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

Just, E.C.

A functional approach to differential indexing

Combining perspectives from typology and corpus linguistics

Published by
LOT
Binnengasthuisstraat 9
1012 ZA Amsterdam
The Netherlands

phone: +31 20 525 2461
e-mail: lot@uva.nl
http://www.lotschool.nl

Cover illustration: Sara Herbst

ISBN: 978-94-6093-405-6
DOI: https://dx.medra.org/10.48273/LOT0620
NUR: 616

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

Proefschrift

ter verkrijging van
de graad van doctor aan de Universiteit Leiden,
op gezag van rector magnificus prof.dr.ir. H. Bijl,
volgens besluit van het college voor promoties
te verdedigen op woensdag 20 april 2022
klokke 10.00 uur
door

Erika Just
geboren te Cham, Duitsland
in 1990

Promotor:	prof. dr. M.A.F. Klamer
Co-promotor:	dr. A. Witzlack-Makarevich (The Hebrew University of Jerusalem)
Promotiecommissie:	prof. dr. E.L.J. Fortuin
	prof. dr. G.L.J. Haig (University of Bamberg)
	prof. dr. M. Haspelmath (Max Planck Institute Jena)
	dr. G.J. van der Wal

Parts of this research were funded by the Federal State Funding
at Kiel University.

Contents

List of Figures

List of Tables

Abbreviations

Glossing follows the Leipzig Glossing Rules, additional abbreviations are:

1SG, ETC.	person and number
1 to 23	Bantu noun classes
ADD.FOC	additive focus
AF	actor focus
ATTR	attributivizer
BM	boundary marker
CJ	conjoint verb form
CONJ	conjunction
EMP	empathy
EMPH	emphatic marking
GER	gerund
HES	hesitation marker
HON	honorific
IND	indicative
IPFV	imperfective
LOC	locative
MID	middle voice
NAR	narrative tense
PREP	preposition
PFV	perfective
QUOT	quotative
REAL	realis
TR	transitive verbalizer

Acknowledgements

In the course of this undertaking of writing a thesis, I have never felt alone. There were always people, either physically or just an email or a video call away, whom I could count on for professional as well as moral support.

The person I am most indebted to is Alena Witzlack-Makarevich. Her expertise, insights, dedication, encouragement, and consistent, honest interest guided me through the last four years. Due to her affiliation with a different university and the spatial distance, the amount of time and energy she spent on me is doubly cherished. The support I received from her went well beyond the needs of supervising a thesis, and I never took any of it for granted. Every conversation with her is a source of inspiration, and her curiosity in new topics and methodologies is contagious. I consider myself very lucky to have her as a mentor and a colleague.

I am also deeply grateful to my supervisor, Marian Klamer, for so kindly taking me and my project in, without having met before, and for always being so supportive and approachable.

Lots of gratitude also goes to my colleagues at the ISFAS in Kiel, for the lively exchange and the opportunities to present my work at its intermediate stages at our colloquia (most notably at the extracurricular, specifically convened ones). Thank you all for your interest, critique and praise, Tobias Weber, Zarina Molochieva, Csilla Kász, Judith Voß, John Peterson, Meg Zellers, Stefanie Berger, Katharina Pommerening, Tina John and Benno Peters. Thank you also for the pleasant and collegial work together at the institute, you all made teaching and writing a thesis at the same time really feasible and hassle-free. And fun.

Special thanks go to Judith and Csilla for discussing and scrutinizing my various thoughts also during lunch breaks and at other occasions outside work-

ing hours, and also for creating a motivating environment, including board and lodge when I could neither work at the office nor from home during the challenging times of the pandemic.

I was lucky to present parts of this research also to experts outside of my immediate environment, most notably to the members of the ”Grammatical Universals” project at the Nikolai-Lab Leipzig. I am very grateful for the helpful comments and the lively exchange, especially to Martin Haspelmath for his kindness, his genuine interest, and for giving me the opportunity to present at the colloquium as well as at the project's final outing.

I am also greatly indebted to the three co-authors of the three case studies which are part of this thesis. The only one who has not been mentioned yet is Slavomír Čéplö, who, besides being utterly enjoyable to work with, also provided a constant wellspring of motivation due to his tireless inquiries on the status quo of my thesis, and the ever resonating “Finish already!”. It is truly appreciated.

Different parts of the thesis were proof-read by different people, and commented on in the most helpful and supporting way. Sincerest thanks go to Meg, Judith, Tobi, Csilla, Jana and Steffi, also for being very helping on the technical side. It goes without saying that all remaining flaws and errors are mine alone. A big thank you also goes to our student assistant Luna Hemmerling, who was a huge help in the annotation of the Gutob data.

At this point I am also indebted to my peer-coaching group, especially to Jana Fischer and Katharina Priewe, who accompanied me during the largest part of this work. Thanks a lot for your sympathetic ears, for advising without giving advise, for helping me through downs and small crises, and especially for celebrating every little success. I am wishing you all the best for your future paths.

Last but not least, my thanks go to my husband Timo, for thinking and reminding me that it would be really classy to have someone with a doctor's degree in the family. That really kept the motivation up.

Chapter 1

General introduction

> [Agreement is] a quite intuitive notion which is nonetheless surprisingly difficult to delimit with precision. (Anderson 1992: 103)

This thesis is concerned with bound markers expressing argument features and mostly found on verbs. I will later cease from using the term *agreement* and use *indexing* instead. However, when this work was initiated, it was intended to be an exploration of agreement domains, i.e. of the syntactic environments in which verbal agreement occurs (Corbett 2006). I was very convinced of the first part of the statement above, but somewhat naive about the second part. Agreement is a widespread phenomenon in the languages of the world, and it has been ascribed a crucial role in the encoding of participants. Almost every grammar or grammar sketch addresses it (even if the language does not have it, in which case it has to be stated that it is not there, in response to some unspoken expectations on the part of the reader).

However, trying to delimit agreement with arguments, even for individual languages, turned out to be difficult. The following quotes are a digest of some of the accounts that raised my awareness in this regard:

1. Under, as yet, unclear circumstances, [subject] agreement *may be suppressed in certain uses* […]. Object marking is, *generally, not* part of the Gta? verbal structure with one exception: the verbal plural marker har- can refer to objects in certain of its uses. (Anderson 2008: 723, emphasis mine)

2. [In Kesawai, t]he object suffixes are *obligatory* for human referents […] and *optional* with other highly animate referents […]. Lower animates such as aine 'fish' and inanimates are *not usually* indicated by object suffixes. (Priestley 2008: 314, emphasis mine)
3. Object marking on the verb is *optional*[1] in Ndengeleko. (Ström 2013: 217, emphasis mine)
4. -PR [the pronominal enclitic] sometimes occurs in the other position, i.e. not where the 'common pattern' puts it […] That discourse considerations affect -PR placement is likely, but what are these? [...] It may have to do with general phonological wellformedness, i.e. how the sentence sounds—in its (pragmatic?) discourse context. [...] -PR can attach to some words […] under extreme (rhetorical) conditions, words they do not attach to in 'ordinary sentences'. (Zide 1997: 325–326, no emphasis necessary)

In light of such descriptions, the question arose of how to account for inter-linguistic variation of the syntactic extent of a phenomenon which exhibited intra-linguistic variation to such a degree. I deemed it appropriate to address the latter, i.e. to start by dealing with language-internal variation, before going into a cross-linguistic overview of language-internal variation.

As some of the quotations 1–4 indicate, agreement in many languages is extremely susceptible to discourse-pragmatic realities and/or referential features. And why shouldn't it be, considering that agreement is not assumed to be a priori associated with the syntactic status of an argument like subject or object, but arises out of topic agreement (Givón 1976): it is employed to facilitate the access towards topical referents, irrespective of the grammatical relations or argument roles (Lehmann 1982, Givón 1983, Siewierska 1997). Two consequences relevant for the present work result from this. Firstly, coding splits in verbal agreement which are triggered by discourse-pragmatic circumstances or referential features are very common in the world's languages; thus, agreement can in many languages best be accounted for by referring to tendencies rather than rules. Secondly, agreement grammaticalizes for A referents (or subjects) more easily than for P referents (or objects), in the sense

[1] This quote is used representatively for all the optional cases of agreement out there.

that it becomes syntactically obligatory (Siewierska 1999), because the subject relation is more strongly associated with topicality than the object relation (Chafe 1976, Li & Thompson 1976, Kibrik 2011: 55).

As will be elaborated on in Section 2.1, dealing with coding splits from a typological point of view is difficult if one maintains the structural implications the term *agreement entails*. Therefore, the more neutral term *indexing* (Haspelmath 2013) will be used in what follows to refer to any kind of bound referent marking, irrespective of the language-specific constructions or criteria relevant for its occurrence. In this context of syntactic impartiality, it is also more fruitful to work with the concept of macroroles instead of grammatical relations of subject and object, as will be outlined below in Section 2.3.

Coding splits involving indexing are labelled *differential indexing*, i.e. instances of *differential marking* (Witzlack-Makarevich & Seržant 2018) in analogy to the more familiar concept of differential case marking (Bossong 1982, Aissen 2003, Hoop & Malchukov 2008). Differential indexing has received substantial attention with regard to objects (or P, T and G arguments), but less regarding subjects (or S and A arguments). However, with differential object indexing, up until now, there are not many investigations which back up qualitative observations with quantitative data (some notable exceptions are, for instance, Seidl & Dimitriadis 1997 and García-Miguel 2015). This fact presents a shortcoming, as the preference for one structure over another based on tendencies rather than grammatical rules can be best accounted for on the basis of extensive annotation of large enough language corpora (cf. Schikowski 2013).

Additionally, the main focus of previous research on this topic has been the presence (for objects) or the absence (for subjects) of indexing, but the examination of *variable placement* of indexes in languages where there seems to be free variation with regard to the host of an index has been relatively neglected (but see Cysouw 2003).

The following objectives for this dissertation arise from these considerations:

1. to showcase examples of the complex interaction of variables (e.g. discourse-pragmatic or semantic) triggering differential indexing, based on quantitatively analyzed corpus data

2. to illustrate out that differential indexing is not merely about the absence and presence of marking but can involve variable placement of the index as well
3. to address the question whether differential indexing is formally and functionally different for different argument roles across languages

This thesis consists of four articles and a chapter providing the theoretical background, as well as a chapter with a summary and conclusions. Two of the articles, *A corpus-based analysis of P indexing in Ruuli* (Chapter 3) and *Differential object indexing in Maltese – a corpus-based pilot study* (Chapter 4), address the first objective and contain corpus-based case studies of two unrelated languages, Ruuli (Bantu) and Maltese (Semitic), which present, at first glance, two similar cases of differential P indexing. In both languages, there is a strong correlation of differential indexing and constituent order, and the phenomenon has been connected to topicality.[2] However, as the descriptive models used in the studies show, the high-order interaction of relevant variables for indexing in the languages are different. The two studies strengthen the claims that have been suggested by previous qualitative investigations and help to gain a deeper understanding of the interplay of the different variables involved. Thus, they also illustrate that it would be inadequate to look for hard and fast rules as to when P indexing occurs in a language which displays some optionality with regard to indexing.

The third paper *Variable index placement in Gutob from a typological perspective* (Chapter 5) also encompasses a corpus-based case study, but it deals not only with the occurrence, but also with the placement of indexes. Gutob (Munda), the language under investigation, displays what is referred to as variable index placement: indexes, for S/A referents in this case, can attach to the predicate, but also to any other constituent in the clause. This behavior has been ascribed to exceptional discourse configurations in previous accounts. However, as this study shows, non-verbal index placement is anything but exceptional in terms of frequency. As for its function, the analysis accounts for the particular discourse effect index placement can have on the host of the index, showing that indexing can not only be sensitive to discourse, but be employed actively to structure it.

[2]To be more precise, this has been done for Maltese (Fabri 1993), as well as for other Bantu languages (e.g. Bresnan & Mchombo 1987, Ngoboka & Zeller 2017, Zerbian 2006 Creissels 2005); the present case study is the first study that addresses differential indexing in Ruuli.

The fourth paper, *A structural and functional comparison of differential A and P indexing* (Chapter 6) addresses the issue that differential A indexing has been somewhat neglected in the typological study of differential marking, as opposed to differential P indexing. It compares differential P indexing with differential A indexing and observes many similarities. It suggests that indexing should be considered to be functionally different from flagging, as it is motivated by referent tracking in discourse, irrespective of the argument role, and this explains the parallel (rather than mirror-image) behaviour of differential A and P indexing.

The remainder of this thesis is structured as follows: Section 2.1 deals more in-depth with the notion of indexing and the considerations that make it preferable over agreement. Section 2.2 elaborates on differential marking, and addresses the phenomena which can be subsumed under what I generally refer to as differential indexing. After that, Section 2.3 briefly explains the choice of argument roles, before Section 2.4 completes the theoretical background with some considerations on information structure relevant for this thesis. Chapters 3–6 provide the four articles, and Chapter 7 provides a general discussion and concludes the thesis.

Chapter 2

Theoretical preliminaries

2.1 Leaving agreement behind

This section serves to briefly justify why the notion of agreement has been abandoned in the course of the writing of this thesis. Although *indexing* has been well established at least since it was taken up and thoroughly defined by Haspelmath (2013) I consider it worthwhile to explain why the concept is preferable to agreement.[1]

Following Haspelmath (2013), an index is any kind of bound person marker. This thesis is exclusively concerned with indexes referring to verbal arguments, ignoring those expressing possessors or adpositional complements. One should not be misled by the deceptively simple appearance given by the term *person marker*. Person has been a well established concept despite its varied character: not only is it a cover term for speech act participants (or locuphoric forms, i.e. first and second person) and non-speech act participants (or allophoric forms, i.e. third persons) (Haspelmath 2013: 211–212); it can also include gender or noun class distinctions as well (Siewierska 2004: 103–104). As for the morphological form of an index, it is irrelevant in the present account whether it is considered a clitic or an affix. The former is often unjustifiably equated with optionality, and the latter with obligatoriness

[1] The term has a longer history, see Haspelmath (2013: 211) for an overview; Iemmolo (2011) uses the term *indexation*, reserving the term *indexing* to refer to the indexing function of case marking on core arguments as defined by Siewierska & Bakker (2008: 292), whereby case marking is considered to index semantic or pragmatic properties of the referent, such as animacy, definiteness, and topicality.

of marking (cf. Haig & Forker 2018: 720), although clitics can be syntactically obligatory, just as affixal indexes can be syntactically optional. Therefore, indexing is more suitable as a comparative concept (Haspelmath 2010, 2013) than agreement.

The suitability of the term indexing is further enhanced by the fact that the notion of indexing does not imply the theoretical load that agreement involves. An important parameter in the definition of agreement has been the presence of a noun phrase (henceforth NP) indexed by an agreement marker within the same clause. This has led to the pervasive distinction between grammatical vs. anaphoric (Bresnan & Mchombo 1987) or pronominal (Siewierska 1999) agreement. The notion of grammatical agreement refers to situations where there is an agreement marker on the verb and a clause-internal co-referential NP at the same time. With pronominal agreement, a co-referential NP is analyzed as clause-external.

This formal dichotomy has functional implications: in grammatical agreement, the NP bears the argument relation to the verb, while the agreement marker is considered to redundantly express referential features. In pronominal agreement, on the other hand, the agreement marker is considered the only true instantiation of the argument, and the coreferential NP then has a non-argument function (Bresnan & Mchombo 1987: 741). Thus, the very same marker is seen as either superfluous, lacking any referentiality, or as being itself the argument. This view has been very influential in subsequent accounts of agreement (e.g. Siewierska 1999, Van Valin 2005, Falk 2006). Also, Corbett (2006: 10) considers grammatical agreement rather than anaphoric agreement as the canonical case.

Assuming that the presence of a referential NP and a verbal marker are mutually dependent is problematic for the cross-linguistic study of agreement. Languages differ with regard to whether and how easily they allow the omission of nominal arguments (cf. e.g. Lambrecht 1994, Bickel 2003), as well as in whether, or under what circumstances, indexing is obligatory. Thus the notion of agreement conflates two parameters which are logically independent (Haig & Forker 2018: 719). The neutral concept of indexing allows for the formal and functional comparison of bound person marking without facing the challenges of simultaneously accounting for other, language-specific syntactic circumstances. That indexing and the expression of a referent by a lexical NP are not only separate means of referential expressions but also functionally dis-

tinct goes without saying. Lexical NPs are usually used for new or contrastive information, topic shifts or for long referential distances (Givón 1983, Ariel 1990, Lambrecht 1994, Kibrik 2011). Non-lexical forms, on the other hand, are used for more accessible information. Indexing in particular is considered a device for keeping track of referents with a certain level of accessibility or topicality (e.g. Givón 1983, Siewierska 1997, Iemmolo 2011).

There are not many accounts which deal with indexing in its own right, acknowledging the logical independence between bound marking on the predicate encoding referential features and the presence of a lexical NP; exceptions are Iemmolo (2011), a typological account of direct object indexing without an a priori assumption that there are different types of the phenomenon based on the behavior of the lexical NP, or Haig & Forker (2018), who give an overview of agreement accounts and strongly advocate, contrary to popular opinion, for not conflating the obligatoriness of an index (being a language-specific question of the exponence of inflectional morphology) and the tolerance of null referential NP (which is also language-specific).

