OPEN ACCESS
Saudi Journal of Engineering and Technology
Abbreviated Key Title: Saudi J Eng Technol
ISSN 2415-6272 (Print) |ISSN 2415-6264 (Online)
Scholars Middle East Publishers, Dubai, United Arab Emirates
Journal homepage: http://scholarsmepub.com/sjet/

Original Research Article

Hausa WordNet: An Electronic Lexical Resource

Amina Imam Abubakar[1*], A. Roko[1], A. B. Muhammad[1], I. Saidu[2]

[1]Department of Mathematics, Usmanu Danfodiyo University, Sokoto, Nigeria
[2]Department of ICT, Usmanu Danfodiyo University, Sokoto, Nigeria

DOI:10.21276/sjeat.2019.4.8.2 | **Received:** 16.08.2019 | **Accepted:** 25.08.2019 | **Published:** 30.08.2019

***Corresponding author:** Amina Imam Abubakar

Abstract

This paper presents Hausa WordNet (HWN), a lexical resource for Hausa language. The HWN extracts knowledge from a conventional Hausa dictionary and adopts a substructure of English and Hindi WordNet as it groups words based on different categories. The HWN Introduces pronunciation and the use of close class categories (CCC) to address the problem of missing pronunciation and coverage from existing WordNets. HWN is evaluated in comparison with existing WordNets (English and Hindi). The performance results show that HWN performs better in terms of pronunciation and CCC.

Keywords: WordNet, Hausa WordNet, Electronic lexical resource.

INTRODUCTION

WordNet is a lexical database for English language which groups English words into sets of synonyms, provides short definitions and usage examples [1, 2]. It is also known as a combination of dictionary and thesaurus [3]. Similarly, it is a source of reference that takes the conventional dictionary to a whole new level. While a dictionary can provide information such as meaning, synonyms, parts of speech (POS) and can organize words in alphabetical order, a WordNet in addition to the features of dictionary can further categorize the words into a set of synonyms (synsets). The synsets are classify into some part of speech known as open class categories such as nouns, adjectives, adverbs and verbs in order to express distinct concepts [4]. Various semantic relationships such as hypernym, hyponyms, meronyms, troponyms, antonyms and entailment provide linkage to the synsets [5]. These relationships enable WordNet to be an ideal tool for word sense disambiguation, semantic tagging and information retrieval [6].

The success of English WordNet motivated the creation of WordNets for some languages such as Hindi [5], Kannada [7], Danish [8] and Sanskrit [9] for the purpose of adopting standard lexical databases. Each of these WordNets uses an open-class categories (OCC) which are linked to each other by various semantic relationships. The class categories generate senses while retaining their original character in other senses. The senses are grouped together according to similarity of meanings in order to remove ambiguity in cases where a single word has multiple meaning. However, these WordNets fail to consider the use of pronunciation of words and also some POS known as close class categories (CCC) such as prepositions, pronouns, conjunctions and interjections are missing and thus, lack coverage.

In this paper, a Hausa WordNet (HWN) is proposed in order to address the aforementioned problems. HWN introduces CCC in order to achieve a wide coverage of words for the language. It also provides pronunciation of words in order to retain the actual word meaning. The proposed HWN has been evaluated using descriptive statistics based on frequency count of categories as done in and the result shows that the HWN performs better compared to the existing WordNets in terms of coverage and pronunciation [2, 5].

The rest of the paper is organized as follows: section 2 presents the proposed HWN, in section 3 evaluation method was discussed. Result was discussed in section 4 and in section 5, we conclude the paper.

