A functional approach to differential indexing: combining perspectives from typology and corpus linguistics

Just, E.C.

Chapter 3

A corpus-based analysis of P indexing in Ruuli[1]

Abstract

Verbs in Bantu languages usually carry an obligatory subject (or S/A) prefix, whereas the presence of transitive object (or P) prefixes depends on various language-specific factors. A number of such factors is well described in a range of studies mainly based on elicited data. In order to examine their interplay in naturalistic texts, we conducted a corpus-based case study of object prefixes (or P indexing in the terminology used in this chapter) in the Bantu language Ruuli (JE103). The corpus of over 15,000 words was annotated for variables such as animacy, identifiability, and textual givenness. The statistically relevant factors for triggering P indexing were identified using conditional inference trees. Unsurprisingly, the results show that the strongest predictor for P indexing in Ruuli is word order. Just as P indexing itself, we assume that word order is a differential pattern expressing the argument's semantic and pragmatic properties. Taking only the latter into account, the analyses reveal that firstly, P indexing seems to be strongly predictable by textual givenness. Secondly, if the referent is given, the probability that it gets indexed is significantly higher if it is human.

[1] This chapter is accepted for publication as: Just, Erika & Alena Witzlack-Makrevich. A corpus-based analysis of P indexing in Ruuli (Bantu, JE103). *South African Journal of African Languages*. Author contributions: EJ and AWM wrote the paper and developed the annotation scheme; EJ summarized previous research and carried out the larger part of the annotation; EJ and AWM performed the statistical analysis.

3.1 Introduction

The topic of this study is P indexing in Ruuli. The phenomenon is known under a range of labels and an in-depth discussion of the notions of P and indexing and our motivation for the use of these labels is provided in Section 3.2. The phenomenon can be exemplified in (3). In clauses in (3a) and (3b), there is an object agreement prefix on the verb, namely *bu-* '14.OBJ', whereas there is no object agreement prefix on the verb in (3c).[2]

(3) Ruuli (Bantu, Uganda, Witzlack-Makarevich et al. 2019)

a. *Obuterega o-bu-maite?*
14.trap 2SG.SBJ-14.OBJ-know.PFV
'Do you know these traps?'

b. *N-bu-maite.*
1SG.SBJ-14.OBJ-know.PFV
'I know them.'

c. *N-a-tung-ire omukali wa-ange.*
1SG.SBJ-PST-marry-PFV 1.woman 1-1SG.POSS
'I married my wife.'

In this chapter, the alternation as in (3) is treated as a case of differential argument marking, i.e. a situation where an argument of a predicate with the same semantic argument role (here patient) is coded differently (Witzlack-Makarevich & Seržant 2018). In the present case, as in many Bantu languages with similar systems, an object prefix can occur on the verb and its presence is determined by certain referential properties of the respective argument, among other conditions (see e.g. Duranti 1979, Morimoto 2002, Ngonyani & Githinji 2006, and Marten & Kula 2012 for some comparative studies). The aim of this study is to identify and quantitatively analyze those properties of arguments which condition the presence of object prefixes in Ruuli. This study is based on a corpus of spoken data and is thus one of the first investigations of the phenomenon in Bantu languages from the corpus-linguistic perspective and on the basis of spoken language data.

The chapter proceeds as follows: After some theoretical preliminaries in Section 3.2, we provide some insight into how the topic of differential P mark-

[2] The glosses follow the Leipzig Glossing Rules, additional abbreviations are as follows: ADD.FOC = additive focus; CJ = conjoint verb form; LOC = locative

ing in Bantu languages has been dealt with in the literature (Section 3.3). Section 3.4 briefly presents the language of the study. Section 3.5 proceeds with our analysis of P indexing in Ruuli. First, we discuss the corpus annotation and the variables we use (Section 4.3.3), we then present the variables that condition P indexing and show how they relate to each other, using conditional inference. Section 6.6 concludes the chapter and discusses further research prospects.

3.2 Terminological considerations

The topic of this study is P indexing. Before we proceed with the study, we first outline how we understand the P argument and indexing and why we prefer these notions over the label object prefix used in Section 6.1 as well as over alternative labels, such as object marking, object or object pronominal agreement commonly used in the literature.

Though yet uncommon in studies of Bantu languages, the terms S, A, and P have been extensively used since the 1970s by comparative and descriptive linguists to compare grammatical relations across languages and describe the properties of verbal arguments in individual languages (see Haspelmath 2011 for an overview of the history of these terms). The major reason for adopting these terms are the various challenges the traditional terms of subject and (direct) object face (see e.g. Witzlack-Makarevich 2019 for an overview). On the one hand, various criteria of subjecthood and objecthood often provide conflicting evidence as to what the 'real' subject or direct object in a language is. On the other hand, traditional grammatical relations are typically identified on the basis of language-specific constructions, i.e. on the basis of different criteria in different languages, and thus suffer from what is called 'methodological opportunism' (Croft 2001: 30).

