Asian Journal of Language, Literature and Culture Studies

Asian Journal of Language, Literature and Culture Studies

4(1): 23-29, 2020; Article no.AJL2C.63924

The Fundamental Approaches to Siswati Lemmatization Process

J. J. Thwala[1*] and N. M. Lusenga[2]

[1]Human and Social Sciences, University of Mpumalanga, Mbombela, Mpumalanga Province, South Africa.
[2]Siswati National Lexicography Unit. (SNLU), Mbombela, Mpumalanga Province, South Africa.

Authors' contributions

This work was carried out in collaboration between both authors. Both authors read and approved the final manuscript.

Article Information

Editor(s):
(1) Dr. Suleyman Goksoy, University of Duzce, Turkey.
Reviewers:
(1) Syazliyati Binti Ibrahim, Academy of Language Studies, Malaysia.
(2) Mohd. Nor Shahizan Ali, Universiti Kebangsaan Malaysia, Malaysia.
(3) Azman Bin Che Mat, Universiti Teknologi MARA Shah Alam, Malaysia.
Complete Peer review History: http://www.sdiarticle4.com/review-history/63924

Original Research Article

Received 01 November 2020
Accepted 06 January 2021
Published 27 January 2021

ABSTRACT

The objectives of this research article are to highlight and explain the significances of the fundamental approaches to Siswati dictionary compilation. The identification and functions of the parts of speech are vital during lemmatization process which involves the arrangement affixation process of semantic process and applications. The prefixes as morphological formatives need to be intensively looked at from singular to plural stance. Prefixation as a process need not to be solely looked at, but it must consider infixation and suffixation if warranted. Although, the semantics approaches and implications are basic in lexicographical products, but the phonological and morphological process are also inevitable. Lemmas are explicated from structural point of view to meanings. To arrive at the meaning of the lemmas, the prior knowledge of prosodic elements, etymological exercise and morph-phonological analysis is needed.

Keywords: Infixation; lemmatization; modulation; suffixation and translation equivalence.

Corresponding author: E-mail: Jozi.Thwala@ump.ac.za;

1. INTRODUCTION

Siswati is a Southern African language spoken predominantly in Eswatini and South Africa. It is the official language of Eswatini. It is one of the nine indigenous languages to enjoy official recognition in South Africa's first post-apartheid constitution [1]. Siswati speakers make up the third smallest official language group in South Africa. In South Africa most of the speakers of this language are situated in the eastern region of the Mpumalanga Province, which borders Swaziland [2]. It falls under the Tekela sub-group, which comprises various dialects such as Bhaca, Lala, Phuthi and Nhlangwini. It is used as the medium of instruction in schools in Mpumalanga Province and Eswatini.

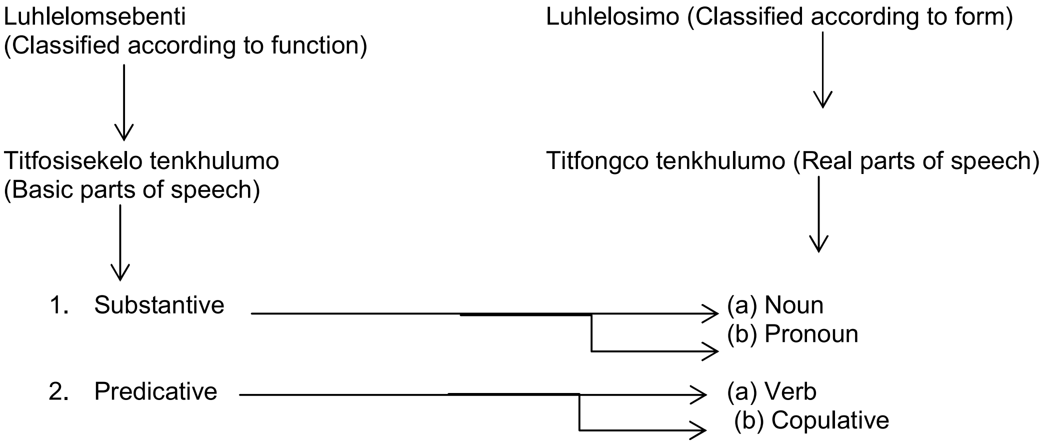
The encoding or decoding aids are used during translation, but the dictionary functions as the only source of semantic information for both the indigenous and the foreign language treated in the dictionary. If anyone needs to know the meaning of a word from the lexicon of indigenous language, he/she consults a bilingual dictionary and interpret the meaning from the given translation equivalents. A bilingual speaker can often use a bilingual dictionary with success to obtain semantic information regarding his indigenous language. For the skillful dictionary user who is competent in both languages, the listing of translation equivalents without the interference of other microstructural entries can be used for instant semantic sing [3].