PROPOSED HWN

Hausa is one of the three national languages of Nigeria and a major language of West Africa, with an estimate of 35-40 million speakers. The language stretches across the northern states of Nigeria and into southern Niger and also of Hausa communities in the

Sudan. Hausa is also spoken as a first language by scattered settlements throughout West Africa and as a second language by millions of non-Hausas in northern Nigeria and in the northern parts of Benin, Togo and Ghana [10]. However, despite the popularity of the language, there is currently no WordNet available that represents its richness in a machine readable form. Therefore, in this paper a HWN is proposed as a standalone WordNet build on monolingual grounds. It adapts a substructure of the English and Hindi WordNet as it group words together based on their meaning, as a result, words with the same similarity are grouped together and disambiguated. Similarly, it gives detailed information on pronunciation, part of speech (both open and close class categories), gender, word senses, usage examples and synsets. The propose WordNet transfers, adjusts and supplement lexical knowledge from a conventional dictionary.

Figure-1 shows HWN framework. The process start with transferring Hausa conventional dictionary into machine readable database. The database is used in building a HWN application with an interface for searching words. When a user type a word, the application will search for the word in the machine readable database and if the word exist then meaning will be displayed using the following categories: pronunciation of the word, POS, gender, various word senses of the word, usage example and synsets (if any) and if the word fails to exist then an error message will be displayed.

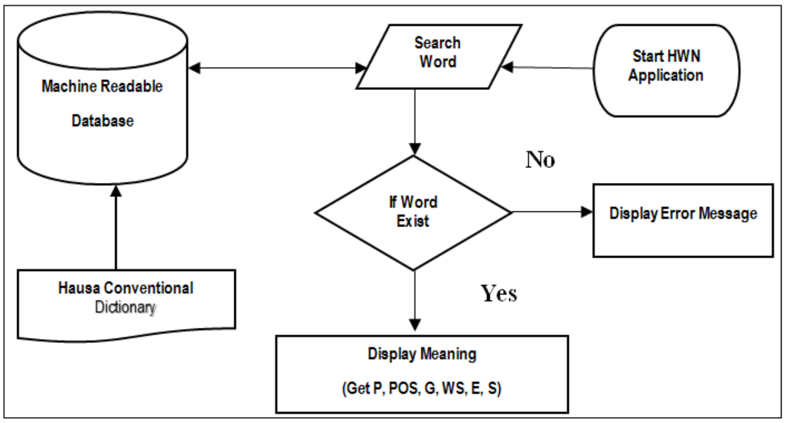


Fig-1: HWN Framework

Keys: p=pronunciation, POS= Part of Speech, G= Gender, WS= Word Senses, E= Usage Example and S= Synset

CONVERTING HAUSA DICTIONARY INTO A MACHINE READABLE DATABASE

The conventional Hausa resource used for the study is Kamusun Hausa [11]. The resource is a monolingual dictionary of Hausa language compiled by department of Hausa of Bayero University Kano (BUK) which was published in the year 2006. It is publicly available as a printed edition and is used heavily in various research purposes such as in building other Hausa resources and applications. In this paper, the dictionary is modified to machine-readable format (xlsx, xml) so as to develop HWN. Table 1 and 2 show some Hausa words as used in both the conventional dictionary and machine readable database respectively.

Table-1: Conventional Dictionary (Kamusun Hausa)

baba *(baba, sn., nj.) (i)* mahaifi ko uba, *d-d* abba (ii) wa ko kanen uba, *d-d* baffa. (iii) sa'a ko abokin mahaifi. (iv) sunan da ake kiran dattijo da shi don girmamawa.(v) sakaya sunan wani mai sunan mahaifi. baba *(baaba, sn., mc.)* (i) mahaifiya ko uwa (ii) ya ko k'anwar mahaifiya, *d-d* inna ko uwa. ba-ba *(baa-ba, har.)* kalmomi masu kore samuwa ko faruwar wani abu: *mis: Tanko ba dan'uwana ba ne.* ba-ba *(baa-ba, har.)* kalmomi masu kore wani aiki mai zuwa: *mis:* ~ zan zo baba. baba *(ba ba, har.)* kalmoni masu kore wani aiki da ya shude: *mis: ~ su zo baba* . baba *(baabaa, sn, nj., jam.,* baabannii) la'ifi, watau mutumin da azzakarinsa ba ya tashi. baba *(baabaa, sn., nj.)* wani tsiro wanda akan sarrafa shi don yin rini.
fatsam *(fatsam, b-fi.,)* kalma mai nima yakuwar abu; *mis: saiwar bishiya tayi fatsam.*
maimaita *(maimaitaa, fi.)* sake yin abu, *d-d* maya.
Kwajaja (Kwaajaajaa, sif, mc/nj) tikeke, musamman ciki, misali yana da ciki kwajaja
Amma (ammaa, har.) kalmar togaciya mai ma'anar sai dai, ko ko da yake; misali, na karanta amma ban gane ba, duba amman ko ammanin.