These kinds of challenges are not uncommon in studies of Bantu languages. On the one hand, some studies have challenged the validity of the notions of subject and direct object for languages of the family, highlighting that not grammatical relations but rather discourse and the pragmatic status of a referent is the most crucial factor in encoding relations via indexing, word order or prosodic features (e.g. Morimoto 2006, Zerbian 2006, Zeller 2008). On the other hand, there are a number of constructions which involve a mismatch between the morphosyntactic behavior of an argument and their semantic role. Among them are the various ditransitive or 'double object' constructions, as well as inversion constructions (see Downing & Marten 2019 for an overview). For instance, Bantu inversion constructions are characterized by the deviation from the prototypical word order with an agent following the verb instead of preceding it, as subjects are expected to do. Furthermore, unlike in the case of typical subjects, these constructions either lack the indexing of the agent on the verb or use expletive indexing.

In light of the above, in the remainder of this paper, we use the term P instead of (direct or transitive) object.[3] We follow Bickel & Nichols (2009) and Witzlack-Makarevich (2019) and understand P as the generalized semantic role of the less agent-like argument of a two-place predicate. Likewise, we use the term S and A to refer to the sole arguments of one-place predicates and to the more agent-like argument of two-place predicates, respectively.

The other terminological convention we follow in this chapter is the use of the terms index and indexing. What motivates this choice? Since at least Bresnan & Mchombo (1987), the status of Bantu object prefixes on the verb as either (incorporated) anaphoric pronouns or (grammatical) agreement markers has received considerable attention and is still a highly contested topic (see Downing & Marten 2019: 278–280 for an overview, Sikuku et al. 2018 for a recent contribution on the topic, see also Creissels 2005: 44–45 for a diachronically-motivated typology of the phenomenon). To avoid committing ourselves to any assumptions concerning the status of the object prefixes as either pronouns or agreement markers, we use the term index (Haspelmath 2013) for any bound markers expressing argument features and attached to the verbal predicate. Indexing is a more neutral term than agreement, as it does not presuppose any syntactic relationship between the marker and the referential NP (Haspelmath 2013). Thus, this concept is detached from the notion of syntactic obligatoriness and the morphological status of the index as either a clitic or an affix.

After the introduction of the terminological framework adopted in this chapter, in Section 3.3 we proceed with the discussion of various approaches

[3] For the sake of readability and comparability with other studies on Bantu argument prefixes, we keep the glosses SBJ and OBJ in the examples, though the use of glosses S/A and P would be more consistent with the terminology adopted here

and explanations to the interaction of the P indexing, word order and referential properties of P in Bantu languages.

3.3 Variation in P indexing in other Bantu languages

A considerable amount of literature has been published on the conditions of P indexing in individual Bantu languages (e.g. Buell 2005 on Zulu, Riedel 2009 on Haya and Sambaa, Downing 2018 on Chewa, Sikuku et al. 2018 on Lubukusu). Only in exceptional cases (most notably, Seidl & Dimitriadis 1997 on Swahili) are these studies corpus-driven (in the sense of e.g. Tognini-Bonelli 2001: 84–85). In fact, Bantu corpus linguistics has only gradually arisen over the course of the last twenty-five years (cf. Kawalya et al. 2014: 61-63, Nabirye 2016). Therefore, the previous analyses of the phenomenon have been largely based on elicited material.

Many of the in-depth descriptions of the phenomenon claim that there are rules that license the co-occurrence of the P index and the respective NP. This might hold for a number of languages, as for instance for Makhuwa. In this language, there are no object indexes except for first and second person and nouns belonging to class 1 and 2. The latter always have to be indexed, irrespective of any referential features of P, its semantics or information structural conditions (van der Wal 2009: 80–85):

(4) Makhuwa-Enahara (Bantu, Mozambique, van der Wal 2009: 84–85)

a. _Ki-ni-ḿ-weha_ _Hamisi/namarokolo/nancoolo?_
1SG.SBJ-1.OBJ-PRS.CJ-1-look 1.Hamisi/1.hare/1.fish.hook

‘I see Hamisi/the hare/the fish hook.’

b. _*Ki-m-weha_ _Hamisi/namarokolo/nancoolo_
1SG.SBJ-PRS.CJ-1-look 1.Hamisi/1.hare/1.fish.hook

‘I see Hamisi/the hare/the fish hook.’

c. _Ki-m-weha_ _nvelo/mikhora/kalapinteero/etthepo_
1SG.SBJ-PRS.CJ-look 3.broom/4.doors/5.carpenter/9.elephant

‘I see the broom/doors/carpenter/elephant.’

d. _*Ki-ni-ḿ-wéham_ _nveló/mikhorá/kalapinteéro/etthepó_
1SG.SBJ-1.OBJ-PRS.CJ-look 3.broom/4.doors/5.carpenter/9.elephant

Int: ‘I see the broom/doors/carpenter/elephant.’

