orthodox Church.

The study of the classification of Ethiosemitic languages is not a new project. There are a number of studies that gratefully contributed to our current knowledge of the classifications of the languages (e.g., Bender, 1971; Cohen, 1961; Hetzron, 1972; Leslau, 1969; Hudson, 2013; Hetzron, 1977; Girma, 2001). However, despite the contributions of several scholars, the classification issue has remained to be a subject of academic discord. To the best of my knowledge, there has not been consensus among Semitists on the issue of the classifications of Ethiosemitic languages (cf: Girma, 2001; Hetzron, 1972; Hudson, 2013; Voigt, 2009). There are multiple reasons for the disagreement; for example, the lack of adequate data on some of the languages, waves of migrations, and a long history of contact among Ethiosemitic and non-Semitic languages (see Goldenberg, 1977; Tekabe, 2020a; Tekabe, 2020b). However, frequently neglected crucial factors are the methodological and theoretical flaws in previous studies. Almost all previous studies attempted to provide a genealogical classification of Ethiosemitic languages, giving a marginal attention to the complex contact situation in the Ethiopian linguistic area, which makes the genealogical classification difficult, if not impossible.

The aspiration to conduct the present study emerged from the methodological and theoretical shortcomings that underpin previous classification studies of Ethiosemitic languages. Almost all previous studies heavily relied on the historical comparative methods, which are the extensions of the Tree metaphor (Schleicher, 1853). The studies, in general, took the assumptions of the Tree model, without adequately recognizing the enduring limitations, especially when they are applied in complex situations such as the Ethiopian linguistic area. The Tree metaphor assumes that language divergence is the outcome of a binary social split. It further speculates that a linguistic community always splits into two and develops independently, without any possibility of

E-mail address: tekabe.l.feleke@uit.no.

https://doi.org/10.1016/j.amper.2021.100074
Received 6 October 2020; Received in revised form 17 April 2021; Accepted 29 April 2021
Available online 20 May 2021

later contact (see François, 2015; Kalyan and François, 2018; Wang and Minett, 2005). Such a contact-free social division may happen in some rare cases, but certainly it is unlikely to be the case in the Ethiopian linguistic area. There is a bulk of literature about centuries of intense contact among Ethiosemitic languages and beyond (see Appleyard, 2011; Crass and Meyer, 2011; Ferguson, 1976; Zaborski, 1991; Leslau, 1945; Leslau, 1952; Leslau, 1957; Meyer, 2005). Given these facts, as also noticed by Aikhenvald and Dixon (2006); François (2015) and Kalyan and François (2018), any attempt to exclude contacted-related features and the subsequent changes can potentially lead to a misleading conclusion.

Therefore, the present study attributes the enduring classification debates partly to the excessive dependency on the genealogical classification since any attempt to classify languages 'genealogically', excluding borrowing, parallel development or drift becomes fruitful only in very exceptional cases. Historical linguists often get a hard time to find out contact-free shared innovations, which are usually fortuitously picked bunch of features. However, always the concern remains to be the number of innovations required to establish a genuine genealogical relationship. The hardest task in establishing a genealogical relationship is ascertaining that the shared innovations are not the outcomes of borrowing, diffusion or parallel development. As Aikhenvald and Dixon (2006, p.4) noted, it is not possible to show exclusively for any particular innovation that it results from genetic inheritance.

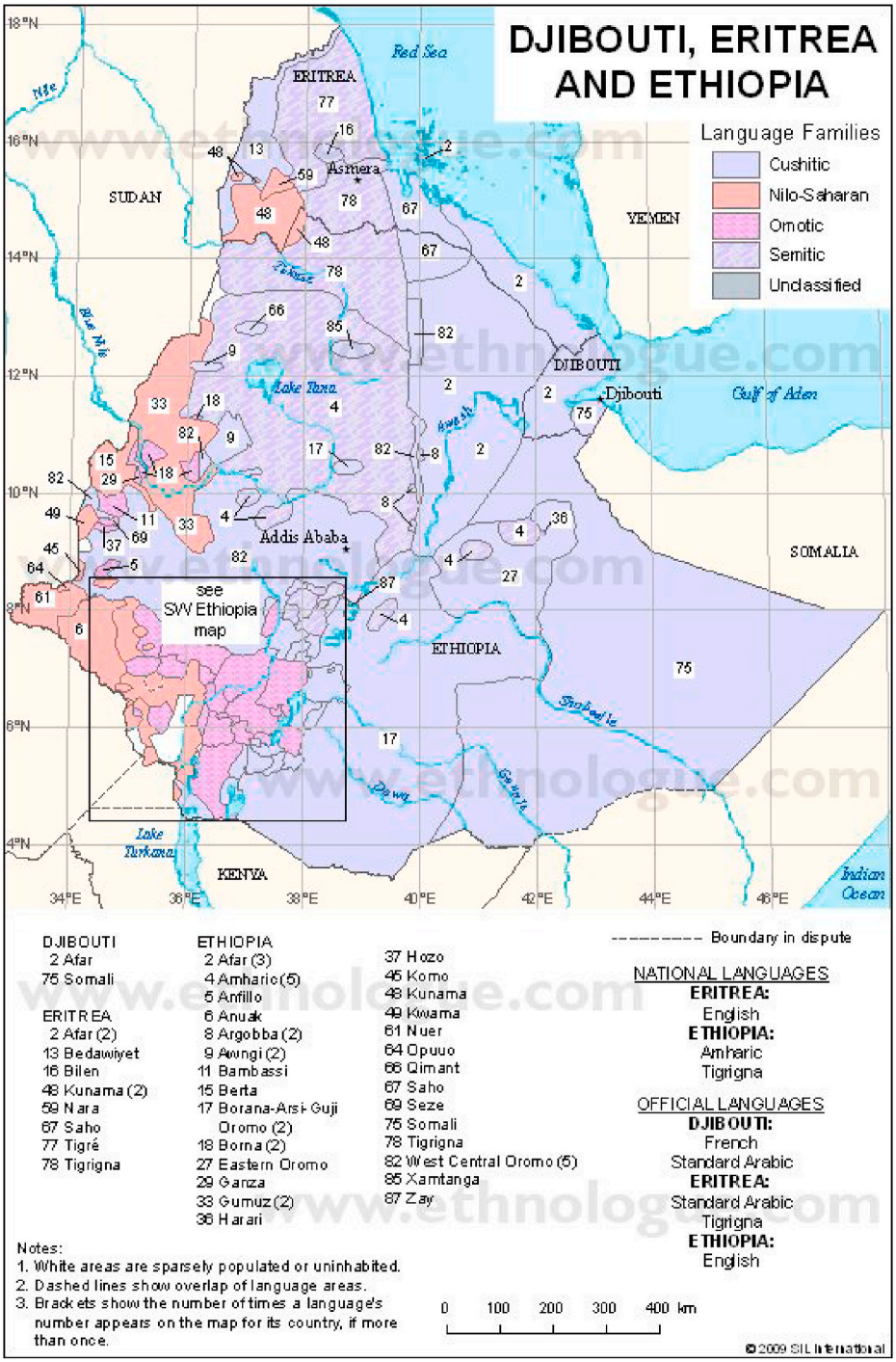


Fig. 1. Geographical distribution of the speakers of Ethiosemitic languages, from Ethnologue.

Given the mounting challenge of finding a contact-free shared innovation and the complex history of contact in Ethiopian linguistic area, the classification of Ethiosemitic languages demands an approach which is different from the traditional comparative methods. As noted by Goldenberg (1977) decades ago, areal classification is one of the alternatives. There are also recently innovated approaches such as Historical Glottometery (François, 2015; Kalyan et al., 2018; Kalyan and François, 2018) which combines the traditional comparative methods with the areal classifications to address classification related issues. The present study opted the areal classification, without making any genealogical claim. It applied the methods of Dialectometry, which quantify the distance among related languages, based on randomly selected aggregate data (Goebl, 2010; Nerbonne and Heeringa, 2001; Nerbonne et al., 2011; Prokic et al., 2013). By employing this approach, the study; (1) provides a complete areal classification of Ethiosemitic languages, (2) argues against previous genealogical claims, and (3) identifies major linguistic and non-linguistic factors that underpin the areal distance among Ethiosemitic languages.

The study addresses these objectives relying on phonetic and lexical measures. For Ethiosemitic languages that do not have native speakers such as Ge'ez, the study relied on the pronunciations of words provided in dictionaries. The phonetic distance was computed by comparing cognates from randomly selected 147 words (see Appendix A.1), using the Levenshtein algorithm. In the present study, the word *cognate* refers to lexical similarity, and it does not necessary refer to genetic relatedness, as is often the case in historical linguistics. The lexical distance was defined as the ratio of non-cognates to the total list of words. To obtain the intended areal classifications, cluster analyses were performed on the phonetic and lexical distance matrices, using Gabmap. Given that clustering is sometimes tricky, the clusters were validated using multidimensional scaling. Among various methods of cluster analysis included in the Gabmap, Weighted Average method was used following Gooskens and Heeringa (2004). Major linguistic and non-linguistic determinants, that is, lexical diffusion among Ethiosemitic languages, the influence of Oromo, geographical distance, and population size were also examined. Due to time and resource constraints, only the influence of Oromo was considered in this study though there could be the influences of other Cushitic and Omotic languages. This is partly because Oromo has relatively a large number of speakers and consequentially its impact could also be relatively large. Lexical diffusion among Ethiosemitic languages was determined using Neighbor-net representation. The influence of Oromo on the Ethiosemitic languages was determined by comparing the phonetic and lexical similarity between Oromo and the Ethiosemitic languages, and by estimating the degree of phonetic and lexical skewing between Oromo and each of the Ethiosemitic languages. With regard to population size, the numbers of residents of the ethnically defined districts (e.g., Ezha Wereda, total population 84,905) was taken as the numbers of speakers of the varieties based on ENSA (2007). The populations of Gura and Ge'ez were not included in this analysis because there is no reliable data about Gura population, and Ge'ez does not have native speakers. The estimated number of Tigre and Eritrean Tigrigna speakers was obtained from Ethnologue (23rd edition, 2020) since there is no recent census data on speakers of the two languages. The geographical distance between each language area was obtained from Google Map. The geographical distance was defined as the length (in km) of the main roads that connects each of the language areas. Multiple Linear Regression analyses were performed to determine the interaction between geographical distance, population size, the influence of Oromo and the phonetic and lexical distances.

2. The origin and classifications of Ethiosemitic

The origin of Ethiosemtic languages has been a subject of academic debate for decades. The debate emerges from the enduring lack of clarity about the origin of the Semitic languages in general. There are two hypotheses with regard to the origin of the Semitic languages: traditional theory and the Africanist view (Girma, 2001, p.59; Fekede, 2015, p.10). The traditional theory assumes that some speakers of Afroasiatic languages initially migrated from Africa to Asia. Through time, after their separation from the rest of Afroasiatic speakers, their language developed its own distinct features in Asia and formed a proto-Semitic language. Then, the speakers of one of the affiliates of the proto-Semitic language migrated from Asia (South Arabia) to Ethiopia (Ehret, 1988, p.641; Fleming, 1968, p.354; Gragg, 1997, p.163; Hetzron, 1972, p.15–19; Hetzron, 1977, p.16). Based on this assumption, today, many scholars believe that Ethiosemitic languages descended from the Proto-Semitic language family, particularly, from the South Semitic branch which was spoken in the South Arabia. According to this theory, these migrants first moved from South Arabia to northern Ethiopia. Then later, some Semitic speakers migrated from the north to the south (Hetzron, 1972, p.36). Those who took the direct route from north to the south are the speakers of Gunnən Gurage and Gafat, which form together Outer South Ethiosemitic (see Hetzron, 1972, p.36). However, those who moved from north Ethiopia to east Ethiopia and then moved to south constitutes the speakers of Amharic, Argobba, Harari and East Gurage (Silt'e, Zay and Wolane) which form together Transversal South Ethiosemitic (Hetzron, 1972, p.36).

