Journal of Research and Academic Writing
Vol. 1 No. 1 (2024): ISSN (Online): 3007-5343
DOI: https://doi.org/10.58721/jraw.v1i1.439
Corpus Construct: A Research Tool in Syntactic Analysis of Bantu Languages

Corpus Construct: A Research Tool in Syntactic Analysis of Bantu Languages

Walter Ochieng Sande
Jaramogi Oginga Odinga University of Science and Technology, Kenya

Article History
Received: 2024.02.20
Revised: 2024.03.27
Accepted: 2024.06.14
Published: 2024.06.16

Keywords
Annotation
Corpus
Construct
Online databases
Parameters

How to cite:
Sande, W. O. (2024). Corpus Construct: A Research Tool in Syntactic Analysis of Bantu Languages. *Eastern African Journal of Humanities and Social Sciences, 1*(1), 1-12.

Abstract
Syntax of languages is understood to be shaped by syntactic universals principles. Despite operating within the constraints of these universals, languages have managed to display their unique syntactic features. This is explained by specificity in the ranking of these universals across languages. To establish the unique features, researchers have to use data collected using a variety of methods some of which are corpus studies, linguistic elicitation, introspection and experimentation. Each of these methods requires a research tool whose development or adoption is dependent on the research question(s) formulated to fill a study gap. The use of corpus construct to generate data, as other research tools, requires an understanding of what type of data is needed to answer which questions on which linguistic features. The construction of a corpus can be done from plain texts or annotated texts. The question is: how can corpus construct be used as a tool in the study of languages whose corpus is yet to be compiled and made available online? This article, therefore, intends to answer this question with biases on Bantu languages. It will be necessary to make databases from the corpus constructs available by building corpora for the languages in question. The findings of this article are deemed important in offering knowledge on building of corpus and how to use the built corpus to investigate a syntactic feature in a Bantu language.

Introduction

Syntax, as one of the six key areas of descriptive study of languages, has been widely investigated in languages with the intent to bring forth unique syntactic features that characterise these languages (Chomsky, 1995; Haspelmath, 2020 & Sigurðsson, 2003). Among the syntactic features researchers have aimed at investigating in languages include but are not limited to word order, word classes, noun classes, tense, aspect and mood marking, verbal extensions and phrases. These features have been analysed within frameworks of different theories of syntax. According to these theories, every well-formed syntactic structure is considered a by-product of competing universal syntactic constraints (McCarthy, 1997). These constraints outdo their counterparts in their competition for optimal outputs (Joâo, 1996; McCarthy, 1997; Prince & Smolensky, 1993). As McCarthy (1997) observed, structural uniqueness in languages is explained by differences in the ranking of universal constraints in each language. Researchers will always need data from target languages to attest to this uniqueness. The kinds of data needed in research are determined by research question(s) understood to be shaped by the study gap the research tends to fill. The study gap in the field of research is considered a reflection of real-world situations, which is, in most cases, informed by existing studies

within the field where the proposed study falls (Creswell & Creswell, 2018). Therefore, A study is meant to come up with a solution to a real-world situation, which can be achieved only through accurate data (Creswell & Creswell, 2018). The accuracy of the data is determined by the correct choice of a research tool or a research tool's improvement adoption to the need of a research question (Adamou, 2019).

It is not unusual to generate data for studying a language without visiting the field where it is practised. This came into play when compiled corpus was discovered to be a useful research tool. The use of already existing corpora as a research tool has been emphasised by Meurers and Müller (2009). According to Meurers and Müller, a corpus construct is generated not only from a spoken discourse but also from already available texts. Since the discovery of the corpus construct as a research tool by scholars such as Adamou (2016), Baker (2009), Fancom (2021) and Hassel (2001) and several online archives have been made available. For instance, Adamou (2019), in her study of corpus linguistic methods, provides an online archive associated with endangered languages and cultures. The corpora include the Endangered Language Archive (ELAR), the DoBeS Archive (Documentation of Endangered Languages), the PARADISEC Archive (Pacific and Regional Archive for Digital Sources in Endangered Cultures) and the Pangloss Collection. Baker (2009), on the other hand, proposes BE06 as the best corpus for general written British English. For a large collection of texts from which examples of real language use can be extracted, Taylor (2008) proposes WebCorp as the best corpus. The archived corpora, as argued by Meurers and Müller (2009), are, in fact, a collection of found real-world data, thus a reliable source of data for the study of natural languages. Not all languages have compiled and archived corpus for linguists to use in their analysis of languages. In such cases, linguists must design a suitable corpus for their research purposes (Fancom, 2021).