As (4a) and (4b) illustrate, nouns belonging to noun class 1 are obligatorily indexed, whereas nouns belonging to other classes cannot be indexed, as in (4c and (4d). Thus, the constraints on P indexing in Makhuwa seem to be purely formal in nature. In other languages of the family, P indexing is licensed by the inherent semantic properties of the referent. For instance, Riedel (2009) demonstrates that in Sambaa, P indexing is in part determined by the animacy hierarchy: it is obligatory for proper names, titles and first and second person referents. It is commonly used with other types of humans, less common with other animates, and rare (but acceptable) with inanimates.

(5) Sambaa (Bantu, Tanzania, Riedel 2009: 45–46)

a. *N-za-mw-ona askofu.*
1SG.S/A-PFV-1.P-see 5.bishop
'I saw the bishop.'

b. **N-za-ona askofu.*
1SG.S/A-PFV-see 5.bishop
Int: 'I saw the bishop.'

c. *N-za-(ji-)ona kui.*
1SG.S/A-PFV-(5.P-)see 5.dog
'I saw the dog.'

d. *N-za-(chi-)ona kitezu.*
1SG.S/A-PFV-(7.P-)see 7.basket
'I saw the basket.'

Riedel (2009) also shows that even in Bantu languages where P indexing is described as obligatory, this obligatoriness is rarely absolute: actually, P indexing in individual languages ranges from obligatory (for certain kinds of referents) to optional (for another group of referents) to ungrammatical (for all remaining P referents). This variation depends on the referent's position on the animacy and definiteness hierarchy (see e.g. Dixon 1979: 85 for a commonly-cited example). The cut-off points within the hierarchies are language specific. Marten & Kula (2012) show in their comparative study of morphosyntactic variation in object marking in Bantu languages that there is a great deal of diversity with regard to the semantic factors that trigger obligatory P indexing.

For several Bantu languages, P indexing is described as depending on the referent's topicworthiness (see Downing 2018: 43–45 for an overview).

In other words, P indexing is often syntactically optional and associated with the pragmatic status of the referent as the topic of the utterance. The P index can be reinterpreted as marking topicworthiness instead of topichood, i.e. it can be sensitive to semantic and/or pragmatic features, such as humanness or definiteness, which are commonly associated with high topicality.

For a number of other Bantu languages P indexing alongside an overt NP figures as one feature in a bundle of structural components such as (non-canonical) word order, disjoint verb forms or intonational cues of dislocation, used to express topicality of a referent (cf. e.g. Bresnan & Mchombo 1987 on Chichewa, Ngoboka & Zeller 2017 on Kinyarwanda or Zerbian 2006 on Northern Sotho). In the Bantu languages that are described to pattern like this, "the same entity is represented by a pronominal marker or by a noun phrase depending on its degree of topicality and recoverability from the context, and the pronominal marker cooccurs with the corresponding noun phrase only if the noun phrase is topicalized in a dislocated construction" (Creissels 2005: 2).

For Chichewa, it has long been claimed that P indexing fulfills a purely resumptive function, and that it always comes along with non-canonical word order and dislocation, to express the topicality of the referent, irrespective of its semantics (Bresnan & Mchombo 1987). However, Downing's (2018) study on modern spoken Chichewa reveals that all the diagnostics for the anaphoricity of the index can be disproven for cases where the referent is human. The study shows that there is a marking asymmetry with respect to P arguments, with humanness being more crucial for indexing than the constituent order. The following sentences in (6), which do not have a prosodic break between the verb and the P NP *aleenje* (2.hunter), were analyzed as being ungrammatical by Bresnan & Mchombo (1987). The same sentences are grammatical in Downing's (2018) re-elicited data, both with and without a prosodic break. She concludes that the P index in Chichewa is a marker for topicworthiness rather than topichood (cf. Dalrymple & Nikolaeva 2011: 51–57).