The alternative hypothesis, contrary to the first one, is that the origin of Ethiosemitic languages is Africa. Therefore, it is called Africanist view. According to this proposal, Ethiosemitic is a descendant of the Afroasiatic language which had been spoken in Africa in pre-Semitic era. The hypothesis implies that Ethiosemitic languages had been spoken in Africa before the expansion of the Semitic languages across Asia and north Africa. According to Girma (2001, p.59–60), two explanations are often provided to support this proposal. First, among six of the Afroasiatic sub-families (Semitic, Cushitic, Omotic, Berber, Chadic and old Egyptian), only one is spoken in Asia while all of them are spoken in Africa. Among these, three of them are spoken in Ethiopia (Semitic, Cushitic and Omotic). Based on this, Ethiosemitic is considered as the mother language of all Semitic languages. Another reason is that more Semitic languages are spoken in Ethiopia (around 16) than in Asia. Hence, based on the 'least moved principle' they assume that the origin of all Semitic languages is Africa, particularly Ethiopia (see Hudson, 2000).

Recent phylogenetic studies seem to support the first proposal. For instance, a study by Kitchen et al. (2009) suggests that Ethiosemitic languages are descendants of South Semitic - a presumptive branch of Semitic language. Furthermore, Fleming (1968) estimates that the South Ethiosemitic languages separated from North Ethiosemitic between 700 BC and 300 BC. The study further estimates that the separation between the North and South Ethiosemitic occurred either in South Arabia or around the Red Sea. Besides, the study predicts that the diversification of the South Ethiosemitic languages began between 300 BC and 100 AD. Moreover, Bender (1966) estimates the date of separation of other Ethiosemitic languages from Ge'ez. The study assumed that Ge'ez is the proto-language of all Ethiosemitic languages, the assumption which is today widely refuted (see Palmer, 1958, p.120). According to his study, North and South Ethiosemitic languages separated around 2000 years ago. The majority of Semitic scholars seem to believe that Ethiosemitic languages descended from the Modern South Arabia. However, it is too early to reject the claim that the origin of all Semitic languages is Ethiopia. Rigorous comparison of Ethiosemtic languages with other Semitic languages and combining this with recent archaeological findings may settle the debate sometime in the future.

The classification of Ethiosemitic languages is as controversial as the origin of the languages. The controversies seem to have several causes. The plausible causes are the complexity of the Ethiopian linguistic area, the type and number of linguistic parameters used and theoretical assumption held by the scholars of Ethiosemitic languages. The Ethiopian linguistic area is extremely diverse and it constitutes a long history contact among Semitic, Cushitic, Omotic and Nilo-Saharan languages (see Ferguson, 1976; Meyer, 2005; Zaborski, 1991). The complex

intermingling among languages of these families makes difficult the proper application of methods of comparative historical linguistics (see Goldenberg, 1977, p.462). Often disparities come from the type and number of linguistic parameters previously employed. Some studies strove to strictly rely on morpho-syntactic shared innovations (e.g., Girma, 2001; Hetzron, 1972; Hetzron, 1977; Leslau, 1969) while others used lexical parameters (e.g., Hudson, 2013; Bender, 1971; Bender et al., 1972; Kitchen et al., 2009). The difference among those that used morpho-syntactic parameter is either the number of shared features used or the degree of precision involved in selecting the 'contact-free shared innovations'. For instance, Hetzron (1972) has been criticized for inconsistently using syntactic and phonetic parameters though he argued that the best classification should consider morphological shared innovations. The assumption in the historical linguistics, in general, is that morphology is less susceptible to borrowing. However, this seems to be just a theoretical stand with little empirical justification. Almost all previous classification proposals relied on the historical comparative approaches, which aim to classify languages genealogically. This has nurtured the existing controversy since genealogy-based classification is practically unachievable, because of contact and heavy borrowing, especially among geographically adjacent languages.

Among numerous classification proposals, Hetzron (1972) is often considered as the most complete in terms of the number of languages included. Hetzron (1972) has also been appreciated for a thorough consideration of morphological features. Hence, the present study considers this work as the main reference, and cites other classification proposals whenever need arises. There is a general consensus among Semitists that Ethiosemitic languages are divided into North and South Ethiosemitic. The North Ethiosemitic languages consist of Tigre, Tigrigna and Ge'ez (see Girma, 2001, p.2; Hetzron, 1972, p.3; Raz, 1983). According to Hetzron (1972), the South Ethiosemitic languages are classified into Outer South and Transversal South (see Fig. 2). The Outer South is further divided into n-Group and tt-Group. The tt-Group consists of Muher and Western Gurage, which in turn is divided into Mesqan and 3 TG (3 Tense Gurage). The 3 TG languages are further classified into CWG (Central West Gurage) and PWG (Peripheral West Gurage) languages. The Central West Gurage languages include Chaha, Gura, Gumer and Ezha while the Peripheral West Gurage languages consist of Endegagn, Inor, Gyeto and Mesmes. According to Hetzron (1972), the Transversal South is further classified into Central Transversal and Eastern Transversal. The Central Transversal consists of Amharic and Argobba while the Eastern Transversal consists of Harari and East Gurage languages, which are further classified into Zay and Silt'e.

Not every scholar agrees on the Hetzron (1972)'s classification (see Goldenberg, 1977); hence, there exist various alternative classification proposals (e.g., Girma, 2001; Hudson, 2013; Kitchen et al., 2009). Nonetheless, none of them could provide a convincing alternative. Girma (2001) and Voigt (2009) are among recent works that aimed at improving Hetzron (1972)'s classification proposal. Girma (2001) and Hetzron (1972) do not differ on the classification of Ethiosemitic languages into North and South. They also do not differ on the classification of Eastern South Ethiosemitic languages. However, unlike Hetzron (1972), Girma (2001) included Wolane under Eastern South Ethiosemitic and presented it as a sister language of Silt'e. Hetzron (1972) did not include Wolane in the classification probably because he thought that it is a dialect of Silt'e. Furthermore, Hetzron (1972) considered Zay as a sister language to Silt'e (classified under East Gurage). Girma (2001), however, considered it as a separate language (as a sister language to East Gurage). The two works differ significantly in the classification of the Outer South Ethiosemitic. Girma (2001) rejected the -n/-tt classifications arguing that these markers are not the features of all the languages that Hetzron (1972) mentioned. Instead, he took the initial letter of some of the languages in the group (Dobbi, Mesqan and Kistane (Soddo)) and formed GMS-group, and divided the Outer South Ethiosemitic into GMS-group and Western Gurage. GMS-group consists of almost all the languages that are included in n-Group in Hetzron (1972)'s classification, except Mesqan. The two scholars differ on the position of Mesqan. Hetzron (1972) classified Mesqan under West Gurage while Girma (2001) classified it under North Gurage. Unlike Hetzron (1972), Girma (2001) dissolved the tt-group arguing that the -tt feature is not a representative of all the languages in the group, and he directly derived Central West Gurage and Peripheral West Gurage from Western Gurage. Clearly, this change primarily affects the position of Muher. Girma (2001) classified Muher under the Central West Gurage, citing Leslau (1992) and Leslau (1969). Girma (2001) also rejected the 3 TG classification of Hetzron (1972), arguing that this tense type does not

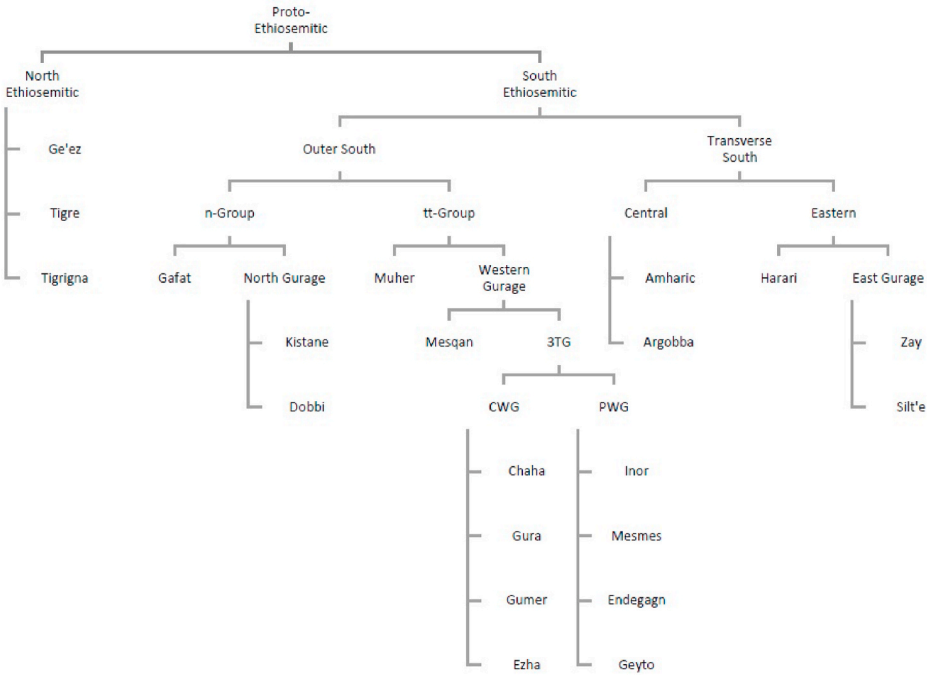


Fig. 2. Classification of Ethiosemitic languages by Hetzron (1972). For Hetzron (1972), Wolane is a dialect of Silt'e.

exist in all the languages. Hence, he divided Western Gurage into Central West and Peripheral West Gurage. This division is the same both for Hetzron (1972) and Girma (2001). However, the two differ on the position of Inor. Unlike Hetzron (1972), Girma (2001) classified Inor under Central West. Quoting Leslau (1996), he argued that Inor is close to the Central West Gurage languages particularly to Chaha.

While there are several studies that employed morpho-syntactic parameters, similar to Hetzron (1972) and Girma (2001), proper classification of Ethiosemitic languages exclusively based on phonetic parameters is just a recent phenomenon (see Tekabe, 2020a; Tekabe, 2020b). Many other studies employed lexical parameters. For instance, Bender et al. (1972) examined 12 Ethiosemitic languages, using a 98-words list from Swadesh (1955). Bender (1971) also compared several Ethiopian languages using the same method. Similarly, Hudson (2013) classified 14 Ethiosemitic languages, using a 250-word list. Likewise, using a list of 255 words, Fekede (2015) investigated lexical similarity among six South Ethiosemstic languages: Kistane, Chaha, Inor, Mesqan, Muher and Wolane. Furthermore, Kitchen et al. (2009) classified 15 Ethiosemitic languages, using lexical parameters. Many of these studies are limited just to the comparison of the percentage of shared lexical items. Relatively more rigorous classification attempts were made by Hudson (2013) and Kitchen et al. (2009). Using the lexical parameter, Hudson (2013) classified Ethiosemitic languages into four parallel groups: North Ethiosemitic - {Tigre, Tigrigna, Ge'ez}, Gunnən Gurage - {Kistane, Mesqan, Muher, Chaha, Inor}, Eastern South {East Gurage (Silt'e and Zay), Harari} and {Amharic, Argobba}. However, Meyer (2018) challenged the reliability of this classification, pointing out three major limitations: the problem of using intelligibility as a criteria to distinguish languages from dialects, pitfalls in the comparison of word lists used for lexical comparison and imprecise cognacy judgment. Kitchen et al. (2009) classified selected Ethiosemitic languages into five major groups: {Ge'ez, Tigrigna, Tigre}, {Harari, Zay, Wolane}, {Kistane, Mesmes}, {Gafat, Argobba, Amharic} and {Chaha, Gyeto, Mesqan, Inor}.