As has been indicated in Chapter 1, there can be language-internal variation with regard to the factors that trigger indexing. In some languages this variation surfaces as a correlation between indexing and some other morphological or syntactic prerequisite, like TAM marking, clause type, or, in fact, the absence or presence of a lexical NP, its part of speech or its position in the clause.

In many cases, however, this correlation is either not perfect, or very weak, with indexing being conditioned by referential features or discourse-pragmatic realities. In such cases, one has to deal with tendencies instead of hard and fast grammatical rules, and more often than not, a number of variables such as animacy, discourse givenness or identifiability simultaneously play their part in directing those tendencies. This thesis concentrates on cases like these.

Examining indexing in its own right and considering the role of the NP as one of many factors with which indexing can potentially be associated, does not only facilitate accounting for the possibly complex relations between variables leading to indexing in a given language, but also unraveling the cross-linguistic reality of those relationships.

2.2 Differential indexing

The term differential indexing refers to variation in indexing, in analogy to the longer-established concept of differential case marking, coined by Bossong (1982). The term differential marking can be used for any argument encoding strategy and is defined by Witzlack-Makarevich & Seržant (2018: 3) as

> Any kind of situation where an argument of a predicate bearing the same generalized semantic argument role may be coded in different ways, depending on factors other than the argument role itself, and which is not licensed by diathesis alternations.

The definition does not entail specification of the "different ways" of coding; it can entail different morphological material, the absence or presence of morphological material, or also variation in its placement in a clause. Differential indexing mainly revolves around the second kind, i.e. it deals with whether a respective index is present or not, as exemplified in (1) and (2) below. However, there are also cases of differential indexing revolving around variability in index placement, i.e. the index is not confined to a particular host, nor to a fixed syntactic position, as in the case of Gutob, presented in Chapter 5.

The factors which can lead to differential marking referred to in the above definition are very diverse. They can relate to the argument itself, to characteristics of a co-argument, to event semantics, or to properties of the predicate, such as clause type, TAM categories, or polarity (Witzlack-Makarevich & Seržant 2018: 12–20). Features relating to the argument itself can be inherent or non-inherent. Inherent lexical argument features are very often associated with implicational hierarchies presenting gradations in animacy, person, and/or empathy (e.g. Dixon 1979, DeLancey 1981 or Croft 2003). Further inherent semantic argument properties relevant for differential marking can be uniqueness (proper vs. common nouns), discreteness (count vs. mass nouns) or number (Witzlack-Makarevich & Seržant 2018: 7). However, it is important to note that it is rarely only one of these factors which licenses the use of a particular marking strategy, but very often an interplay of several of them. This will be laid out in detail in the case studies on Ruuli and Maltese, which showcase the complex high-order interaction of different factors lying at the heart of P indexing in these languages.

Inherent argument features can also be morphological in nature, such as the part of speech of the argument NP, gender/noun class, or inflectional class assignment (Witzlack-Makarevich & Seržant 2018: 7–9). Whereas morphological features are relatively straightforward to account for, semantic properties like a referent's position on the animacy hierarchy are more difficult to investigate due to their gradient character. The same is true for non-inherent argument features conditioned by discourse, as well as the whole discourse setting and the information structure of an utterance in context. Although there is a basic consensus of what falls under such notions as topic, comment or focus, they are not only marked by very diverse means in the languages of the world, but even similar means (like differential marking, for instance) can have different effects in different languages. Moreover, even if a certain structure is identified as being reserved for, say, a topic, the reverse statement that every topic in the language is marked by this structure would be problematic. A concept like topic, which has been considered as lying at the heart of many a differential marking phenomenon (e.g. Taylor 1985: 78, 91, Fabri 1993: 92, Macaulay 1996: 139–140, Iemmolo 2010, Ivanov 2012, or Virtanen 2014: 404) is actually an accumulation of different discourse effects (Ozerov 2018, 2021), of which givenness and identifiability (which are in turn also quite difficult to measure) are only two. Section 2.4 further deals with the effect of information structure on indexing.

As the definition of differential marking used here does not include any syntactic prerequisites, various phenomena can be classed as differential indexing. It can be encountered, for instance, as "clitic doubling" (e.g. Jaeggli 1981, Aoun 1999, Preminger 2009, Arkadiev 2010, Sikuku et al. 2018), "object reduplication" (e.g. Friedman 2008, Čéplö 2014), "optional agreement" (e.g. Zwicky & Pullum 1983, Muxí 1996), "agreement suspension" (Iemmolo & Witzlack-Makarevich 2013), or "agreement asymmetry" (e.g. Bolotin 1995). The "lack of subject-verb agreement" as described by Lambrecht & Polinsky (1997) as one of several constructions used for propositions with sentence-focus also falls under differential indexing.

The terms "doubling" or "reduplication" suggest that firstly, these phenomena are defined on the basis of a co-referential NP, i.e. that there are two instantiations of an argument (in the form of an NP and of an index), and secondly, that what is differential in these cases is the exceptional addition of the respective index. The use of terms like "suspension" and "lack", on the other

hand, suggests that an index which would usually be expected is omitted. It will be shown in Chapter 6, however, that such a differentiation is not really appropriate, as indexing, irrespective of the argument role, can be employed (or not) for particular referents who continue to hold a particular status with regard to a relevant referential features.

Most of the accounts mentioned above deal with differential indexing of objects (or P arguments). This phenomenon has received particular interest, not only with regard to individual languages, but also from a family perspective (e.g. De Cat & Demuth 2008, Riedel 2009, or Klamer & Kratochvíl 2018), from an areal perspective (e.g. Friedman 2008 or Souag 2017), as well as from a typological perspective (Iemmolo 2011).

There are also studies dealing with differential subject (or S and/or A) indexing, but either for particular languages (e.g. de Cat 2004 on French) or on a small-scale typological basis (Ouhalla 1993, Lambrecht & Polinsky 1997). However, the use of the notions of subject and object are problematic for language comparison: grammatical relations are typically identified on the basis of language-specific constructions (Bickel 2011). Thus, different criteria are used in different languages to identify them, a fact referred to as "methodological opportunism" by Croft (2001: 30). I will thus refrain from using these notions and use the generalized semantic argument roles (or macroroles) instead, which is also in accordance with the definition of differential marking provided by Witzlack-Makarevich & Seržant (2018). The following Section 2.3 will briefly elaborate on the choice of framework followed in the present work.

2.3 Generalized semantic argument roles

For differences in marking patterns to be characterized as differential, they may not involve a change of the argument's generalized semantic argument role. The present notion of generalized semantic argument roles follows the approach brought forward by Bickel & Nichols (2009), Bickel (2011), Witzlack-Makarevich (2011), and Witzlack-Makarevich (2019), based on the numerical valency of a predicate (see Haspelmath 2011 for an overview of different interpretations of the terms S, A, P, T and R, or G respectively).

To base the definition of differential indexing on the notions of subject and object would not be expedient for the present purpose. That grammatical relations are construction-specific, and, by consequence, language-specific (Dryer 1997, Croft 2001), has been accepted in linguistic typology and, to some extent, in language description (cf. Witzlack-Makarevich 2019: 4). Indexing has often been considered a specific constructional means used to code a subject or an object of a language, and as being reserved for privileged arguments (e.g. Næss 2007: 17). However, even if a language groups S and A arguments together through indexing in the majority of cases (and in fact, there is a strong cross-linguistic tendency for A and S to align with regard to indexing, see Bickel et al. 2013: 33 and Siewierska 2013), splits based on factors such as verb class or the referent's affectedness are not uncommon (e.g. Næss 2007: 58-61). Therefore, to speak of differential subject indexing can present challenges for an individual language, let alone for the purpose of language comparison.

Similarly, to speak of differential object indexing has its drawbacks, considering firstly, how differently languages go about aligning P, T and G in terms of indexing, and secondly, the cases of language-internal splits with regard to these arguments. For instance, in the Alor-Pantar language Teiwa, there is differential indexing based on animacy. However, only animate P or G referents can be indexed, whereas T cannot be indexed in bivalent predicates (Klamer 2010: 176–177). In Alaaba (Cushitic), P indexing is similarly sensitive to animacy (i.e. only animate Ps can be indexed), but it is also sensitive to information structure and definiteness (i.e. not every animate P is indexed) (Schneider-Blum 2007: 90, 142). However, unlike in Teiwa, the same index can refer either to T or G in ditransitive predicates, provided that it is animate (Schneider-Blum 2007: 179). So even though for Teiwa, one could describe indexing on the basis of the grammatical relation of secondary object (Dryer 1986), it does not prove helpful for languages like Alaaba (neither does a direct/indirect object distinction).

Therefore, for the case studies on Ruuli and Maltese (Chapters 3 and 4), only Ps were considered. Actually, for the Maltese study, the term 'object' was used, as it is well established in Maltese linguistics; however, what was looked at de facto were the P arguments of instances of the verb *nagħmlu* 'we do/make'.

For the case study on Gutob, indexes referring to S as well as A arguments were considered. These are the only arguments that can be indexed in Gutob and they behave identically with respect to indexing. For the fourth paper (Chapter 6), I focused on A and P arguments only, as semantic opposites in bivalent predicates, excluding their possible alignment with other roles.

2.4 Handling information structure

It has often been stated that there is a strong relationship between indexing and the topicality of the referent (Givón 1976, 1983, Lehmann 1982, Siewierska 1997, Iemmolo 2011 inter alia). The notions of topic and focus are used quite frequently, in comparative work as well as in the description of individual languages (in the domain of referent encoding as well as elsewhere), often without further clarification of which information-structural properties are subsumed in the respective use of those labels. Topic and focus have been assumed to be universal categories (Ozerov 2018); generally, topicality has been considered to be connected to factors like givenness and a high degree of identifiability, whereas focus has been considered to imply new, emphasized or contrastive information. But despite their intuitiveness, the actual pragmatic effects of constructions or markers ascribed to focus or topic can vary dramatically from language to language, as well as within a given language system from usage to usage.

The assumption that information-structural categories are universal has led them to be used as umbrella terms for different discourse effects, which, in turn, has led to theoretical or typological biases (Ozerov 2018: 78). This can blur the realities of the actual usage of a certain construction. I will give two concrete examples from the domain of indexing: both in Babine Witsuwit'en, example (1) as well as in Maltese (2), differential P indexing is described as being linked to the topicality of the P referent:

(1) Babine-Witsuwit'en (Athabascan, Gunlogson 2001: 374)

a. *Dinï hida nïlh'ën.*
man moose look.at.it.3SG

'The man is looking at a moose.'

b. *Hida dinï yi-nílh'ën.*
moose man 3SG.P-look.at.it.3SG
'The moose is looking at the man.'

(2) Maltese (Semitic, Fabri 1993: 92)

a. *Jien nara l-programm.*
I see:1SG.IPFV DET-program(M)
'I am watching the program.'

b. *Jien nara-h il-programm*
I see:1SG.IPFV-3SG.M.P DET-program(M)
'The program, I am watching it.'

In (1b) and (2b), there is P indexing (*yi-* in Babine-Witsuwit'en and *-h* in Maltese), but there is none in (1a) and (2a). Although for both languages, the authors mention topicality as underlying cause for this alternation, the variables which comprise topicality in each case differ. For Babine Witsuwit'en, Gunlogson states that first, the presence of the index correlates with a definite interpretation: indexing is obligatory with proper names, demonstratives and possessed objects (2001: 378). What also plays into topicality here is anticipation management: indexing informs the addressee that more discussion of the introduced topic is to be anticipated (Gunlogson 2001: 393). In contrast, for Maltese, it was found that indexing is strongly associated with specificity (rather than definiteness), as well as with the part of speech of a referential NP. Thus, topicality can be related to diverse pragmatic or semantic features (another factor which is often crucial is animacy, see e.g. Riedel 2009) and even if the relevant factors can be identified, they can interact in complex ways.

Considering differential A indexing, it has been suggested that the absence of topicality of the A referent can result in the omission of indexing (e.g. Lambrecht & Polinsky 1997, Mereu 1999, Malchukov & Ogawa 2011). Additionally, Siewierska (2004: 159–163) has noted that the omission of indexing can be attributed to the referent being in focus. But similarly to differential P indexing, loss of topicality or focality on the part of the A referent should not be overgeneralized to different languages. For instance, in colloquial French (see example 11 in Chapter 4), indexing for A referents is omitted if these are focal: a lexical A in focus cannot co-occur with the person proclitics; however, there is an exception, namely if the lexical A is a pronoun, it is

obligatorily indexed (Culbertson 2010). So although A indexing in colloquial French is sensitive to certain discourse effects related to the focus category, the discourse-structural associations of pronouns (such as identifiability or givenness) seem to prevail and trigger indexing.

Nevertheless, in descriptive work one has to deal with the terms focus and topic as they have been applied by the respective authors, based on their intuitions and expertise in the languages. Just as with any descriptive category, typologists often have to interpret the data and sometimes adapt it to the comparative concepts they use (cf. Haspelmath 2010).

The situation is different when carrying out case studies based on corpora of individual languages. One can put more effort and attention into finding (probably) relevant proxies for information-structural categories (such as new vs. given, or definite vs. specific vs. non-specific), morphosyntatctic circumstances (such as noun class) and referent semantics, and annotate the corpus accordingly. For the case studies on Ruuli and Maltese, relevant variables were selected based on previous findings reported from the literature on differential indexing in general, as well as some language-specific structural factors. The analysis was carried out using conditional inference trees, which present a non-parametric alternative to multiple regression. They are non-parametric models, which means the structure of the model is not predefined but develops through the data. Conditional inference trees make predictions through recursive testing, based on repeated significance tests (at an α level of .05). Therefore, conditional inference trees provide stronger predictive performance than simple decision trees (cf. Hothorn et al. 2006). The latter can show high variance and can be prone to overfitting. The model accounts for how strongly each variable is associated with the outcome, which is binary in this case (index vs. no index). The analysis was carried out using the ctree() function in the party package (Hothorn et al. 2006) in the R environment (R Core Team 2020).

It has to be mentioned, however, that even this methodology can probably never account for all the subtleties that underlie any construction which is somehow sensitive to discourse-pragmatics, nor for the nuanced effects its use can achieve on the part of the hearer. In the realm of information structure, one has to deal with abstractions, which are very often hard to fully grasp conceptually. By using proxies such as givenness, identifiability, or a measurement of referential distance (Givón 1983), one preselect factors one con-

siders relevant, and although working with naturalistic corpora seems pretty bottom-up, one implements top-down reasoning based on particular choices one makes for annotation. Nevertheless, such an approach can back up previous findings from descriptive work and at the same time raise awareness of the interactions of the different factors that can be involved.

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

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

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

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.

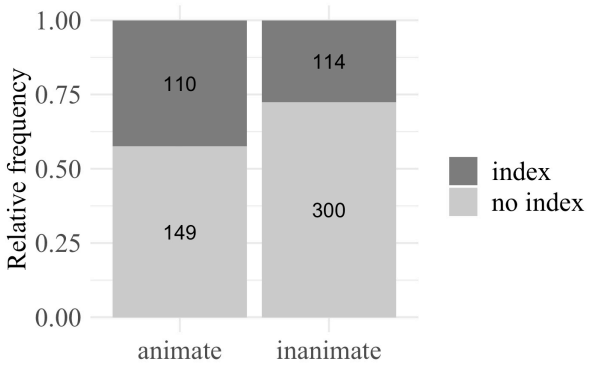

Figure 3.1: Indexing and animacy in Ruuli

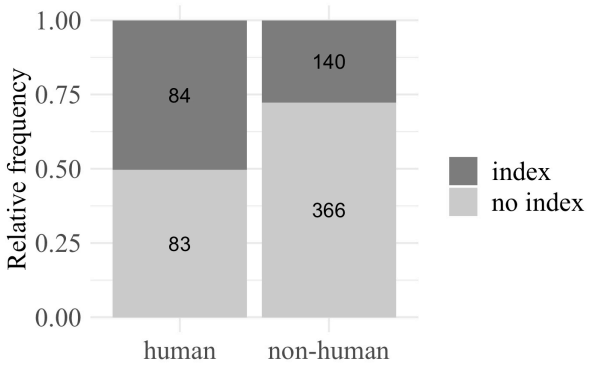

Figure 3.2: Indexing and humanness

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 ($p = 0.005$).