(6) Chichewa (Bantu, Malawi, Downing 2018: 48, re-elicited from Bresnan & Mchombo 1987)

a. *Njúuchí zi-na-wá-lúma aleenje.*
10.bee 10.SBJ-PST-2.OBJ-bite 2.hunter
'The bees bit the hunters.'

b. *Zi-na-wá-lúma aleenje njúuchi.*
10.SBJ-PST-2.OBJ-bite 2.hunter 10.bee
'The bees bit the hunters.'

As has been shown by comparative studies, Bantu languages attest a lot of variation with respect to features licensing differential indexing (Riedel 2009, Marten & Kula 2012). Taking the relevant factors of P indexing identified in other studies as a point of departure, our study aims at revealing which of these factors have the strongest association with P indexing in Ruuli, the language of our case study. The next section introduces briefly the language of the study and its relevant morphosyntactic properties before turning to the description of our corpus annotation and its statistical evaluation in Section 3.5 with the goal of gaining deeper insights into the interplay of the relevant variables.

3.4 Language background

Ruuli (ISO 639-3: ruc, also known as Ruruuli-Lunyala) is a Great Lakes Bantu Language, spoken in Uganda in the districts of Nakasongola and Kayunga in the area around Lake Kyoga. It is the language of the Baruuli and the Banyara people. The ethnic groups of the Baruuli and Banyala are estimated to be about 160,000 (140,000 Baruuli, 21,000 Banyala; Uganda Bureau of Statistics 2016). Two main varieties can be distinguished. Until recently, this mainly orally used language has been undescribed. Only recently, the language came into focus of an ongoing documentation project, which resulted in several publications including Namyalo et al. (2021). The compilation of the corpus of primarily naturalistic spoken Ruuli is currently in progress. As of 2020, this corpus consists of 200,000 words and serves as the database for the present study.

Ruuli is a typical Bantu language. The dominant constituent order with transitive verbs is AVP. Each noun in singular and plural belongs to one of the 21 noun classes which are numbered in correspondence to the reconstructed Proto-Bantu noun classes (Van de Velde 2019: 238–241). Ruuli does not have the correspondences of the noun classes 19, and 21. The nominal prefixes on the nouns are not segmented in the examples, the gloss indicates the class followed by a fullstop before the respective noun gloss, as e.g. in *obuterega*

‘14.trap’ in (3a) above. Ruuli nouns regularly carry an augment, also known as pre-prefix or initial vowel (cf. Van de Velde 2019: 247). The augment appears before the noun class prefix and has the forms a-, o-, or e-, determined by the vowel of the noun class prefix. The augment in Ruuli is not determinative (cf. Blois 1970: 152) and there is no correlation between its presence and an index on the verb. It is neither segmented nor glossed in the examples in this paper for the sake of space, as e.g. the augment o- in *obuterega* in (3a). Like many other Bantu languages, Ruuli is a tonal language. Currently, the Ruuli tone is still under investigation and the examples in this chapter are provided following the practical orthography, which does not indicate tone. The way the research question is operationalized in this study, tone is not relevant for the present analysis, though tone and more generally prosody are invoked in arguing for the dislocated status of some P arguments (see Section 3.3). The simplified structure of the finite verb in Ruuli is given in (Table 5.2). Arguments are indexed in the obligatorily filled S/A (or subject) position, and the optionally filled P (or transitive object) position. Tense and aspect categories are expressed as either prefixes or suffixes.[4] The scheme in (5) does not list the extensions (passive, applicative, causative, reciprocal and reflexive, which all follow the root except for the reflexive). The final vowel -a, which follows the verb stem unless there is the subjunctive suffix -e or the perfective suffix -ire, is neither segmented not glossed in the examples below.

TA1-	(NEG-)	S/A-	(NEG2-)	TA2-	(P-)	ROOT	(-TA3)	-FINAL	-POST-FINAL

Table 3.1: The morphological structure of the Ruuli verb

The indexes on the verb always correspond to the noun class of the argument, i.e. there is never a mismatch, such as the one found in some languages of the family, for instance in Sambaa, displayed in (5a), where the noun *askofu* ‘bishop’ belongs to noun class 5, but triggers an index of noun class 1, which is the noun class usually reserved for human beings in singular. As the examples in (3) show, to express the P argument in Ruuli there can be an index alongside an overt P NP, such as *obuterega* ‘14.trap’ in (3a), an index only, as

[4] NEG1 is the standard negation prefix, while NEG2 is only used in prohibitive, and negative hortative and jussive constructions. The post-final slot can only be occupied by the habitual suffix.

in (3b), or a NP only, as in (3c). P indexing in Ruuli is realized via a verbal prefix, which expresses referent features, specifically, noun class in case of third person referents and person and number in case of first and second person. In the dominant constituent order, P follows the predicate, as in (7a) and (7b), but the inverse order is also possible, as in (7c) and (7d).