Regardless of these attempts, the issue of classification of Ethiosemitic languages has remained a subject of further inquiry. As also noticed by Girma (2001, p.61), Fleming (1968, p.354) and Faber (1997, p.3–4), so far there is no convincing evidence with regard to the genealogical relationship among many of the Ethiosemitic languages. In previous classification proposals, for instance, Silt'e is closer to Harari than to the Gunnən Gurage languages (also see Fig. 2). Furthermore, Kistane is closer to Gafat than to other Gurage varieties. Moreover, there is still a controversy with regard to the position of Mesqan. Hetzron (1972) and Hetzron (1977) classified Mesqan under West Gurage while Girma (2001) classified it under North Gurage. Muher also does not have a settled position in the classification of Ethiosemitic languages. While Hetzron (1972) classified it under the tt-Group, Girma (2001) placed it under Central West Gurage. This evidence together show that the classification of Ethiosemitic languages is not a finished business (also see Girma, 2001, p.61; Hetzron and Bender, 1976, p.5; Hudson, 2000, p.75–76; Goldenberg, 1977, p.461; Leslau, 1969, p.97; Leslau, 1992, p.12).

3. Methods

This section presents the methods used to address the objectives stated in §1. It describes the data sources (§3.1), methods of computing the phonetic and lexical distances (§3.2), methods of clustering and cluster validation (§3.3), techniques of identifying the determinants of the linguistic distance and examining the relationship between the determinants and the linguistic similarities (§3.4).

3.1. Data sources

The 147 word lists (see Appendix A.1) examined in the present study were collected through own fieldwork and published sources. Tigre list of words was obtained from published sources, mainly from Leslau (1982), Elias (2014), Leslau (1948), Leslau (1945a) and Leslau (1945c). Argobba list of words was collected from Zelealem and Siebert (2001), Leslau (1973), Mohammed et al. (2014), Leslau (1957) and Leslau (1997). Gyeto, Zay, Wolane and Dobbi lists of words were collected from Leslau (1979). Ge'ez list of words was obtained from Leslau (1987) and Leslau (2010). Harari list of words was obtained from Leslau (1963). Amharic and Oromo lists of words were from the author who is a balanced bilingual of the two languages. Tigrigna list of words was obtained from a native Tigrigna speaker who is also a linguist. Chaha, Endegagn, Ezha, Gumer, Gura, Inor, Kistane, Mesqan, Muher and Silt'e lists of words were collected during field works conducted from February to May 2018. There are some cautions in order here; given that data from the published sources were collected from informants recruited by the researchers of the previous studies, it was impossible to identify the specific dialects of the informants. Furthermore, since the data from published sources were collected at different time by different linguists, it was difficult to control variations that emerge from individual differences and time variation. Moreover, there are different Argobba varieties, mainly the Tollaha (Wetter, 2006, 2010; Getahun, 2009, 2018) and Showa Robit variety (Leslau, 1997; Zelealem and Siebert, 2001). The data used in the present study was obtained largely from studies conducted on the latter variety.

3.2. The phonetic and lexical distances

The phonetic and lexical distances among the Ethiosemitic languages were computed based on the data obtained from the fieldwork and published sources. With regard to the data from the fieldwork, the author phonetically transcribed the randomly selected 147 word lists, based on the knowledge of three native speakers recruited from each language area. Cognates were identified by the investigator based on form similarity. In almost all Semitic languages, sequence of consonants form the basic word meaning (root). Hence, the similarity of the consonantal roots was taken as a core parameter. False friends were excluded using the semantic parameter - the attested meaning similarity. For instance, the word *gaz* has different meanings in Amharic and Chaha. In Amharic it refers to 'gas' while in Chaha it stands for 'war', suggesting that form similarity alone is not sufficient to identify the cognates. Phonetically transcribed list of words in the published sources were adopted from the works stated in §3.1. The adoption also involved minor standardization processes to maintain uniformity of phonetic representations across the languages. For instance, variations in the representation of the mid-central vowel was standardized by uniformly using [ə]. In some previous works it was presented as [ə] while in some works it was presented as [ä]. Then, shared cognates in the lists of words were identified by the author, based on two parameters: form and meaning similarity. For meaning equivalence, the author relied on dictionary definitions. After cognate identification, cognates that are shared among eleven (11) of the twenty (20) languages were considered for the phonetic comparison. In other words, the phonetic distance was computed on cognates that are shared at least by 55% of the languages. No especial attention was given to borrowed words because the study has no intention to make any genealogical claim.

The phonetic distance was computed just between the cognates, based on Kessler (1995, p.65)'s recommendation that there is no need of computing the distance among words that do not have any form similarity. To compute the phonetic distance of the concerned languages, first the cognates were aligned. Then, the distance among the cognates was computed using Levenshtein algorithm in Gabmap, based on the number of phones which was inserted, deleted or substituted. The distance computation was made based on the simplest cost assignment. The simplest cost assignment assigns equal cost for each operation, 1 point for each operation. Levenshtein algorithm provides absolute and relative (normalized) string distance. In the present study, the normalized phonetic distance was employed to minimize the variation due to

differences in the length of the cognates. Table 1 presents a sample phonetic distance computed between Kistane and Chaha, based on the word 'cloud'. In this particular example, the Levenshtein distance is 2. This distance should be divided by the number of segments (six in this example) to obtain the normalized distance. Hence, the relative or normalized phonetic distance between the two cognates is 0.33 (2/6). This value can be converted into percentage and presented as 33%.

The lexical distance among the Ethiosemitic languages was determined by computing the ratio of the non-cognates to the total lexical items. After the cognates and non-cognate words in the pairs of languages were identified, the lexical distance was computed by dividing the number of non-cognate words by the total lexical items in the pair of languages. Then series of cluster analyses were performed on the phonetic and lexical distance matrices. The classification results obtained from the analyses were compared to selected previous genealogical classifications (Bender, 1971; Girma, 2001; Hetzron, 1972; Kitchen et al., 2009), based on the cophonetic distance between each node in the classification trees. The cophonetic distance between any two terminal nodes in a tree was defined as the number of nodes one has to go up from, let say, language A to the lowest common node shared between the member of the pairs, and then down to language B (see Gooskens and Heuven, 2018).

3.3. Clustering and cluster validation

The present study employed Gabmap to classify the languages and to validate the classifications. Gabmap is a web-based dialect classification and visualization software, developed by computational linguists at the University of Groningen. It provides several statistical alternatives to perform cluster analysis (Ward's Method, Complete Link, Group Average and Weighted Average). The classification alternatives were included in Gabmap, based on the findings of Prokić and Nerbonne (2008) which evaluated the stability of various clustering techniques. In general terms, based on the sample size and the type of data, one can choose any of the four alternatives. Relying on Gooskens and Heeringa (2004), the Weighted Average method was applied to the selected Ethiosemitic languages. Clustering can sometimes be tricky - a small variance in a distance matrix can result in entirely different groupings. To enable a proper management of this, Gabmap provides two cluster validation techniques - multidimensional scaling and fuzzy clustering. The present study employed multidimensional scaling to make sure that the clustering results are reliable. Gabmap provides the result of multidimensional scaling plot, together with the corresponding dialect map. The multidimensional scaling plot displays the distance among the languages in an n-dimensional space. In other words, it takes the full site by site distance matrix as an input and creates a representation in the n-dimensional space where distances are approximations of the original linguistic distances (see Leinonen et al., 2016; Nerbonne et al., 1999; Snoek, 2014). The results of the multidimensional scaling can be plotted in a Cartesian coordinate system. On the plot, similar data points are placed close to each other. The cluster validation part of the Gabmap provides options that can be utilized to refine the classifications, by excluding distinct groups and narrowing the analysis to the languages that do not precisely group together. The multidimensional scaling results can also be used to inspect the amount of variance explained by each dimension. The first dimension of the multidimensional scaling usually explains much of the variance in the data, and additional dimensions are added for the precision of the approximation of the distance, but each additional dimension explains less of the variance than the first one. In multidimensional scaling, data points with similar values are always shown close to each other. Gabmap also automatically creates dialect map of the languages, using Google Earth and the linguistic distance as inputs.

Table 1
Phonetic distance, using Levenshtein algorithm.

Kistane - Chaha 'cloud'					
d	a	m	ə	n	a
d	a	b	ə	r	a
		1		1	
Absolute				2	
Relative				0.33	

3.4. Determinants of the linguistic distance

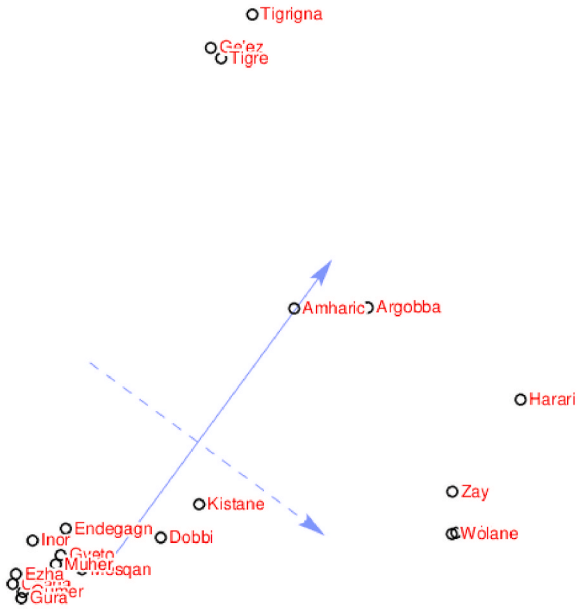
The present study employed the Neighbor-net representation to determine whether the similarity among the Ethiosemitic languages was due to the horizontal diffusion of features or the outcome of genealogical inheritance. Neighbor-net algorithm provides two types of outputs based on the nature of the input or linguistic distance. If the input distance reflects a vertical relationship, it gives a tree-like structure, but if the input distance reflects a horizontal relationship, it produces net-like structure (see Huson and Bryant, 2010). In addition to the horizontal diffusion of features, the degree of the influence of Oromo on the Ethiosemtic languages was examined using two approaches: by determining the degree of phonetic and lexical similarity between Oromo and each of the Ethiosemitic languages, and by computing the lexicostatistical skewing between each pair of Ethiosemitic languages with respect to Oromo. The lexicostatistical skewing, for example, between language 'A' and language 'B' with respect to language 'D' was defined as the similarity between language 'A' and language 'D' minus the similarity between language 'B' and language 'D'. For instance, if the lexical similarity between Argobba and Oromo is 45%, and the lexical similarity between Amharic and Oromo is 40%, the 5% lexical difference between Argobba and Amharic with respect to Oromo is considered as the lexical difference between Argobba and Amharic because of the influence of Oromo on Argobba (see Tekabe, 2020a; Heine, 1974; Minett and Wang, 2003; Wang and Minett, 2005). In the present study, both phonetic and lexical skewing were computed based on phonetic similarity (Appendix A.2.2) and the lexical similarity (Appendix A.2.3) indices respectively. The similarity indices were obtained by subtracting the phonetic and lexical distance between each pair of languages from 100 ($s_{index} = 100 - d$) respectively. The phonetic skewing between two related languages, let say, language 'A' and language 'B' with respect to Oromo was defined as the phonetic similarity between language 'A' and Oromo minus the phonetic similarity between language 'B' and Oromo. The lexical skewing between two related languages, for example, language 'A' and language 'B' with respect to Oromo was defined as the lexical similarity between language 'A' and Oromo minus the lexical similarity between language 'B' and Oromo. The absolute values of the phonetic and lexical skewing were used for the Regression analysis since some of the skewing values were negative.