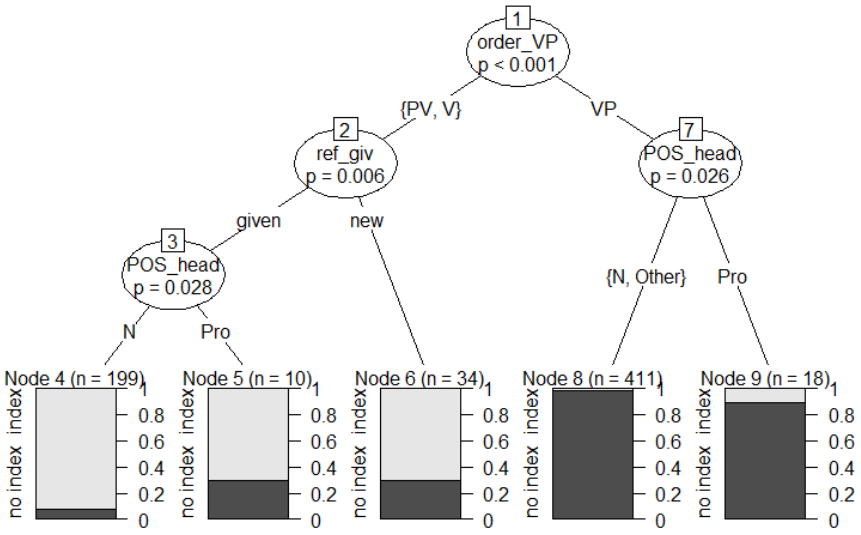

Figure 3.3: Conditional inference tree with all possible predictors for indexing

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 or-

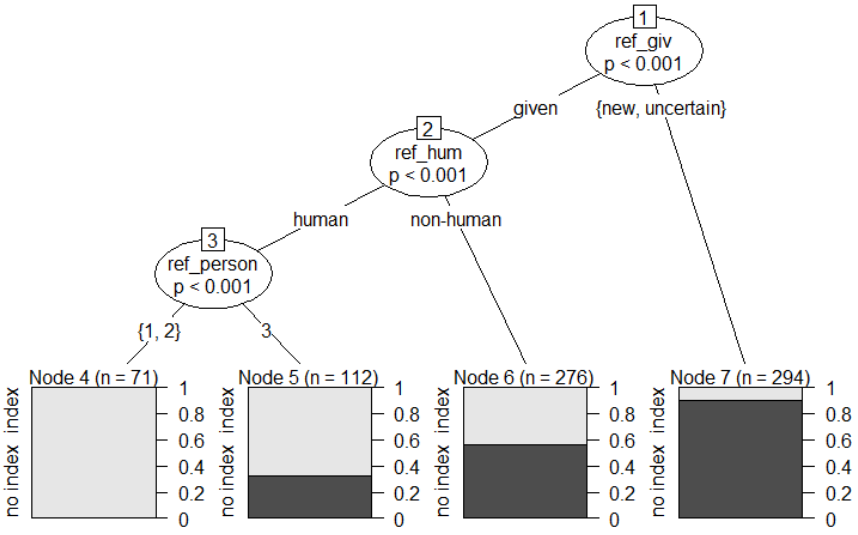

Figure 3.4: Conditional inference tree with indexing as response variable, excluding word order as a predictor

der, 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 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.

Chapter 4

Differential indexing in Maltese[1]

Abstract

This chapter presents the first corpus-based study of DOI in Maltese.[2] In this pilot study, the potential triggering factors were tested as predictors in a descriptive model. The results show that the strongest predictor for object indexing in Maltese is word order, but when taking only semantic referential features into account, the analyses reveal that DOI seems to be strongly predictable by definiteness, as well as by the part of speech of the head of the NP. Our study therefore supports observations from previous investigations, both on Maltese and typological; furthermore, the analysis gives insight into the combined effects of the relevant factors.

4.1 Introduction

[1]This chapter will appear as: Just, Erika & Slavomír Čéplö. Differential object indexing in Maltese – a corpus based pilot study. In Przemysław Turek & Julia Nintemann (eds.), *Maltese: Contemporary changes and historical innovations*, Berlin: De Gruyter. Author contributions: EJ and SČ developed the annotation scheme and carried out the annotation; SČ compiled the sub-corpus for the annotation; EJ provided for the typological background, performed the statistical analysis and wrote the paper.

[2]As indicated in 2.3, for the present study, the term '(direct) object' was used, as it is well established within Maltese linguistics. But P arguments where de facto considered.

4.1.1 Maltese

Maltese is a Semitic language and the national and co-official language of the Republic of Malta. There are several standard works on Maltese linguistics. The most comprehensive work is the reference grammar by Borg & Azzopardi-Alexander (1997). As for the lexicon, the fullest account can be found in the Maltese-English Dictionary by Aquilina (1987).

The language is exceptional in a number of ways. Due to the country's history, a large number of loanwords as well as syntactic and phonological features from Romance (mostly Old Sicilian, but also Italian) and English add to the Semitic supra-stratum which constitutes the basis of the phonology, morphology and lexicon (Borg & Azzopardi-Alexander 1997: xii). English loan words still continue to find their way into the language, mainly due to the fact that English has retained its prestigious status and is an official language of Malta. Code-switching is also very common, and a nonnative variety of English has been developing (Stolz 2011: 241-242).

According to Comrie & Spagnol (2016), around 35% of the lexicon is borrowed, and Maltese has even been referred to as a mixed language by Aquilina (1958). The complex historical contact situation of Maltese gave rise to different word formation strategies, most notably in the verbal system, where there is a distinction between templatic verbs and concatenative verbs. Whereas templatic verbs are formed on the basis of root-and-pattern morphology typical for Semitic languages, concatenative verbs are built by the combination of syllabic roots with verbal suffixes (Spagnol 2011). The class of templatic verbs contains not only verbs of Semitic origin, but also a few hundred verbs from Romance and a couple from English, in which consonants have been reanalyzed as consonantal roots and embedded in one or more verbal patterns. The class of the concatenative verbs consists, for the most part, of verbs with Romance or English origin, with very few from Semitic (Spagnol 2011: 49).

Maltese verbs are inflected for mood, tense-aspect (perfective, imperfective), as well as person, number, and gender (for the third person singular) of the S or A argument (or subject). As for object (or P) indexing, what inflectional class a verb belongs to does not play a role, as it is concatenative

in all paradigms.[3] The same set of indexes is also used for possessor marking on inalienable nouns and complements of prepositions. P indexes also express person and number, as well as gender for third person singular of the referent.

4.1.2 Background and terminology

In the literature on Maltese, the phenomenon under discussion appears under various different names: it is known as "optional direct object agreement" (Fabri 1993), "suffixed object pronoun" (Borg & Azzopardi-Alexander 1997), "pronominal clitic" (Fabri & Borg 2002, Vella 2009), or "object reduplication" (Čéplö 2014). In the literature on Arabic, Romance or Balkan languages, the term "clitic doubling" frequently occurs (see e.g. Aoun 1999, Kallulli & Tasmowski 2008b or De Cat & Demuth 2008); this term also appears to be preferred by linguists of the generative bend and, whether with good reason or not, it is oftenused as the default.

For our analysis, we use the more neutral term differential indexing, with indexes being defined as bound markers on the verbal predicate expressing argument features, most commonly person and number. Indexing (Haspelmath 2013) is a more neutral term than agreement, as it does not presuppose any syntactic relationship between the marker and the referential noun phrase (Haig & Forker 2018). Later in this chapter, we will make comparisons between differential indexing in Maltese and other languages, Semitic as well as from other families. This would be rather difficult if we chose to discuss the phenomenon under the umbrella of agreement, due to the various presuppositions associated with that term, the expected relationships between controller and target, let alone of the theoretically loaded terminology. Also, the morphological status of the index as a clitic or an affix is considered irrelevant. The latter is often equated with obligatoriness of marking, which is unjustified (Haig & Forker 2018: 720), as clitics can be obligatory, just as affixes can be grammatically optional.

The term *differential marking* was coined by Bossong's (1982) work on Sardinian and New Iranian languages, originally referring to variation in ob-

[3]There is no consensus in the literature as to the morphological status of the P, T and G indexes in Maltese; for instance, Fabri & Borg (2002), Spagnol (2011) (and others) call them (en)clitics, whereas Fabri (1993: 101) speaks of affixation, and also Borg & Azzopardi-Alexander (1997) call them suffixed pronouns.

ject case marking. However, differential marking patterns are not restricted to case or adpositional marking, but include indexing as well, as it is likewise a means of encoding arguments. Differential argument marking (DAM) can broadly be defined as any situation where an argument of the predicate with the same semantic argument role is coded differently (Witzlack-Makarevich & Seržant 2018). The argument role in question in the current study is the P argument, i.e. the less agent-like argument of a two-place predicate.[4]

4.1.3 Differential indexing in Maltese

Whereas subject indexing is typically obligatory in Maltese, be it with or without a co-occuring referential noun phrase, object indexing alongside an overt noun phrase has been considered optional and triggers indexing only if the referential noun phrase is the topic of the clause (Fabri 1993: 92). In Maltese, both the direct and the indirect object can be indexed, but the present study limits itself to the investigation of direct objects in monotransitive clauses. The following sentences exemplify how the presence of the object index for the third person singular, masculine in (9b) does not change the propositional content of the clause compared to (9a):

(9) a. *Jien nara l-programm.*
I see:1SG.IPFV DET-programme(M)
'I am watching the programme.'

b. *Jien nara-h il-programm.*
I see:1SG.IPFV-3SG.M DET-programme(M)
'The programme, I am watching it.'

c. *Jien nara-h.*
I see:1SG.IPFV-3SG.M
'I am watching it.' (Fabri 1993: 92)

The direct object index in Maltese can also have a purely pronominal function, as shown in (9c), but this does not constitute a differential pattern, so it is the difference between the structures in (9a) and (9b) which interests us in the present study. DOI in Maltese has been given some attention in the literature, but whereas Sutcliffe (1936) and Aquilina (1940) only give descriptions of the phenomenon, it is Fabri (1993) who is the first to associate it with

[4]This definition follows the generalized argument roles framework as proposed by Bickel & Nichols 2009, Bickel 2011, and Witzlack-Makarevich 2011.

information structure, describing the pragmatic inacceptability of indexed object noun phrases which are in focus, exemplified in (10) below: Helgard, who is the information asked for in (10a) is new information in (10b), therefore in focus and cannot be indexed, which renders sentence (10c) unacceptable in this context. The preposition *lil* is obligatory with human or human-like direct objects, as well as indirect objects, and has by itself nothing to do with indexing the object.

(10) a. *Lil min rajt.*
PREP who see.PFV.2SG
'Whom did you see?'

b. *Rajt lil Helgard.*
ses.1SG PREP Helgard
'I saw Helgard.'

c. **Rajt-ha lil Helgard.*
see.1SG-3SG.F PREP Helgard
int. 'I saw Helgard.' (Fabri 1993: 145)

Later studies have extended Fabri's work and put differential object indexing into the broader context of sentence information structure. Fabri & Borg (2002) investigate the relationship between topicality, focality and word order. The basic word order in Maltese is SVO, but it is by no means the only option. The authors state that also with indexed object noun phrases, other configurations are possible and thus both postverbal and preverbal object noun phrases with and without indexes are possible. Borg & Azzopardi-Alexander (2009) add phonological aspects to these analyses, stating that indexed preverbal object noun phrases form an intonationally separated unit from the remainder of the clause. Unfortunately, OV(S) is the only order they consider in their investigation.

All the studies on differential object indexing in Maltese come to the conclusion that a full account of the phenomenon is a difficult task to accomplish, and that no referential feature such as humanness, animacy or definiteness alone triggers the construction, but that the role of information structure lies at the core of it. Indexing preverbal objects has been claimed to be the main topicalisation strategy in Maltese (Borg & Azzopardi-Alexander 1997: 126), whereas non-indexed preverbal object noun phrases are usually in focus (Fabri & Borg 2002: 359-360).

The studies concerned with differential indexing mentioned so far were all based on the authors' intuitions as native speakers. The first one to go about the matter based on examples from real texts or spoken data is Čéplö (2014) who investigates DOI in Maltese against the background of the different clitic doubling phenomena described in the languages of the Balkan and Romance languages. The study explores various examples of DOI in Maltese, with their various contexts and different word orders. Amongst other things, the findings affirm that preverbal indexed objects are not necessarily marked for definiteness. There are examples of quantified as well as bare nouns which are preverbal and indexed (Čéplö 2014: 206-207), which suggests that it is rather specificity than definiteness which makes it likely for a referent to be indexed in this position.

As for postverbal noun phrases, Čéplö (2014: 219) states that indexing them seems to be "optional", as his investigation shows that, in terms of intonation, the indexed object can form an intonational unit with the verb, like the non-indexed one and that a pause between the verb and the noun phrase is possible but not obligatory, which shows that an indexed postverbal object noun phrase is not necessarily dislocated. He further observes that these indexed in situ objects, although rarer than non-indexed ones, occur frequently in exclamations, exhortations and especially in questions (Čéplö 2014: 219), which can be seen as an indicator for their information-structural prominence. As noted earlier, there is no one-to-one correspondence between word order and the information-structural status of the referent (Fabri & Borg 2002: 359-360).

So it seems that in Maltese, neither the placement of the noun phrase, nor intonational cues, nor indexing can on their own be regarded as a means to express the information-structural status of an object referent. Conversely, it is not very fruitful to try and account for DOI in Maltese on the basis of the rather fuzzy labels topic and focus only (cf. Čéplö 2014: 213), even if looking at corpus data reveals that undeniably, indexed objects are highly prominent and to analyze them as topics seems justified and obvious. Also, an important finding from Borg & Azzopardi-Alexander's (2009: 73–74) study is that with preverbal indexed objects, the preposition *lil*, which has to be used for any referent which can be referred to with a personal name, is no longer obligatory. This is an indicator that indexing is a way of marking highly accessible referents.

But then again, not every object referring to a topical referent is indexed, neither in Maltese nor in other languages with DOI. Dalrymple & Nikolaeva (2011: 51–57) in their study on differential indexing phenomena in Romance and Bantu languages propose to speak of *topicworthiness* rather than *topichood* when accounting for the referential features which trigger differential indexing. In order to account for a referent's status with regard to topicality, different versions of hierarchies have been proposed,[5] at the core of which lie argument properties such as animacy, givenness, identifiability and the full noun vs. pronoun distinction. Clearly, for some languages topicality scales are more important than for others, and the thresholds therein, i.e. from which position in the scale does a referent require special (differential) marking, have to be considered language specifically for each property. In other words, variables such as givenness, animacy or identifiability contribute to topicworthiness to different degrees in each language. One of the goals of this chapter is to set the stage for a fine-grained analysis of the triggering factors, as it is seemingly a highly complex interplay of different variables which renders object referents in Maltese worthy of indexing.

4.2 DOI crosslinguistically

4.2.1 Beyond Semitic

Differential indexing is not restricted to objects, but can affect subjects as well. But with differential subject indexing, it is very often the absence of an otherwise present index that codes the deviating scenario. Differential subject indexing has up until now been somewhat neglected in the study of differential marking phenomena (but see Just submitted). Consider spoken French, where subject indexing (with a verbal proclitic) has become quite common (cf. Culbertson 2010), but it is not possible for all types of subjects. For instance, it is not possible with quantified and indefinite subjects or with noun phrases which provide answers to wh-questions (features typically associated with focus). The latter case is as exemplified in (11a).

(11) French (Indo-European, France)

[5] See Witzlack-Makarevich & Seržant 2018 for a detailed overview.

a. Ceux du groupe A (*ils) ont fini leur travail.

'[Q: Who finished their work?] Those in group A have finished their work.' (de Cat 2004: 6)

b. Moi *(je) l'ai appelé.

'[Q: Who called Jean?] I called him.' (Culbertson 2010: 115)

In contrast to differential subject indexing, much attention has been devoted to differential object indexing, in language-specific studies (e.g. Muxí 1996, Béjar 1999 on Selayarese or Downing 2018 on Chichewa), family or area specific studies (e.g. Friedman 2008 on languages of the Balkans, Riedel 2009 on Bantu or Klamer & Kratochvíl 2018 on Alor-Pantar) as well as typological studies (e.g. Arkadiev 2010 or Iemmolo 2011). Differential object indexing is very often associated with topic-related argument properties, an indexed P being the more marked construction as opposed to a non-indexed P. This is not surprising given the fact that, whereas the A role is usually occupied by referents bearing topic features and is located high on the animacy and definiteness scale, the P argument is generally associated with the opposite, often serving to introduce new information (DuBois 1987). Therefore, a deviating scenario seems to call for a distinct marking pattern.

Before going into some examples, a few words will be said on the notion of topic and focus and how they are used here. The terms involve a great deal of vagueness due to the extent of linguistic diversity, both in form and function of information-structural features involved. Topic is generally connected to the notions of givenness, a high degree of identifiability and also to a high ranking in the person hierarchy, and is assumed to relate to the hearer's knowledge. Focus, on the other hand, brings about an information update and is associated with such notions as newness or contrastiveness.

But as a matter of fact, the meanings conveyed by similar constructions in different languages labelled as "focus construction" or "topic construction" are so manifold that topic and focus should often be considered as interpretive effects of such constructions, and not as being at their core (cf. Matić & Wedgwood 2013). Also, comparable and recurrent structures in different languages do not necessarily convey the same information-structural status of a referent (cf. Skopeteas & Fanselow 2010 for an investigation of non-canonical word order, clefts and focus). This also holds for structurally similar DOI constructions in different languages.

With DOI, it is very often the case that an otherwise per-default non-present index shows up if the referent is high in saliency or topicality, i.e. bearing features which are usually not associated with objects, or the P role, respectively (DuBois 1987). Sometimes, the situation is quite straightforward and object indexing can even become mandatory for subclasses of nouns, such as for all nouns referring to animates, as in the Alor-Pantar language Teiwa spoken in Indonesia (Klamer 2010) or all nouns referring to humans in the Madang language Kesawai spoken in Papua New Guinea (Priestley 2008). In other languages, the line between indexing and non-indexing is not that easily drawn. One such example is the Bantu language Sambaa, spoken in Tanzania, where object indexing also has to do with animacy: it is obligatory for proper names, titles – as in (12a) and (12b) – and first and second person pronouns; the index can therefore not be omitted with these referents. It is described as common (but not obligatory) with other types of human referents, less common with animals – as in (12c) – and rare but acceptable with inanimates (Riedel 2009: 45–46).