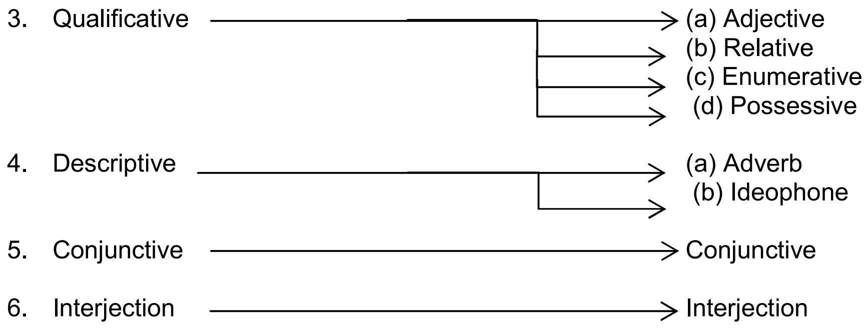
Bilingual dictionaries are employed as polyfunctional sources of semantic information. Their main function is not a transfer of meaning. They are aids in Interlingua translations and must focus on a treatment that enables the user to render a good and sound translation. The main aim of the dictionary should not only be the establishment of a relation of semantic equivalence between source and target language, but a lexicographer must endeavor to reach communicative equivalence.

When people consult a bilingual dictionary, they seldom realize that the information given is not essentially a statement about meaning but a list of translation equivalents Louw [4]. The functional status of these translation equivalents is that they may be used in certain contexts to substitute the source language item. Where the specific contexts in which translation equivalents can be used to substitute the lemma are not given as part of the lexicographical treatment, it is possible that the creation of semantic equivalence can lead to the establishment of communicative equivalence. A relation of equivalence exists between lemma and the translation equivalent paradigm.

1.1 Research Methodology and Theoretical Underpinning

The functional approach is selected for this research work and broadly looked at from various perspectives where they explicitly depict the representational function as the use of language to convey statements, knowledge and facts to represent reality. The interactional function is the use of language that establishes social contacts. The communicative functions are looked at in relation to micro and macro structural process. For Siswati language, only the following parts of speech are vital. Doke's classification of words into word categories gives this schematic representation:





Flow chart. Doke [5]'s classification of words into word categories

A definition of a dictionary is a reference work that lists words found in a language in alphabetical order and gives their meanings. It is a basic knowledge that a dictionary lists words and their meanings. The significance of a definition in a dictionary cannot, therefore, be overemphasized. It is an intrinsic component of a dictionary. It is considered as one of the most important aspects of the dictionary-making process. As a result of this rigorous process, users can interact with the definition and know what a specific word means as explained in it.

Defining formats are pertinent in so far as they assist the lexicographers or editors to write the systematic and culturally relevant definitions that make dictionaries user-friendly. These are formats that provide the editor with certain guidelines to follow in constructing a definition, hence they contribute to consistency in the whole process [6].

This is so because they further help to illuminate the definition. With their aid the editor can write a clear, specific definition, which is culturally and contextually relevant. According to Sinclair [6], "a corpus is a collection of naturally occurring language texts, chosen to characterize a state or variety of a language". If the usage from the corpus is studied, it is found that the surrounding words and phrases help considerably in determining the meaning. Each word, therefore, occurs in its own special environment, called its *context,* and by studying the context it can often be established which meaning is intended.

A definition is an intrinsic component of any dictionary. It is itself a statement of the meaning of a word which is distinct in outline. It explains the meaning and the use of a word in a speech community. Two facts result from these observations: firstly, that a mother-tongue speaker is usually best equipped formulate effective definitions of words in a particular language; and, secondly, that it is a rigorous process that follows set principles so as to provide a concise, distinct and conceptually appropriate statement of the meaning of a particular word.