As indicated in §1, the present study also examined the influence of population size and geographical distance on the linguistic distance among Ethiosemitic languages. The population size of the speakers of each language was obtained from ENSA (2007). ENSA (2007) report does not directly provide the number of the native speakers. Nevertheless, it reports the number of residents in each district. Given that the administrative divisions of the districts are based on the ethnic background of the residents, the number of residents of the ethnically defined districts (e.g., Ezha Wereda, total population 84,905) was taken as the numbers of speakers of the varieties. Travel distance (the length of the main road which public transportation (buses) uses) between each of the language sites was taken as the measure of the geographical distance. The travel distance was obtained from Google Map. The influences of population size and geographical distance were determined based on the hypothesis of the Gravity model (Trudgill, 1974) which predicts that linguistic similarity is directly proportional with population size, but inversely proportional with the square of the geographical distance. To determine which of the factors (the influences of Oromo, population size and geographical distance) influence the distance among Ethiosemitic

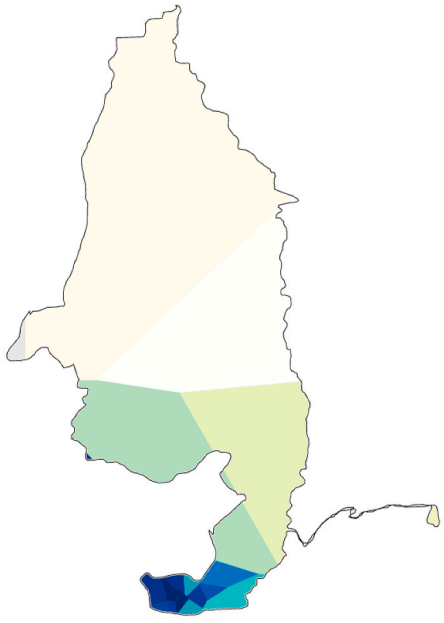
languages, Multiple Linear Regression was performed in R, using geographical distance, population size and the influence of Oromo as independent variables, and phonetic and lexical similarities as dependent variables. For the sake of easy illustration, the Ethiosemitic languages were divided into three groups based on the degree of phonetic and lexical similarity: *similar* (if the languages share 80% or more phonetic/lexical similarity), *partially similar* (if the languages share 70%–80% phonetic/lexical similarity) and *not similar* (if the languages share less than 70% phonetic/lexical similarity).

4. Results

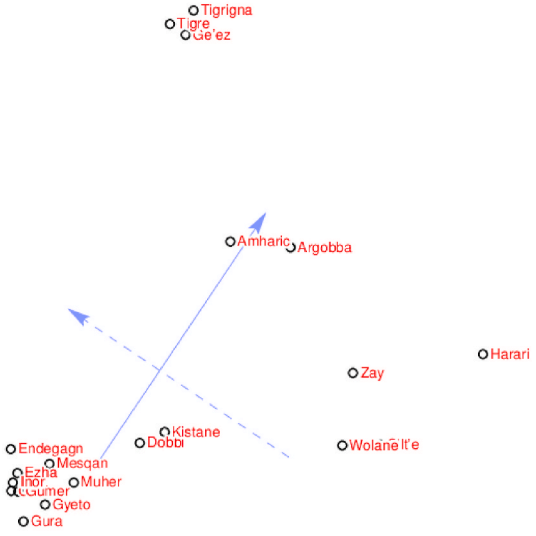
This section presents results obtained using the methods discussed in §3. In §4.1, the phonetic and lexical classifications are presented, combined with the results of multidimensional scaling. Only the classification trees are presented in the section; the phonetic as well as the lexical similarity matrices are presented in Appendix A.2.2 and A.2.3 respectively. Section 4.2 compares the areal classifications reported in §4.1 to the classifications previously proposed by historical linguists. Section 4.3 reports the results of the analyses of the determinants of linguistic distance.



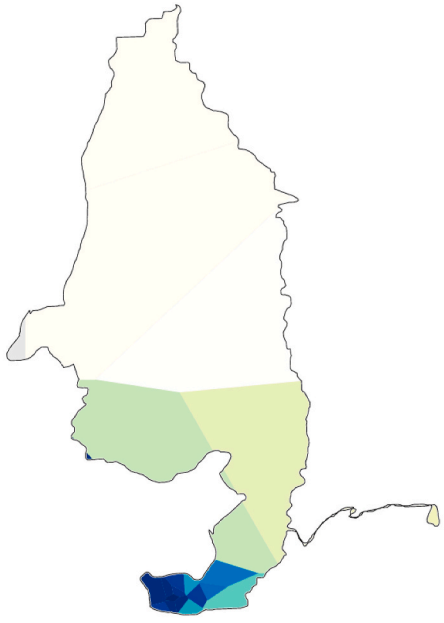
(a) Plot of multidimensional scaling in two-dimensional space for the phonetic distance. Since Silt'e and Wolane have very close phonetic similarity, their locations overlap on the Figure.



(b) Multidimensional scaling map for the first dimension of the phonetic distance; light color shows areas with the highest phonetic distance



(c) Plot of multidimensional scaling in two dimensional space for the lexical distance. Some languages overlap since they have very close similarity.



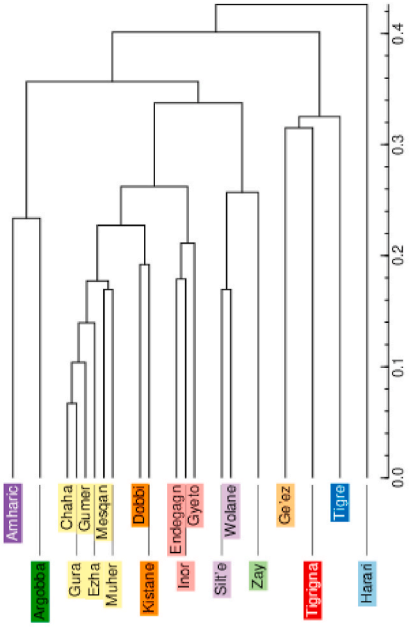
(d) Multidimensional scaling map for the first dimension of the lexical distance; the light color shows areas with the highest lexical distance

Fig. 3. Multidimensional scaling maps and plots of Ethiosemitic languages.

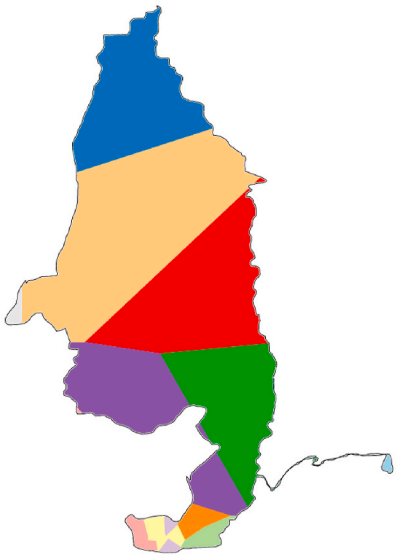
4.1. The Classification of Ethiosemitic languages

In line with the explanation in §4, the classification results are supplemented by multidimensional scaling results. Gabmap provides one map for each dimension of the multidimensional scaling. Only the map of the first dimension is reported here, and readers are referred to Appendix A.2.4 for the map of the second dimension. Fig. 3(a) shows the multidimensional scaling plot of the phonetic distance in two-dimensional space. The first dimension is represented by a solid arrow and the second dimension by a dashed arrow. In the figure, the first dimension shows that Chaha, Gura, Gumer and Ezha have the lowest phonetic distance values while Tigrigna has the highest distance value - the arrow points from low to high values. The values of the other languages are between these two extremes.

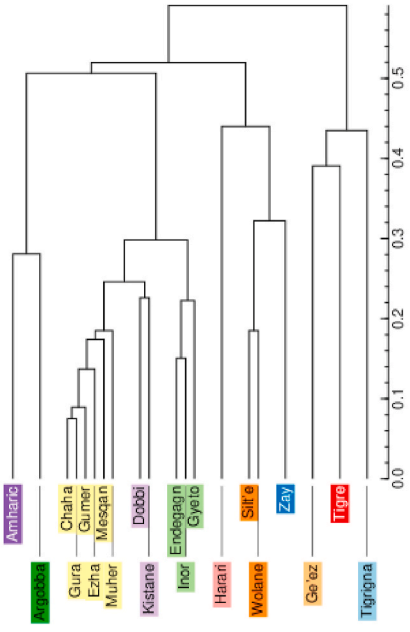
The second dimension (dashed arrow) indicates that Chaha, Gura, Gumer and Ezha have the lowest distance values while Zay, Wolane and Silt'e have the highest values. The multidimensional scaling results presented in 3(a) clearly illustrates that {Ge'ez, Tigrigna and Tigre} form a group. Similarly, {Amharic and Argobba} as well as {Silt'e, Wolane and Zay} form independent groups. The figure further shows that {Harari} is a separate language. Moreover, {Kistane, Dobbi} form a group. The classification of the remaining languages cannot be clearly seen from Fig. 3(a). Further inspection, after excluding those that clearly form independent groups, revealed that {Chaha, Gura, Gumer, Ezha, Mesqan, Muher} form a group. In the same manner, {Endegagn, Inor, Gyeto} form a group. Hence, based on the multidimensional scaling results, the whole Ethiosemitic languages are classified into seven groups. The classification results revealed a surprising position for Harari, which contradicts its usual position within the Ethiosemitic languages as the sister of East Gurage languages (Silt'e, Wolane and Zay)



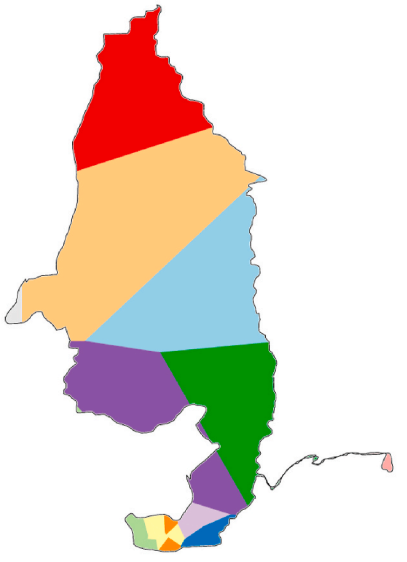
(a) Classification of Ethiosemtic languages based on phonetic distance



(b) Ethiosemitic linguistic map based on phonetic distance



(c) Classification of Ethiosemitic languages based on lexical distance



(d) Ethiosemitic linguistic map based on lexical distance

Fig. 4. Phonetic and lexical classifications of Ethiosemitic languages.

(see Girma, 2001; Hetzron, 1972). Fig. 3(b) shows the map of the first dimension of the multidimensional scaling for the phonetic distance. The light color shows the area with the highest phonetic distance, which is the Tigrigna area.

The multidimensional scaling plot of the lexical distance is illustrated in Fig. 3(c). As the figure depicts, the first dimension is indicated by a solid arrow and the second dimension by a broken arrow. The first dimension shows that Tigrigna has the highest value, and Gumer, Gura, Ezha and Chaha have the lowest values. The values of other languages are somewhere between these two extremes. The second dimension shows that Wolane and Silt'e have the lowest values while Endegagn has the highest value. The figure further illustrates that {Ge'ez, Tigre and Tigrigna} form a group, {Amharic, Argobba} also form a group. Similarly, {Zay, Wolane, Harari, Silt'e} form a group. {Dobbi, Kistane} are also closely related. Unlike the phonetic classification, Harari forms a group with {Zay, Wolane, Silt'e} in the lexical classification. The classifications of the remaining languages might not be clearly seen from Fig. 3(c). Further inspection, after excluding those that clearly form groups, showed that {Endegagn, Inor, Gyeto} form a group. Moreover, {Chaha, Gura, Gumer, Ezha, Mesqan, Muher} form a group. However, Mesqan and Muher do not perfectly fit into their groups, compared to the other group members. In general, the multidimensional scaling results of the lexical distance show that there are six groups of Ethiosemitic languages. Fig. 3(d) shows the map of the first dimension of the multidimensional scaling for the lexical distance. The light color shows the area with the highest linguistic distance which is, once again, the Tigrigna area.