Table-1 displays the word sense(s) for the words: *baba, fatsam, maimaita, kwajaja, amma*. Each line in the table describe a word and each word is followed by its pronunciation, POS, gender (where applicable) which are enclosed inside bracket. The word sense comes after the bracket and a Roman numerals is used in displaying word senses. The dictionary uses‘*d-d*’ or ‘*duba*’ (daidaitun ma’ana) to indicate a synsets and ‘mis: ~’ (misali) for usage example.

However the machine readable format uses ‘*dd*’ instead of ‘*d-d’ and ‘duba’* to indicate synsets and ‘*misali*’ instead of ‘*mis: ~*’ to indicate a usage example. In addition, new Hausa vocabularies and abbreviations are included using crowdsourcing and domain human expert. Similarly, in order to avoid too much redundancy, word senses are compiled together so that words can be read as a whole as shown in Table-2. Table-2 contains two (2) columns; the first column represent Hausa words and the second column represent the meaning of words. The Meaning of each word is categorized into two parts, the first is the bracket and the second is the word sense

Table-2: Machine Readable Database

Word	**Meaning**
baba	(baba baaba baa-ba baabaa, sn har, nj mc); (i) mahaifi ko uba, *dd* abba. (ii) wa ko kanen uba, *dd* baffa. (iii) sa'a ko abokin mahaifi. (iv) sunan da ake kiran dattijo da shi don girmamawa.(v) sakaya sunan wani mai sunan mahaifi.(vi) mahaifiya ko uwa. (vii) ya ko k'anwar mahaifiya, *dd* inna ko uwa.(viii) kalmomi masu kore samuwa ko faruwar wani abu: misali Tanko ba dan'uwana ba ne.(ix) kalmomi masu kore wani aiki mai zuwa: misali zan zo baba. (x) kalmoni masu kore wani aiki da ya shude: misali su zo baba . (xi) la'ifi watau mutumin da azzakarinsa ba ya tashi. (xii) wani tsiro wanda akan sarrafa shi don yin rini.
Fatsam	*(fatsam, b-fi.,);* kalma mai nuna yakuwar abu *misali: saiwar bishiya tayi fatsam.*
Maimait	*(maimaitaa, fi.);* Sake yin abu, *dd* maya.
kwajaja	(Kwaajaajaa, sif, mc/nj); tikeke, musamman ciki, misali yana da ciki kwajaja.
Amma	(ammaa, har.); kalmar togaciya mai ma'anar sai dai, ko ko da yake, misali, na karanta amma ban gane ba, *dd* amman ko ammanin.

