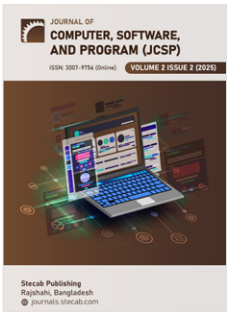

Journal of Computer, Software, and Program (JCSP)

ISSN: 3007-9756 (Online)

Volume 2 Issue 2, (2025)

https://doi.org/10.69739/jcsp.v2i2.1141

https://journals.stecab.com/jcsp

Published by
Stecab Publishing

Research Article

Sautex: A Language-Specific Phonetic Matching Algorithm for Resolving Spelling Variations in Hausa Personal Names

*[1]Bernard Ephraim, [2]Ajah A. Ifeyinwa

About Article

Article History

Submission: October 02, 2025
Acceptance : November 04, 2025
Publication : November 16, 2025

Keywords

African Language Technology, Hausa Natural Language Processing, Low-Resource Languages, Name Matching, Phonetic Encoding, Record Linkage, Soundex Algorithm

About Author

[1] Department of Computing Sciences, Admiralty University of Nigeria, Ibusa, Delta State, Nigeria

[2] Department of Computer Science, Ebonyi State University Abakaliki, Ebonyi State, Nigeria

Contact @ Bernard Ephraim
ephraim1989ben@gmail.com

ABSTRACT

Spelling variations in personal names pose significant challenges for information retrieval and record linkage, particularly in low-resource languages such as Hausa. This paper presents a phonetic encoding algorithm, *Sautex*, specifically adapted to the phonological structure of Hausa, derived from the English Soundex system. *Sautex* was evaluated using a dataset of 17,591 Hausa name spelling attempts with edit distances ranging from 0 to 4. The system achieves a phonetic match accuracy of 77.20% and 66.86% recall for the positive class for the H* variant and 83.02% and 75.32% correspondingly for the H** variant, outperforming the baseline English Soundex by up to 11.53 and 16.76 percentage points in accuracy and recall for the positive class. These results demonstrate the viability of phonology-aware, language-specific encoding systems for African languages. Further studies might be undertaken to evaluate the performance of this algorithm on English names and its generalisation for other Nigerian names. The research aligns with two United Nations Sustainable Development Goals (SDGs), notably SDG 10 (Reduced Inequalities) by ensuring equitable digital representation of Hausa names, and SDG 9 (Industry, Innovation, and Infrastructure) by advancing localized NLP innovations.

Note:

i. *Sautex* is the contraction of the word Sauti, which means Sound in Hausa, and Soundex.

ii. H* indicates values for the *Sautex* code with the first character included

iii. H** indicates values for the *Sautex* code with the first character excluded

iv. English is abbreviated as Eng.

Citation Style:

Bernard, E., & Ajah, A. I. (2025). Sautex: A Language-Specific Phonetic Matching Algorithm for Resolving Spelling Variations in Hausa Personal Names. *Journal of Computer, Software, and Program, 2*(2), 25-33. https://doi.org/10.69739/jcsp.v2i2.1141

1. INTRODUCTION

Sounds are a means of conveying meaning, thus establishing successful communication between two or more parties. Distinct sounds, also called phonemes, which are the units of sound (Haruna, 2023), are blended to form spoken words used in day-to-day interactions. Each language has its particular set of sounds, which may share some similarities with the sounds in other languages. These sounds are grouped into two major categories based on their voicing. On the one hand, there are the vowel sounds, which are produced without obstructions to the flow of air as it passes through the larynx to the lips (Ambalegin, 2021), and on the other, the consonant sounds, which are produced by the partial or total obstruction of the airstream (Haruna, 2023).

Despite the distinctiveness of phonemes, two words might sound the same but vary slightly in their spellings and meanings. These kinds of words, which have the same pronunciation but different spellings and meanings, are called homophones (Simon, 2023; Bhatti *et al.*, 2014). When dealing with information search and retrieval, the definition of homophones might exclude the condition of having different meanings. For instance, these two strings, "*retrieval*" and "*retreival*", might have the same "*pronunciation*" and "meaning", but one is spelt correctly and the other wrongly (Manning *et al.*, 2008). Hence, it is expedient to have a mechanism for matching words that sound alike but are spelt slightly differently. This is what Soundex is known for. Soundex systems assume that it is most likely that the first letter in a string and the consonants in the string are correct, and the error (difference) resides more or less in the associated vowels. Thus, it is possible to generate a "*phonetic hash*" that retains the first letter and the characteristics of the consonants in a string that matches the "phonetic hash" for misspellings of the phonetic type. The Soundex algorithm solves this problem by assigning a digit to a group of similar-sounding letters (Bhatti *et al.*, 2014). The Soundex algorithm generates a code that contains the first character of the string and a three-digit code which is representative of the phonetic sound (Raghavan & Allan, 2004). Thus, words with similar sounds and little variations in their homophones can be retrieved together (Bhatti *et al.*, 2014).