Defining can be considered as one of the most significant processes in dictionary making. It is the process through which users may become more aware of the meanings and explanations of words. It is important for users to know the meanings of words so that they can use them in a prescribed manner in both spoken and written media. A dictionary is prescriptive in so far as it distinguishes the formal and colloquial uses of any specific word. A set of defining principles is therefore of paramount importance during the defining state so that the editors can write clear, brief, consistent, culturally relevant and specific definitions.

Defining formats are crucial in that they provide the editor with a scheme to follow when defining a specific word. It is a framework from which an editor can construct a definition. They are important in that they contribute to consistency in the treatment of related sets of headwords. This view is successfully accentuated by Gellerstam's [7] observation that "the development of definition formats is advantageous in that once a particular format has been agreed upon, a number of words can be defined with." The goal of any defining guideline is to enable an editor to write what is widely referred to as a *user-friendly* definition.

1.2 Attributes or Value Combination

The text selection criteria consist of several attributes or value combinations. The attributes we discuss are:

- Language
- Time
- Mode
- Medium
- Domain

The attributes or value combinations enable the readers to classify any text and situate it in a specific part of the sampling frame. The parameters described, provided a valuable model for building a lexicographic corpus.

- **Language** – is vital in the corpus. It is determined that the corpus is monolingual, bilingual or multilingual. It represents the varieties of the languages and reveal how far the corpus accounts for regional dialectical variations.
- **Time** – is essential as it can either be synchronic or diachronic. In a synchronic corpus, the constituent texts come from one specific period, whereas the texts making up a diachronic corpus come from an extended period. Essentially, corpus-builders must decide 'how diachronic' their corpus needs to be in order to support the kind of lexicography they will be doing. While a historical dictionary requires a fully diachronic corpus, dictionaries designed for learners deal mainly with contemporary language, so they need a synchronic corpus which provides a snapshot of the language as it is used at the time of compilation.
- **Mode** – reflects the written, spoken or both texts. It is an evidence-based phenomenon for communication equivalence. Newer 'hybrid' forms of text associated with the web complicate the matter further. Conversations are in real time displays many of the characteristics of spontaneous spoken dialogue. Spoken text is for several reasons – more difficult to collect than written, so practical issues will often influence your corpus design model.
- **Medium** – refers to the 'channel' in which the text appears. A simple classification would distinguish print media and spoken media. The former includes the books, newspapers, magazines, learned journals, dissertations, movie scripts, government documents, and legal status. Spoken media include face-to-face conversations, broadcasts and post casts, public meetings, and educational settings (seminars, lectures).
- **Domain** – refers to the subject matter of a text: what the text is about. Domain is not a 'universal' parameter because not all kinds of text can be classified in these terms.

1.3 Dictionary Information

A dictionary comprises the words of a language that are arranged in a strict alphabetical order. If it is regularly used, it becomes user friendly. It contains a valuable and informative guide.

Two words, called headwords, usually appear in bold lettering at the top of each page.

- The word on the left indicates the first word of that page.
- The word on the right is the last word of that page.
- This is a helpful, time-saving device in finding a word.

It provides you with the following information:

(a) The meaning of the word.
(b) The pronunciation of the word.
(c) The different parts of speech derived from the word, for example, nouns and adjectives.
(d) The origins/source of the word or also referred to as etymology.

Where relevant, the following may also be given:

(a) The abbreviation of the word.
(b) The plural of the word.
(c) Other parts of speech, for example, nouns, verbs or adverbs.
(d) A phrase conveying the meaning of the word.
(e) Idiomatic use of the word.

In a dictionary, language may be formal or informal depending on its function between the user and the recipient. For formal language, the vocabulary expresses accuracy and coherence, while in informal language, the vocabulary is less accurate and less sophisticated. It is advisable to have a special marker for slang, jargon and colloquialism in the dictionaries.

1.4 Prosodic Elements

Tone is distinctive pitch level of a syllable used to distinguish word meaning or to convey grammatical distinctions. On the other hand, there is a stress that goes hand in hand with tone. Stress is the degree of force used in producing a syllable [8].