The Syntax of Bantu languages, according to the existing literature, is very rich and complex in comparison with other language groups (Basweti, Achola, Barasa & Michira, 2015; Guthrie, 1967; Troyer, 2007; Waweru, 2011). The richness of these languages is marked by their agglutinative nature (Guthrie, 1967; Sande, 2019 &Waweru, 2011). It is important to note that these grammatical features operate within universal syntactic constraints (McCarthy, 1997). In fact, from McCarthy's (1997) point of view, the grammar of a language is constructed from language-particular rankings of these constraints. Language particularity in the ranking of universal constraints explains the uniqueness of the syntactic structures of languages. The structural uniqueness of Bantu languages, although constructed from the same universal constraints, has attracted a lot of descriptive studies of the structure of syntax across Bantu languages (Harford & Malembe, 2017; Kager, 1999; Kula, 2002; Prince & Smolensky, 2004). Data have to be consulted to describe a language's syntactic structures. These data have to be extracted from particular sources, some of which include datasets freely available online (Just & Witzlack-Makarevich, 2022; Parvess, 2023); native speaker's input (Devitt, 2006; Wasow & Arnold, 2004), conversation and sociolinguistic recordings (Podesva & Zsiga, 2013; Schleef & Meyerhoff, 2010) and secondary sources (Himmelmann, 2012). This article intends to add to the already existing knowledge of Corpus Linguistics. The article, using three Bantu languages, Olusuba, Gikuyu and Ekegusii, justifies using a corpus construct from a dataset as an alternative research tool in analysing the syntactic structures of Bantu languages.

Methodology

Data for this study existed in plain, unannotated texts sourced online from the archived secondary materials of three Bantu languages, namely Olusuba, Gikuyu and Ekegusii. All the PDF downloads were first converted into Microsoft Word format for easy annotation. The downloaded texts were then loaded into the corpus using the web-based tool sketch engine. The challenge faced during the loading was the lack of proportionality in the corpus; some loaded texts were longer than others, which may

lead to targeted syntactic features being more common in particular texts than in others. In this case, the corpus balance was achieved by the researcher working within a sample size set by the corpus software employed. The software set limits on the concordance output lines; the search automatically stopped when the corpora got to the set limit. It is worth noting that the study focused only on a syntax-oriented corpus. Given that syntactic structures are rule-governed, the issue of data representativeness was taken care of – a sample of the corpus would be used to generalise the outcome because syntactic shaping of the targeted texts works within universal constraints.

The compiled corpus was annotated and segmented to retrieve as much relevant information as possible. The annotation divided the corpus into two sections: the header and the body. In this case, the header contained metadata comprising discourse type, the title of the discourse, the role of speakers in the discourse and the context of the discourse. Coding information for this study was formed around syntactic features such as tense, number, person, mood, aspect, noun class, topicality metric, word and affix order, and verb valence. Despite the representation of grammatical features by different bound morphemes in the three languages, syntactic behaviours of these features cut across the three languages, forming the backdrop against which the current study was conducted. The behaviours included order and fixity in the slotting of bound morphemes within the matrix of the host root and morphological modification of bound morphemes to accommodate features such as number, tense and aspect, as well as noun class. Noun class was involved in instances where the interlink of verbals and/or adjectives with nouns formed sentences.

During the corpus construction, the researcher involved native speakers of the three languages for the interpretive judgment and refinement of the codes, thus eliminating incompleteness and ambiguity in the corpora. For easy analysis, the compiled corpora were translated into English in three tiers: the first was the original corpus, and the second was morpheme by morpheme glossing. The third tier was the English equivalences of the corpus. The compiled corpora from each of the three languages were subjected to content analysis to test the potential impact the compiled corpora had on the grammar modelling tasks.