1. Inside the bracket are word pronunciation, POS (OCC or CCC), gender and many other categories, but in this work, the most frequently used categories are considered as follows.
   a. Pronunciation (P) comes first in every bracket such as ‘*baba baaba baa-ba baabaa*’ from Table-2. Hausa language heavily depend on pronunciation in determining word senses. Therefore, the paper introduces pronunciation using word stress to differentiate between word senses. Word stress uses vowel repetition to make a stress on a syllable and a single vowel to make it unstress. Stressed syllable is higher in pitch, longer in duration and generally a little louder than unstressed syllable [12]. Example, from Table-2 the word ‘*baba*’ can be pronounce in various ways such ‘*baba*’ (father) has 2 syllables which are all unstressed because it has no any vowel repetition while ‘*baaba*’ (mother) also have 2 syllables, the first syllable ‘*baa*’ is stressed because of vowel repetition and the second ‘*ba*’ is unstressed as there is no vowel repetition.
   b. Part of Speech (POS) comes second which consist of both open class categories (OCC) and close class categories (CCC). OCC are noun (*suna*), verb (*fi’ili*), adverb (*bayanin fi’ili*) and adjective (*sifa)* which are donated using the following abbreviations *sn, fi, b-fi,* and *sif* respectively while CCC are sometimes identified as stop words in some language processing such as word sense disambiguation. Stop words are usually refers to the most common words in a language and are mostly filtered out before or after processing in order to improve performance as they carry less important meaning and reduce processing time.
      However, this paper introduces CCC such as pronoun (*wakilin suna or wsn*), preposition (*madanganta or mdg*), conjunction (*mahada or mhd*) and interjection (kalmomin *motsin rai or m-r*) so as to capture all Hausa POS. This paper uses both OCC and CCC to get a wide coverage of Hausa words in order to make HWN an ideal tool for some language processing such as multilingual classification and sentiment analysis. Some categories appears in quite a few number of times and as such we used NA (not available) to indicate their absent on a word. Table 3 and 4 categorizes both OCC and CCC respectively.
   c. Gender (G) comes third in the bracket, the two grammatical Hausa genders are used: masculine (*namji*) and feminine (*mace*) which are donated by ‘nj’ and ‘mc’ respectively. A combination of both is used if the word is used in describing both genders, example from Table 3, the word ‘*kwajaja*’ have both ‘mc’ and ‘nj’ as gender. Thus, CCC words have no gender and synsets.

Table-3: Classification of an OCC words

Categories Words	Pronunciation	OCC	Gender	Word Sense	Example	Synsets
Baba	**Baba**	**sn**	**nj**	**(i) mahaifi ko uba (ii) wa ko kanen uba (iii) sa'a ko abokin mahaifi. (iv) sunan da ake kiran dattijo da shi don girmamawa (v) sakaya sunan wani mai sunan mahaifi.**	**NA**	**Abba, baffa**
Fatsam	**Fatsam**	**b-fi**	**NA**	**kalma mai nuna yakuwar abu.**	**Saiwar bishiya tayi fatsam**	**NA**
Maimaita	**Maimaitaa**	**fi**	**NA**	**Sake yin abu**	**NA**	**Maya**
kwajaja	**Kwaajaajaa**	**sif**	**mc/nj**	**Abu mai girma, musamman ciki**	**Yana da ciki kwajaja**	**Tikeke**

Table-4: Classification of a CCC words

Categories Words	Pronunciation	CCC	Word Sense	Example
Ta	**ta**	**wsn**	**Wakilin suna na mace mai daukar lokacin aiki wanda ya wuce ko wanda ake sa ran faruwarsa.**	**Ta Audu sabuwa ce.**
A	**a**	**Mdg**	**Madanganci wato harafi mai bayyana wurin da wani aiki ya faru.**	**An haife shi a Kano.**
Amma	**ammaa**	**mhd**	**Mahada wato kalmar togaciya mai ma'anar sai dai, ko ko da yake.**	**na karanta amma ban gane ba.**
Af!	**af!**	**m-r**	**Kalmar motsin rai da ake amfani da ita wajen tunawa da abin da aka manta.**	**Af! Sam na manta da maganar da mukayi.**