(7) Ruuli (Witzlack-Makarevich et al. 2019)

a. *Iswe tu-li-ire bunyonyi na obusolo.*
1PL 1PL.SBJ-eat-PFV 14.bird COM 14.animal
'As for us, we have eaten birds and animals.'

b. *Ni-a-ba-iryaku abairaange*
NAR-1.SBJ-2.OBJ-marry.again 2.friend:1SG.POSS
'and he took my friends as other wives.'

c. *Amaani mu-ta-ire = mu.*
16.strength 2PL.SBJ-put-PFV = 18.LOC
'You have put in a lot of strength.'

d. *Naye nje eisumu n-a-li-zw-ire = ku.*
but 1SG 5.spear 1SG.SBJ-PFV-5.OBJ-abandon-PFV = 17.LOC
'But as for me, I abandoned the spear.'

If an overt P argument follows a verb with a P index, it is separated by a pause,[5] in elicited examples as well as in corpus examples. However, there are no phonological cues separating preverbal P NPs (neither pause nor penultimate lengthening).[6] There are, however, syntactical cues for structural difference of preverbal P: the A argument can intervene between the P argument and the verb which carries a P index, as in (8):

(8) *Obwo-te njee-na ti-n-bu-maite leero.*
14.DEM-FOC 1SG.PRO-ADD.FOC NEG-1SG.SBJ-14OBJ-know.PFV today
'I also don't know them this time.'

The fact that indexed P NPs following the verb are separated from the rest of the clause by a pause, whereas those preceding the verb are not speaks for an asymmetrical phrasing pattern in Downing's (2011) typology of prosodic

[5] The preliminary study of the relationships between intonational and syntactic phrasing in Ruuli in Zellers et al. (2020) does not differentiate between phrases with unindexed and infrequent P arguments

[6] Lengthening of a phrase penult vowel is a common salient cue to prosodic phrasing in Bantu (cf. Downing 2011).

phrasing in Bantu dislocation: Only right dislocation is phrased separately in Ruuli, the correlate for dislocation being a pause.

The examples in (7) show that P can be indexed or not, i.e. differentially marked, in a preverbal as well as in the postverbal position. There are no obvious differences as to the functions of these distinct forms. Also, investigation of the corpus data revealed that the use of the index does not correlate with TAM distinctions or other properties of the clause. Instead, we are dealing with a case of the so-called argument-triggered differential argument marking, as defined by Witzlack-Makarevich & Seržant (2018: 17). This means that based on some referential properties of the P argument, Ruuli speakers make a choice between nearly synonymous constructions, either indexing the referent or not. Possible triggering factors for differential argument marking can be inherent (e.g. animacy) of a referent or non-inherent (e.g. identifiability), but often enough one faces complex combinations of argument inherent and non-inherent factors (Witzlack-Makarevich & Seržant 2018: 4), as differential argument marking systems can be multidimensional (Aissen 2003). Thus, the choice of a certain marking strategy may depend on more than one or two variables, which interact in such a way that the impact of one of them may depend on another. The relevant variables for differential P indexing in Ruuli and the nature of their interaction is treated in the following section.

3.5 A case study of differential P indexing in Ruuli

To reveal the nature of such a high-order interaction of three or more variables as described in Section 3.4, one is bound to work with large corpora. Schikowski (2013) in his work on differential object marking in Nepali showed in an impressive way how a fine-grained annotation of ample data can reveal the impact of individual variables on a certain form and assess for each relevant variable how much of the variation they can explain. Although the use of either nominative or dative case marking for the same semantic argument roles in Nepali had been connected to referential properties, such as animacy or definiteness, and information-structural distinctions in earlier studies, no previous account of the phenomenon was able to consider the relevant variables to full extent, let alone to describe how they interact. On the basis of annotated corpus data, Schikowski (2013) uncovered in his quantitative analysis about a dozen statistically relevant variables, both inherent referential properties (such as person or humanness) and non-inherent properties (such as identifiability or givenness), as well as structural features (such as distance from the predicate or co-argument's case).

As mentioned in Section 3.3, there are few corpus-based studies of the phenomenon in Bantu (the only one we are aware is Seidl & Dimitriadis 1997 on Swahili). In this section we apply a methodology similar to the one suggested in Schikowski (2013) to analyze factors which lead to the variation in P indexing in Ruuli.