It is important to note that name misspellings are common in multilingual societies, affecting everything from identity verification to digital search. While phonetic matching algorithms such as Soundex are widely used in English and European languages, their effectiveness diminishes in tonal and agglutinative languages like Hausa. Moreover, to the best of our knowledge, only one research effort, by Oketunji (2024), has been made to adapt the existing Soundex algorithm to the Hausa language, and it is inadequate: it only considered the consonant combination of /sh/ sound, which he replaced with /s/ whereas there are a lot more such consonant combinations to cover.

To address this gap, we designed a Hausa-adaptation of the American Soundex algorithm (American Soundex, n.d.; Soundex System, 2024) which we code-named, *Sautex* - a contraction of the words Sauti, which means sound in Hausa and Soundex. Its accuracy was evaluated using a real-world dataset of name misspellings collected via a large-scale spelling survey.

2. LITERATURE REVIEW

Previous research has explored phonetic algorithms such as Soundex (American Soundex, n.d.; Soundex System, 2024) in which a phonetic approximation (known as phonetic hash) is generated for a name or word. The algorithm assumes that the first letter of string and the consonant sounds therein are correct. Hence, a special attention was given to the consonant sounds. These consonant sounds were grouped into nine categories based on the similarity of their pronunciation. Each category is assigned a number between 1 and 9. All vowels, and the letters y and w are assigned the number 0. A phonetic hash is generated by retaining the first character of that word, replacing each consonant character in the string with the category number, and removing consecutive duplicate occurrences of this number. Afterwards, all vowels are removed. The final set of numbers are appended to the first character and trimmed to four characters. Where the outcome is less than four characters, it is then padded with zeros to make up four characters. The Metaphone, and Double Metaphone algorithms approach this by generating approximations based on a more complex consideration beyond grouping sounds into categories. Their approach rather considered how certain character combinations would sound in various contexts hence, generating a "*primary*" and an optional "*secondary*" code to account for possible ambiguities in the pronunciation of a word (Lawrence, 2000). However, these are optimised for European languages and often fail in non-Western contexts. An attempt to design a Soundex algorithm for the Hausa language was undertaken by Oketunji (2024) to bridge this gap. In his work, Oketunji enhanced the Daitch-Mokotoff Soundex algorithm, which supports Slavic and Yiddish surnames, to support Hausa, Igbo, Yoruba, Hindi, and Urdu. His implementation, however, only considers the consonant combinations of /sh/, where this is replaced with /s/ in the string, which is inadequate for the Hausa language, as it has a myriad of other consonant combinations.

Limited work exists on phonetic modelling for African languages, and even less on phonetic error correction for Hausa names. Hence, this work contributes to this underexplored area by applying linguistic features of Hausa phonology to inform phonetic coding.

2.1. Existing soundex algorithm for english names

According to American Soundex (n.d.) and Soundex System (2024), the following algorithm is used for generating the Soundex code for any given English name.

Step 1: Retain the first letter of the name and drop all other occurrences of a, e, i, o, u, y, h, and w.

Step 2: Replace consonants with digits as follows (after the first letter):

- b, f, p, v ⟶ 1
- c, g, j, k, q, s, x, z ⟶ 2
- d, t ⟶ 3
- l ⟶ 4
- m, n ⟶ 5
- r ⟶ 6

Step 3: If two or more letters with the same number are adjacent in the original name (before Step 1), only retain the first letter; also, two letters with the same number separated by '*h*', '*w*', or

'*y*' are coded as a single number, whereas such letters separated by a vowel are coded twice. This rule also applies to the first letter.

Step 4: If there are a few letters in the word to assign three numbers, append zeros until there are three numbers. If there are four or more numbers, retain only the first three.

This algorithm will generate "*R163*" as the Soundex Code for "*Robert*" and "*Rupert*" and "*R150*" for "*Rubin*". "*Ashcraft*" and "*Ashcroft*" both produce "*A261*". "*Tymczak*" yields "*T522*", not "*T520*" (the chars '*z*' and '*k*' in the name are coded as 2 twice since a vowel lies in between them). "*Pfister*" yields "*P236*", not "*P123*" (the first two letters have the same number and are coded once as '*P*'), and "*Honeyman*" yields "*H555*".

3. METHODOLOGY

3.1. Hausa phonemes and phonetic representations

According to Haruna (2023), the Hausa language has forty-seven (47) phonemes, thirty-four (34) of which are consonants and thirteen (13) vowels. However, other researchers, as Mohammed (2001) found, maintain that there are between thirty-two (32) and thirty-five (35) Hausa consonant phonemes. Since Soundex systems emphasise consonants found in words, attention was given to only Hausa consonant phonemes. Consequently, the following Hausa consonant phonemes, found in Table 1, were derived from works by Mohammed (2001), Malah & Rashid (2015) and Erbasi (n.d.). In all, thirty-seven (37) consonant phonemes were obtained. The phonetic symbols for the Hausa consonant phonemes are not yet standardised. For instance, Mohammed (2001) used the symbol /~/ to represent the phoneme /ɓ/ as in the word "*ɓera*", different from /ɓ/ used by Erbasi (n.d.). In another case, Malah & Rashid (2015) used the symbols /kj/ and /gj/ to represent the phonemes expressed in orthography as "*ky*" and "*gy*" respectively. These are, however, defined as /c/ and /ɟ/ by Erbasi in credible repositories.