Unannotated Corpus: An Overview

The study focused on plain texts in Microsoft Word format. The plain texts were sourced from archived secondary materials with different linguistic contents in Olusuba, Gikuyu, and Ekegusii. The targeted corpora in this study had neither linguistic annotation nor linguistic segmentation. The current study deals with unannotated data whose annotation and segmentation require the researcher's insight into the universal constraints behind shaping a language's syntax. As Beck (2023) explained, data annotation is the labelling of the data to show the outcome one wants their research machine to predict. Therefore, unannotated data is a composition of excerpts without such labelling. Using the knowledge of universal constraints, a researcher can annotate and segment texts from any language to study and develop syntactic patterns alongside establishing situations where the patterns occur.

The researchers' abilities to study and establish syntactic patterns under what syntactic conditions have enabled them to develop corpus constructs usable in the syntactic analysis of any language in the world. The developed corpora are made available online for consultation by the subsequent researchers. For the matter of online available corpora, a researcher does not have to go to the field for data collection; they will only be required to formulate research question(s) concerning the study gap and, with a search tool, get an already constructed corpus to use in answering the research question(s). The following are the study's findings under specific linguistic features common in Bantu languages.

Word Order

Word order is a syntactic feature linguist worldwide have studied in languages. Established in these studies, words follow particular patterns in the formation of sentences, and these patterns must adhere to them, or else the sentence crushes at the logic level – at the level of interpretation. In most cases, word order refers specifically to the order of subject, object and verb (Dryer, 2000; Barasa, 2022). In most cases, these arguments are by affixes (Guthrie, 1967; Sande, 2019 &Waweru, 2011). Therefore, word order in Bantu languages can be viewed as a sequence of argument-related affixes concerning the host verb. The default order of arguments concerning host verbs in Bantu languages is given by Derek and Gérard (2014) as shown in the formula S (Aux) VO (Adjuncts). Consider data from the three languages justifying Derek and Gérard's argument:

Gikuyu

1 (a) *ndĩragũrire thimũ*
N-ra-gũr-ir-e thimũ
1sg-NR.PST-buy-COMPL-FV phone
'I bought a phone (recently)'
(SVO)
(Englebretson, 2015: 124)

(b) *ũramũmatwarĩire*
ũ-ra-mũ-ma-twar-ĩ-ir-e
2sg-NR.PST-3Osg-3Opl-take-APPL-COMPL-FV
'You took him/her to them'
(SVOO)
(Englebretson, 2015: 124)

Olusuba

2 (a) *aagala omuoyo*
a-agal-a o-mu-oyo
1sg-want-FV AUG-3-soul
'He/she wants a soul'
(SVO)
(Sande, 2019: 80)

(b) *afuumbira owusera*
a-fuumb-ir-a o-wu-sera
1sg-cook-APPL-2Osg AUG-14-porridge
I cook for myself porridge
(SVOO)
(Sande, 2019: 108)

Ekegusii

3. *agosoma ebibilia*
a-go-som-a e-ø-bibilia
3sg-NR.PST-read-FV AUG-9-bible
'He/she read the bible'
(SVO)
(Basweti, 2005: 33)

The data above shows that the three Bantu languages have the common SVO word order. This justifies the universality of the SVO word order pattern. In analysing the word order in Bantu languages, a

researcher may use SVO as a canonical pattern–any word order patterns are treated as a transformation of the canonical pattern. Therefore, in establishing unique syntactic structures concerning word order patterns, a researcher is expected to use SVO as a basis of analysis. The examples provided in the three languages provide researchers with the corpus constructs they can consult in their study of any Bantu languages, especially in the syntactic behaviour of affixes in those languages.

Noun Classes

Classifying nouns into their respective classes is one of the most common features in Bantu languages. The classes are differentiated by noun class prefixes (Demuth, 1988; Denny &Creider, 1976). Derek and Gérard (2014) observed that Noun class prefixes are an extensive system of concord. That means that the prefixation of words, both complements and specifiers, in Bantu languages follows patterns in concordance with the class noun prefixes. Established from the samples from the three Bantu languages is the prefixal marking of noun classes. Each class is identified by a specific prefix. In Olusuba, as found by Sande (2019), 20 classes of nouns exist; Gikuyu has 17 noun classes (Englebretson, 2015), and Ekegusii as well as 17 noun classes (Basweti, 2005). These class markers prefixes take their specific slots within the matrix of a noun. Provided is the default structure of nouns in Bantu languages: Augment + Noun Class Prefix + Root (Guthrie, 1967; Meeussen, 1967 &Meinhof, 1932). Any other structure of a noun different from the default structure is considered a derivative of the default. Data given in this study provides a researcher interested in any Bantu language with an insight into the morphophonological behaviour of nouns from any Bantu language. See the data given from the three languages:

Olusuba

4 (a) *omugaka*
o-mu-gaka
AUG-1-parent
'Parent'
(Sande, 2019: 108)

(b) *awagaka*
a-wa-gaka
AUG-2-parent
'Parents'
(Sande, 2019: 108)

(c) *omutwe*
o-mu-twe
AUG-3-head
'Head'
(Sande, 2019: 108)

(d) *emitwe*
e-mi-twe
AUG-4-head
'Heads'
(Sande, 2019: 108)

Gikuyu

5 (a) *Mũndũ*
Mũ-ndũ
1-person
'Person'
(Englebretson, 2015: 18)

(b) *andũ*
a-ndũ
2-person
'Persons'
(Englebretson, 2015: 18)

(c) *muoyo*
mu-oyo
3-heart
'Heart'
(Englebretson, 2015: 18)

(d) *mioyo*
mi-oyo
4-heart
'Hearts'
(Englebretson, 2015: 18)

Ekegusii

6 (a) *omomura*
o-mo-mura
AUG-1-boy
'Boy'
(Basweti, 2005: 20)

(b) *abamura*
a-βa-mura
AUG-2-boy
'Boys'
(Basweti, 2005: 20)

(c) *eriiso*
e-ri-iso
AUG-3-eye
'Eye'
(Basweti, 2005: 21)

(d) *amariso*
a-ma-iso
AUG-4-eye
'Eyes'
(Ombati, 2005: 21)

From the data shown above, it can be concluded that noun classes in Bantu languages are morphologically identified by prefixes whose slots in the matrix of the noun are strictly immediately before the root.

Noun class, as observed in the three Bantu languages, was also discovered to be influential in the morphological shaping of other items, particularly verbs and adjectives, that may form part of the sentence or the phrase a noun is a component of. Therefore, any researcher interested in the study of grammar and morphophonology of Bantu languages may consult corpus constructs from the three languages as these may enable the researcher to predict possible patterns of affixes within the structure of a noun. See the examples given:

Olusuba

7 (a) *enyundo ewiri*
e-ny-iindo e-wiri
AUG-9-nose AUG-two
'Two noses'
(Sande, 2019: 124)

(b) *iriiso rino*
i-ri-isori-no
AUG-5a-bottle 5a-this
'This eye (proximal)'
(Sande, 2019: 124)

(c) *awaana wano waria*
a-wa-anawa-no wa-ri-a
AUG-2-child 2-this 2-eat-FV
'These children eat'
(Sande, 2019: 128)

Gikuyu

8 (a) *rĩithorĩu*
Rĩ-ithorĩ-u
5-eye 5-that
'That eye'
(Englebretson, 2015: 59)

(b) *kacũcũ kaathugumĩra*
ka-cũcũ ka-athugumĩra
12-grandchild 12-young
'Young grandchild'
(Englebretson, 2015: 59)

(c) *ũrĩanĩmũathĩki*
ũ-rĩanĩmũ-athĩki
1-DEM (Distal) COP 1-obedience
'That one is obedient'
(Englebretson, 2015: 58)

Ekegusii

9 (a) *ebinto ebio*
e-bi-nto e-bio
AUG-8-thing AUG-this
'Those things'
(Basweti, Achola, Barasa & Muchira, 2015: 98)

(b) *abanto abange*
a-ba-nto a-ba-nge
AUG-2-person AUG-2-quantifier
'Many persons/people'
(Basweti, Achola, Barasa & Muchira, 2015: 98)

(c) *emeteyare konyogera*
o-mo-te o-re ko-nyeger-a
AUG-4-tree Aux(PST) PROG-sway-FV
'Trees were swaying'
(Basweti, 2005: 33)

Verbal Structure

As agglutinative languages, Bantu languages have verbs that can subcategorise for several grammatical features, including person, number, tense, mood, extension and aspect. The majority of these features are marked by morphemes affixed to the root. According to Meeussen (1967), Bantu verbal words have a similar structure whose default is presented as:

Initial – Negative -Subject– Tense/Aspect – Object ≠ Root – Extension – Final Suffix

The data provided in this study are the derivatives of the default structure given above.