3.5.1 Corpus annotation and relevant variables

In order to track the relevant variables for Ruuli P indexing and account for their individual impact, parts of the corpus described in Section 3.4 (15,324 words) have been annotated. The annotated data come from six free conversations, with a total of 13 speakers (six women and seven men, aged between 38 and 64). The speakers were encouraged to discuss various topics, such as education, politics, culture and traditions. Based on the relevant factors we identified in the Bantu literature and the literature on differential argument marking in general, we annotated independent transitive clauses for whether the P argument is indexed, whether the corresponding NP is overt, for constituent order, the noun class of the head of the NP, the semantics of the referent, textual givenness (i.e. whether the referent had been mentioned in preceding discourse, irrespective of how many utterances where between the last mention and its resumption), and for their identifiability, i.e. whether the referent is definite, specific or non-specific. Table 3.2 displays the annotated variables, and their respective values (also see Appendix A for more details).

Before diving into the high-order interactions of the different variables, we briefly discuss some basic statistics of our data. These will serve as a quantitative description, as well as a clarification of the choices we made with regard to the architecture of our model.

Our annotated part of the corpus yields 754 tokens of transitive clauses, both with and without overt P NP. In 430 (57%) of these clauses, an overt P NP follows the verb, and only in 98 (13%) of the cases, the NP precedes

Variable	**Values**
indexing	index, no index
overtness of NP	overt NP, no overt NP
constituent order	VP, PV, V
PoS of the head	noun, pronoun
modification	modified, none
semantics of P	human, animal, object, abstract, event, organization
identifiability	definite, specific, non-specific
textual givenness	given, new

Table 3.2: Annotated variables with their respective values

the verb; in all other clauses, there is no overt P NP. Of the 145 observations without overt NP, 17 tokens also have no index on the verb, i.e. there is no realization of P at all and it has to be inferred from the context. Looking into indexing across persons reveals that first and second person Ps always have to be indexed, whether they are additionally realized as free pronouns or not. As for the frequency of indexing in general, no P indexing seems to be more common than P indexing, as the counts presented in Table 3.3 suggests. It is therefore more promising to look at the third person only in order to investigate the factors that cause differential P indexing in Ruuli. The rest of this section deals with the third person only. We also found that neither number nor modification of the NP were relevant for indexing P. Also, the

person	**index**	**no index**	**total**
first person	47	0	**47**
second person	25	0	**25**
third person	224	448	**672**
total	**296**	**448**	**744**

Table 3.3: Absolute numbers of indexed and non-indexed Ps in the Ruuli dataset

part of speech of the head did not turn out to be of significance. As for the semantics of P, animates are slightly more likely to be indexed; the relative frequency of indexing increases if the semantics are further specified to human vs. non-human. This is shown in Figure 3.1 and Figure 3.2.

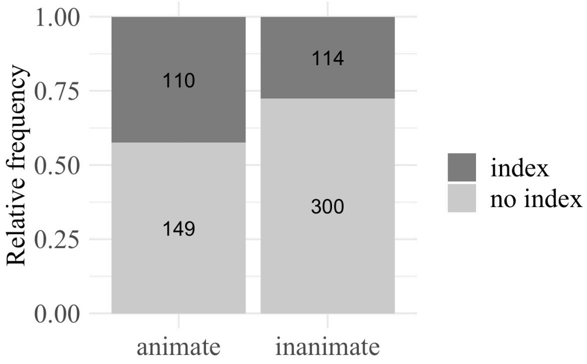

Figure 3.1: Indexing and animacy in Ruuli

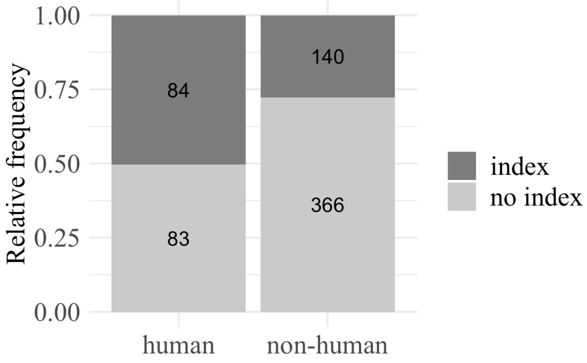

Figure 3.2: Indexing and humanness

The last two factors considered relevant are givenness as well as identifiability. We used the former as a proxy for the information structural status of a referent. Due to the various notions associated with the term “givenness” and the apparent fuzziness of subdividing categories (Baumann 2012) we decided to code only for the two values “new” and “given” within the preceding discourse, as the most basic concepts of information structure.