Table 1. Hausa phonemes and their proposed *sautex* codes

SN	Phoneme	Usage	Sautex Code	SN	Phoneme	Usage	Sautex Code
1	/b/	bargo	1	20	/dd/	hadda	3
2	/ɓ/ or /~/	ɓera, ɓargo	1	21	/l/	larabci	4
3	/bb/	babba	1	22	/n/	nama	5
4	/ɸ/	fanka	1	23	/ŋ/	naŋ	5
5	/ɸj/	fyaɗe	1	24	/m/	mudu	5
6	/z/	zakara	2	25	/ʏ/ or /ɹ/	rama, larabci	6
7	/dʒ/	ja	2	26	/r/	birni	6
8	/tʃ/ or /č/	can	2	27	/kw/	kwali	7
9	/ʃ/	shanu	2	28	/k'w/ or /ƙw/	ƙwarƙwata	7
10	/s/	sarki	2	29	/gw/	gwarzo	7
11	/s'/	s'aba	2	30	/kj/ or /c/	kyau	8
12	/ts/	tsoho	2	31	/k'j/ or /ƙj/	ƙyale	8
13	/k/	kada	2	32	/gj/ or /ɟ/	gyara	8
14	/k'/ or /ƙ/	Kabili or k'abila	2	33	/w/	waka	9
15	/g/	gari	2	34	/ʔ/	'anjima	0
16	/gg/	gaggawa	2	35	/j/	yaro	#
17	/t/	mutane	3	36	/ʔj/	'ya'ya	#
18	/d/	daga	3	37	/h/	harbi	@
19	/ɗ/	ɗaiɗai	3				

3.2. Sautex design

This Hausa Soundex (*Sautex*) system modifies the traditional Soundex rules to accommodate Hausa phonemes, digraphs, and emphatic consonants. For instance, additional phonemes (such as /gw/, /kw/, /h/, /w/, etc), missing in Oketunji (2024), are supplemented from works by Mohammed (2001), Malah & Rashid (2015) and Erbasi (n.d.). Occurrences of the consonant combination /ph/, /ƙw/, /ƙy/,/k'w/, /k'y/, /ƙ/, and /k'/ were replaced with /f/, /kw/, /ky/, /kw/, /ky/, /k/, and /k/, respectively. As seen in Table 1, the thirty-seven (37) consonant phonemes were grouped into twelve (12) Soundex clusters based on how closely the phonemes sound. A *Sautex* Code (comprising the numbers 0 to 9 and the letters '#' and '@') was given to each cluster. The vowels a, e, i, o, and u, the phoneme /ʔ/ and the characters ', " ", -, and any other special character were assigned the *Sautex* code 0. The phonemes /b/, /ɓ/, /bb/, /ɸ/, and /ɸj/ were assigned the *Sautex* code 1. The phonemes /z/, /dʒ/, /tʃ/, /ʃ/, /s/, /s'/, /ts/, /k/, /ƙ/, /g/, and /gg/ were assigned the code 2. The phonemes /d/, /dd/ and /ɗ/ were assigned the code 3. The phoneme /l/ was the code 4. The code 5 was assigned to

the phonemes /n/, /ŋ/, and /m/. The phonemes /ɹ/ and /r/ were assigned the code 6. The code 7 was assigned to /kw/, /ƙw/, and /gw/. The phonemes /kj/, /ƙj/, and /gj/ were assigned the code 8. The code 9 was assigned to the phoneme /w/. The code '#' was assigned to /j/ and /?j/, and the code '@' was assigned to /h/.

Given the likely confusion that may arise in the way the following consonant combinations can be pronounced, they were transformed as follows: iy → i, dj → j, dg → g, np → p, mp → p, nf → f, mf → f, gb → b, nb → b, mb → b, nd → d, md → d, ng → g, ph → f, p → f, q → k, ch → s, sh → s, nk → s, and ts → s. Also, the Hausa language lacks the x and v characters; hence, they are transformed as x → s and v → f in line with the English Soundex system. The consonant combinations dh and th, which are of Arabic origins, are pronounced as z; hence, they are transformed as dh → z and th → z.

It was observed that the consonants l, h, r and y are usually silent (depending on the accent of the speaker) except when used as either the first character in the word string or placed in between two vowels. It was found that replacing them with their Soundex character code in positions where they are silently pronounced is problematic. For example, Sadi was spelt by a respondent in the survey conducted during this research as Sahdee, yielding the Soundex codes S300 and S@30, respectively, which do not match. Thus, when such characters appear either next to a consonant or are not the first character in the word string, they are assumed to be silent (vowel in the case of y) and given the code 0 so that Sadi and Sahdee can now have the same value S300.