(12) Sambaa (Benue-Congo, Tanzania, Riedel 2009: 45-46)

a. *N-za-mw-ona askofu.*
1SG.SBJ-PFV-1.OBJ-see 5.bishop
'I saw the bishop.'

b. **N-za-ona askofu.*
1SG.SBJ-PFV-see 5.bishop
int.: 'I saw the bishop.'

c. *N-za-(ji)-ona kui.*
1SG.SBJ-PFV-(5.OBJ)-see 5.dog
'I saw the/a dog.'

In another Bantu language, Nkore-Kiga (Uganda), the indexing morphology is quite similar to that of Sambaa: a verbal prefix agreeing with the noun class of the object noun phrase. As a rule, objects are not indexed if the coreferential noun phrase is overt. However, they always have to be indexed if the object is "topicalized" and the noun phrase shifts to the preverbal "topic-position" (Taylor 1985: 78, 91)[6]. In (13a), the object *enkoni* is postverbal and there is no index on the verb. In (13b), the pronominal object is indexed (*gi-*).

[6]Bantu languages commonly have SVO as the basic word order.

Finally, in (13c), there is again an overt NP referring to the object, now in the preverbal position, and it is also obligatorily indexed:

(13) Nkore-Kiga (Benue-Congo, Uganda, Taylor 1985: 91)

a. *Omuntu a-kwata enkoni*
1.person 1.SBJ-hold 9.stick
'Someone is holding a stick.'

b. *Omuntu a-gi-kwata*
1.person 1.SBJ-9.OBJ-hold
'Someone is holding it.'

c. *Enkoni omuntu a-gi-kwata.*
9.stick 1.person 1.SBJ-9.OBJ-hold
'As for the stick, someone holds it.'

A third, again different, Bantu example comes from Ruuli, also spoken in Uganda. Looking at (14a) and (14b) might suggest that DOI in Ruuli works quite similar to that in Nkore-Kiga, however, it is not restricted to preverbal object noun phrases, although it is in fact more commonly found with such than with in situ objects (Erika & Witzlack-Makarevich accepted), exemplified by (14c). (14d) shows a non-indexed preverbal object which indicates that Ruuli is different from Nkore-Kiga in this regard: placing an object noun phrase in a preverbal position does in itself not trigger indexing.

(14) Ruuli (ruc, Benue-Congo, Uganda, 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. *Obuterega o-bu-maite*
14.obuterega 2SG.SBJ-14.OBJ-know.PFV
'Do you know the obuterega traps?'

c. *Naye we o-bi-maite ebyo?*
but 2SG 2SG.SBJ-8.OBJ-know.PFV 8.DEM
'But you, do you know that?'

d. *Amaani mu-ta-ire-mu.*
6.strength 2PL.SBJ-put-PFV-LOC
'You have put in a lot of strength.'

These examples from Ruuli suggest that DOI is related to topicality of the referent. However, this is hard to tell without context, and on the basis of

selected examples only. A quantitative study of the triggering factors for differential indexing in Ruuli (Erika & Witzlack-Makarevich accepted) examined the interplay of these factors. Based on a corpus annotation for variables such as noun class, PoS, animacy, identifiability and textual givenness, the statistically relevant factors were identified using conditional inference. The results of the analyses show that the strongest predictor for DOI in Ruuli is in fact word order, with preverbal noun phrases being more likely to be indexed than postverbal ones; but taking only semantic properties of the referent into account, the analyses reveal that DOI seems to be strongly predictable by textual givenness and humanness. This is not really surprising with regard to the assumption that topicality is involved in DOI in Ruuli, as both givenness and humanness relate to high accessibility and thus topicality. However, as not every human or every given object is indexed, it shows that we are dealing with probabilistic rules, which can be adequately described using descriptive models.

We have just seen three examples of languages belonging to the same family, where object indexing is quite similar with regard to its morphological realisation. Also, object indexing can be labelled differential in Sambaa, Nkore-Kiga and Ruuli, as in all three languages, it is the same macrorole, the P argument, which becomes indexed only under certain conditions. Although these conditions can be traced back to referential features which are usually not associated with this macrorole, they are not the same from language to language. Also, DOI has grammaticalized to some extent in one of the languages, namely Sambaa, where it has become obligatory for certain nouns.

A similar situation can be found in the languages of the Balkans, where DOI has been integrated into the different language systems to various degrees. Starting off as a pragmatic phenomenon expressing the topicality of a referent, it has grammaticalized to various extents (or not at all) in the different languages (Friedman 2008). But unlike the situation found with the Bantu languages, the languages of the Balkan belong to different language families, which indicate the areality of this phenomenon. Consider the following examples from Macedonian (Slavic), Albanian, and Romanian (Romance); in Macedonian (15), object indexing is obligatory for specific direct objects (Franks & King 2000: 115). In Greek, exemplified in (16), DOI is never obligatory but preferred if the referent is topical: the sentence in (16) can be the answer to the questions ‘Who read the book?’ or ‘What did Ana do to/with the book?’,

but not to a question like 'What did Ana read?', where 'the book' would be in focus (Kallulli 2000: 219). In Romanian, shown in (17), DOI with post-verbal objects is dependent on the presence of the special preposition *pe*, which in turn is conditioned by the semantics as well as by definiteness of the referent (Cojocaru 2004: 34).

(15) Macedonian (Slavic, Northern Macedonia, Franks & King 2000: 115)
*Marija *(go)=poznava učenikot/Vlado/toj učenik/nego.*
Marija 3SG.M.ACC=know.3SG pupil.DEF/Vlado/that pupil/3SG.M
'Mary knows the pupil/Vlado/that pupil/him.'

(16) Greek (Greece, Kallulli 2000: 219)
I Ana to=diavase to vivlio.
DEF Ana 3SG.N.ACC=read DEF book
'Ann did read the book/read the book.'

(17) Romanian (Romance, Romania, Cojocaru 2004: 34)
O=aştept pe ama.
3SG.F.ACC=wait.1SG PREP mom
'I'm waiting for mom.'

These examples illustrate what the hitherto largest typological study of DOI, Iemmolo (2011), already suggested, namely that DOI across languages is systematically associated with signalling high salience or prominence of the object referent. Furthermore, the findings from Ruuli (Erika & Witzlack-Makarevich accepted) as well as other usage based studies of differential marking phenomena (such as Schikowski 2013 on differential object case marking in Nepali or Schnell 2018 on subject indexing in Vera'a) suggest that every time a structure in any language is described as optional, this should be read with a grain of salt. Before we have another look into DOI in Maltese and our analysis based on bulbulistan corpus data (Čéplö 2018a), we will briefly look into differential indexing in other Semitic languages.

4.2.2 DOI in Semitic languages

Before turning back to DOI in Maltese and our investigation of the factors responsible, we will briefly discuss DOI constructions in other Semitic languages. Amberber (2009) provides an overview of Amharic differential case marking, including an account of differential indexing. In Amharic (Ethiopia),

the basic word order is SOV, and direct objects are differentially case marked, with definite objects only receiving the case marker *-n* (Amberber 2009: 745-746). If a direct object is case marked in this manner, it can also be indexed. By implication, only definite referents can be indexed, but they don't have to be; the indexing is thus described as "optional" (Amberber 2009: 745). See (18) for an overview of the situation in Amharic:

(18) Amharic (Semitic, Ethiopia, Amberber 2009)[7]

a. *ləmma t'ərmus-u-n səbbər-ə-(w)*
Lemma bottle-DEF-ACC break.PFV-3.M.SBJ-3.M.OBJ
'Lemma broke the bottle.'

b. *ləmma and t'ərmus səbbər-ə-(*w)*
Lemma one bottle break.PFV-3.M.SBJ-3.M.OBJ
'Lemma broke one bottle.'

In example (18a), the direct object t'ərmus is definite, and is thus obligatorily case marked. Additional to the subject, which is always indexed in Amharic, the direct object too can be indexed, as indicated by the brackets. In the second example, (18b), the direct object is indefinite, not case marked, and thus cannot be indexed.

As for varieties of Arabic, DOI is not only attested for Maltese but also for varieties of the Levant, including northern Iraq, and Central Asia (seeSouag 2017 for an overview). Like with Bantu languages, the indexes within the Arabic languages are similar in form. Examples (19a)-(19c) show that in Lebanese Arabic, the direct object index – which can also be used pronominally without co-occuring noun phrase – can be absent (19b) and present (19b) alongside the object noun phrase, a circumstance which is called "optional" by Aoun (1999: 17):

(19) Lebanese Arabic (Semitic, Lebanon, Aoun 1999: 14–17

a. *kariim ʔakal suuʃi*
Karim eat.3SG.M.PFV sushi
'Karim ate sushi.'

b. *kariim ʃeef-o*
Karim see.3SG.M.PFV-3SG.M.ACC
'Karim saw him.'

[7]Upper and lower case in all examples throughout the thesis are adopted from the original sources.

c. *kariim ʃeef-o la-saami*
Karim see.3SG.M.PFV-3SG.M.ACC PREP-Sami
'Karim saw Sami.'

d. *kariim ħəkee-lo la-saami ħkeeye*
Karim tell.3SG.M.PFV-3SG.M.DAT PREP-Sami story
'Karim told Sami a story.'

e. *kariim ʃeef kteeb-o la-saami*
Karim see.3SG.M.PFV book-3SG.M PREP-Sami
'Karim saw Sami's book.'

f. *kariim raaħ maʕ-o la-saami*
Karim go.3SG.M.PFV with-3SG.M PREP-Sami
'Karim went with Sami.'

However, in contrast to Maltese, there is not only differential direct object and differential indirect object indexing (as in 19d) in Lebanese Arabic, but also differential marking involving bound forms of possessors, shown in (19e) and of prepositional complements, as in (19f). Both these examples would also be possible without the bound person form. Another contrast to Maltese is that in Lebanese Arabic, a co-referential lexical noun phrase – be it the object, possessor or prepositional complement –has to be preceded by the preposition *la-*. This "dummy case assigner" (Souag 2017: 49) can never occur with these noun phrases if there is no additional bound person form, with the exception of indirect objects (Aoun 1999: 17). Differential indexing in the varieties of the Levant has also been associated with information-structural properties of the referent, both topicality (e.g. Cowell 1964 and Brustad 2000 on Syrian Arabic) and focality or emphasis (Levin 1987 for northern Palestinian varieties). Souag (2017) clarifies that although one might be tempted to trace DOI in Arabic varieties back to their shared heritage, the DOI constructions differ greatly from one language to another, and show more similarities with contact languages than with their genetically more closely affiliated languages. This claim has been made before for different regions of the Arabic speaking world, but Souag (2017) provides the first microtypological investigation of the clitic doubling phenomena, which include differential object indexing. He looks at each region where these constructions have been reported, making comparisons and pointing at the links with adstrates languages. For Maltese, he demonstrates the similarities in differential indexing to Sicilian and the dissimilarities to Levant Arabic, concluding that "Maltese clitic doubling is thus

better explained as the result of Sicilian superstratum influence than as a retention from some early stage of Arabic" (Souag 2017: 61). This suggests that a closer study of this contact phenomenon is a desideratum, contingent on our full understanding of DOI in Maltese and all the relevant variables.

4.3 Factors licensing DOI in Maltese: a corpus based pilot study

4.3.1 Research questions

As noted in Section 4.1.3, DOI in Maltese appears in contexts where the referent can be described as being topical. This is not surprising, as crosslinguistically, DOI is generally and systematically associated with signalling high salience or prominence of the object referent (Iemmolo 2011). However, in Section 4.2 we've shown that saliency and topicality are quite nebulous concepts, and although we find recurring factors being involved (such as identifiability, givenness or animacy), the different variables have a different weight and interact to different degrees across different languages. From these observations, the following research questions arise for Maltese:

1. What are the factors which license differential object indexing in the presence of a co-referential overt NP in Maltese?
2. Which of these factors are the strongest predictor(s), i.e. , which make it more probable for an object to be indexed?
3. Are the factors hierarchically ordered?

Research question 1) aims at identifying the variables which are basically relevant for topicworthiness in Maltese, as the number one candidate out of the variety of all potential factors discussed in the literature on Maltese, other Semitic languages, and differential argument marking in general. Research question 2) seeks to compare the effect of each of the relevant factors, and with 3), we set out to find out more about how the factors interact; for instance, whether one variable, such as animacy, is only statistically relevant if another feature (e.g. givenness or specificity) is involved, too. These three research questions logically build on one another. In the next section, we describe how we went about the corpus annotation which lead to answering

question number 1) which in turn forms the starting point for questions 2) and 3).

4.3.2 Variables

In order to find out which variables connected to topicworthiness of a direct object are relevant for Maltese, a special layer of annotation was added to a sample of sentences taken from the bulbulistan corpus[8] (BC; Čéplö 2018a). Based on previous findings on the phenomenon of DOI, transitive and ditransitive clauses where annotated for the following formal and semantic variables, which are displayed in Table 4.1, along with their values. Below, some of the variables will be briefly discussed in more detail. Clauses with overt object noun phrases, as well as without (i.e. with pronominal reference), were annotated in order to see whether - or for which variables - there are differences with regard to the topicworthiness of overtly expressed and non-overtly expressed referents.

It has been mentioned before that the preposition *lil* has to be used for referents which can be referred to by a proper name, i.e. for human referents or human-like referents. This feature was included as being potentially relevant for indexing due to the finding by Borg & Azzopardi-Alexander (2009: 73–74) that with preverbal indexed objects, *lil* is no longer obligatory for referents which would usually require its presence. Part of speech of the coreferential noun phrase was included due to the findings from other languages, such as Sambaa or French (cf. examples 11 and 12 in 4.2.1), where word classes are directly relevant for indexing. Whether the object noun phrase is modified or not was included for two reasons: firstly, to account for the length of the phrase, although this is admittedly not a very precise way of measurement. And secondly, referents which require modification can be considered less identifiable by the addressee than referents which do not require modification.

This brings us to the next variable which requires a brief explanation: identifiability. Together with givenness it was used as a proxy for the information-structural status of the referent. With identifiability we aim at capturing the extent to which a referent can be explicitly identified by the speaker and the

[8]The whole corpus consists of approx. 160 million PoS-tagged tokens and is composed of online newspaper articles, parliament records, fiction (imaginative literature and blogs), and non-fiction (academic texts, popular science, sermons, Wikipedia entries)

Variables	**Values**
index	present, absent
presence of *lil*	present, absent
word order	SVO, VOS, OVS, SOV, OSV VSO
person, number and gender	1SG, 2SG, 3SG.F, 3SG.M, 1PL, 2PL, 3PL of the referent
referent semantics	human, kinship, animal, anthro- pomorphic,physical object, event abstract entity
PoS of the object	noun, pronoun, NA (i.e. non-overt)
subcategory of the head noun	proper noun, common noun, personal pronoun, impersonal use of personal pronoun, possessive pronoun, demons- trative pronoun, interrogative pronoun
modification of the noun phrase	modified, not modified
subcategory of modification	adjective, numeral, determiner relative clause, possessive, multiple
identifiability	definite, specific, non-specific
textual givenness	given, new
clause type	main clause, relative clause, adverbial clause, complement clause
clause polarity	positive, negative
sentence type	declarative, imperative, interrogative, exhortative

Table 4.1: Features for the quantitative analysis of DOI in Maltese

hearer. The concept of definite as it is 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, and referents we labelled non-specific are not identifiable, neither by the speaker nor the hearer (Lyons 1999). Due to the various notions associated with the term givenness and the apparent fuzziness of subdividing categories (cf. Baumann 2012), we decided to code only for the two values new and given within the preceding discourse. So in total, it was decided to add a layer of annotation of fourteen variables, thirteen of which were considered possible independent variables, and the presence vs. the absence of the index as the response variable.

4.3.3 Corpus, challenges and preliminary solutions

Immediately upon starting the annotation, we stumbled across a significant problem: it turns out that in the existing corpora of Maltese, DOI is extremely rare. We initially set out to annotate the Maltese Universal Dependencies Treebank (MUDT, Čéplö 2018b) which, like the larger bulbulistan corpus, is composed of four text types – newspaper, fiction, non-fiction and parliament, i.e. parliamentary debates (Čéplö 2018a: 58–62, 172–176) – which fall into two categories based on their origin: written (newspaper, fiction and non-fiction) and spoken (parliament). When selecting the text samples to conduct analysis, we quickly found that in the fictional texts in MUDT (which we selected for annotation so as to avoid boredom), DOI hardly ever occurs. One explanation for this would be that this phenomenon is much more prevalent in spoken Maltese. To confirm this would require a sufficiently large corpus of spoken Maltese, still a desideratum in Maltese linguistics. This would be quite a finding, since such a split had not been mentioned in previous literature, but would not be surprising given that this is similarly the case in other DOI systems which have not yet fully grammaticalized (cf. Section 4.2.1). The other explanation, of course, is that DOI in Maltese is actually quite rare across the board (cf. Čéplö 2018a: 235), contrary to assessments such as that provided by Borg & Azzopardi-Alexander (1997: 126) who describe DOI with preverbal objects as “a wide spread characteristic of Maltese”.