With identifiability we aim to capture the extent to which a referent mentioned by the speaker can be explicitly identified by them and the hearer. Definiteness is not morphologically expressed in Bantu NPs. The concept of “definite” as used in our annotation is based on the notions of uniqueness and familiarity and describes that a referent can be identified by both the speaker and the hearer. A “specific” referent, in turn, is unambiguously identifiable by the speaker only, referents labeled “non-specific” are not identifiable, neither by the speaker nor the hearer (cf. Lyons 1999).

3.5.2 Predicting the probability of P indexing in Ruuli

As mentioned above, a number of factors have been identified as being relevant for P indexing in Bantu languages. The annotation of several of these factors was added to the Ruuli corpus in order to examine their interplay and the individual impact of each one of them.

Like a logistic model, a decision tree makes a prediction of an outcome based on given variables. In our case, the outcome is binary, which means we have two alternative responses: indexed P and not indexed P. Tree-based methods have some advantages over other statistical models. Their visualization makes them interpretable in a straightforward way, as the prediction process can be followed quite easily.[7]

The order of interactions is mirrored in the trees’ nodes, where the splits occur. Also, tree- based methods can handle missing data quite well and are especially robust in cases with a relatively high number of variables compared to the sample size of the data. The recursive partitioning of conditional inference trees, as used in the present study, is based on repeated significance tests, providing better predictive performances than simple decision trees (cf. Hothorn et al. 2006). The latter can show high variance and can be prone to overfitting. Once the variable with the strongest association with the response variable is identified, the algorithm makes a binary split and subdivides the dataset into two subsets; this is then repeated with the next variable. As stated above, all instances of first and second persons show indexing on the verb, whether there is an accompanying free pronoun or not. Therefore, we included 3rd person referents only.

[7] For recent linguistic studies using conditional inference trees see, e.g. Tagliamonte & Baayen (2012), Klavan et al. (2015), Rezaee & Golparvar (2017), Hundt (2018) or Just & Čéplö (to appear); for discussion and criticism of tree based models in corpus linguistics see Gries (2020).

Figure 3.3 shows a conditional inference tree for P indexing in Ruuli, if all relevant factors are considered. All splits are significant at the level of 0.05. The first split at the first node at the top divides the dataset into two, based on word order. The variable word order also includes the value V, for instances without overt P NP. The first subset, with VP word order, branches to the right, and the second subset, which entails PV and V, branches to the left. This means that this variable has the strongest association with indexing. The strongest predictor for the subset of PV/V is givenness (node 2); the probability for discourse-new referents to be indexed lies here at about 70% (node 6). For given referents within the subset, part of speech (node 4, p = 0.028) can trigger indexing, with pronouns being more likely to be indexed. The last split (node 7), occurring within the VP subset, is also induced by the part of speech; we can see that overt Ps following the verb are very unlikely to be indexed, and although the difference between proper nouns and pronouns seems to be small at first glance, it is significant (p = 0.005).