At the end of a word string, certain consonants might appear to be silent; hence, special consideration was given to the following consonants: l, h, y, r, m, and n. When they appear at the end of a word string (which is short enough to consider all its consonants), their sounds tend to be less significant to the listener. They may be altogether ignored, which may lead to a slight but significant variation in the final *Sautex* code. For instance, the names Zara and Faisal were spelt as Zarah and Faysa by two respondents in the survey. These yield the values Z600 for Zara and Z6@0 for Zarah (F240 for Faisal and F200 for Faysa), which are entirely different! To solve this problem, the instance of any of these characters ending a word string is removed so that Zara and Zarah both have the Soundex value Z600 (and Faisal and Faysa the Soundex code F200).

Given that the traditional Soundex algorithm retains the first character before performing any character transformation, potential candidates might be inadvertently excluded from the search result when the misspelt consonant is the first character, even if it is within the same phonetic group as the correct consonant (Celko, 2014). For instance, Jibrịn (J100) was spelt as Kibrin (K100) by one of the respondents. Though the characters J and K belong to the same Soundex group, they do not match here. Hence, *Sautex* considered performing character transformation to a consistent representative character in a group first before retrieving the first character from the output of the transformation operations. Thus, all associated characters (or consonant combinations) in a group are consistently replaced by one character (or consonant combination) in the group. For instance, J and K belong to the ch, sh, ts, c, k, g, j, s, s', z group and are replaced by S. With this, Jibrịn and Kibrin would have S100 as their *Sautex* code.

3.3. Sautex algorithm

The following *Sautex* algorithm is designed to adapt the existing Soundex algorithm for the English language to the Hausa language.

Step 1: Normalise the input string (this converts all language diacritics into normal romanised letters, converts the input string to lowercase, validates that the input string does not contain numbers or emoticons, and removes all instances of the characters ', -, _, ".","" and " " (space) and other special characters in the input string).

Step 2: Carry out the following replacement operation in order from top to bottom:

- dj → j → s
- dg → g → s
- np, mp → p → f
- nf, mf → f
- gb, nb, mb → b → f
- nd, md → d
- ng → g → s
- ph, p, v → f
- x → s
- dh, th → z → s
- q, nk → k → s
- iy → i

Note that the operation ph → f must be carried out first before the operation p → f to avoid having an orphaned h.

Step 3: Replace the following consonant phonemes with their closest homophones in the following order (top to bottom, left to right):

- ch, sh, ts, c, k, g, j, s',z → s
- y' → y
- k'w, ƙw, gw → kw
- k'y, ƙy, gy → ky
- k', ƙ → k → s
- ɓ → b → f
- ɗ, t → d
- ŋ, m → n

Step 4: Retrieve the first character from the output of Step 3 and store it in a variable, say first_char.

Step 5: Inspect the instances of the characters: l, h, r and y in the output of Step 3. In each case, replace the instances of the character with the equivalent *Sautex* character code (l → 4, h → @, r → 6, and y → #) if found between two vowels, else replace with 0.

Step 6: Retrieve all the characters from the output of Step 5 except the first character and store them in another variable, say rest_of_word.

Step 7: Remove the instance of l, h, y, r, m, or n ending the string in rest_of_word.

Step 8: Replace the following phonemes in rest_of_word with their *Sautex* Values. Note: the order of replacement should be top to bottom and left to right of the list:

- fy, f → 1
- ky → 8
- kw → 7
- s → 2

- d ⟶ 3
- n ⟶ 5
- w ⟶ 9

Step 9: Replace the instances of the following vowels a, e, i, o, u (in rest_of_word, the output of Step 8) with the Sautex code 0.
Step 10: If a *Sautex* code (that is, numbers 1-9 and # or @) in the output of Step 9 has consecutive multiples, replace that *Sautex* code with one instance. However, if the repetition of such code is non-consecutive or separated by the *Sautex* code 0, then leave the code unchanged.
Step 11: Remove every instance of the *Sautex* code 0 in rest_of_word, the output from Step 10.
Step 12: If the output string (rest_of_word) from Step 11 is more than three (3) characters in length, then take the first three (3) characters and drop the rest. Otherwise, pad the string to the right with 0s to get a character length of three (3).
Step 13: For H*, change the case of the character stored in first_char (obtained from Step 4) to uppercase and prepend it to the output string rest_of_word, obtained in Step 12. For H**, append the character 0 to rest_of_word.
Table 2 illustrates sample outputs from the proposed variants of the *Sautex* algorithm.