Lexicographers, as dictionary compilers, experience some difficulties during their word. A good user-friendly dictionary should clearly give the exact meaning of a word without leaving the user in doubt of the given meaning. It should carry all necessary characteristics that a dictionary must have. However, as a matter of fact, compilation of dictionaries is rule governed. Space is one element that should be considered. Unlike, in grammar books, dictionaries should not include detailed information. They do not analyze words and phrases but give the meaning of a word to maintain precision. Nevertheless, there are vital characters that are mostly left out in dictionaries of indigenous African languages such as tone and stress. Some of these languages use diacritic marks to indicate the tone's peripheral status.

On the other hand, there is tone. Tone has to do with pitch. Pitch depends on the rate of vibration of the vocal cords, [9]. In Siswati, like other tone languages, tone has a predominantly lexical function. One can notice that the exact speech sound that carries the tone marker is mainly the vowel of that specific syllable since all vowels in the Siswati language are voiced thus vocal cords of the speakers vibrate. Some consonants become voiced because they assimilate the feature voiced from an adjacent speech sound, is known as contiguous or contact assimilation [8].

There are three prosodic elements that are evident in indigenous languages. They are stress, length and tone.

1. Length is a period taken to utter a word.

Long vowels occur in the penultimate syllable of a word in isolation.

- Kubo : na (to see)
- Ku : va (to hear)

Half long vowels occur in the penultimate syllable inside a sentence or phrase.

- Bafa :na badla : la ibho: la (The boys are playing soccer)
- Umu :ti wa:mi uya: sha (My house is burning)

Short vowels occur in certain ideophones and monosyllabic demonstratives

- Monosyllabic demonstrative lo. this one
- Ideophone chu!

Abnormal length occurs in emotional speech

- Lapha *: ya* (Right over there!)

2. Stress is a force existed in uttering a syllable in speech. It is characterized by dynamic accent. It makes one syllable is more prominent than others. The stress is on the penultimate syllable

- Bo : na (see) Boni : sa (make to see) Bonaka: la (be seen)

When the word is lengthened by the addition of suffixes, the stress continually moves forward.

3. Tone is the musical modulation of the voice in speech while intonation is tonal pattern used in speech. Tone distinguishes a language group. It further detects the nationality of a person.

- (LLH) *Bomake* and *bobabe* words are grammatically correct but tonologically wrong
- (HHL) *Bomake* and bobabe are correct grammatically and tonologically.

Semantic tone – differences of meaning in wrong

(HHH) *inyanga*	(moon)
(HHL) *inyanga*	(traditional doctor)

(HLL) *litsanga*	(pumpkin)
(LLL) *litsanga*	(thigh)

Grammatical tone – grammatical function of words.

(HL) *bona*	(look)	*Thoko ubona likati*
(LL) *bona*	(they)	*Bona boThoko bahambile*

Tone may be high or low, high-low, low-high, falling and rising. A variety of diacritic signs are employed.

Tone is mainly morphosyntactic, meaning that the tonology is far more complex than that of a noun.

Verb tone is influenced by the syntactical consideration.

Noun tone_is influenced by the phonological rules.

Tone languages have lexically significant, contractive and relative pitch on each syllable.

- ✓ Lexical *kusindza* (to escape)
  kusindza (to smear on the floor)
- ✓ Grammatical contrastiveness

(HHL) *akanatsi* (he is not drinking)
(HLLH) *akanatsi* (he has not drunk)

Verb

Monosyllabic verb	*dla* (eat)	*fa* (die)
Disyllabic verb	*bita* (to call)	*buta* (to question)
Trisyllabic verb	*kuthula* (to be quiet)	
	Kutfula (to unload)	

Realization process

a	*ngi*	*lwi*	
L	H	L	
a	*ngi*	*yi*	*lwi*
L	H	H	L

2. DISCUSSION

Moon [10] and Hanks [11] assert the following:

- The headword should appear in the definition,
- The headword should be shown in a typical context of common usage because “context disambiguates” [11],
- Definitions should consist of full sentences,
- Explanations should consist of two parts, with the first part showing the word in use and the second part explaining the meaning, and
- The format should suggest a preference rather than a restriction.

A good definition should be short, user-friendly, culturally relevant and consistent with others of its type and should use superordinate terms [11]. When a definition has been made simple and familiar, it helps the dictionary to be user-friendly. Evident characteristics of dictionary of defining formats are that:

- Definitions should be as natural as possible,
- Definitions should project the typical usage of a headword,
- Definitions should be easy to understand, and
- a “dictionary as prose” [11] should be created, hence there should be done away with the use of parentheses.