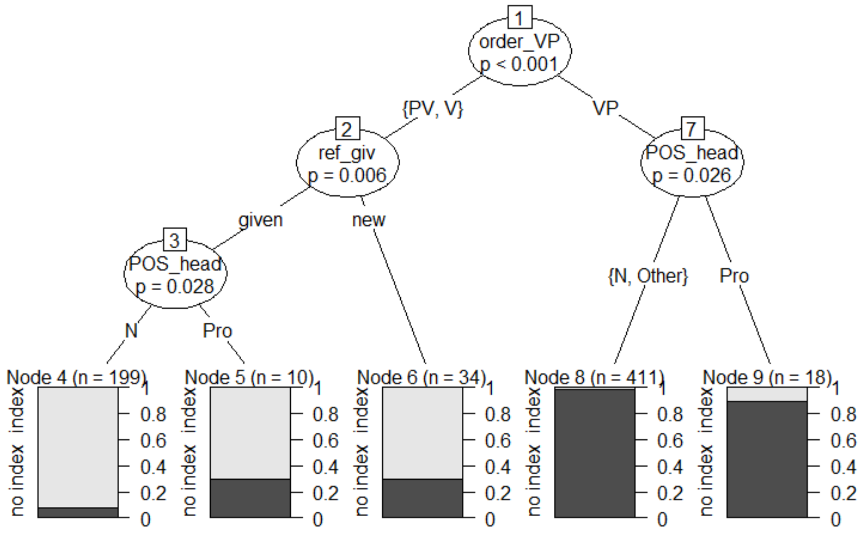

Figure 3.3: Conditional inference tree with all possible predictors for indexing

This result shows that the strongest predictor for P indexing in Ruuli seems to be word order; but just as P indexing itself, we assume that word order is a differential pattern reflecting the argument's semantic properties. Therefore, we built another tree model, without word order as a potential predictor. For the tree in Figure 3.4, we only considered the semantic and pragmatic variables (person, number, humanness, identifiability and givenness). This second model shows that without word order, givenness is the strongest predictor for indexing, dividing the dataset into given and new referents. The second split (node 2) is caused by humanness of P, with human referents displaying a higher probability of being indexed as compared to non-human referents. New referents are overall less likely to be indexed (node 2). These findings show that P indexing is in fact strongly correlated with word order, with P arguments outside their canonical postverbal position being more likely to be indexed. However, this correlation is not absolute, as there are exceptions in our corpus, of preverbal Ps being not indexed, and postverbal ones being indexed. It can be assumed that word order and indexing both are structural means to express the discourse status of a referent, which are commonly combined.

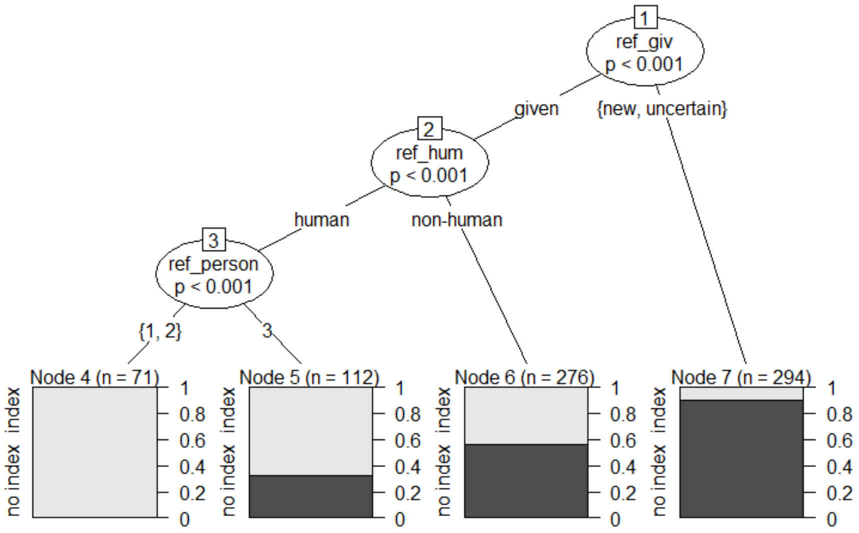

Figure 3.4: Conditional inference tree with indexing as response variable, excluding word order as a predictor

Figure 3.4 shows what happens to the second model if we take word order as a response instead of indexing. As one might expect, the splits are identical. In addition, the subsets of the response variable word order are nearly identical to the configuration of the first split in Figure 3.3.

3.6 Conclusion

Our analysis of the annotated corpus data suggests a few conclusions. First, Ruuli does not have any restriction of the co-occurrence of the P index and the corresponding NP, as reported for other Bantu languages (Riedel 2009, Downing 2011, Marten & Kula 2012). Also, indexing in Ruuli is not restricted to the referent's semantic properties such as animacy or humanness, although the latter plays a major role in triggering it. Neither is there an absolute obligatoriness for P indexing with referents in any syntactic or pragmatic context. P indexing in Ruuli is therefore an instance of differential argument marking as defined by Witzlack-Makarevich & Seržant (2018), i.e. that this marking strategy is not caused by the referent's argument role, but other factors connected to it. In the case of Ruuli, the factors are textual givenness and humanness, with given human referents displaying the highest probability of becoming indexed. Word order, i.e. whether the coreferential NP precedes or follows the verb, is apparently caused by the same conditions. These findings are neither surprising nor new. But they confirm what has been said about not only the differences between different Bantu languages, but also the inadequacy to try and find hard and fast rules as to when P indexing occurs in a Bantu language which displays some optionality with regard to this marking strategy (Riedel 2009: 89). Our approach shows that the findings of previous studies are in accordance with the outcome of a quantitative corpus study, and that the latter can help to get a deeper understanding of the interactions of the different variables involved. Further research with similar methodology could also be conducted focusing on other argument roles such as S/A, investigating the relevant factors which trigger deviations from the usual indexing pattern, or T and G in ditransitive predicates: constructions involving more than one object have been explored thoroughly in the Bantu literature (e.g. Marten & Kula 2012, Diercks et al. 2015 on Kuria, Zeller 2015 on Zulu or Ranero 2019 on Luganda), revealing the variation, in the family as well as language internally, with regard to word order or indexing.