3.4. Hausa personal names collection and preprocessing

To effectively conduct this research, an authoritative dataset of Hausa personal names was needed. This dataset is expected to express the standard orthography of the Hausa language. Many online sources of Hausa personal names exist, but most lack adherence to the standard orthography of the Hausa language. Notwithstanding, a list of 869 authentic native Hausa names and 132 Arabic/Islamic names domesticated by the Hausas was obtained from Adamu & Muhammad-Oumar (2023). Over 150 names native to the Hausa people of Nigeria and Niger were retrieved from Hausa Submitted Names (n.d.), KasarHausa24 (2024), and Apindi & Simwa (2022). Seventeen (17) Christian Hausa names were obtained from Kperogi (2022). Another set of over 55 native Hausa and Islamic Hausa names was retrieved from Adamu (2020). The ninety-nine (99) names of Allah were obtained from 99 Names of Allah (n.d.) and My Islam (n.d.), and the derivatives of these names were obtained by adding the prefix "*Abdul*", which means "*servant of*" to each of the 99 names. More names were identified in a report by Lawson *et al.* (1965), submitted to the CIA, which included over 390 personal names native to the Hausa people of Nigeria, forming the dataset of names used in this work. Names from the sources mentioned were used in a complementary manner. The contributors to this dataset included government linguists and language experts from the academic and religious communities.

3.5. Name spelling survey

A Hausa name spelling survey was conducted over two months (from June 2, 2025, to July 27, 2025). The survey consists of audio files of curated Hausa names recorded by the researcher, who speaks Hausa as a second language. A total of 2,061 names were curated. Of this total, 1,705 unique names were identified, and 356 were alternative orthographic forms of some of the 1,705 unique names. Hence, a total of 1,705 audio files were recorded for this survey.

Table 2. Sample Output for H* and H** Variants of the *Sautex* Algorithm

SN	Name	H*	H**
1	'Yā'ya	Y#00	#000
2	Mūstafā	N231	2310
3	Mustapha	N231	2310
4	Tsōfo	S100	1000
5	Sumā'ila	S540	5400
6	Samā'ilu	S540	5400
7	Sayyjadī	S230	2300
8	Yar bakā	Y120	1200
9	Aisha	A200	2000

The survey aims to collate the spelling attempts of respondents based on the content of the audio file. Respondents were required to play audio files presented randomly to them and then attempt to spell the words. The survey was carried out using a custom data collection tool designed by the researcher and hosted at (Name Spell, n.d.).

3.6. Participant

There were a total of 259 respondents, the majority of whom are students in the researcher's university community, representing various ethnic groups in Nigeria, and predominantly do not speak Hausa as either a first or second language.

3.7. Hausa name spelling dataset

A total of 25,832 spelling attempts were collected from the Hausa name spelling survey. Spelling attempts having more than four (4) edit distances were considered bad samples, hence discarded. Then, a manual inspection for unrelated spelling attempts was carried out on the trimmed dataset, reducing the total to 17,591 spelling attempts. For each case, the intended name, the spelling attempt, the *Sautex* code of the intended name, the *Sautex* code of the spelling attempt, and the edit distance between the intended name and the spelling attempt were considered. The intended names were recorded into audio files (by the researcher, who speaks Hausa as a second language) and played by the respondents, who then listened and made spelling attempts. Spelling errors range from zero errors (edit distance = 0) to minor typos (edit distance = 1) and more complex deformations (edit distance = 4). A copy of this evaluation dataset and the necessary evaluation Python script can be found at (Bernard & Ifeyinwa, 2025).

3.8. Evaluation criteria

A match is considered successful when the *Sautex* code of the spelling attempt matches that of the correct (intended) name. A true positive (TP), represented in the confusion matrix (Figure 1) as Match-Match, implies that the spelling attempt was correct and the *Sautex* codes for the attempt and the intended name matched. A false negative (FN), represented as Match-No Match, implies that the attempt and the intended name matched,

but the *Sautex* codes generated for both did not match. A false positive (FP), represented as No Match-Match, implies that the attempt and intended name did not match, but the generated *Sautex* codes matched. The last group, the true negative (TN), represented by No Match-No Match, implies that the attempt and the intended name, together with the generated *Sautex* codes, did not match.

The following metrics where used in the evaluation of the developed algorithm:

Given that

- TP = True Positives
- TN = True Negatives
- FP = False Positives
- FN = False Negatives

then

i. Overall Accuracy (Confusion Matrix Formulation)Overall Accuracy = (TP + FP)/(TP + FN + FP + TN) ....(1)

ii. The recall for positive class (*Sautex* code matches) within 1 and 4 edit distances:

$$\text{Recall}_{1\text{-}4} = \frac{\text{TP}}{\text{TP} + \text{TN}} \quad \text{....(2)}$$

4. RESULTS AND DISCUSSION

A total of 25,832 spelling attempts were obtained from the survey conducted. Spelling attempts with edit distances above four (4) were considered bad and discarded. Afterwards, a manual inspection was performed to identify unrelated spelling attempts in the dataset. These operations reduced the dataset to 17,591 attempts.