2. The second part is the meaning that display the following:
   a) Word sense (WS) are meaning of a word or various meaning of a word. The senses are grouped together according to their similarity of meanings in order to remove ambiguity in cases where a single word has multiple meaning.
   b) Synset (S) is a set of one or more synonyms that are interchangeable in context. From Table-2, all synset of a word are attached to the word sense of that word and is donated using the word ‘*dd*’, for example the word ‘*baba*’ with the following word senses (i) *mahaifi ko uba, dd abba* (ii) *wa ko kanen uba, dd baffa* which means the synsets of the word ‘*baba*’ with the above word senses are ‘*abba*’ and ‘*baffa*’ respectively. Construction of comprehensive Hausa synsets will have to be done together with dialects. Hausa language due to its geographic spread have various kinds of dialects, marked by differences in pronunciation, grammar and vocabulary. However, the use of dialect is out of the scope of this paper.
   c) Usage Example are example of the word as described by the word sense. From Table 2, usage example is denoted using the word ‘*misali*’, example the word ‘*fatsam*’ has a usage example as ‘*saiwar bishiya tayi fatsam.*

ALGORITHM THAT DISPLAYS WORDS WITH THEIR CATEGORIES

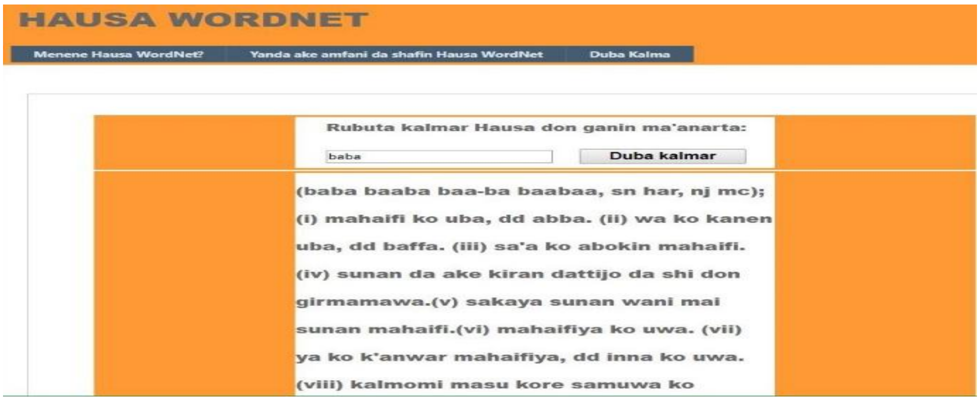
An algorithm that utilizes the machine readable database and accept word as input and display meaning (pronunciation, POS, gender, word sense, usage example and synset) as output is shown below.

Algorithm: HWN

```
Definitions: POS : Part of Speech
             P : Pronunciation
             G : Gender
             WS : Word Sense
             E : Usage Example
             S : Synset
             BP: Bracket part
             DB : Database

Input:       W : Word

Output:      P, POS, G, WS, E, S

Open DB
Type the keyword
Select * from table 2 where W= keyword
if   Recordset != null
     W=Recordset.W
     meaning=Recordset.meaning
     meaning2=meaning.split (';')
     BP=meaning2[0].split(',')
     WS=meaning2[1].split('.')
     P=BP[0].split(" ")
     POS=BP[1].split(" ")
     G=BP[2]. split(" ")
   for each WS in meaning {
       S=WS.split('dd')[1]
       S.append (S)
       }
   for each WS in meaning {
       a=WS.split ('misali')
       WS=a[0]
       E=a[1]
       E.append(E)
       meaning.append(meaning)
       }
display P, POS, G, WS, E, S
else
display keyword not in DB
```

DEVELOPING A WEB BASED APPLICATION

A database engine was created using Microsoft SQL server management studio (2008 R2) for storing and retrieving lexical information. However, In order to give users access to this information, a usable user interface that facilitates the mapping of words to their corresponding meaning was developed as shown in Figure-2. This Interface enable end users to retrieve the data and display it via a web-based tool. The interface tool used for HWN is Microsoft ASP visual studio (2010 professional).