To go about this challenge of rare occurrences of the phenomenon under investigation, we abandoned any attempts to annotate MUDT and instead focused on the larger and more general bulbulistan corpus, while limiting ourselves to the parts of it that come closest to naturalistic speech, namely

the parliamentary debates transcripts.[9] To control for variations in semantics, valency etc., we decided to limit our investigation to a single verb; to ensure there would be enough material for the analysis, we selected the verb *għamel* 'to do/make', since one of its forms, *nagħmluha*, 'we do/make it.F', is the most frequent verb with a direct object index in the parliament text type, with 4896 occurrences in total. We therefore extracted all (orthographic) sentences (cf. Čéplö 2018a: 63–64 containing the keyword *nagħmlu* 'we do/make' (without index) and *nagħmluha* 'we do/make it(F)' (with an index); the preceding and the following 1000 characters were extracted as well to be able to account for the context. This, incidentally, had the positive side effect to control for verb semantics at the same time. We randomized their order and then proceeded to annotate all the relevant clauses, meaning all the clauses which contained *nagħmlu* with an overt object NP, and all the clauses containing *nagħmluha* with or without an overt object NP. We excluded clauses which contained *nagħmluha* where the direct object has a pronominal function, more specifically that of an expletive pronoun (something rarely discussed in the literature), such as the very frequent *(ħalli) nagħmluha ċara* '(let's/in order to) make it clear'. In this manner, we ended up with a sample subcorpus of the bulbulistan corpus of 73555 words of parliamentary debates which contain 286 relevant clauses for annotation, with 133 instances of *nagħmlu* plus noun phrase, and 153 occurrences of *nagħmluha* plus noun phrase.

Due to the nature of a text type such as parliamentary debates, we had to face some cutbacks with regard to our variables. Because of the topics discussed in parliament, as well as the limitation to one verb *nagħmlu* 'we do/make', the referent semantics were restricted to non-animate referents. As a consequence, there were also no objects accompanied by *lil* (see Appendix B for details). Therefore, a more in-depth investigation of referent semantics as described in 3.2 as well as the interplay of object indexing and the presence of the preposition *lil* are two of the objectives reserved for a follow-up study of DOI in Maltese, which should be based on naturalistic data of spoken Maltese.

[9] We refer to these transcripts as "coming close to naturalistic speech" and not "naturalistic speech", as a comparison of randomly selected transcripts with their audio recordings has made it clear that some editing was executed (Čéplö 2018b:58).

4.3.4 Findings

As laid out above, DOI in general as well as in Maltese cannot be explained by hard and fast rules, but has to be explained on the basis of a set of variables. This chapter presents the starting point for a quantitative analysis of these variables. In the following, we provide some descriptive statistics of object indexing in Maltese on the basis of our corpus annotation of clauses containing *nagħmlu* and *nagħmluha* of those variables we found to be relevant, before presenting the evaluation of the variable interplay based on conditional inference.

The first variable we found to have an impact on indexing is givenness, with given referents, i.e. referents which where mentioned within the preceding 1000 characters, being more likely to be indexed. Figure 4.1 shows that the proportion of given referents in clauses with indexed object noun phrases (*nagħmluha*) is much larger than the proportion of new referents with this verb form. Consequently, with *nagħmlu* (no index) we find more new than given referents.

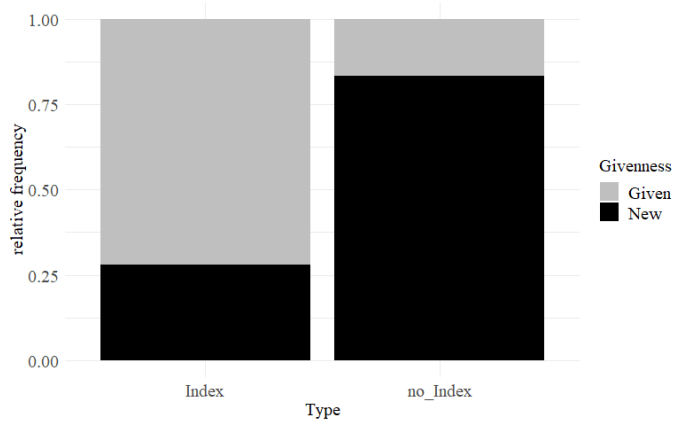

Figure 4.1: Indexing and givenness, clauses with overt objects only

Similarly, turning to our second proxy for information structure, identifiability, Figure 4.2 reveals the distribution of indexed and non-indexed objects over definite, specific and non-specific referents. We see that non-specific referents, i.e. referents which are neither identifiable by the hearer nor the speaker, are never indexed. The proportion of definite and specific referents

is fairly equal among the non-indexed referents, whereas for indexed referents, the definite referents outweigh specific ones by far.

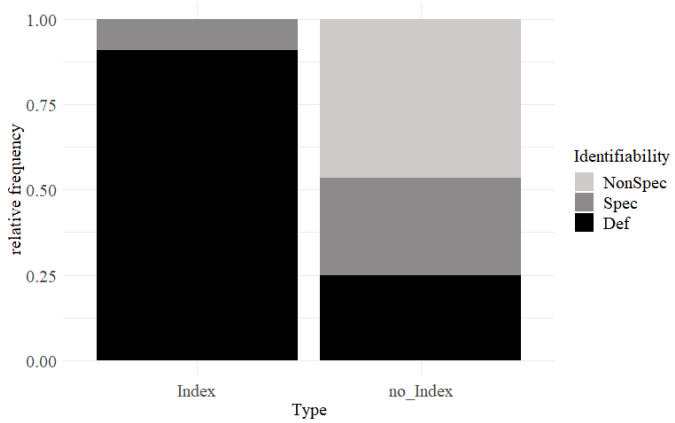

Figure 4.2: Indexing and identifiability, overt objects only

The third variable we found to have a fair impact on indexing is the part of speech of the object noun phrase. In Figure 4.3 we see that if the referent is realized pronominally, it is very likely to be indexed: the number of non-indexed pronouns is very low.

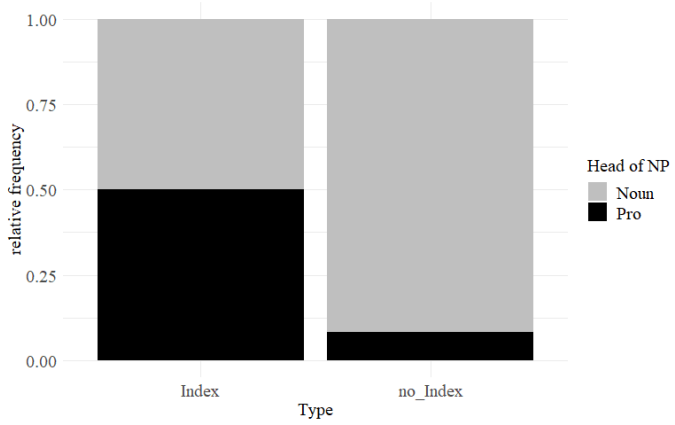

Figure 4.3: Indexing and PoS of the object noun phrase

Lastly and unsurprisingly, word order also had a strong impact on indexing. In our data we found no occurrences of preverbal objects which were

not indexed, so all preverbal objects are indexed. Additionaly we can see in Figure 4.4 that postverbal objects are less likely to be indexed.

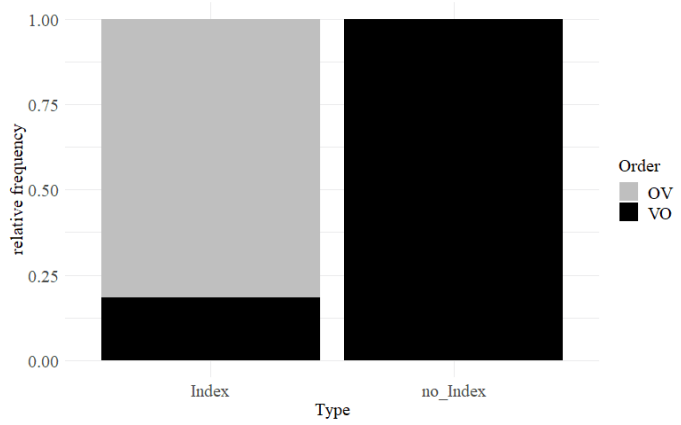

Figure 4.4: Indexing and order of verb and object

All other variables, modification of the noun phrase, clause and sentence types and polarity, have not (yet) shown to be of relevance. However, the data base for the present pilot study is still small, and, as has been mentioned above, a corpus of naturalistic spoken Maltese would be ideal. A follow-up study will hopefully not only provide us with more discourse contexts to work with, but also provide the opportunity to look into more variables such as a fine grained analysis of referent semantics, predicate semantics, the presence of *lil* and register. Although our database is not yet ideal for a thorough investigation on DOI in Maltese, we nevertheless want to get an idea of how the variables described above are weighted against each other.

A nice way to go about a probabilistic distribution of a response variable (indexing in our case) is using conditional inference trees (Tagliamonte & Baayen 2012, Levshina 2015). Like a logistic model, a decision tree makes a prediction of an outcome based on given variables. In our case, the outcome is binary, that 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. 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, as in our case. The recursive partitioning of conditional inference trees, as used in the present study, is based on repeated significance tests, providing better predictive performance 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.

The first tree we present in Figure 4.5 shows the effects of all possible predictors mentioned above: givenness, identifiability, PoS of the noun phrase and order of verb and object. 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 first subset, with OV (object-verb) word order, branches to the left, and the second subset, which entails clauses with VO (verb-object) order, branches to the right. This means that the variable word order has the strongest association with indexing, and we here see again that in our data, all preverbal objects are indexed (node 2). The strongest predictor for the subset of VO is discourse accessibility (node 3); the probability for given referents (node 4) to be indexed lies at 100% if they are also pronominal (node 5), and at around 20% if the object is a proper noun (node 6). For discourse-new referents, a split occurs (at node 7) on the basis of identifiability. As the splits are always binary, the two values definite and specific are grouped together and opposed to the value non-specific, which is a 100 percent indicator of non-indexing in the data. In a nutshell, the tree shows that the strongest predictor for object indexing in Maltese is word order, followed by discourse accessibility, with given referents being more likely to be indexed. Within the given referents, PoS is the strongest predictor for indexing, and for the new referents, it is identifiability, with no significant split between definite and specific referents.

Figure 4.5 shows that the strongest predictor for DOI in Maltese seems to be word order; but just as object indexing itself, we can assume that word order variation is a differential pattern depicting the argument's referential properties. Therefore, we built another tree model, without word order as a potential predictor.

This second model in (Figure 4.6) shows that without word order, identifiability is the strongest predictor for indexing (split at node 1). This time,

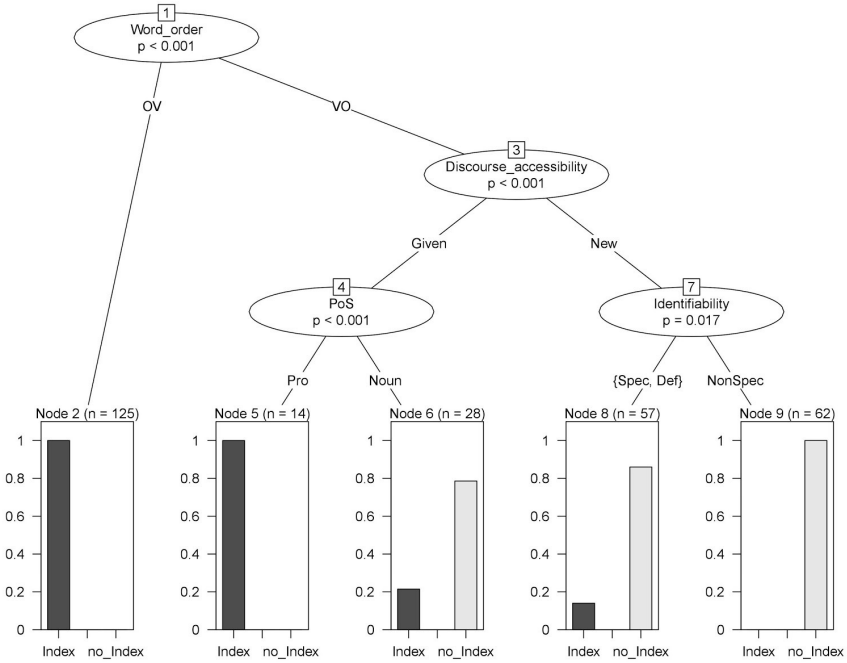

Figure 4.5: Conditional inference tree with all potential predictors for indexing

definiteness outweighs the other two values. The second split within the specific/ non-specific subset then again occurs on the basis of this feature, which means that also the difference between specific and non-specific referents is significant, with specific referents having a likelihood of just over 20% of being indexed (node 3). As for the right branch, PoS subdivides the definite referents (nodes 6 and 7). In sum, disregarding word order, identifiability (definite vs. specific or non-specific) is the strongest predictor for indexing in our data, followed by PoS of the object NP, with pronouns (which are inherently definite) always being indexed.

4.3.5 Summary

Looking into systems where object indexing is not grammaticalized, and has often been labelled optional (as is the case in Maltese), it is impossible to account for it on the basis of rules. All one can do is describe it bottom-up and investigate the variables which might correlate with it to various degrees. In

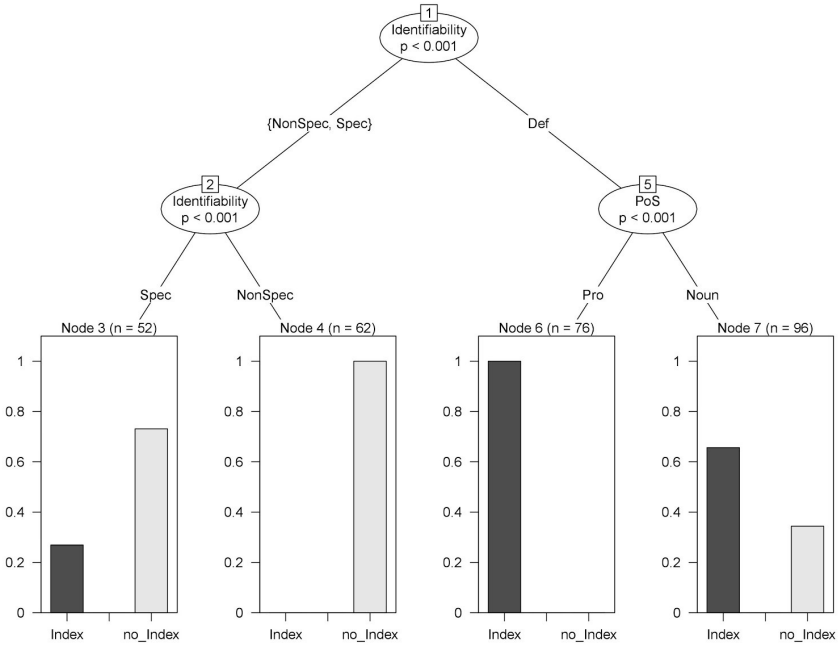

Figure 4.6: Conditional inference tree without word order as predictor

this chapter, we set out to do just that for Maltese using corpus data. Due to the realities of the available Maltese corpora, we had to deal with several cutbacks with regard to our database of factors: indexing overt objects is nearly non-existent within the written parts of the corpus, and rare in the quasi spoken transcripts of parliamentary debates. In order to find a comparable number of clauses with indexed and non-indexed referents, we extracted a random sample of clauses containing either *nagħmlu* ('we do/make' without index) or *nagħmluha* (with an index) from the parliament text type, and added an additional layer of annotation of variables thought potentially relevant for topicworthiness and thus indexing. From the variables which were left, accessibility, identifiability, PoS, and word order turned out to be significant within our subcorpus.

Using conditional inference trees, we were able to identify how the variables interrelate. Considering all four variables, word order is the strongest predictor for direct object indexing: in our data, all preverbal objects were indexed. The second strongest predictor among the postverbal objects is acces-

sibility, with discourse-given-referents being much more likely to be indexed than discourse-new ones. However, leaving out word order and only looking at the referential properties, we are facing identifiability as being the strongest predictor for DOI in Maltese, followed by PoS of the noun phrase.

In summary, DOI shows a strong correlation with identifiability of the referent. However, the results also show that indexing is not restricted to specific or definite referents, neither is there an absolute obligatoriness for these referents to be indexed. DOI in Maltese 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 referents' argument role, but other factors connected to it. These findings are neither surprising nor new. But they confirm what has been said about the inadequacy to try to find hard and fast rules to account for DOI where the phenomenon has not grammaticalized. Our approach shows that the findings of previous studies are in accordance with the outcome of a quantitative - though provisional - corpus study, and that such studies can help at getting a deeper understanding of the interactions of the different variables involved.

4.4 Outlook

A complete investigation of a phenomenon such as DOI has to be done on the basis of spontaneous spoken data, as the triggering factors are often related to topicality or salience of the referent. To this day, there is no corpus of such data for Maltese, so for a preliminary investigation, we settled for a subcorpus of parliamentary transcripts to see what this pilot study might reveal with regard to the different factors triggering DOI. Our next step will be a more in-depth investigation of the matter once a suitable corpus is available. This will make it possible to account for more variables which might be of importance, too, such as animacy of the referent or modification of the noun phrase. Also, different verb semantics have to be investigated along with referential features in order to be able to make more profound statements about the nature of DOI in Maltese.