Olusuba

10 (a) *ngatu mukubiranga*
nga-tu-ø-mu-kub-ir-ang-a
NEG-1Spl-PRES-2Opl-kick-APPL-PROG-IND
'We are not kicking for you'
(Sande, 2019: 143)

(b) *tukasawa*
tu-ka-saw-a
1Spl-PST(Remote)-pray-IND
'We prayed (long ago)'
(Sande, 2019: 149)

Gikuyu

11 (a) *nĩaaiyithia*
nĩ-a-a-iy-ithi-a
FOC-3Ssg-PST-steal-CAUS-FV
'Caused to steal from'
(Englebretson, 2015: 91)

(b) *mburanĩyaura*
ø-mburanĩ-ĩ-a-ur-a
14-rain FOC-3S-PST(Near)-rain
'Rain has fallen'
(Englebretson, 2015: 97)

Ekegusii
12 (a) *omoite*
o-mo-it-e
2Ssg-3Osg-beat-PST(Near)
'You beat him/her'
(Basweti, 2005: 33)

(b) *bamora mire*
βa-mo-ram-i-re
3Spl-3Osg-abuse-PST(Near)-ASP(Perfective)
'They have abused him/her'
(Basweti, 2005: 33)

As shown in the data above, Bantu languages' verbal matrix is a composition of affixes, each assigned grammatical role(s) by the verb roots that host them – each affix is slotted in a specific position within the matrix of the verbals. This aspect of subcategorisation for grammatical feature markers by the verb root is common in the Bantu language. Therefore, corpora like the ones provided in this study may be used as a point of departure for such studies in the study of the verb structure of any Bantu language.

Discussion

The development of a corpus-based research tool in studying the syntax of a Bantu language requires the establishment of research gap(s) in a study. From the research gap, research question(s) are formulated to generate syntactic possibilities, thus spelling out paradigms of a syntactic structure. Various corpus constructs can be developed depending on what syntactic feature a researcher aims to investigate in a Bantu language. This construction is done using samples from languages. It is from these samples that generalisations are made. Sometimes, the gathered samples from a language, according to Zipf's law, have few occurrences in a corpus (Meurers & Müller (2009). According to the understanding of this article, fewness occurrences in a corpus should not be an issue given the similarity in structure of most Bantu languages. Not all syntax features are addressed in this study, but a few basic ones are common across Bantu languages.

In terms of word order, the patterning of words, according to McCarthy (1997), is governed by syntactic constraints whose ranking is language based. Through the interaction of these syntactic constraints, a sentence is formed. The sentences formed from the corpus constructed from the three languages of reference display definable word orders, either canonical or non-canonical. In constructing a corpus for the study of word order in Bantu languages, the researcher is first advised to formulate research question(s) that will generate a canonical word and later provide other non-canonical word orders whose patterning may be language specific. Through search engines and with the guide by the research question(s), the canonical word order is identified and given annotation through tagging. The annotation includes coding the subject argument and object argument markers in a sentence, which, in Bantu languages, are mostly marked by affixes within a verbal matrix. The positioning of the argument markers concerning the default order of constituents across sentences in the text gives information on the possible derivatives of the canonical word order.

It is important to note that tagging as a codification process must be done so that the code used gives the user extra information on word orders. In cases of non-canonical word orders, the researcher is expected to run the engine on the text and tag words with codes to identify the classes of the tagged words. After the coding, the researcher is advised to pick a word via corpus software, run over the text and identify the number of times the picked word has been used in sentences and its positional changes against its immediate constituents in the same sentences. This may allow for the restructuring of word order from the canonical word order. Corpus constructed on word order can be used in studying simple and complex sentence structure of Bantu languages, the system of concord displayed by Bantu languages, and word or affix order as a study topic.

From the three samples of Bantu languages, it is important to note that specific class noun prefixes are also identified and compiled into larger corpora representing a subset of the Bantu language's potential through annotation and segmentation. From the established class marker prefixes, more tagging is done with codes which provide the user of the corpus with more information on relationships among classes of nouns with information on grammatical constraints on the prefixal changes in Bantu noun classes; with knowledge on how a noun class in a Bantu language controls morphological shapes of words in the sentence within which that class of noun forms part of. These pieces of information on structural constraints are used to spell out paradigms of the noun classes against the basic classes. Constructs from noun classes can be used in the study of morphological features of noun classes of Bantu languages; agreement in Bantu language sentences; concord patterns in sentences of Bantu languages; number marking in sentences from Bantu languages and inflectional patterns of words in sentences in Bantu languages.