The dendrograms obtained from the phonetic and lexical measures are presented in Fig. 4(a) and (c) respectively. In the dendrograms, varieties that are very similar (dialects) are indicated by the same color (e.g., Inor, Gyeto and Endeagn) while those that are not dialects of one another are indicated by different color (e.g., Argobba and Amharic), though they are in the same group. The dendrograms also show that the classifications obtained from the phonetic and lexical measures are almost the same. Nonetheless, {Harari} is phonetically different from all the Ethiosemitic languages. Including {Harari}, there are seven groups of Ethiosemitic languages, based on the phonetic parameter. The position of Harari reported in the present study contradicts the one previously proposed by historical linguists. For instance, Girma (2001), Hetzron (1972) and Hetzron (1977) classified Harari under Eastern South Ethiosemitic, along with {Wolane, Silt'e and Zay}. The positions of Mesqan and Muher reported here are also slightly different from the ones reported by historical linguists (cf: Grima, 2001; Hetzron, 1972). Fig. 4(b) shows the linguistic map of the Ethiosemiic languages, based on the phonetic parameter. In the map, very similar languages are represented by the same color.

The lexical classification of Ethiosemitic languages is illustrated by Fig. 4(c). The figure shows that {Amharic, Argobba} form a group. Similarly, {Gura, Gumer, Ezha, Chaha, Mesqan, Muher} form a group. Similar to the phonetic classification, {Kistane, Dobbi} form a group. Moreover, {Endegagn, Inor, Gyeto} form a group. Likewise, {Silt'e, Wolane, Harari, Zay} form a group. Contrary to the phonetic classification, {Harari} form a group with the East Gurage languages. This classification perfectly replicates the position of Harari previously proposed by Girma (2001), Hetzron (1972) and Hetzron (1977). Moreover, {Ge'ez, Tigre, Tigrigna} form a group, consistent with the phonetic classification. Fig. 4(c), in general, shows six groups of Ethiosemitic languages. Fig. 4(d) presents the linguistic map of the Ethiosemitic languages, based on the lexical parameter. On the map, different languages are represented by different color; similar languages are illustrated by the same color.

The lexical and the phonetic classifications are in a perfect contrast with regard to the position of Harari. Harari does not form a group with the East Gurage languages: {Silt'e, Wolane, Zay} in the phonetic classification. The extreme difference between Harari and the East Gurage languages is possibly because of the influence of Oromo and other Cushitic languages such as Somali on Harari (see Leslau, 1956; Leslau, 1957; Leslau, 1952; Leslau, 1945). It could also be that Harari has been under the influence of Arabic (see Lesalu, 1956; Lesalu, 1957). The geographical isolation of Harari from the rest of Ethiosemitic languages could be another factor. In addition to Harari, the positions of Muher and Mesqan contradict the proposals of Hetzron (1977) and Hetzron (1972). Fig. 4(a) and (c) further illustrate a strong similarity between Mesqan and Muher, and the Central west Gurage languages. This similarity was previously reported by Fekede (2015), Tekabe (2020a) and Tekabe (2020b). Regardless of the inconsistent positions of Harari, Mesqan and Muher, there is a general tendency of similarity between the current areal classifications and previous genealogical classifications by historical linguists. Full understanding of the factors underlying the inconsistency among these languages merits further investigation.

4.2. Similarity among the classification proposals

This section illustrates the degree of similarity between the classifications presented in §4.1 and previous classifications by historical linguists, based on the *cophonetic* distance between nodes in the classification trees. The cophonetic distance between any two terminal nodes in a classification tree was defined as the number of nodes one has to go up from, let say, language A to the lowest common node shared between the member of the pairs and then down to language B (see Gooskens and Heuven, 2018). For instance, in Fig. 4(c), the cophonetic distance between Tigre and Ge'ez is two: (1) from Tigre one node up to the mother node, (2) from the mother node down to Ge'ez. After determining the cophonetic distance between each node (see Appendix A.2.1), the Pearson's correlation was computed to illustrate the relationship between the areal classifications reported in §4.1 and previous classifications by the historical linguists. It is important to mention here that different number of language varieties were included in the classifications of previous studies. Hence, the number of languages for which the cophonetic distance was computed varies from study to study.

Among several previous classification proposals, the cophonetic distance of Hetzron (1972), Girma (2001), Bender (1971), Hudson (2013) and Kitchen et al. (2009) were compared in Table 2 since they dealt with the classification of the majority of Ethiosemitic languages. The Pearson's correlation coefficient analyses show that the cophonetic distance of all the classifications correlate strongly to each other. The strong correlation between the phonetic and lexical classifications ($r = .850$) shows a high degree of substitutability between the two measures (see Tekabe, 2020a; Tekabe, 2020b; Gooskens et al., 2008; Tang et al., 2009). There is a relatively lower correlation between Girma (2001)'s classification and the lexical classification. Likewise, the classifications of Hetzron (1972) and Hudson (2013) correlate more strongly to the phonetic areal classification as compared to the lexical classification. However, the classifications of Bender (1971) and Kitchen et al. (2009) correlate more strongly to the lexical classification, as compared to the phonetic classification. As will be discussed in the proceeding section, compared to the lexical classification, the phonetic classification is more sensitive to the areal diffusion and contact. Hence, the general tendency of a strong correlation among the classifications in Table 2 seems to imply that previous classification proposals may not be genealogical, as many scholars of Semitic languages may think.

4.3. Determinants of the linguistic distance

As indicated in §1, four determinants, i.e., contact-induced lexical diffusion among Ethiosemitic languages, the influence of Oromo, population size and geographical distance were examined. The degree of the influence of lexical diffusion was estimated using Neighbor-net representation. Neighbor-net algorithm produces classifications of different structures based on the nature and type of the distance data; if the input data are circular (due to a horizontal transmission), it returns collections of circular splits, networks. In other words, it provides net-like

Table 2
Correlations among selected classifications of Ethiosemitic languages.

	Bender	Girma	Hetzron	Hudson	Kitchen	Phonetic	Lexical
Bender (1971)	–	.755	.827	.839	.864	.766	.749
Girma (2015)	–	–	.819	.806	.703	.634	.580
Hetzron (1972)	–	–	–	.937	.807	.655	.611
Hudson (2013)	–	–	–	–	.824	.677	.751
Kitchen (2009)	–	–	–	–	–	.681	.674
Phonetic Class.	–	–	–	–	–	–	.850
Lexical Class.	–	–	–	–	–	–	–

classifications of languages. However, if the input data are non-circular, it returns a binary tree. This tendency of the Neighbor-net representation renders an opportunity to examine if the distance data is tree-like or net-like (Prokić et al., 2013). Hence, the Neighbor-net algorithm can be used to inspect whether the lexical and phonetic classifications presented in §4.1 reflect a genealogical relationship or just an areal similarity.

Fig. 5 displays approximately six groups of Ethiosemitic languages: {Ge'ez, Tigrigna, Tigre}, {Amharic, Argobba}, {Harari, Silt'e, Wolane, Zay}, {Kistane, Dobbi, Muher}, {Inor, Endegagn, Gyeto}, {Mesqan, Gura, Chaha, Ezha and Gumer}. The net-like structure of the Neighbor-net presentation shows the influence of across-language lexical diffusion on the distance among the languages. The across-language lexical diffusion could be attributed to the geographical proximity among the languages. The classical example is the lexical affinity among the Gurage varieties. Except for Harari and the East Gurage languages (Silt'e, Wolane and Zay), the remaining Gurage varieties form a net-like structure, reflecting a strong tendency of lexical contact and diffusion. Given that these varieties are spoken in a small geographical area, their close similarity can be attributed to the geographical proximity.

The influence of Oromo on the Ethiosemitic languages was examined by comparing the phonetic and lexical similarity between Oromo and each of the Ethiosemitic languages, and by performing Multiple Linear Regression analysis on the phonetic and lexical skewing between each pair of Ethiosemitic languages with respect to Oromo. The phonetic similarity between Oromo and the Ethiosemitic languages was obtained by comparing all the 147 words (mass comparison), not based on a comparison just between cognates. As Fig. 6 depicts, almost all Ethiosemitic languages share about 50% phonetic similarity with Oromo. The Figure further illustrates that Harari and the three East Gurage languages (Silt'e, Wolane and Zay) are among the Ethiosemitic languages that are highly influenced by Oromo. This influence must be due to the geographical proximity between Oromo and the four languages (see Tekabe, 2020a and Tekabe, 2020b for similar argument).

Fig. 7 illustrates the percentage of lexical similarity between Oromo and the Ethiosemitic languages. As can be seen from the Figure, Oromo shares about 20% lexical similarity with many of the Ethiosemitic languages. The figure also shows that Argobba, Silt'e, Wolane and Zay are among the Ethiosemitic languages that are highly influenced by Oromo.

The influence of Oromo on Ethiosemtic languages was further analyzed using the lexical and phonetic skewing between each of the Ethiosemitic language with respect to Oromo. As pointed out in §3.4, the analysis was performed using Multiple Linear Regression in R. In this analysis, the influence of the population size and the geographical distance were also included as independent variables, together with the phonetic and lexical skewing. Hence, in total, four independent variables were analyzed which consist of phonetic skewing, lexical skewing, geographical distance and population size. The lexical and phonetic similarities between each pair of Ethiosemitic languages were considered as dependent variables. As there were two dependent variables (phonetic similarity and lexical similarity), the regression analysis was performed twice, once for each dependent variable.

First, the regression model was fit with three independent variables (geographical distance, population size and phonetic skewing) and with

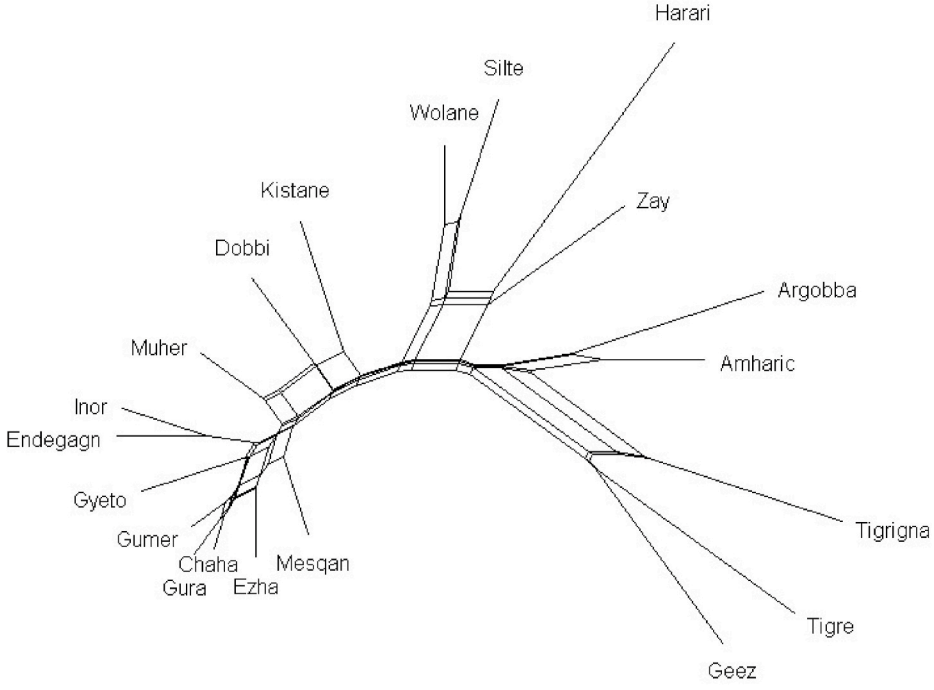


Fig. 5. Neighbor-net representation of Ethiosemitic languages.