Chapter 5

Variable index placement in Gutob from a typological perspective[1]

Abstract

Gutob (Munda, India) displays a special kind of differential indexing in that S/A indexes can attach to other hosts apart from the verb, unconstrained by syntax. Previous studies have described non-verbal index placement in Gutob as exceptional, establishing verbal indexes as the default; however, a corpus based analysis has still been owing until now. Comparative studies on variation in the placement of indexes show that there is not only inter-linguistic variation with regard to index placement, but in some cases also intralinguistic variation. Against this background, we present a case study on index placement in Gutob based on quantitative corpus data. Our analysis shows that although index placement in Gutob is in fact conditioned by discourse effects, non-verbal clitics cannot be considered particularly exceptional. Strikingly, we observe that indexing does not succumb to discourse, but can itself be used to structure it, marking the hosts as particularly noteworthy.

[1]This chapter is submitted as: Just, Erika & Voß, Judith. *Variable index placement in Gutob from a typological perspective.* Author contributions: EJ and JV worked out the annotation scheme together; EJ wrote the paper; JV provided the expertise on Gutob, as well as the corpus; annotation was largely carried out by student assistant Luna Hemmerling under the supervision of EJ and JV.

5.1 Introduction

In this study we investigate a special type of differential indexing, i.e. variation in the encoding of referents through bound person marking (traditionally referred to as agreement). Whereas differential indexing as described in previous studies (e.g. Iemmolo 2011) typically refers to conditions under which an index is present or absent, we are concerned with the position of indexes, in the following referred to as variable index placement.

In this section we will first elaborate on the concepts relevant for our analysis, including information-structural concepts. The remainder of the chapter is structured as follows: Section 5.2 will give an overview over variable index placement in various languages, first providing some theoretical background in Section 5.2.1, then illustrating these findings with examples from some languages where indexing is not confined to one position only but still syntactically determined (Section 5.2.2). We will then turn to syntactically unconstrained index placement in Section 5.2.3. Section 5.3 will start off with a summary of referent indexing in Munda languages, followed by our Gutob case study in Section 5.4. After a brief introduction to the language, we present the formal properties of S/A indexes in the language in Section 5.4.2, before presenting our corpus-based findings in Sections 5.4.3 and 5.4.4. Section 5.5 will elaborate on the discourse effects of index placement in Gutob, and Section 5.6 concludes the study with some final remarks.

5.1.1 Differential Indexing

Indexes are defined as bound markers expressing argument features, most commonly person and number, and most commonly attached to the verbal predicate. Indexing (Haspelmath 2013) is a more neutral term than agreement, as it does not presuppose any syntactic relationship between the marker and the co-referential NP, nor whether the latter is obligatorily expressed (Haig & Forker 2018). Also, the morphological status of the index, as a clitic or an affix, is irrelevant, as the latter is often unjustifiably equated with obligatoriness of marking when it comes to agreement or indexing (Haig & Forker 2018: 720). This chapter discusses differential indexing as a type of differential marking, which in turn refers to any situation where an argument of a predicate bearing the same generalized semantic role may be coded in differ-

ent ways, depending on factors other than the argument role itself (Witzlack-Makarevich & Seržant 2018: 16).

This definition of differential marking captures changes in marking patterns, but it does not imply any conditions on the differences in coding. Thus, the term differential marking can refer to differential flagging, i.e. case marking and adpositions, as well as to differential indexing. Also, it can involve the uses of different markers, the general presence of marking, or, as in the present case, the position of a marker in a clause. The definition also includes both differential marking due to predicate properties (such as TAM, polarity or clause type) as well as argument properties (both inherent as well as non-inherent).

The study of differential indexing has largely been focusing on the presence vs. the absence of indexes, both in language or family specific studies as well as in typological ones (e.g. Arkadiev 2010 or Iemmolo 2011). There has not been much cross-linguistic work on variable index placement (but see Cysouw 2003 and Forker 2016). This type of differential indexing that is not characterized by whether there *is* an index or not, but *where* the respective index is placed, is illustrated by (20). It is a minimal pair from a conversation, showing how the S/A index can attach to different constituents: the noun specifying what was brought by the guests in the first sentence (palm wine), and the amount of it (one goria) in the third sentence.

(20) Gutob

Speaker A: *iŋɋi? solop=nen gor-ek riŋ-o? ɖugu*
HES palm.wine=3PL goria-one bring-CVB AUX.PST
'Eh, they had brought one goria of palm wine.'

Speaker B: *riŋ-o?=nen ɖugu*
bring-CVB=3PL AUX.PST
'Had they brought it?'

Speaker A: *ũ solop gor-ek=nen riŋ-o? ɖugu*
yes palm.wine goria-one=3PL bring-CVB AUX.PST
'Yes, they had brought one goria full of palm wine [...].'

Examining different studies on the placement of indexes (e.g. Capell 1972, Barbosa 1996, Harris 2000, Baker 2002 or Dixon 2002) shows that there is both interlinguistic and intralinguistic variation with regard to index placement. Not only do different languages prefer different positions for referent

indexing, but very often a given language has several potential positions for an index. And although Cysouw's (2003) study already provides an insightful description of the cross-linguistic variation with regard to the placement of indexes, there is to our knowledge no corpus-driven study to account for the internal variation in a language in which the placement of the index is not predictable by grammatical rules but is sensitive to pragmatic and/or semantic factors.

Such lack of hard coding for differential marking phenomena makes them especially liable to what is often called "optionality" in reference grammars and other language specific studies. This usage of this term somewhat blurs the fact that the choice of one marking strategy over another might be well motivated, albeit not necessarily syntactically. Usage based studies have shown that even though certain constructions might not be put down to a single factor which is easy to discern, the respective form serves an intentional communicative goal on the part of the speaker (see Schikowski 2013 on DOM in Nepali, Erika & Witzlack-Makarevich (accepted) on differential P indexing in Ruuli, or Just & Čéplö (to appear) on the same phenomenon in Maltese). Section 5.1.2 now briefly elaborates on information-structural categories, especially *topic* and *focus*, and how they are dealt with in the present study.

5.1.2 A note on information-structural categories

Unlike morphological or semantic referential properties (such as gender or animacy, respectively), information-structural phenomena are often difficult to identify as the cause of differential marking patterns (Witzlack-Makarevich & Seržant 2018: 10–11). The reason for that is the variety of discourse phenomena associated with the traditional categories topic and focus, and the linguistic diversity both in form and function of the features involved. It is therefore questionable to conceive of topic and focus as language external universal categories reflected in cross-linguistically stable categories (Matić & Wedgwood 2013, Ozerov 2018, Ozerov 2021).

It has been generally accepted that topic is associated with givenness, a high degree of identifiability, and is assumed to relate to the hearer's knowledge. Focus, on the the other hand, brings about an information update, and is thus associated with notions such as newness or contrast. And although there are comparable constructions in different languages which are ascribed to the

concepts of topic or focus (such as left-dislocation or clefts), they are used to map different types of interactional management (Ozerov 2018) and there is no one-to-one relation between recurrent structures and their pragmatic effects in the respective languages (e.g. Gómez González 1997 or Skopeteas & Fanselow 2010).

Following Matić & Wedgwood (2013), Ozerov (2018) convincingly argues that topic and focus are not universal categories but rather constitute umbrella terms for a pool of different features such as – in the case of focus – contrast, correction, or an answer to a content question (also see Mushin 2006: 292–293). This clustering of features under a single term like focus has lead to the application of testing methods which constrain the interpretation of a linguistic form: once a form has been ascribed to expressing one of the prototypical features (e.g. if it marks contrast or the new piece of information in an answer), some of its other functions might be overlooked, resulting in a biased or incomplete picture of its actual contribution to information management[2].

Differential indexing entailing the absence of an otherwise present index has often been considered to be such a structure: referent indexing (often, but not exclusively, of S/A) can be suspended if the respective referent is in focus (see e.g. Ouhalla 1993, Lambrecht & Polinsky 1997, Mereu 1999 and Siewierska 2004: 159–162). But differential marking manifested in the *placement* of an index rather than its absence has been ascribed to focus of the *host* constituent (Cysouw 2003, Forker 2016).

Considering example (20) from Gutob above, it seems tempting to assume the index marks focus, as it can be interpreted as emphasizing its host constituent: *what* was brought by the guests in the first sentence in (1) and, upon further request on the part of the dialogue partner, *the amount* of what has been brought in the last one. However, due to the reasons just outlined, we avoid an *a priori* establishment of a focus category for Gutob in the present study, and instead give a bottom-up description of the motivations for the shift of the index from the verbal predicate. Nevertheless, the notions of focus and topic have played a central role in the description of differential indexing phenomena, and especially focus has been considered very important in accounts

[2]The term information management is used as an alternative to information structure, bypassing the challenges of the traditional notions associated with the latter (Ozerov 2018, 2021)

of index placement. Thus, the terms are used in Section 5.2, whenever we adopt them, as they have been used in the respective studies.

5.2 Variable index placement

In the following section we consider the distinction between indexes which have a dedicated host and those which have variable hosts. As for the latter, one can further differentiate between those which occupy a fixed position syntactically and those which do not, as is the case in Gutob.

5.2.1 Typological overview

Siewierska (2004: 26–32) gives an extensive description of indexes in various languages which are not always attached to a particular type of stem (and which are therefore not "bound" in her terminology) but which have a designated syntactic position. Designated syntactic position means that there is some variation with regard to the part-of-speech of the host word, however the position of the index within the clause is nevertheless grammatically determined and not flexible. Following Anderson (1993: 74), she uses the typology of specialized positions listed in (21) for indexes which do not always attach to the same stem. She also mentions languages in which two of the positions in Anderson's (1993: 74) typology are possible (Siewierska 2004: 26-32).

(21) a. verb phrase initial position
b. verb phrase final position
c. second position
d. penultimate position
e. pre-head position
f. post-head position

Example (22) from Kharia and (23) from Vera'a illustrate indexes showing variability with regard to their host, but occupying syntactically fixed positions. In Kharia, the S/A index is an enclitic to the verb in affirmative clauses, as in (22a), whereas it attaches to the negative particle in negated clauses, as in (22b). According to Anderson's (1993) typology, indexes in Kharia are therefore either in a pre-head or in a post-head position within the

verb phrase,[3] conditioned by polarity, and no other position is possible. In Vera'a, S/A referents are only indexed in the aorist, and not all person markers in this paradigm are phonologically bound. However, the ones which are bound (as the non-singular marker *=k* in (23b)) attach to whatever element precedes the predicate, which in most cases is the last word of the S/A NP, or an adverb of a closed class, which can intervene between the S/A NP and the predicate (Schnell 2018: 750–752).

(22) Kharia (Munda, Peterson 2011: 58)

a. *kayom=ta=ɲ*
speak=PRS=1SG
'I speak.'

b. *um=iɲ kayom=ta*
NEG=1SG speak=PRS
'I don't speak.'

(23) Vera'a (Oceanic, Schnell 2018: 759)

a. *dē=k van 'ō' di mē=n sisidiñ*
1PL.INCL=AOR:NSG go carry 3SG DAT=ART RDPL.bird.trap
'and we will go bird catching with him.'

b. *gidu=k van=ēk traem*
1DU.INCL=AOR:NSG go=AOR:NSG try
'"Let's go try!"' [lit: "We two will/shall go, (we) will/shall try."]

Example (23) from Vera'a shows that the typology of index positions in (21) is not exhaustive: the index is not part of the verb phrase (therefore neither in position a., b., e. or f.), nor is it in second or penultimate position with regard to the whole clause. It is "detached" (Bickel & Nichols 2007: 176) from its predicate, but unlike positions c. and d., its position is fixed in relation to the predicate, always directly preceding it.

5.2.2 More than one option for the position of an index

As outlined in the previous section, even in systems where indexing has grammaticalized in a sense that it becomes obligatory (even if sometimes only in parts of the paradigm), the index can still be flexible in selecting a host, which

[3] These indexes have also been called anticipatory clitics (Peterson 2011: 61–62, see also Dixon & Aikhenvald 2002: 46).

once again demonstrates that there is no equation of obligatoriness with the status (as an affix or clitic) of marking (see Haig & Forker 2018: 720).

There are, however, languages which display an intermediate position between a grammatically conditioned, fixed syntactical position for an index on the one hand, and great freedom of position on the other hand: in such languages, the index has a default position, often the head of the predicate, but can leave it and appear in an alternative, syntactically fixed position. One of these languages, also reported by Siewierska (2004: 27–29), is Nganhcara, where indexes can occur in two syntactic positions. Smith & Johnson (1985: 104) state that the favored position is that encliticized to the last element before the verb (as in 24a), though indexes also occur encliticized to the verb itself (as in 24b).

(24) Nganhcara (Pama-Nyungan, Smith & Johnson 1985: 102, 106)

a. *Nganhca nga'a-nhca yenta*
1PL.EXCL.NOM fish-1PL.EXCL.NOM spear
'We speared the fish.'

b. *Nganhca nhingu nga'a waa-ngu-nhca*
1PL.EXCL.NOM 3SG.DAT fish give-3SG.DAT-1PL.EXCL.NOM
'We gave him a fish.'

As the placement of the index in these examples is not determined by phonological, morphological or syntactic factors, it probably serves a communicative goal of the speaker. The discourse function of index placement is somewhat under-studied, but Cysouw (2003) investigates the attraction of indexes to various positions of discourse prominence in a sample of 40-odd languages. He aligns the different positions of index placement with the status of its host with regard to focus. Indexes which are not confined to the head of the predicate most often attach to elements which are considered to be inherently focused, such as question words and negation markers. Next in the focus hierarchy are constituents with intended focus, i.e. their focus status arises out of a particular situation, as, for example, NPs in contrastive or emphatic contexts. The two other focused contexts that play a role in this hierarchy are stage setting (clause linkers and adverbs), and modified (indefinite and quantified) NPs.

As an explanation for this attraction of indexes towards focal elements, Cysouw (2003) proposes that as indexes themselves are highly topical, and

therefore non-focal, this combination is a "juncture of opposites", the less focal element binding itself on to the most focal one. He also demonstrates that clause-second position is very frequently used for indexes, either as default (such as in some Pama-Nyungan languages like Yingkarta, Wajarri, Ngiyambaa or Warlpiri as well as in the Uto-Aztecan language Yaqui[4]) or as an alternative to a position within the verb phrase (e.g. in Suleimaniye Kurdish or Cypriot Greek).

An example for clause-second position as an alternative to verbal position comes from Kuuk Thaayorre. In this language, there is differential indexing in two ways: i) the index is not (yet) grammatically obligatory (Gaby 2006: 342–343)[5], as exemplified by (25a), which would be equally grammatical without *=ay*, and in (25b) where the third person singular accusative *=unh* features twice; and ii) the position of the index alternates between clause-second position, as in (25c) and (25d), and verb-final position, as in (25a), which is preferred (Gaby 2006: 216–217).

(25) Kuuk Thaayorre (Pama-Nyungan, Gaby 2006: 217)

a. *ngay ii-rr-kuw Darwin-ak yat=ay*
1SG.NOM there-towards-west Darwin-DAT go.PFV=1SG.NOM
'I went west to Darwin.'

b. *thil=unh koow rathirr=eln=unh*
again=3SG.ACC nose.ACC chop-PFV=3PL.ERG=3SG.ACC
'They slashed his nose once more.'

c. *inh'nhul=ay yik, kuuk inh'nhul*
this.one=1SG.NOM say-NPST word this.on
'I'm telling this story.'

d. *ngul=ul=unh man.pert-e theerka-n-r*
then=3SG.ERG=3SG.ACC shoulder-ERG return-TR-NPST
nhaknkath-an
camp-DAT
'And he carried it back home on his shoulder to camp.'

Second position clitics, also called Wackernagel-clitics, are cross-linguistically quite common and person indexes are not the only elements which are

[4]For more information on some Uto-Aztecan languages where the index can shift away from the verb to the clause-second position see e.g. Press (1979: 77) on Chemehuevi or Wistrand Robinson & Armagost (1990: 250–252) on Comanche.

[5]Gaby (2006) argues that the index is in the early stages of grammaticalizing from the respective free pronouns into indexes.

prone to this position, but also other inflectional material, such as TAM or evidentiality markers, or inflected verbs (Anderson 1993, Mushin 2006).[6] There is a considerable amount of literature addressing the syntactic and phonological properties of Wackernagel-clitics, but less on the functions of this position (but see Anderson 1993, Mushin 2006, Mushin & Simpson 2008). Two rather contrary motivations that draw elements to the clause-second position have been proposed. On the one hand, it was suggested that the clause-second position elements (indexes or other inflection) are actually targeting a clause initial position, but are blocked from occurring there due to language-specific phonological or morphosyntactic constraints, and therefore shift to clause-second position (Anderson 2005: 142–152). On the other hand, it was suggested that the elements in question are "bare-bones grammatical information" (Mushin 2006: 296) and thus are attracted to elements in the first position, which, in turn, is generally recognized to be associated with focal effects (e.g. Mithun 1992, McConvell 1996 or Cysouw 2003).