Rubuta kalmar Hausa don ganin ma'anarta:

baba

(baba baaba baa-ba baabaa, sn har, nj mc); (i) mahaifi ko uba, dd abba. (ii) wa ko kanen uba, dd baffa. (iii) sa'a ko abokin mahaifi. (iv) sunan da ake kiran dattijo da shi don girmamawa.(v) sakaya sunan wani mai sunan mahaifi.(vi) mahaifiya ko uwa. (vii) ya ko k'anwar mahaifiya, dd inna ko uwa. (viii) kalmomi masu kore samuwa ko

Fig-1: The word “*baba*” as shown in HWN Interface

EVALUATION

The evaluation of the proposed HWN is perform using descriptive statistics to get the frequency counts of the categories. Frequency count discover the number of occurrences of various categories present in a contexts of a language used (such as words in texts) in order to provide statistical basis for the categories. The resulting count is compared to the existing WordNets (English and Hindi). The paper uses SQL query in finding frequency of categories. Below is an example of a query that find the frequency count of all synsets from the machine readable database.

```
SELECT Count (meaning) As Synset
FROM Table2
WHERE meaning like '% dd%';
```

The data for the proposed HWN was extracted from a conventional dictionary consisting of 27,414 words with 28,992 pronunciation spread across 14,990 open class categories that produces 33,878 senses and 17,725 synset.

RESULTS AND DISCUSSION

The proposed HWN is composed of a total of 25,101 words with 31,441 pronunciation that spread across 15,007 open class categories and 530 close class categories, were 33,927 senses are created from both categories with 17,775 synsets as shown in Table-5.

Table-5: Proposed HWN Frequency Table

Categories	**Frequency Count**
Pronunciation	31,441
Part of speech (CCC)	530
Part of speech (OCC)	15,007
Number of Words	25,101
Synsets	17,775
Word senses	33,927

Figure-3 demonstrates the frequency count of the newly created categories. HWN performs better in terms of pronunciation and CCC and this is achieve by introducing pronunciation and CCC words to the proposed HWN. Even though the Frequency count of CCC is quite small (530), it is however add some coverage for the new created language resource as it now captures all the part of speech for Hausa language (*suna, fi'ili, bayanin fi'ili, sifa, madanganta, mahada, wakilin suna and kalmomin motsin rai*) and will be an ideal tool for multilingual classification and sentiment analysis.

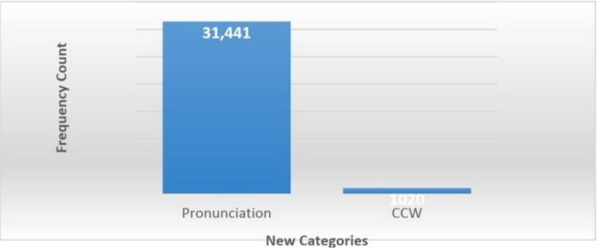


Fig-3: Frequency count of new categories

Table-6 compare the frequency count of the proposed HWN with the existing English [2] and Hindi WordNet [5].

Table-6: Comparison with English and Hindi WordNets

WordNets **Categories**	**English WordNet**	**Hindi WordNet**	**Hausa WordNet**
OCC	**155,287**	**40,463**	**15,007**
Number of Words	**155,287**	**105,409**	**25,101**
Synsets	**117,659**	**40,463**	**17,775**
Word senses	**206,941**	**40,463**	**33,927**

Figure-4 shows the frequency count of categories in proposed HWN, English and Hindi WordNets. The result shows that HWN have a smaller frequency counts compared to the existing WordNets. The reason for this can be attributed to the fact that English language is a global language that is rich in vocabulary than Hausa language. It can also be attributed to the failure to use various Hausa dialects in the development stage to capture various Hausa words, word senses and synsets.