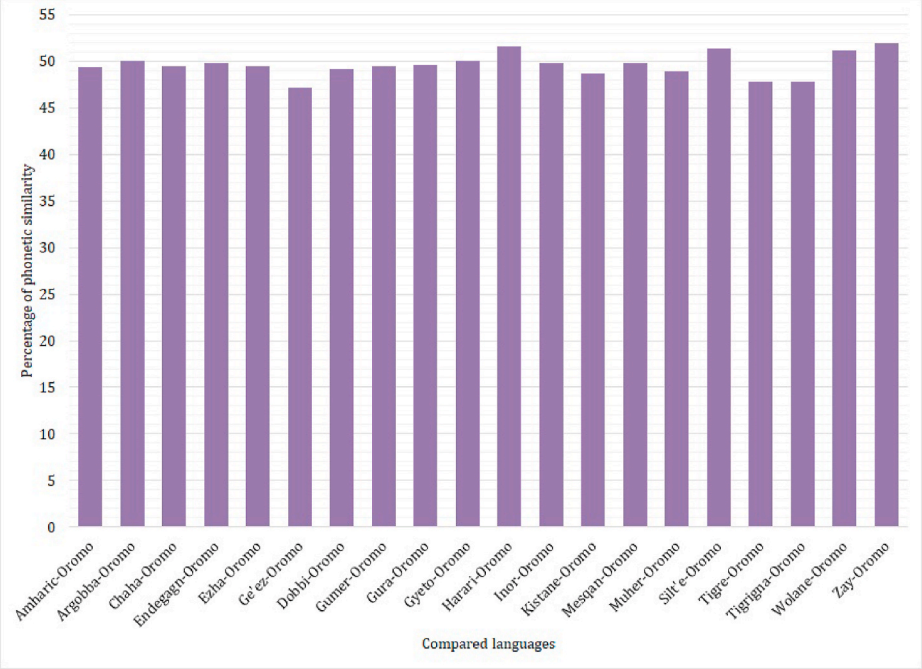


Fig. 6. Phonetic similarity between Oromo and Ethiosemitic languages.

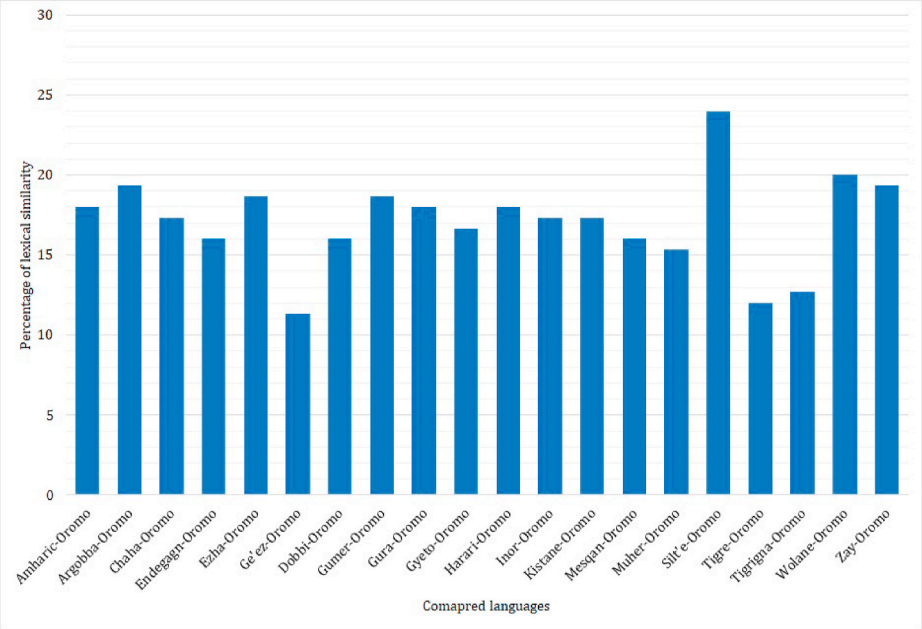


Fig. 7. Lexical similarity between Oromo and Ethiosemitic languages.

one dependent variable (phonetic similarity). The Regression analyses performed on these variables showed a statistically significant effect of geographical distance (β = -.620, t = −10.218, p < .001), and similarly a significant effect of phonetic skewing on the phonetic similarity (β = -.247, t = −4.129, p < .001). However, only a marginal effect of the population size was obtained (β = .592, t = .592, p = .555). The influences of the geographical distance and phonetic skewing on the phonetic similarity among Ethiosemitic languages are further illustrated in Fig. 8. The figure shows that as the geographical distance between the language sites decreases, the phonetic similarity among the languages increases. This result substantiates the prediction of the Gravity model (Trudgill, 1974). Similarly, the figure shows that as the population size decreases, the phonetic similarity among the languages decreases. However, this result contradicts the assumption of the Gravity model, which predicts that population size is directly proportional with linguistic similarity. The figure further illustrates that as the percentage of phonetic skewing decreases, the phonetic similarity among the languages increases, implying a significant influence of Oromo on the phonetic similarity among the languages.

In the second analysis, the Regression model was fit with three independent variables (geographical distance, population size and lexical skewing) and with one dependent variable (lexical similarity). The regression analysis performed on these factors showed a significant effect of geographical distance on the lexical similarity (β = -.555, t =

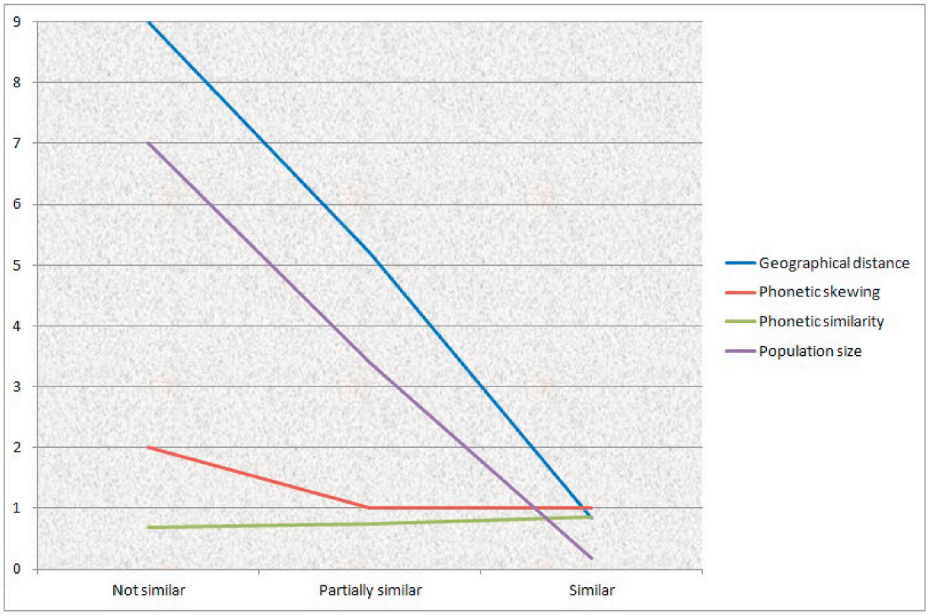


Fig. 8. The influence of geographical distance, population size and phonetic skewing.

−6.16, p < .001), and a marginal impact of the population size (β = .056, t = .683, p = .496). Moreover, the result showed a marginal influence of lexical skewing on the lexical similarity (β = -.063, t = -.704, p = .483). The two regression analyses confirm a significant influence of geographical distance, the influence of diffusion of phonetic features from Oromo to the Ethiosemitic languages. However, the results do not confirm the influence of population size and the impact of lexical borrowing between Oromo and the Ethiosemitic languages.

Fig. 9 shows the relationship among the three independent variables (geographical distance, population size and lexical skewing) and the lexical similarity among the Ethiosemitic languages. The figure reveals that as the geographical distance decreases, the lexical similarity among the Ethiosemitic languages also decreases. This result, once more, confirms the prediction of the Gravity model. The population size also increases as the lexical similarity between the languages decreases. This result contradicts the Gravity model, which assumes a direct relationship between population size and linguistic similarity. Moreover, as the lexical skewing decreases, the lexical similarity among the languages increases. However, this relationship is not statistically significant. Fig. 9 further illustrates that the influences of geographical and lexical skewing are more prominent among languages that are not similar, compared to languages that are similar from lexical point of view.

5. Discussion

The phonetic and lexical classifications of Ethiosemitic languages presented in §4.1 show that {Argobba and Amharic} form a group, but they are not similar to the extent that makes one the dialect of another. This confirms a similar conclusion by Getahun (2018), but contradicts the intelligibility claims by Zelealem and Siebert (2001) and Mohammed et al. (2014). Argobba data analyzed in the present study was largely

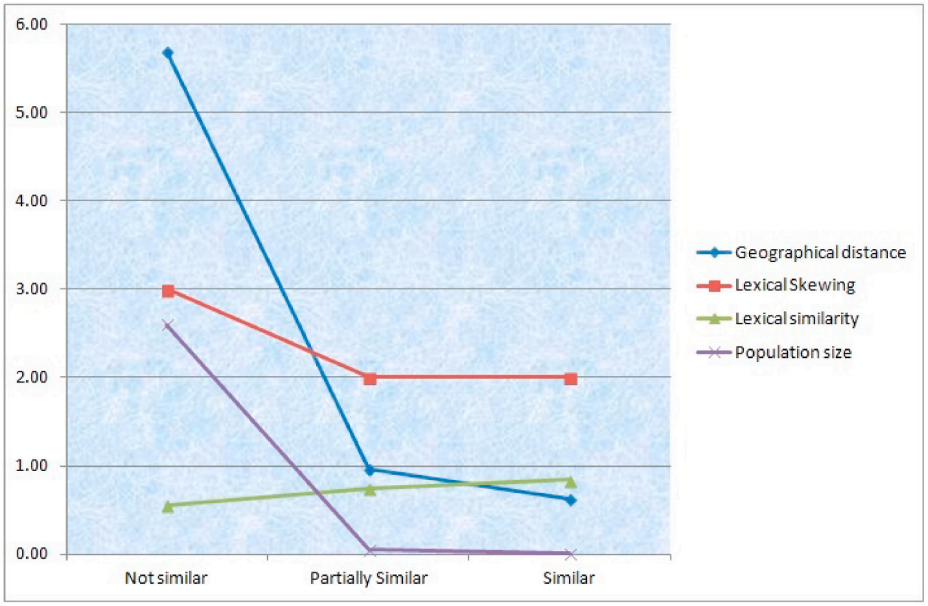


Fig. 9. The influence of geographical distance, population size and lexical skewing.

obtained from Zelealem and Siebert (2001); Mohammed et al (2014) and Leslau (1957) and Leslau (1973), which focus on Argobba of Shewa Robit area. However, the result is similar to Getahun (2018) which was conducted on Argobba of Shonke and T'olaha. Furthermore, both the phonetic and lexical classifications show that {Wolane, Silt'e and Zay} are closely related languages. While Silt'e and Wolane are very similar (see also Meyer, 2014), Zay seems a bit different from the two varieties. This result confirms the strong similarity between Silt'e and Wolane previously reported in several studies (e.g., Girma, 2001; Meyer, 2006; Hetzron, 1972; Hetzron, 1977). Likewise, both the phonetic and lexical classifications indicate that {Chaha, Gura, Gumer, Ezha, Muher and Mesqan} form a group. This classification largely matches the classification of Hetzron (1972). The only difference is on the position of Mesqan which is classified under West Gurage languages in the classification of Hetzron (1972) and the position of Muher which is classified under tt-Group in the classification of Hetzron (1972). The current position of Muher perfectly fits into the position proposed by Girma (2001). Girma (2001) grouped Muher and Central West Gurage (Chaha, Gura, Ezha and Gumer) together. Contrary to Muher, the position of Mesqan reported in §4.1 contradicts the position proposed by Girma (2001). Girma (2001) classified Mesqan under North Gurage, along with Dobbi and Kistane. In both phonetic classifications, {Dobbi and Kistane} form a group, similar to the classification proposals of Hetzron (1972) and Girma (2001).