The latter idea of a syntactic "beacon" (Mushin 2006: 296) attracting markers or constituents with low pragmatic impact resulting in Wackernagel-clitics or clause-second position verbs in many languages is compliant with Cysouw's (2003) observation that indexes as "less focal elements" bind themselves "on the most focal element". It also goes well with the fact that there are languages where markers of modality or evidentiality attach either to the verb or to any other focused constituent (Facundes 2000, Aikhenvald 2003).[7]

In a nutshell, there is evidence that indexes (as well as other grammatical markers) are often attracted to constituents that the speaker wants to highlight – and these constituents are frequently found clause initially. This can lead to indexes being found in clause-second position, either exclusively or as an alternative to attaching to the verb. We will now turn to languages where indexes are not confined to one or two syntactic positions but can be attracted to any constituent in the clause.

[6] Clause-second position sounds more straight forward than it is, as there is variation with regard to the rules of attachment and whether the clitics attach to the first (prosodic) word or first (prosodic) constituent of a clause.

[7] As outlined in Section 5.1.1 no two languages have identical categories of focus, i.e. focus marking in language A does not have the exact same pragmatic effect as focus marking in language B; however, there is undoubtedly a set of communicative functions which can be ascribed to this traditional notion, and which overlap to various degrees from language to language (see Ozerov 2018 for a discussion and overview of these features).

5.2.3 Syntactically unconstrained index placement

The placement of indexes may be even less constrained by morphosyntactic criteria than having several alternative syntactic slots. Some language descriptions suggest that the position of an index cannot always be lead back to a hard and fast grammatical rule, and is therefore sensitive to information management. For Mutsun, an extinct Utian language of California, Okrand (1977: 171) reports that indexes are usually second position clitics, following the first word of a sentence, whatever this word may be. However, there are a few exceptions to this rule where the index attaches to other constituents. The motivation for this cannot be discerned based on his data. Also, the distribution of the indexes and respective independent pronouns, which cannot be used together, remains unclear.

The situation seems to be even more complex in Ute (Uto-Aztecan, Givón 2011: 170–192). Here, the same set of indexes can in principle be used for either S, A or P, and compete with free pronouns as well as zero anaphora. Zero anaphora has been identified as means of tracking of S/A if it persists as "agentive" subject; if participants start interacting, indexes are used for the absolutive (S/P) argument. As for the position of the index, Givón (2011: 170) states that it can attach "not only to the verb, but to any first word in the clause". However, in his count of host positions in the clause, he finds that although 81.9% of all non-verbal indexes are in fact in clause-second position, nearly 20% of non-verbal indexes are not clause-second; unfortunately he does not enlarge upon these cases.

The conditions of index placement are described more transparently for Sanzhi Dargwa (Nakh-Daghestanian).[8] The default position of the index is postverbal, but it can also be conditioned by the focal status of the host of the index, a phenomenon called "floating agreement" by Forker (2016): the index leaves its postverbal position (exemplified in 26a) and floats off to constituents which are focal or contribute new information, thus serving to emphasize its host, as e.g. 'the dishes' in (26b) or 'I' in (26c). The emphasized constituents are underlined in the translation. According to Forker, the host can be any other constituent without fixed syntactic position. However, these examples are restricted to elicited sentences (Forker 2016: 1). Which role is indexed

[8] See (Bickel & Nichols 2007: 176–177) quoting (Kibrik 1997) for discussing a similar phenomenon in Tsakhur, another Nakh-Daghestanian language.

in Sanzhi-Dargwa is governed by the person hierarchy 2 > 1 > 3 (Forker 2016: 4).

(26) Sanzhi Dargwa (Nakh-Daghestanian, Forker 2016: 2)

a. *du-l hana t'alaʕħ-ne ic-an = da*
1SG-ERG now dishes-PL wash.IPFV-PTCP = 1
'Now I will / have to wash the dishes.'

b. *du-l hana t'alaʕħ-ne = da ic-an, c'il t'ult'-e*
1SG-ERG now dishes-PL = 1 wash.IPFV-PTCP then bread-PL
d-uc'-an = da
NPL-bake.IPFV-PTCP = 1
'Now I will / have to wash the dishes, later I will make bread.'

c. *du = da Sanijat-li-j χabar b-urs-an*
1SG = 1 Sanijat-OBL-DAT story N-tell.PFV-PTCP
'I will / have to tell Sanijat the story.'

Forker (2016: 20) also mentions Polish, Paez (isolate, Colombia) and Zargulla (Omotic) as further examples for index placement conditioned by information management. In Zargulla, the situation is quite intriguing and deserves to be elaborated on: first of all, S/A indexing is described as "optional", which is in itself an interesting fact worth to be studied in further detail;[9] Amha (2007) mentions that identifiability and animacy play a role to some extent. However, a prerequisite for indexing is the presence of the focus marker *-tte*, i.e. indexing cannot occur on its own, at least not in declarative clauses. The focus marker, on the other hand, can be used on its own without an index, as in (27a), and it can shift to various constituents to mark them for focus, and the index (if present) always moves along, as exemplified in (27b)-(27d). The index can, however, be attached to question words without the focus marker, as in (27e) (Amha 2007: 200–202).

(27) Zargulla (Omotic, Amha 2007: 201–202)

a. *s'úho ʔíndó-y ʔúkkó-tte-ínne*
Tsuho:GEN mother-NOM be.close-FOC-PST
'Tsuho's mother moved closer.'

b. *naʔá-z-í kátsa bays-í, maaʔó = tte-s sang-í,*
child-M-NOM grain.ABS sell-CVB cloth.abs = FOC-3SG.M buy-CVB

[9] In contrast to the indexes in declarative clauses, which can be used only in focused constructions, S/A indexes in the negative interrogative and imperative/optative are obligatory and seem to be well entrenched and older (Amha 2009: 215).

ɗum-us-í *yeénne*
be.dark-CAUS-CVB come.PFV

'The boy sold grain, bought cloth, and came late.'

c. *naʔá-z-í* *kátsa* *bays-í,* *maaʔó* *sang-í=tte-s,*
child-M-NOM grain.abs sell-CVB cloth.ABS buy-CVB=FOC-3SG.M
ɗum-us-í *yeénne*
be.dark-CAUS-CVB come.PFV

'The boy sold grain, bought cloth, and came late.'

d. *naʔá-z-í* *kátsa* *bays-í,* *maaʔó* *sang-í,*
child-M_NOM grain.ABS sell-CVB cloth.ABS buy-CVB
ɗum-us-í=tte-s *yeénne*
be.dark-CAUS-CVB=FOC-3SG.M come.PFV

'The boy sold grain, bought cloth, and came late.'

e. *ʔas'o-y* *ʔánna-s* *yene*
man-NOM where-3SG.M exist.NPST

'Where is the man?'

Question words can be considered inherently focal, and this seems to be the reason why additional focus marking is blocked from them in Zargulla, but indexing is nevertheless possible. Before discussing index placement in Gutob, which is also syntactically unconstrained but sensitive to information management, we will first provide a short overview of referent indexing in Munda languages more generally in the following section. This is worthwile as S/A indexing in Gutob is quite exceptional compared to the other members of the family.

5.3 Referent indexing in Munda

The Munda languages belong to the Austroasiatic phylum and are spoken in Central and Eastern India, surrounded by Indo-Aryan and Dravidian languages. The internal classification of Munda languages is still a matter of debate, but there is some consensus that Gutob is most closely related to Remo and Gta' (Anderson 2008: 1–4; Sidwell 2015: 194–197).

The verbal complex in Munda languages exhibits a range of inflectional categories, including indexing for person and number. However, indexing in the languages of the family differs with regard to three aspects. There is variation among the languages concerning i) the morphological form of the indexes;

ii) the argument roles which can be indexed (see Cysouw (2004) and Anderson (2007: 64) for an overview); and iii) the position of the indexes. While object indexes are mostly suffixes, S/A indexes are either prefixes or enclitics/suffixes[10]. Table 5.1 shows an overview of Munda S/A indexing. Korku is an exception among the Munda languages, and is not listed in the table, as it lacks indexing altogether, except for of some locational copular expressions (Anderson 2007: 64). The variation in form and function of indexes in Munda languages has caused considerable debate with regard to their historical development (cf. Pinnow 1966, Anderson 2001, Anderson & Zide 2001 and 2007, and Donegan & Stampe 2004). In this section we will mainly pay attention to enclitic S/A indexing in Munda languages, as this is the only indexing that is found in Gutob, the language of our primary focus.

[10]Even for individual languages there is sometimes no consistency in labelling the indexes, cf. e.g Osada (2008) (suffixes) and Anderson (2007) (clitics) on the Mundari indexes.

	1SG	1DU.INCL	1DU.EXCL	1PL.INCL	1PL.EXCL	2SG	2DU	2PL	3SG	3DU	3PL
Santali	=ɲ	=laŋ	=liɲ	=bon	=le	=m	=ben	=pe	=e	=kin	=ko
Mundari	=ñ	=laŋ	=liŋ	=bu	=le	=m	=ben	=pe	=e(ʔ)/ =i(ʔ)	=kin	=ko
Kharia	=[i]ɲ/ŋ	=naŋ	=dʒar	=niŋ	=le	=[e]m	=bar	=pe		=kijar	=ki, moj
Remo	=niŋ	=naŋ		=naj		=no	=pa	=pe			
Gutob	=niŋ			=nei		=nom		=pen			=nen
Juang	-V_1-	ba-		nV_1-		mV_1-	ha-	(h)V_1-		-kia	-ki
Gorum	ne-			le-		mo-		bo-			gi-
Sora	-aj			-be	ə-…-aj			ə-…-ɛ			-dʒi
Gta’	n-			ni-	næ-/nɛ-	na-	pa-	pe-			-har-

Table 5.1: S/A indexes in Munda languages (adapted from Anderson (2007: 76), supplemented by data from Neukom (2001) on Santali and Osada (2008) on Mundari

)

Santali and Mundari, as well as Kharia, Remo and Gutob display S/A indexing as enclitics. In most of these languages, S/A indexes either attach to the main verb or to the constituent immediately preceding it, though the factors which determine the placement can vary. In Mundari, for instance, indexes obligatorily attach to the element immediately preceding the predicate, except if the clause consists of a predicate only in which case they attach to the verb (Hoffmann 1903: 12–13). The situation is similar in Santali where the S/A indexes (which are obligatory for animates, and possible for inanimates) either attach to the verb itself, or, more commonly, to the immediately preceding element if there is one. However, the shift of the S/A index away from the verb does not seem to be obligatory, but more of a strong tendency as there are exceptions (which are not elaborated on, see Neukom 2001: 113–115). The following examples illustrate the acceptability of the index attached to the constituent preceding the verb, to an independent pronoun in (28-a), or to the affirmative particle in (28-b), as well as attached to the verb itself, as in (28-c). Example (28-d) is ungrammatical, as animates always have to be indexed:[11]

(28) Santali

Q *cala-k'a-m?*
go-IND-2SG
'Will you go?.'

A a. *hɛ̃, iɲ-iɲ cala-k'a.*
yes 1SG-1SG go-IND
'Yes, I shall go.'

b. *hɛ̃-ɲ cala-k'a.*
yes-1SG go-IND
'Yes, I shall go.'

c. *hɛ̃, cala-k'a-ɲ.*
yes go-IND-1SG
'Yes, I shall go.'

d. * *hɛ̃, iɲ cala-k'a.*
yes 1SG go-IND
'Yes, I shall go.'

[11] Also P and G, as well as possessors can be indexed if they are animate. However, the position of these indexes is confined to the verb, i.e. non-S/A indexes cannot attach to other constituents (Neukom 2001: 115–117), although they are morphologically identical.

Gutob, as will be shown in the following sections, seems to be the only language of the family where the position of the index is not determined on purely syntactic grounds.

5.4 Case study: S/A indexing in Gutob

Section 5.3 shows that S/A indexing in Munda languages is not uniform. The South Munda languages Kharia, Gutob and Remo all have enclitic indexes which are formally similar. Gutob and Remo clearly form a subgroup of the family, but with regard to indexing, Gutob shows more similarity with Kharia and the North Munda languages Mundari and Santali.

The Remo S/A indexes are identified as clitics, however, nothing about hosts other than verbs is mentioned in the descriptions (Fernandez 1983, Swain 1997 and Anderson & Harrison 2008). In Kharia, exemplified in (22) above, the subject index is enclitic to the verb in affirmative clauses and attaches to the negative particle in negative clauses, and its placement is therefore rule governed. The situation is different in Gutob, where the position of the S/A index is syntactically unconstrained. This will be described further in Section 5.4.2, after a brief introduction of the language, including some relevant information on its morphology and syntax.

5.4.1 Language and data

The Gutob language (ISO gbj, sometimes referred to as Bodo Gadaba) is mainly spoken in the Koraput District in the highlands of the state of Odisha and in neighbouring districts in the state of Andhra Pradesh in Eastern India. In the present study, the name Gutob is used as it is the name the speakers themselves use both for their language and for themselves as a social group. There is little reliable information on the number of speakers of Gutob. The census of 1991 counts around 28,000 Gadabas, but does not distinguish between Gutob Gadaba and the Dravidian Ollari Gadaba. Estimates range from 5.000 to 20.000 speakers (Rajan & Rajan 2001, Griffiths 2008, Berger 2015). Our study is based on a corpus collected during a recent language documentation project (Voß 2018) between 2016 and 2018. The whole corpus contains 18.5 hours of transcribed audio and video recordings consisting of fictional stories, conversations and interviews, personal narratives and elicitations.

The Gutob people in the Koraput district live in a multilingual setting and Desia, the regional Indo-Aryan *lingua franca*, is present on a daily basis. In the village of Jalarhanzar, where the data for this study were obtained, most younger speakers have shifted to Desia. According to Griffiths (2008: 635–636), there are at least two dialects of Gutob, which he calls Koraput Gutob and Andhra Gutob. Most of the previous research as well as the present analysis is based on Koraput Gutob.

The prevalent constituent order in Gutob is APV, although A and P are sometimes reversed. The clause-final position of the verb is more fixed. In afterthought constructions A and more commonly P may follow the verb, but are clearly set off prosodically. Adjectives, adverbs, demonstratives and quantifiers usually precede their head. With regard to morphosyntactic alignment, Gutob is a nominative-accusative language.

As is typical for Munda languages, Gutob displays complex verbal morphology. Gutob has basic voice, which is marked by TAM/voice portemanteau suffixes. Voice marking closely correlates with transitivity and most verbs are either always in the middle or always in the active voice. A small set of voice-alternating verbs exists, such as verbs with a causative alternation, e.g. 'break' or 'tear', or motion verbs in which voice marks directionality. A change of voice from active to middle can be employed to reduce transitivity of normally transitive verbs, e.g. in reflexive constructions, although this is rare. To increase the transitivity of an otherwise intransitive verb it has to take the causative marker, a change of voice alone is not sufficient in this case. Further categories marked on the verb are negation, reality status and honorifics. The following template illustrates the morphological structure of a finite verb in Gutob. It indicates that verbal indexes are not obligatory on the verb, but if they do occur, they have a fixed slot within the verbal morphology.

NEG-	CAUS-/<CAUS>	ROOT	-TAM.VOICE	(=S/A)	-PRS	-HON

Table 5.2: The morphological structure of the Gutob verb

For our investigation of the behavior of S/A indexes in Gutob, we annotated 2318 finite main clauses for overt S/A reference through an index and/or a pronoun in a subcorpus of 32669 words, comprised of 12 narratives and stories from everyday life (approx. 360 min) by 7 speakers (20-70 years,

all female). Clauses were annotated for person and number, the position of the index, and the part of speech of the host of preverbal indexes; with verbal indexes, we also annotated whether non-verbal placement would have been syntactically possible (see Appendix C for details).

5.4.2 Formal properties of indexes in Gutob

As for S/A reference in general, third person NP arguments may be omitted if the referents can be inferred from the context. First and second referents, however, are usually expressed by a full pronoun and/or a bound index. The person indexes are formally identical to the free personal pronouns, except for the third person, which is zero marked. The clitic used for third person plural *=nen* is a general plural marker, which, apart from marking reference to a third person plural S/A argument, also attaches to NPs to mark them for plural, as well as to imperatives with second plural reference. Table 5.3 illustrates the identity in form of the free pronouns and the indexes for all but the third persons.

	SG		PL	
	free	bound	free	bound
1	niŋ	=niŋ	nei/naj	=nei/=naj
2	nom	=nom	pen	=pen
3	maj	∅	maj=nen	=nen

Table 5.3: Free pronouns and bound indexes

Despite the fact that person indexes are crosslinguistically commonly derived from pronouns, and that similarity between indexes and their free counterparts is more common in first and second person than in third person, identity of the two paradigms as found in Gutob has to be viewed as exceptional (Siewierska 2004: 251). The formal identity between the two paradigms as well as the fact that referents can be marked by either free forms or bound forms or both at the same time (see examples in 32 below) has induced Voß (2015) to deal with the issue of whether the two paradigms can actually be

distinguished. She finds that they can: clause-initial person markers like in (29) are unambiguously pronouns, as they can host clitics which are reserved for nominals, such as the additive marker *=sa* in (29b). Also, they are not repeated in coordinated clauses, as in (29c). As for indexes as part of the verbal predicate, it is clear that they are part of the morphology, as they have their fixed slot within the verbal template (see Table 5.2 above). Also, they are often repeated in coordinated clauses with the same S/A referent (29c).