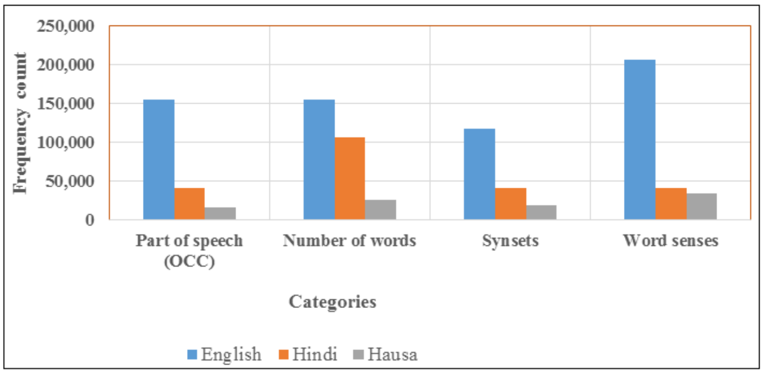


Fig-4: Frequency count of categories in proposed HWN, English and Hindi WordNets

CONCLUSION AND FUTURE WORK

HWN is proposed (in its early stage) as a new lexical resource for Hausa language which transfers, adjusts and supplement lexical knowledge from a conventional dictionary. The HWN Introduces pronunciation and the use of CCC to address the problem of missing pronunciation and coverage. The propose HWN is evaluated by comparison with existing WordNets (English and Hindi) using descriptive statistics to get the frequency count of categories. The performance results shows that the proposed HWN performs better in terms of pronunciation and coverage. The proposed HWN not only adds to the sparse collection of machine readable Hausa resource, but also gives new insights into Hausa vocabularies and applications. Future work will focus on improving HWN general structure, pronunciation structure, establishing semantic relationships and the use of various Hausa dialects for further refinement and completion.

REFERENCES

1. Fellbaum, C. (1998). WordNet: An Electronic Lexical Database (p. Cambridge, MA: MIT Press).
2. Miller, G. A. (1995). WordNet: A Lexical Database for English. *Communications of the ACM*, *38*(11), 39-41.
3. Navigli, R., & Ponzetto, S. P. (2012). BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, *193*, 217-250.
4. Kilgarriff, A. (2000). Wordnet: An electronic lexical database. *Language*, *76*(3), 706.
5. Bhattacharyya, P., Jha, S., & Narayan, D. (2001). A WordNet for Hindi. *International Workshop on Lexical Resources in Natural Language Processing, Hyderabad, India*, 1–8.
6. Morato, J., Marzal, M. Á., Lloréns, J., & Moreiro, J. (2004). WordNet Applications. In *Proceedings of the 2nd International Conference of Global WordNet*, 270–278.
7. Sahoo, K., & Vidyasagar, V. E. (2003). Kannada WordNet – A Lexical Database. In *IEEE Conference publication*.
8. Pedersen, B. S., Nimb, S., Asmussen, J., Sørensen, N. H., Trap-Jensen, L., & Lorentzen, H. (2009). Dannet: The challenge of compiling a wordnet for Danish by reusing a monolingual dictionary. *Language Resources and Evaluation*, *43*(3), 269–299.
9. Kulkarni, M., Dangarikar, C., Kulkarni, I., Nanda, A., & Bhattacharyya, P. (2010, January). Introducing sanskrit wordnet. In *Proceedings on the 5th Global Wordnet Conference (GWC 2010), Narosa, Mumbai* (pp. 287-294).
10. Newman, R., & Newman, P. (2001). The Hausa lexicographic tradition. *Lexikos*, *11*(1), 263-286..
11. Said, B. (2006). kamusun Hausa na Jami'ar Bayero. (A. M. Maikudi karaye, Lawan Danladi Yalwa, Habib Ahmad Daba, Abdu Yahya Bichi, Abba Rufa'i, Abdullahi Umar kafin Hausa, Sammani Sani, Ed.). Kano: Center for the Study of Nigerian Languages.
12. Aslam, M., & Amin, A. (2011). Introduction to English Phonetics and Phonology (pp. 60–68). Cambridge University Press.