These guidelines are instrumental in organizing, formulating and presenting definitions that are user-friendly. Dictionaries are regarded as containers of knowledge and more specifically as comprehensive and fundamental sources of linguistic information. This has definite consequences for the inclusion of a representative variety of information types as microstructural entries in the treatment of each lemma. Despite all the different linguistic categories accommodated in each dictionary, the average user still perceives dictionaries primarily as reference works aimed at the transfer of information on the meaning of words. This attitude influences the lexicographer when making decisions on the inclusion and presentation of information.

The comprehensive transfer of linguistic information in dictionaries is often impeded by the presence of a semantic bias that dominates the microstructural presentation. In monolingual dictionaries the focus on the definition and the lack of an extensive treatment of other categories such as grammatical information often illustrates this bias. A less explicit but often more aggressive application of the semantic bias can be identified in bilingual dictionaries. This procedure is most probably motivated by the typical usage patterns identified in the average dictionary user’s utilization of a bilingual dictionary. Users focus their lexicographical inquiries on semantic aspects in their search for “the target language meaning of a source language item”. Having found a target language item, no further attention is paid to additional information which might be of paramount importance for the correct comprehension and the use of the specific item.

The status of bilingual dictionary as a source of semantic information is especially evident in a multilingual society. In a monolingual society, bilingual dictionaries are not used as reference works in the day to day linguistic needs of the average member of a speech community. Monolingual descriptive dictionaries are employed for this function. In a multilingual society the use of bilingual dictionaries forms an integral part of the daily communication process.

3. CONCLUSION

The dictionaries are the lexicographical products that reflect the developed sound strategies and procedures for planning. Most structural components of macrostructure, microstructure and medio structure are employed to the certain levels. The dictionaries are, however, the instruments of communicative and linguistic empowerment that are based on sound theoretical and practical knowledge and skills. The strategies and approaches of lemmatization process do not follow a rigid and fixed rule for user-friendly cross-references. The varieties of meaning such as conceptual, connotative, reflective, thematic, stylistic and effective are reflected to various lemmas.

4. RECOMMENDATIONS

It is recommended that various types of dictionaries must be defined and analyzed to bring about the full knowledge of these lexicographical products. The significances of various formatives, such as prefixes, infixes, roots and suffixes need emphasis and practical application. Meaning needs to be highlighted through a precise explanation and synonyms. Due to the development of the language through civilization, acculturation and enculturation, adoptive are completely inevitable.

COMPETING INTERESTS

Authors have declared that no competing interests exist.

REFERENCES

1. Loubser M, Puttkammer MJ. Viability of neural networks for core technologies for resource-scarce languages. Information. 2020;11(1):41.
2. Mzamo L, Helberg A, Bosch S. Introducing XGL-a lexicalised probabilistic graphical lemmatiser for isiXhosa. In 2015 Pattern Recognition Association of South Africa and Robotics and Mechatronics International Conference (PRASA-RobMech). IEEE. 2015;142-147.
3. Dibitso MA, Owolawi PA, Ojo SO. Part of speech tagging for Setswana African language. In 2019 International Multidisciplinary Information Technology and Engineering Conference (IMITEC). IEEE. 2019;1-6.
4. Louw B. Ivory in the text or insincerity in the writer? Amsterdam: Benjamin's 157-176; 1985.
5. Doke CM. Textbook of Zulu Grammar. Johannesburg: Longmans Southern Africa; 1974.
6. Sinclair J. Collins Cobuild English Usage London: Harper Collins; 1990.
7. Gellerstam M. *EURALEX 96 Proceedings.* Gothenburg: Gothenburg University; 1993.
8. Crystal D. A dictionary of linguistics and phonetics. Cambridge: Blackwell Publishers; 2003.
9. Katamba F. An introduction to phonology; 1989.
10. Moon RE. 'The Analysis of Meaning' in Sinclair. 1987;86-103.
11. Hanks PW. Definitions and explanations in Sinclair. 1987;116-136.

Peer-review history:
The peer review history for this paper can be accessed here:
http://www.sdiarticle4.com/review-history/63924