Furthermore, both the phonetic and lexical classifications confirm that {Endegagn, Gyeto and Inor} form a group. This classification perfectly fits the classification of Hetzron (1972), Hetzron (1977), Tekabe (2020b) and Tekabe (2020b). The present study could not confirm the grouping of Inor with the Central West Gurage, which was previously proposed by Girma (2001). A rather surprising result in the present study is the position of Harari. In the phonetic classification, Harari does not form a group with any of Ethiosemitic languages. None of the previous studies reported similar position of Harari. Hetzron (1972), Hetzron (1977) and Girma (2001) classified Harari under Eastern South Ethiosemitic, together with East Gurage languages (Silt'e, Wolane, and Zay), which is also the case in the lexical classification presented in §4.1. There are three plausible causes of the phonetic deviation of Harari. First, Harari might have been under extreme influence of the neighboring Cushitic languages, mainly of Oromo and Somali. This is a tentative suggestion and full understanding of the contact situation requires further investigation. In §4.3, it was also illustrated that Harari is one of the languages that have been phonetically affected by Oromo. It was further illustrated that phonetic skewing among Ethiosemitic languages, with regard to Oromo, is one of the determinants of the distance among Ethiosemitic languages. Second, Harari has been geographically isolated from the remaining Ethiosemitic languages (see Fig. 1). This isolation might have resulted in Harari-internal dynamics that has not been common among the rest of the Ethiosemitic languages. To add more, Harari has a long history of contact with Arabic (see Leslau, 1945; Leslau, 1956; Leslau, 1957). This might have significantly affected the phonology and the sound system of Harari. Unlike Harari, the phonetic and lexical classifications of the three North Ethiosemitic languages: Tigre, Tigrigna and Ge'ez perfectly match the classifications previously proposed by Girma (2001), Hetzron (1972), Hetzron (1977) and Bender et al. (1972). The only difference in the present classification, with regard to North Ethiosemitic languages, is that Ge'ez is closer to Tigre than to Tigrigna, both in the phonetic and lexical classifications.

In §4.2, it was further illustrated that the phonetic and lexical measures strongly correlate to each other. This implies that the two measures can be used interchangeably in order to determine the linguistic distance among closely related languages (see also Tekabe, 2020a; Tekabe, 2020b; Gooskens et al., 2008; Heeringa, 2004; Tang et al., 2009). Nevertheless, it was also illustrated that phonetic classification is more sensitive to areal diffusion and contact, compared to the lexical classification. This implies that lexical classification is more appropriate, especially, in situation where genealogical classification is preferred over areal classification. Furthermore, the strong correlation between the current phonetic and lexical classifications, and the classifications previously provided by the historical linguists hints that, like morphological and syntactic features, phonetic and lexical parameters can be used as reliable parameters for the classification of related languages. A similar conclusion was previously reported in Tekabe (2020a), Tekabe (2020b), Gooskens and Heuven (2018) and Tang et al. (2009). Nevertheless, such a strong correlation between the areal and genealogical classifications arises a serious theoretical concern. Given that the areal and the genealogical classifications rely on extremely different theoretical assumptions, any significant correlation between the two classification types makes the reliability of the genealogical classification suspicious (see Aikhenvald & Dixon, 2006, François, 2015; Kalyan and François, 2018 for similar argument).

The results presented in §4.3 also show that the distance among Ethiosemitic languages, in general, is the outcome of three major factors: the contact among Ethiosemitic languages, the influence of Oromo, and the geographical distance among the language areas. The results indicate the involvement of two opposing phenomena; a convergence among the Ethiosemitic languages because of the phonetic and lexical diffusion on the one hand (see also Tekabe, 2020a; Tekabe, 2020b) and divergence among the languages as the result of the influence of Oromo on the other hand. It is important to stress that only Oromo lexical items are analyzed in the present study though the influence of Cushitic on Ethiosemitic can also come from other Cushitic languages such as Agaw. The phonetic and lexical features that are shared between Oromo and the Ethiosemitic languages might have also Afroastic origin. This is an area that needs future rigorous investigation. The geographical distance has both the pulling and pushing effect. The phonetic difference between Harari and the rest of the Ethiosemitic languages is an illustration of the effect of the divergence phenomenon. Given that many of the Ethiosemitic languages are spoken in a small geographical area, and that they have a long history of contact, the convergence effect is stronger than the divergence effect, especially in the Gurage area (see Tekabe, 2020a; Tekabe, 2020b; Crass and Meyer, 2009). Many of the Ethiosemitic languages share a great deal of phonetic and lexical similarity with Oromo due to the geographical proximity; however, it is the East Gurge languages (Zay, Wolane and Silt'e) and Harari that are highly influenced by Oromo (see also Leslau, 1957; Leslau, 1945; Tekabe, 2020a). In general terms, with regard to the similarity among Ethiosemitic languages, geographically closer languages are more similar, compared to the languages that are geographically far apart. This means that the similarity/difference among the Ethiosemitic languages is dependent on the geographical proximity. Similar results were previously reported in Tekabe (2020a), Heeringa (2004), Heeringa et al. (2011), Nerbonne et al. (2005) and Gooskens (2005). Results in §4.3 indicate a marginal role of population size. As discussed in Tekabe (2020a), this could be because of several factors, for example, a conservative social tradition of the Ethiosemitic speakers, restricted contact among the speakers, sparsely distributed population and hostility among Ethiosemitic speakers, due to the present and past political antagonism and conflicts. The absence of the influence of population size was also previously reported in Nerbonne et al. (2005).

As pointed out in the introduction, the present study aimed at providing the areal classification of Ethiosemitic languages, without any effort to control borrowings and contact-induced features. The reported influence of Oromo, geographical distance and contact among Ethiosemitic languages also substantiate that the classifications reported in §4.1 are not genealogical. However, in §4.2, a strong correlation was reported between the areal classifications and previous genealogical classifications. Given that the genealogical classification focuses on the vertical diffusion of features while the areal classification focuses on the horizontal diffusion, the two classifications should not have been similar to a great extent (see Aikhenvald & Dixon, 2006; François, 2015; Geisler and List, 2013; Kalyan and François, 2018). This correlation hints that previous classifications of Ethiosemitic languages may not be 'genealogical'

as some scholars of Ethiosemitic languages may think. Contrary to the scholars' assumption, previous classifications might not be based on shared innovations; it is very likely that the scholars erroneously used contact-induced features for the genealogical classification of Ethiosemitic languages (see Aikhenvald & Dixon, 2006; François, 2015; Kalyan et al., 2018). More importantly, given the Ethiopian linguistic area that has been under extensive language contact, contact-free binary division of languages is logically almost impossible, contrary to the assumption of the historical comparative methods.

Finally, regardless of the controversies that always exist in the attempt to make a distinction between dialects and languages, based on the classifications presented in §4.1, one may conclude that some Ethiosemitic languages can function as variants of some common languages. For instance, Chaha, Ezha, Gura, Gumer, Mesqan and Muher can be considered dialects of one another. The intelligibility testes conducted on these varieties also show the same result (see Tekabe, 2020b). Similarly, Inor, Endegegn and Gyeto seem dialects of one another (see Tekabe, 2020a). Silt'e and Wolane are also dialects of one another (see Fekede, 2015). Kistane and Dobbi can also be considered dialects of one another (Meyer, 2005). The remaining languages seem to be independent languages. These conclusions are largely based on the phonetic and lexical similarity among the languages. The author is cognizant of the complexity of the matter, which is often not even linguistic, but largely political, psychological and social issues. Determining whether these varieties are dialects or not may constitute a significant contribution to the attempts that have been made to standardize some of the languages. For example, in the standardization process, Amharic and Argobba can be treated as separate languages (see also Getahun, 2018; Girma, 2015). This conclusion is against the recommendation of Bender et al. (1972), Girma (2001), Mohammed et al. (2014) and Zelalem (2004) which argued that the two languages are dialects of one another. Those that have involved in the standardization of the Gurage varieties (see Meyer, 2018; Fekede, 2015) may also benefit from the four group of Gurage varieties proposed in the present study: {Chaha, Gura, Gumer Ezha, Muher, Mesqan}, {Dobbi, Kistane}, {Inor, Endegan, Gyeto} and {Silt'e, Wolane, Zay}; given that the political, ethnic and social issues associated with the standardization of the Gurage varieties are properly solved (see Meyer, 2018).

Declaration of competing interest

The article has been written solely by the corresponding author, and I have no conflicts of interest to disclose.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.amper.2021.100074.

References

Aikhenvald, A., Dixon, R., 2006. Introduction. In: Aikhenvald, A., Dixon, R. (Eds.), Areal Diffusion and Genetic Inheritance: Problems in Comparative Linguistics. Oxford University Press, Oxford, pp. 1–23.

Appleyard, D., 2011. Semitic-Cushitic/Omotic relations. In: Stefan, W. (Ed.), The Semitic Languages: an International Handbook. De Gruyter Mounton, Berlin, pp. 38–53.

Bender, M., 1966. Notes on lexical correlations in some Ethiopian languages. J. Ethiop. Stud. 4 (1), 5–16.

Bender, M., 1971. The languages of Ethiopia: a new lexicostatistic classification and some problems of diffusion. Anthropol. Ling. 13 (5), 165–288.

Bender, M., Cooper, R.L., Ferguson, C., 1972. Language in Ethiopia: implications of a survey for sociolinguistic theory and method. Lang. Soc. 1 (2), 215–233.

Cohen, D., 1961. Le vocabulaire de base sémitique et le classement des dialectes du sud. Semitica 11, 55–84.

Crass, J., Meyer, R., 2009. Introduction. In: Crass, J., Meyer, R. (Eds.), Language Contact and Language Change in Ethiopia. Deutsche Nationalbibliothe, Cologne, pp. 1–16.

Crass, J., Meyer, R., 2011. Ethiosemitic-Cushitic language contact. In: Stefan, W. (Ed.), The Semitic Languages: an International Handbook. De Gruyter Mounton, pp. 1266–1275.

Ehret, C., 1988. Social transformations in the early history of the Horn of Africa: linguistic clues to developments of the period 500 BC to AD 500. In: Proceedings of Eighth International Conference of Ethiopian Studies. Institute of Ethiopian Studies, Addis Ababa, Ethiopia, pp. 639–651.

Elias, D., 2014. The Tigre Language of Ginda, Eritrea: Short Grammar and Texts. Brill.

ENSA, 2007. *Ethiopian Statistical Report of the 2007 Population and Housing Census: Population Size by Age and Sex, South Nation Nationalities*. Addis Ababa: Federal Democratic Republic of Ethiopia. Population Census Commission.

Faber, A., 1997. Genetic subgrouping of the Semitic languages. Semit. Lang. 15, 3–15.

Fekede, M., 2015. Intergroup Communication Among Gurage: A Study in Intelligibility, Inter-lingual Comprehension and Accommodation. Lambert Academic Publishing, Saarbucken.

Ferguson, C., 1976. The Ethiopian language area. Language in Ethiopia. In: Bender, M., Bown, J., Cooper, R., Ferguson, C. (Eds.), Language in Ethiopia. Oxford University Press, Oxford, pp. 63–76.

Fleming, H., 1968. Ethiopic language history: testing linguistic hypotheses in an archaeological and documentary context. Ethnohistory 15 (4), 353–388.

François, A., 2015. Trees, waves and linkages: models of language diversification. In: Charless, B., Brown, E. (Eds.), The Routledge Handbook of Historical Linguistics. Routledge, London, pp. 161–189.

Geisler, H., List, J.-M., 2013. Do languages grow on trees? The tree metaphor in the history of linguistics. In: Fangerau, H., Geisler, H., Halling, T., Martin, W. (Eds.), Classification and Evolution in Biology, Linguistics and the History of Science. Ranz Steiner Verlag, Stuttgart, pp. 111–124.