The first step in developing corpora for the study of verbal structure is identifying verbal words using search engines. The identified verbal words are then tagged using codes that differentiate the verbal words from other possible words within. Once the researcher identifies and tags the verbal words, the next step is the establishment of the possible morphemes the identified words are likely to contain. This is done against the default structure. At this point, the researcher will be required to mark the features within the default structure that each of the tagged words contains, giving them different codes based on the grammatical functions of the features. In cases where feature(s) in the default are not represented in the tagged word, the researcher is advised to put a dash at the space slotted for the absent feature. This is done to take care of the principle of greed by Chomsky (1995), who states that every candidate in a sentence formation must guard its slot with greed, or else the sentence structure crashes at the LF level. Therefore, the order of features in the default structure should not interfere with whether the feature is present in the tagged word. This will also guide future researchers using the compiled corpora in the syntactic analysis of any Bantu languages worldwide. Constituting the constructs, therefore, were the spellouts of the paradigms of the verbal structures. Corpus constructs on verbal structures in Bantu languages are very important in the study of areas such as verbal extensions; tense, mood and aspect features; morphology of verbals in a Bantu language; transformation in sentences; movement as a syntactic process in Bantu languages; and negation in Bantu languages.

Conclusion

Bantu languages are some of the world's languages, and they are unannotated compilations of electronic corpora. This article has addressed scenarios where data exist in form of plain texts. Discovered in this article is that from plain text, a corpus can be compiled within a line of syntactic features. Compilation involves using a search engine for relevant corpora for an intended construct; the relevant corpora are annotated and/or segmented through tagging and codification. After

identifying a default structure, the search can only be done in Bantu languages. It is from the default structure that the possible derivatives are identified and described.

This article also establishes that a corpus construct can be used to study more than one syntactic feature of a Bantu language. Important in such studies is a researcher's good formulation of research question(s) and understanding of what exactly the question(s) need(s). With the researcher's formation of the research question(s) and understanding of the question(s), even an already constructed corpus can be improved to take care of the need(s) of the research. It is worth noting that Bantu languages have common syntax features. These features may be common in terms of their ordering in a word, morphological markers, patterns of movement, especially in the formation of transforms from kernels, concordance in forming sentences, and many others. Based on the commonness of these features across Bantu languages, findings on one Bantu language can be used to predict the syntactic behaviours of any Bantu language.

References

Adamou, E. (2019). Corpus linguistics methods. In J. Darquennes; J.C. Solomons; W. Vandenbuss-che. *Language contact: An International Handbook*. De Gruyter, 638-653, 2019, Handbooks of Linguistics and Communication Science series (HSK).

Baker, P. (2009). *Contemporary Corpus Linguistics.* Continuum.

Barasa, D. (2022). Pronouns and pronominal alignment in Ateso. *Arusha Working Papers in African Linguistics, 4*(1), 100-114.

Basweti, N. O. (2005). A Morphosyntactic Analysis of Agreement in Ekegusii in the Minimalist Program. MA. Dissertation.

Basweti, N. O., Achola, E. A., Barasa, D., & Michira, J. N. (2015). Ekegusii DP and its Sentential Symmetry: A Minimalist Inquiry. *International Journal of Language and Linguistics, 2(2),* 93-107.

Beck, J. (2023). Quality aspects of annotated data: A research synthesis. *Springer Nature Switzerland*, 1-23. https://doi.org/10.1007/s11943-023-00332-y.

Chomsky, N. (1995). *The Minimalist Program.* MIT Press.

Creswell, J. W., & Creswell, J. D. (2018). *Research Design: qualitative, quantitative, and mixed methods (5th ed.).* SAGE.

Demuth, K. (1988). Noun classes and agreement in Sesotho acquisition. In: M. Barlow and C. Ferguson (eds), *Agreement in natural language: approaches theories and descriptions.* University of Chicago Press.