Out of the 17,591 attempts, 5,486 are true positives, that is, they matched the intended names (albeit string normalisation), and the *Sautex* algorithm generated corresponding matching *Sautex* codes. A total of 8,094 attempts are at least 1 edit distance (and at most 4) from the intended name but have matching *Sautex* codes, hence are grouped as false positives. A total of 4,011 attempts did not match the corresponding intended names, nor did their *Sautex* codes, hence grouped as true negatives. No false negatives were recorded. In all, the *Sautex* algorithm successfully matched 13,580 attempts (Figure 1).

Comparatively, the English Soundex Algorithm, as implemented by Turk (2011), was used to generate corresponding Soundex codes for the same set of 17,591 attempts used for the Hausa Soundex algorithm proposed here. Out of the 17,591 attempts, 5,486 are true positives, 7,089 are false positives, 5,016 are true negatives, and there were also no false negatives. In all, the English Soundex algorithm successfully matched 12,575 attempts (Figure 2).

The confusion matrices for the dataset is given in Figure 1, 2, & 3, and the comparative prediction scores are given in Table 3.

Table 3. A comparative prediction scores on hausa names using *sautex* and english soundex algorithms

Metric	Sautex		English
	H*	H**	
True Positive (TP)	5,486	5,486	5,486
False Negative (FN)	0	0	0
False Positive (FP)	8,094	9,118	7,089
True Negative (TN)	4,011	2,987	5,016
Total Soundex Matches (TP+FP)	13,580	14,604	12,575
Total Sample (TP+FN+FP+TN)	17,591	17,591	17,591

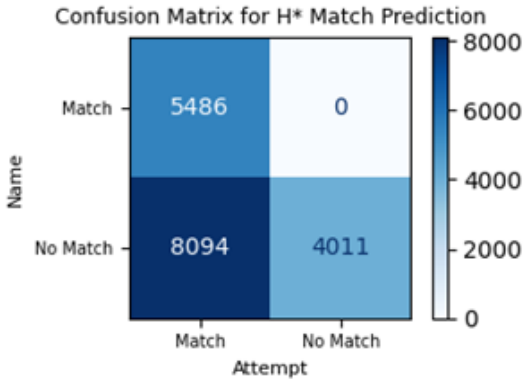

Figure 1. Confusion Matrix for H*

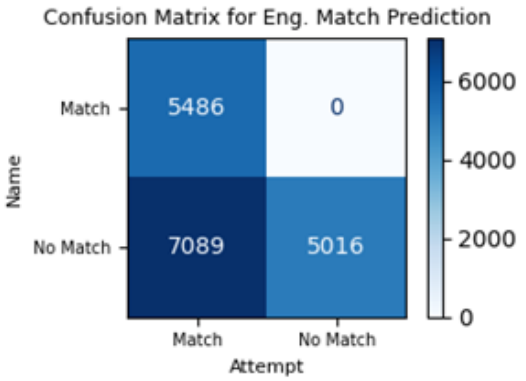

Figure 2. Confusion Matrix for Eng

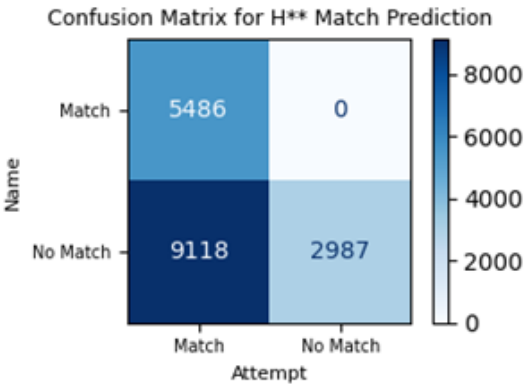

Figure 3. Confusion Matrix for H**

Comparative confusion matrices for predicting hausa names using *sautex* and english soundex algorithms

4.1. Discussion

The data in Table 4 and Figure 4 reveal that the accuracy of the

Sautex algorithm (H*) in predicting a valid match for a spelling attempt reduces with increasing edit distance. When the edit distance was zero (0), the *Sautex* algorithm predicted the intended names with 100% accuracy. This prediction accuracy was reduced to 76.95% for one (1) edit distance, 68.27% for two (2), 58.66% for three (3), and 43.72% for four (4) edit distances. The algorithm achieves an overall prediction accuracy of 77.20% and a 66.86% prediction accuracy on the subset of data having at least one edit distance. These are seen to be better than the results from the English Soundex algorithm, with an accuracy difference of 2.26% at one (1) edit distance, 12.9% at two (2), 12.11% at three (3) and 9.51% at four (4) edit distances. The difference in the overall accuracy of the *Sautex* algorithm (H*) and the English Soundex algorithm (Eng.) is 5.71%. The recall for positive matches within 1 and 4 edit distances were 66.86% for H* and 58.56% for Eng (Table 5).