Getahun, A., 2009. Argobba Verb Morphology and Syntax: A Documentation with Head-Driven Phrase Structure Grammaras Analytic Framework.

Getahun, A., 2018. The structure of Argobba nominal phrase. J. Afr. Lang. Ling. 39 (2), 127–151.

Girma, A., 2001. The Ethio-Semitic languages (Re-examining the classification). J. Ethiop. Stud. 34 (2), 57–93.

Girma, A., 2015. Argobba Speech Varieties: Comparison. Red Sea Press, Trenton.

Goebl, H., 2010. Dialectometry and quantitative mapping. In: Lameli, A., Kehrein, R., Rabanus, S. (Eds.), Language and Space. An International Handbook of Linguistic Variation, vol. 2. De Gruyter Mouton, Berlin, New York, pp. 433–457.

Goldenberg, G., 1977. The Semitic languages of Ethiopia and their classification. Bull. Sch. Orient Afr. Stud. 40 (3), 461–507.

Gooskens, C., 2005. Traveling time as a predictor of linguistic distance. Dialectol. Geolinguistica (DiG) 13, 38–62.

Gooskens, C., Heeringa, W., 2004. Perceptive evaluation of Levenshtein dialect distance measurements using Norwegian dialect data. Lang. Var. Change 16 (3), 189–207.

Gooskens, C., Heeringa, W., Beijering, K., 2008. Phonetic and lexical predictors of intelligibility. Int. J. Humanit. Arts Comput. 2 (1–2), 63–81.

Gooskens, C., Heuven, V., 2018. How well can intelligibility of closely related languages in Europe be predicted by linguistic and non-linguistic variables?. In: Linguistic Approaches to Bilingualism, 4, pp. 1–29, 8.

Gragg, G., 1997. Old south Arabian phonology. In: Keye, A. (Ed.), Phonology of Asia and Africa, vol. 1. Winona Lake, Indiana, pp. 163–165.

Heeringa, W., 2004. *Measuring Dialect Pronunciation Differences Using Levenshtein Distance* (Unpublished Doctoral Dissertation). University of Groningen.

Heeringa, W., Nerbonne, J., van Bezooijen, R., Spruit, M.R., 2011. Geografie en inwoneraantallen als verklarende factoren voor variatie in het Nederlandse dialectgebied. Tijdschr. Ned. Taal Lett. 123 (1), 70–82.

Heine, B., 1974. Historical linguistics and lexicostatistics in Africa. J. Afr. Ling. 3 (11), 7–20.

Hetzron, R., 1972. Ethiopian Semitic: Studies in Classification, No. 2. Manchester University Press.

Hetzron, R., 1977. The Gunnän-Gurage Languages, vol. 12. Istituto Universitario Orientale, Naples.

Hetzron, R., Bender, M., 1976. The Ethiosemitic languages. In: Bender, L. (Ed.), Languages in Ethiopia. Oxford University Press, London, pp. 2–29.

Hudson, G., 2000. Ethiopian Semitic archaic heterogeneity. In: 14th International Conference of Ethiopian Studies. Addis Ababa. Addis Ababa University.

Hudson, G., 2013. Northeast African Semitic: Lexical Comparisons and Analysis. Harrassowitz, Wiesbaden.

Huson, D.H., Bryant, D., 2010. User Manual for SplitsTree4 V4. 13.1 (Unpublished).

Kalyan, S., François, A., 2018a. When the waves meet the trees: a response to Jacques and List. J. Hist.Ling. 9 (1), 167–179.

Kalyan, S., François, A., et al., 2018b. Freeing the comparative method from the tree model: a framework for historical glottometry. Senri Ethnol. Stud. 98, 59–89.

Kessler, B., 1995. Computational dialectology in Irish Gaelic. In: Proceedings of the Seventh Conference on European Chapter of Association for Computational Linguistics. Morgan Kaufmann, Burlington, pp. 60–66.

Kitchen, A., Ehret, C., Assefa, S., Mulligan, C.J., 2009. Bayesian phylogenetic analysis of Semitic languages identifies an early bronze age origin of Semitic in the near east. Proc. R. Soc. Lond. B Biol. Sci. 276, 2703—2710.

Leinonen, T., Çöltekin, Ç., Nerbonne, J., 2016. Using Gabmap. Lingua 178, 71–83.

Leslau, W., 1945a. Grammatical sketches in Tigré (north Ethiopic): dialect of Mensa. J. Am. Orient. Soc. 27, 164–203.

Leslau, W., 1945b. The influence of cushitic on the Semitic languages of Ethiopia, a problem of Substratum. Word 1 (1), 59–82.

Leslau, W., 1945c. The verb in Tigré (North-Ethiopic): dialect of Mensa. J. Am. Orient. Soc. 1–26.

Leslau, W., 1948. Supplementary observations on Tigré grammar. J. Am. Orient. Soc. 11, 127–139.

Leslau, W., 1952. The influence of Sidamo on the Ethiopic languages of Gurage. Language 28 (1), 63–81.

Leslau, W., 1956. The Arabic loanwords in Gurage (southern Ethiopia). Arabica 3 (3), 266–284.
Leslau, W., 1957. Arabic loanwords in Argobba (south Ethiopic). J. Am. Orient. Soc. 77 (1), 36–39.
Leslau, W., 1963. Etymological Dictionary of Harari. University of California, Berkeley.
Leslau, W., 1969. Toward a classification of the Gurage dialects. J. Semit. Stud. 14 (1), 96–109.
Leslau, W., 1973. Argobba vocabulary. Rassegna Studi Etiopici 26, 21–43.
Leslau, W., 1979. Etymological Dictionary of Gurage (Ethiopic): English-Gurage Index, vol. 2. Harrassowitz, Wiesbaden.
Leslau, W., 1982. North Ethiopic and Amharic cognates in Tigre. In: Annali. Istituto Orientale di Napoli Roma, 42, pp. 1–86, 2.
Leslau, W., 1987. Comparative Dictionary of Geez (Classical Ethiopic): Geez-English, English-Geez, with an Index of the Semitic Roots. Otto Harrassowitz Verlag.
Leslau, W., 1992. Gurage Studies: Collected Articles. Otto Harrassowitz, Wiesbaden.
Leslau, W., 1996. Chaha and Ennemor: an analysis of two Gurage dialects. In: Hudson, G. (Ed.), Essays on Gurage Language and Culture. Harrassowitz, Wiesbaden, pp. 111–122.
Leslau, W., 1997. Ethiopic Documents: Argobba Grammar and Dictionary, vol. 47. Otto Harrassowitz Verlag.
Leslau, W., 2010. Concise Dictionary of Ge'ez. Otto Harrassowitz Verlag.
Meyer, R., 2005a. Gogot. In: Siegbert, W. (Ed.), Encyclopaedia Aethiopia. Wiesbanden, Harrassowitz, pp. 829–830.
Meyer, R., 2005b. Riddles as indicator of cultural and linguistic convergence in the Gurage region. Scrinium 1 (1), 174–196.
Meyer, R., 2006. Wolane: Descriptive Grammar of an East Gurage Language (Ethiosemitic). Köppe Verlag, Köln.
Meyer, R., 2011. Gurage. In: Weninger, S. (Ed.), The Semitic Languages: an International Handbook. De Gruyter Mouton, Berlin, New York, pp. 1220–1257.
Meyer, R., 2014. Zay. In: Alessandro, B. (Ed.), Encyclopaedia Aethiopia. Wiesbanden, Harrassowitz, pp. 160–162.
Meyer, R., 2018. Language standardization efforts in Gurage. In: Conference Paper Presented on Ethiopian Studies (October, 1–5). Mekelle University, Mekelle.
Minett, J., Wang, W., 2003. On detecting borrowing: distance-based and character-based approaches. Diachronica 20 (2), 1–80.
Mohammed, H., Jordan, L., Boone, R., Netzley, J., 2014. Sociolinguistoc Survey of Argoba. SIL International.
Nerbonne, J., Colen, R., Gooskens, C., Kleiweg, P., Leinonen, T., 2011. Gabmap-a web application for dialectology. Dialectologia: Revista electrònica 2 (2), 65–89.
Nerbonne, J., Heeringa, W., 2001. Computational comparison and classification of dialects. Dialectol. Geolinguistica (DiG) 9, 69–83.
Nerbonne, J., Heeringa, W., Kleiweg, P., 1999. Edit distance and dialect proximity. In: Sankoff, D., Kruskal, P. (Eds.), Time Warps, String Edits and Macromolecules: the Theory and Practice of Sequence Comparison, V–Xv). CSLL, Stanford.
Nerbonne, J., Van Gemert, I., Heeringa, W., 2005. A Dialectometric View of Linguistic "Gravity". University of Groningen, Groningen.
Palmer, F., 1958. Comparative statement and Ethiopian Semitic. Trans. Philol. Soc. 57 (1), 119–143.
Prokić, J., Nerbonne, J., 2008. Recognizing groups among dialects. Int. J. Humanit. Arts Comput. 2 (1–2), 153–172.
Prokić, J., Nerbonne, J., et al., 2013. Analyzing dialects biologically. In: Heiner, F., Hans, G., Thoresten, H., William, M. (Eds.), Classification and Evolution in Biology, Linguistics and the History of Science. Franz Steiner, Suttgart, pp. 147–159.
Raz, S., 1983. Tigre Grammar and Text. Undena Publications, Malibu.
Schleicher, A., 1853. The First Divisions of the Indo-European Indigenous People [English Translation]. General Monthly Bulletin for Science and Literature, pp. 786–787 [English Translation].
Snoek, C., 2014. Review of Gabmap: doing dialect analysis on the web. In: Language Documentation & Conservation, 8, pp. 192–208.
Swadesh, M., 1955. Towards greater accuracy in lexicostatistic dating. Int. J. Am. Ling. 21 (2), 121–137.
Tang, C., van Heuven, V., et al., 2009. Mutual intelligibility of Chinese dialects experimentally tested. Lingua 119 (5), 709–732.
Tekabe, L., 2020a. A Combined Approach Towards Mesauring Linguistic Distance: A Study on South Ethiosemitic Lnaguages. Unpublished doctoral dissertation. University of Verona.
Tekabe, L., 2020b. Mapping the dimensions of linguistic distance: a study on South Ethiosemitic languages. Lingua 243, 102893.
Trudgill, P., 1974. Linguistic change and diffusion: description and explanation in sociolinguistic dialect geography. Lang. Soc. 3 (2), 215–246.
Voigt, R., 2009. North vs. South Ethiopian Semitic. In: Addis Ababa. Proceedings of the 16th International Conference of Ethiopian Studies, vol. 4. Addis Ababa University, pp. 1375–1387.
Wang, W., Minett, J., 2005. Vertical and horizontal transmission in language evolution. Trans. Philol. Soc. 103 (2), 121–146.
Wetter, A., 2006. The Argobba of Tollaha–a comparative overview. In: Proceedings of the Xvth International Conference of Ethiopian Studies, Hamburg July (20-25). Harrassowitz, Weiesbaden, pp. 899–907.
Wetter, A., 2010. DasArgobba: Eine deskriptive Grammatik der Varietät von Schonke und T'ollaha (Zentraläthiopien). Rüdiger Köppe Verlag.
Zaborski, A., 1991. Ethiopian language subareas. In: Stanisław, P., Eugeniusz, R. (Eds.), Unwritten Testimonies of the African Past. Wydawnictwa Uniwersytetu Warszawskiego, Warszaw, pp. 123–134.
Zelalem, L., 2004. The Amharic dialects revisited. *From beyond the Mediterranean*. In: Akten des, 7, pp. 449–480.
Zelealem, L., Siebert, R., 2001. Sociolinguistic survey report of the Argobba language of Ethiopia. Addis Ababa: SIL Int.