Denny, J. P., & Creider, C. (1976). The semantics of noun classes in Proto-Bantu. In: C. Craig (ed.), *Noun classes categorisation*. Benjamins.

Derek, N., & Gérard, P. (2014). *The Bantu Languages. (2nd ed.).* Routledge.

Devitt, M. (2006). Intuitions in Linguistics. *The British Journal for the Philosophy of Science, 57(3),* 481-513. https://wwwjstor.org/stable/3873480.

Dryer, M. S. (2000). Word Order. *Shopen Anthology* (2nd ed.).

Englebretson, R. (Ed.). (2015). A Basic Sketch Grammar of Gĩkũyũ. *A Special Issue of Rice Working Papers in Linguistics*. https://creativecommons.org/licences/by/3.0/us/.

Fancom, J. (2021). *Corpus Studies of Syntax.* Cambridge University Press.

Guthrie, M. (1967). *Comparative Bantu: an introduction to the comparative linguistics and prehistory of the Bantu languages*. Gregg Press.

Harford, C., & Malambe, G. (2017). An Optimality Theoretic Perspective on Perfective Imbrication in siSwati. *Nordic Journal of African Studies, 26(4)*, 277-291.

Hassel, M. (2001). Internet as Corpus: Automatic Construction of a Swedish News Corpus. *Online Proceedings of NODALIDA 2001.*

Himmelmann, N. P. (2012). Linguistic Data Types and the Interface between Language Documentation and Description. *Language Documentation and Conservation, 6,* 187-207.

Joâo, C. (1996). *Word Order and Constraint Interaction*. Holland Institute of Generative Linguistics.

Just, E., & Witzlack-Makarevich, A. (2022). A corpus-based analysis of P indexing in Ruuli (Bantu, JE103). South African Journal of African Languages, 42(2), 234-242.

Kager, R. (1999). *Optimality Theory.* Cambridge University Press.

Kula, N. (2002). *The phonology of verbal derivation in Bemba*. LOT.

Lehmann, C. (2004). Data in Linguistics. *The Linguistic Review 21*(3/4), 275-310. https://www.christianlehmann_data_in_linguistics .

McCarthy, J. J. (1997). *Process-specific constraints in Optimality Theory*. Linguistic Department Faculty Publication Series. University of Massachusetts, Amherst.

Meeussen, A. E. (1967). Bantu grammatical reconstructions. *Africana Linguistica, 3*, 79-121.

Meinhof, C. (1932). *Introduction to phonology of Bantu languages*. Reimer, Vohsen.

Meurers, W. D., & Müller, S. (2009). Corpora and syntax. In AnkeLüdeling&MerjaKytö (eds.), *Corpus linguistics*, vol. 2 (Handbooks of Linguistics and Communication Science 29), 920–933. Berlin; New York: De Gruyter Mouton.

Parvess, J. (2023). BantuBERTa: Using Language Family Grouping in Multilingual Language Modeling for Bantu Languages. Unpublished MA Dissertation.

Prince, A., & Smolensky, P. (1993). *Optimality Theory: Constraint Interaction in Generative Grammar*. Blackwell.

Prince, A., & Smolensky, P. (2004). *Optimality Theory. Constraint Interaction in Generative Grammar*. Blackwell.

Podesva, R. J., & Zsiga, E. (2013). Sound recordings: acoustic and articulatory data. *Research Methods in Linguistics,* 169-194.

Sande, W. (2019). A Description of Olusuba Morphophonology: Towards Preservation of an Endangered Language. Unpublished PhD Thesis.

Schleef, E., & Meyerhoff, M. (2010). *Sociolinguistic methods for data collection and interpretation*. Routledge.

Sigurðsson, H. A. (2003). The Silence Principle. In L.-O. Delsing, C. Falk, G. Josefsson, & H. Á. Sigurðsson (Eds.), *Grammar in focus: festschrift for Christer Platzack 18 November 2003, 2,* 325-334.

Taylor, C. (2008). What is corpus linguistics? What the data says. *ICAME Journal, 32,* 179-200.

Wasow, T., & Arnold, J. (2004). Intuition in linguistic argumentation. *Lingua 115*, 1481-1496.

Waweru, M. M. (2011). Gĩkũyũ Verbal Extensions: A Minimalist Analysis. Unpublished PhD Dissertation.