Table 4. Edit distance vs prediction accuracy for the hausa and the english soundex algorithms

Edit Dist.	Attempts	Soundex Matches			Accuracy [equ. 1]			Recall for Edit Dist. > 0 [equ. 2]		
		Sautex		Eng.	Sautex		Eng.(%)	Sautex		Eng. (%)
		H*	H**		H* (%)	H** (%)		H* (%)	H** (%)	
0	5,486	5,486	5,486	5,486	100.00	100.00	100.00			
1	4,603	3,542	3,902	3,438	76.95	84.77	74.69			
2	3,816	2,605	2,942	2,113	68.27	77.10	55.37			
3	2,245	1,317	1,535	1,045	58.66	68.37	46.55	66.86	75.32	58.56
4	1,441	630	739	493	43.72	51.28	34.21			
Total	17,591	13,580	14,604	12,575	77.20	83.02	71.49			

Upon examining the true negatives in the dataset (generated using the *Sautex* algorithm, H*), it was observed that approximately 1,024 of the 4,011 true negatives could be classified as false positives if the first characters in the generated *Sautex* codes are removed (Table 6). This is so because, traditionally, the English Soundex algorithm assumes that the first letter or sound in a name is always spelt correctly (Celko, 2014). However, the dataset used here shows that this may not always be the case. Due to differences in accent, names could be mispronounced, and this mispronunciation could be at any position in the string.

In light of this, excluding the first character from the generated *Sautex* code for the true positives and false positives does not affect the outcome of a spelling prediction, since the four characters in the code already match. Hence, the *Sautex* algorithm (H*) was modified to H**, which excludes the first character in the already generated *Sautex* code (H*). With this modified *Sautex* algorithm (H**), 1,024 of the above set of true negatives were successfully predicted as false positives, hence increasing the overall system accuracy by 5.82% and by 11.53% and the recall for the positive class to 16.76% when compared to the English Soundex algorithm (Eng.) (Table 5).

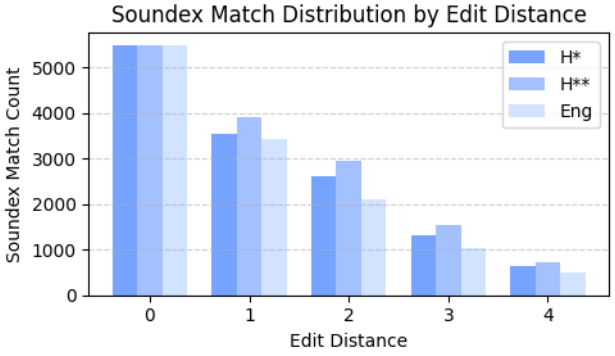

Figure 4. Soundex match distribution by edit distance

Table 5. Comparison of the accuracies of the *sautex* and english soundex algorithms

Edit Dist.	Accuracy (%)			Difference (%)		
	Sautex		Eng.	H**-H*	H*-Eng.	H**-Eng.
	H*	H**				
0	100.00	100.00	100.00	0	0	0
1	76.95	84.77	74.69	7.82	2.26	10.08
2	68.27	77.10	55.37	8.83	12.9	21.73
3	58.66	68.37	46.55	9.71	12.11	21.82
4	43.72	51.28	34.21	7.56	9.51	17.07
Overall	77.20	83.02	71.49	5.82	5.71	11.53

Table 6. Sample true negatives that differ only at the first character of their *sautex* codes

SN	Name	Attempt	Name Soundex (H*)	Attempt Soundex (H*)	Name Soundex (H**)	Attempt Soundex (H**)
1	Matawalle	Otawale	N394	O394	3940	3940
2	Yatsohi	Giatsohi	Y2@0	S2@0	2@00	2@00
3	Hone	Une	H500	U500	5000	5000
4	Dattijo	Hatidjo	D320	H320	3200	3200
5	Iko	Eco	I200	E200	2000	2000

5. CONCLUSION

This research successfully adapted the standard English Soundex algorithm to the Hausa language, leading to the design of a more comprehensive Soundex algorithm code-named *Sautex*. *Sautex* is a contraction of Sauti, which means Sound in Hausa, and Soundex. The *Sautex* algorithm has two variants, H*, which considers the first character in the Soundex code and H**, which does not. Both variants showed up to an 11.53% increase in prediction accuracy and 16.76% recall for the positive class when compared to the English Soundex algorithm, which is a significant improvement in the accuracy of spelling predictions for Hausa names. This was achieved by considering the various consonant combinations that are possible in the Hausa language and their positional characteristics.

Areas to explore further are evaluating the performance of this algorithm on English names and its generalisation for other Nigerian names.

REFERENCES

99 Names of Allah. (n.d.). *99 Names of Allah*. Retrieved May 28, 2025, from https://99namesofallah.name/

Adamu, A. U. (2020). Suna Linzami: Hausa Names as Ethnographic Identifiers. *Algaita: Journal of Current Research in Hausa Studies, Bayero University, 13*(1), 1-15. https://auadamu.com/phocadownload/Encyclopedia/2020_Suna_Linzami_Hausa_Names_as_Ethnographic_Identifiers.pdf

Adamu, A. U., & Muhammad-Oumar, A. A. (2023). 1001 Traditional Hausa Names. 10.13140/RG.2.2.25251.73768

Ambalegin, A. (2021). Phonological Analysis of English Vowel Pronunciation. *Annual International Conference on Language and Literature (AICLL)*, 28-45. KnE Social Sciences. https://doi.org/10.18502/kss.v5i4.8665

American Soundex. (n.d.). Wikipedia. Retrieved April 10, 2025, from https://en.wikipedia.org/wiki/Soundex

Apindi, C., & Simwa, A. (2022, August 22). *150 Hausa names and their meanings: List for boys and girls*. Legit.ng. Retrieved May 26, 2025, from https://www.legit.ng/1117187-hausa-names-meanings.html

Bernard, E., & Ifeyinwa, A. (2025). Hausa Soundex dataset with evaluation script for reproducibility. Zenodo. https://doi.org/10.5281/zenodo.16812500

Bhatti, Z., Waqas, A., Ismaili, I. A., Hakro, D. N., & Soomro, W. J. (2014). Phonetic based SoundEx & ShapeEx algorithm for Sindhi Spell Checker System. *AENSI-AEB, 8*(4), 1147-1155. Retrieved 3 9, 2025, from https://arxiv.org/pdf/1405.3033

Celko, J. (2014). *Joe Celko's SQL for Smarties: Advanced SQL Programming* (5th ed.). Elsevier Science. https://doi.org/10.1016/C2013-0-18881-2

Erbasi, B. (n.d.). *Sounds of Hausa. General Phonetics Final Project*. Retrieved 4 8, 2025, from https://sail.usc.edu/~lgoldste/Ling415/Final_Project/LanguageX/languageX.pdf

Haruna, S. (2023). A Phonological Study of Consonants and Vowels Phonemic Merger in Hausa. *British Journal of Multidisciplinary and Advanced Studies, 4*(3), 45–59. https://doi.org/10.37745/bjmas.2022.0196

Hausa Submitted Names. (n.d.). *Behind the Name*. Retrieved April 29, 2025, from https://www.behindthename.com/submit/names/usage/hausa

KasarHausa24. (2024, December 6). *See Real Hausa Native Names and their Meanings*. KasarHausa24.com. Retrieved May 26, 2025, from https://www.kasarhausa24.com/see-real-hausa-native-names-and-their-meanings/

Kperogi, F. A. (2022). *Wonders of Northern Christian Names. Facebook: Esan People's Blog*. Retrieved 5 23, 2025, from https://www.facebook.com/Esanpeopleblog/posts/wonders-of-northern-christian-namesby-farooq-a-kperogihausa-speaking-christians-/360221082775201/

Lawrence, P. (2000). The double metaphone search algorithm. *C/C++ Users Journal*, 38-43.

Lawson, E. D., Sheil, R. F., & Rogers, P. A. (1965). The Onomastic Treasure of the CIA. Central Intelligence Agency. 10.13140/2.1.1735.0722

Malah, Z., & Rashid, S. M. (2015, 6). Contrastive Analysis of the Segmental Phonemes of English and Hausa Languages. *International Journal of Languages, Literature and Linguistics, 1*(2), 106-112. https://doi.org/10.7763/IJLLL.2015.V1.21

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval* (P. Raghavan & H. Schütze, Eds.). Cambridge University Press. https://www-nlp.stanford.edu/IR-book/

Mohammed, U. A. (2001). *Aspects Of Segmental Phonology Of Hausa*. Retrieved from https://www.researchgate.net/

publication/369481573_ASPECTS_OF_SEGMENTAL_PHONOLOGY_OF_HAUSA

My Islam. (n.d.). *Learn The 99 Names of Allah* (With Meaning and Benefits). My Islam. Retrieved May 28, 2025, from https://myislam.org/99-names-of-allah/

Name Spell. (n.d.). Retrieved from https://name-spell.thrinkle.com

Oketunji, A. F. (2024, 4). *Enhancing and Applying Daitch-Mokotoff Soundex Algorithm on Ethnic Names*. https://doi.org/10.5281/zenodo.11009946

Raghavan, H., & Allan, J. (2004). *Using Soundex Codes for Indexing Names in ASR documents*. ACL Anthology. Retrieved from https://aclanthology.org/W04-2905.pdf

Simon, C. (2023). A critique on English homophones and homographs. *African Journal of Social Issues, 6*(1), 106-113. 10.4314/ajosi.v6i1.7

Soundex System. (2024, January 9). *National Archives*. Retrieved on April 10, 2025, from https://www.archives.gov/research/census/soundex

Turk, J. (2011). *Jellyfish*. Retrieved 8 6, 2025, from https://jamesturk.github.io/jellyfish/