(29) a. *uraʔ kunig, nom maŋɖem piŋ-loŋ*
NEG old.woman 2SG why come-FUT
'No old woman, why should you come.'

b. *niŋ=sa ɖoɲ-tu=niŋ*
1SG=too cook-FUT=1SG
'I, too, will cook.'

c. *nom ca iɖ-tu=nom, lai som-tu=nom maʔ som-tu=nom*
2SG tea drink-FUT=2SG rice eat-FUT=2SG curry eat-FUT=2SG
'You drink tea, eat rice and eat curry.'

Referent expression can either be in the form of a free pronoun only, as in (29a), or an index, as in (29c), or both, as in (29b), and also in (32b) below, where there are even three realizations of the same referent in one clause.

If indexes are part of the verbal complex, a distinction has to be made with regard to the form of the predicate. Complex predicates in Gutob distinguish between explicator verb constructions and auxiliary constructions (Voß 2015: 225–226). Explicator verb constructions consist of a main verb carrying the semantic load, plus a second inflected verb. Explicator verbs are homophonous with lexical verbs but have undergone extensive semantic bleaching and express aktionsart distinctions (cf. Butt & Geuder 2003: 330–331). In these constructions, the clitic tends to attach to the last element, the explicator verb, as (30a) shows, although there are exceptions. In auxiliary constructions conveying TAM distinctions, however, the index is more likely to attach to the main verb, as shown by example (30b).

(30) a. *nom dapre moɽ-gu piŋ-gi=nom*
2SG quickly rise-MID.CVB come-MID.PST=2SG
'You got up quickly.'

b. *bezri aɽo=bo? ui-gi=pen ɖugu*
tomato garden=LOC go-MID.CVB=2PL AUX.PST
'You went to the tomato garden.'

There can be further variation in complex predicates when it comes to negation. Whereas in simple predicates, standard negation is marked by a prefix (see Table 5.2 above) which does not affect index placement, auxiliary constructions are negated by means of the negative particle *ura?*. In present tense, *ura?* usually replaces the auxiliary, whereas in the past tense, the auxiliary follows the negative particle. In our subcorpus, out of 62 clauses which are negated by means of *ura?*, there are 19 clauses where S/A reference is either expressed by a pronoun or a preverbal index (which will be further elaborated on in 5.4.4). In the remaining 43 clauses which have an index in the negated complex predicate, we find six instances (14%) where the index attaches to the negative particle *ura?*, while it attaches to the verb in the remaining clauses. So while this particle may host the index, as in (31a), this is by no means the mandatory position. More commonly, the index attaches to the lexical verb, as in (31b). This is different from the situation in Kharia, where a negative particle becomes the mandatory host of the person index.

(31) a. *buzei ura?=niŋ*
INF.understand NEG=1SG
'I didn't understand.'

b. *sa?mel ri~riŋ=niŋ ura?*
millet INF~bring=1SG NEG
'I didn't bring millet.'

The verbal position has been considered default in previous studies, although the placement of the indexes varies considerably. What has caused some debate in the formal analyses of the Gutob person marking system are the indexes which are not part of the verb complex, in the following called preverbal indexes. Earlier accounts have ascribed the non-verbal placement of indexes to the inherent features of the hosts, with certain adverbs, adverbials and interrogatives being preferred as hosts (Zide 1997). Like pronouns and verbal indexes, preverbal indexes can often be the only realization of a referent in a clause (see (20) and (36)). Voß (2015) found that preverbal indexes frequently co-occur with verbal indexes, as in (32b), which suggests a

functional similarity to free pronouns. On the other hand they usually do not host nominal morphology like the additive clitic and are often repeated in coordinated clauses, as in (32a), which makes them more similar to the verbal indexes than to the pronouns.

(32) a. *o-maj lai=niŋ beɖ-o? ma?=niŋ beɖ-o?*
OBJ-3SG rice=1SG give-PST curry=1SG give-PST
'I gave him/her rice and I gave (him/her) curry.'

b. *naj maŋɖem=naj gisiŋ=nen bon-o?=naj ɖutu*
1PL why=1PL chicken=PL raise-CVB=1PL.S/A AUX.PRS
'Why have we raised chicken?.'

In Section 5.4.3, we present the analysis of our corpus annotation regarding index placement in Gutob, before proposing an account of possible contexts in which indexes shift to preverbal hosts in Section 5.4.4.

5.4.3 First and second vs. third person reference

As has been mentioned in Section 5.4.2, there are no indexes for third person singular in Gutob and the referents are very often not overtly expressed at all. In our annotated subcorpus, there were only five instances of a third person singular S/A referent expressed by a pronoun. Third person referents can only be indexed in the plural by the use of the general plural marker *=nen*. This marker, however, behaves quite differently from the indexes expressing person and number of first and second person referents. As with the non-third person indexes, also *=nen* in its indexing function does not have to attach to the verb, as in (33) where it attaches to a preverbal adverb. However, our data show that third plural indexes attach to the verb more often than non-third person indexes.

(33) *a?=nen bana-gu beɖ-o?*
now=3PL forget-MID.PST give-ACT.PST
'Now they forgot.'

Table 5.4 shows the distribution (with absolute numbers in brackets) of indexes in affirmative clauses with third person plural vs. non-third persons, comparing the numbers for clauses with only verbal indexes, only preverbal indexes, as well as both options within a single clause. It shows that for third plural referents, preverbal indexes are uncommon, 96.7% of the clauses have

a single index in the predicate. In contrast, for non-third persons, only 59% of the clauses have verbal indexes only. In both cases a small amount of clauses also has indexes in both verbal and preverbal position. These numbers already show that the preverbal placement of an index is much more common for non-third persons, whereas for the marker *=nen* '3PL' the verbal position appears to be a default.

	verbal	preverbal	both	total
1st & 2nd	59.0% (646)	32.4% (354)	8.6% (94)	1094
3PL	96.7% (1089)	2.00% (23)	1.2% (14)	1126

Table 5.4: Verbal and non-verbal indexing for non-3rd person vs. 3PL referents

The difference between third and non-third person indexes is even more striking if one excludes those clauses that consist of a predicate only, thus making a preverbal placement impossible. In this subset, given in Table 5.5 the picture does not change much for third person: the verbal placement remains by far the most common. For non-third persons, however, the preverbal position is now much more prevalent than the verbal position. The majority of clauses, namely 54.1%, have a single index in preverbal position whereas only 31.5% have a single index in the predicate. In the remaining 14.4% of clauses there are indexes in both positions.

	verbal	preverbal	both	total
1st & 2nd	31.5% (206)	54.1% (354)	14.4% (94)	654
3PL	93.7% (552)	3.9% (23)	2.4% (14)	589

Table 5.5: Verbal and non-verbal indexing for non-3rd person vs. 3PL referents, excluding clauses comprised of verbs only

Not only do indexes for third person plural and non-third persons display very distinct distributions, but there are functional differences as well. The enclitic *=nen* can also mark plurality in nouns and functions as an associative plural marker even with uncountable nouns, as in (34). At the same time

plural NPs in Gutob do not have to be overtly marked for plural. Therefore, the use of *=nen* as an index for third person plural referents attaching to a preverbal (object) NP can cause ambiguity. Furthermore, object NPs do not have to be marked overtly by the non-pronominal object enclitic *=lai*, which would follow the nominal plural marker and eleviate the ambiguity. Consider example (35), where, similarly to (32a) above, the index could in principle also attach to *paʈai*, but this would result in the plural reference being unclear. Due to these circumstances, it was hard for several cases in our subcorpus to decide whether *=nen* is in fact an index or a marker of a nominal plural. Therefore, in the analysis of the peculiarities of preverbal indexes in Section 5.4.4 we focus on first and second person referent indexes only.[12]

(34) *tonda=nen baŋ-to=nei*
lemonade=PL send-HAB=1SG
'We send lemonade and such things.'

(35) *o-maj paʈai beḍ-tu=nen*
OBJ-3SG dress give-ACT.FUT=PL
'They will give her a dress/dresses.'

5.4.4 Preverbal indexes in Gutob

This section deals with a more detailed analysis of indexes in preverbal position. We have already shown in Tables 5.4 and 5.5 above that the assumption that the verbal position for indexes in Gutob is the default one (Zide 1997, Anderson 2007: 70–71 and Griffiths 2008: 643, 653), and that anything else is an exception, has to be reconsidered at least for non-third person indexes. This assumption might be issuing from three facts: firstly, many clauses consist of a predicate only, and do thus not provide an alternative for verbal placement of the index. Secondly, previous investigations have not differentiated between third and non-third person indexes. Finally the indexing behavior in other Munda languages might have biased previous analyses for Gutob.

Table 5.6 shows the kinds and positions of (non-third) reference in more detail. The most frequent position is, indeed, verbal, but not by a great deal: taken together, clauses with preverbal indexes (either as sole reference or

[12]For a discussion of conceptual characteristics of first and second vs. third person forms see e.g. Benveniste (1971: 195–205) or Dahl (2000).

in combination with another index or a pronoun) make up 41% of all the clauses.

verbal only	44,70% (523)
preverbal only	26,84% (314)
pronoun & verbal	10,43% (122)
preverbal & verbal	7,18% (84)
pronoun only	6,50% (76)
pronoun & preverbal	3,50% (41)
pronoun & preverbal & verbal	0,85% (10)
total	1170

Table 5.6: Type and position of reference in non-3rd persons

The preverbal elements indexes can attach to can be locative (as in 36a) or temporal (as in 36b) adverbials, object NPs (as in 36c), interrogatives (as in 32b) above, but also adjectives or demonstrative pronouns (as in 36d).

(36) a. *pen=nu=bo?=nom ui-a=be*
2PL=ATTR=LOC=2SG go-MID.IMP=HON
'You (SG) go to your (PL) place.'

b. *usoŋ muiro? gisiŋ=naj sir-o? som=be kina*
today one chicken=1PL roast-ACT.CVB eat=HON please
'Let's roast and eat one chicken today, please.'

c. *maŋ ma?=nom ɖoɲ-tu=be*
why sauce=2SG cook-ACT.FUT=HON
'Why did you cook curry sauce?'

d. *ito?=o? dinke=niŋ olai-o? ɖuloŋ*
like.this=EMPH daily=1SG hang-CVB be.FUT
'Like this I will daily hang it [the calabash] up.'

Zide (1997) in his analysis of Gutob person reference finds himself quite puzzled in view of this variation of positions and hosts, and ascribes it to "extreme (rhetorical) conditions" and something which would not come up in

"ordinary sentences" (Zide 1997: 326); he admits, however, that his data is too scarce to come up with a satisfying analysis. In his data, verbs are also the most preferred hosts, followed by certain adverbials as well as wh-words; among this class of constituents there are some members that are even more favored, namely *eke?* 'here', *a?* 'now', *begi* 'quickly', *dapre* 'afterwards', *ũdoj* 'when', *mono?* 'where' and *maŋ* 'why' (Zide 1997:313-327).

However, even though these words have a strong tendency to be the host of the S/A index, it would be likewise grammatical if the verb carried the index in their presence, as exemplified in (37). And it should be remembered that Table 5.5 above also provided quite a substantial number of cases where the verb was preceded by one or more suitable candidates for hosts but nevertheless carried the index.

(37) *a? keʈei-gu=nom*
now arrive-MID.PST=2SG
'Did you arrive just now?'

Table 5.7 provides the numbers for the different preverbal types of hosts; we can see that adverbials (including adverbial phrases like locative phrases) are indeed very frequent hosts, followed by NPs, interrogatives and object pronouns.

ADV	39% (189)
OBJ.NP	37% (176)
interrogatives	17% (83)
OBJ.PRO	3% (14)
NUM	2% (8)
DEM	2%(8)
total	480

Table 5.7: Different hosts for non-3rd person preverbal indexes

Thus, no lexical item or class in Gutob can be said to serve as the default host of an index. As a consequence, the index placement regarding the position

in the clause is variable as well and several examples (20, 32b and 38) show that the idea that preverbal indexing is limited "to the word immediately preceding the verbal complex" (Anderson 2007: 70) has to be re-evaluated. What is more, there are also a few tokens in the corpus used for this study where there are two preverbal indexes in one clause. In (38), there are not only two non-verbal indexes attaching to preverbal constituents (an interrogative and a manner demonstrative), but also a clause initial pronoun and a verbal index. Examples like this suggest that person indexes are not merely referential in Gutob and that their placement is not merely sensitive to discourse structure, but that indexes themselves have a discourse structuring function, which we will come back to in Section 5.5.

(38) *nom maŋɖem=nom ito?=nom ɖe~ɖem piŋ-gu=nom*
2SG why=2SG like.this=2SG INF~do come-MID.PST=2SG
'Why did you come this way?'

Finally, it should be mentioned that although index placement is variable in Gutob, it is not completely without constraints. For one thing, an S/A pronoun cannot host an index, see (39). Also, postverbal constituents in afterthought constructions do not host indexes, whereas free pronouns can appear there.

(39) **niŋ=niŋ sun-tu*
1SG=1SG speak-FUT
intended: 'I will speak.'

5.5 The discourse effect of index placement in Gutob

Having described the properties and frequency of preverbal indexing in Gutob, we now turn to illustrate how it contributes to information management. We argue that the placement of S/A indexes in Gutob is not merely sensitive to discourse, but that it is deployed judiciously as an information management marker.

This becomes especially evident in the light of examples where an index attaches to a preverbal constituent although there is already S/A reference in the clause, either by a verbal index and/or a pronoun. Thus, an index

would not be required if it would simply be needed for the sake of referencing (examples 32b and 38, also see Table 5.6). In these cases, the index is used as a device for marking a piece of information (or several pieces of information) as particularly noteworthy.

Marking a constituent as particularly noteworthy means that the speaker assumes that the interlocutor might not meet it with the right level of attention. Answers to content questions, question words or negators, which have been considered as being inherently focal in the literature, can host an index in Gutob, but they often don't. So they do not attract them by default, but only in contexts where the constituent needs more attention than the speaker expects it to receive purely on the grounds of it being new or unexpected information.

Consequently, there is also variation in sentences which contribute several pieces of new information: the speaker can use indexing to mark one or several constituents as being particularly noteworthy. In (40), where the speaker is telling about preparations that are done for the rituals for the dead, they want to highlight the objects of the actions. The predicates contribute new information, too, and could have been just as well the hosts of the indexes, but to a different effect. Indexing the objects as well as the verbs would have equated them in terms of how much attention they are to receive.

(40) *suŋol=nei goi-tu peɳɖom=nei ɖoɲ-tu*
firewood=1PL cut.down-ACT.FUT rice.liquor=1PL cook-ACT.FUT
'We will cut down firewood, we will cook rice liquor.'

Indexing can coincide with contrast marking, but it rarely does so. In (41), an index attaches to a constituent marked for contrast by the clitic *=o?*. In such a case, an index can increase the effect of this marker and draws even more attention to the host than enhanced by *=o?* alone.

(41) *eke=o?=niŋ lai ma? som-tu=niŋ*
here=EMPH=1SG rice curry eat-ACT.FUT=1SG
'I will eat rice and curry here. (i.e. not at home)'.

That indexing is not confined to contrastive or new information is also reflected in the fact that also discourse-given information like object pronouns can be host of an index (see Table 5.7). This is in line with the observation that the allocation of attention is logically independent from other discourse

effects which have been ascribed to the notions of topic or focus (Ozerov 2021).

The commitment towards a certain constituent on the part of the speaker can also shift within the scope of only a few utterances. This can be shown quite nicely in examples like (20) and (42) which constitute clausal minimal pairs, of which there are actually quite a few in our data. In each consecutive utterance, a different element is considered of more importance by the speaker and thus hosts the index. Examples (42a) and (42b), produced by the same speaker within the same text, illustrate this nicely. A delegation from one village had gone to visit another village for wedding negotiations and stayed over night. In the morning they decide to leave, uttering (42a) with the index attached to the predicate 'stay'. When asked by the people from their own village why they had come back so early, they reply with the sentence in (42b), but this time the index does not attach to the verb, but to the interrogative, thus drawig the attention towards the 'why'.

(42) a. *nei* *maŋɖem* *ɖu-loŋ=nei*
1PL.EXCL why stay-MID.FUT=1PL.EXCL
'Why should we stay?'

b. *nei* *maŋɖem=nei* *ɖu-loŋ*
1PL.EXCL why=1PL.EXCL stay-MID.FUT
'Why should we stay?'

The targeted use of the index can also cause variation regarding its placement within a complex verb phrase (cf. Section 5.4.2). In the explicator verb construction in (43b), the index attaches to the second verb, the explicator verb conveying the aktionsart of the predicate; this is the most frequent placement of indexes within explicator verb constructions. In (43a), a few utterances in advance, however, the index attaches to the lexical verb, thus highlighting the semantics of the event to a stronger degree.

(43) a. *goj-gu=niŋ* *ui-loŋ* *ɖio?* *[...]*
die-MID.PST=1SG go-MID:FUT QUOT
'I will die [...]'

b. *oh eno?=niŋ goj-gu* *ui-loŋ=niŋ=be* *ɖio?*
oh here=1SG die-MID.PST go-MID:FUT=1SG=HON QUOT
'Oh, here is it I will die.'

An index does not attach to a particular host in a clause because it has to, casually speaking, go *somewhere*, but is placed in a manner best suitable for the discourse-oriented need on the side of the speaker, namely to ensure the desirable amount of attention. That verbs still make up the majority of hosts can be traced back to two facts: firstly, many clauses in Gutob consist of (simple or complex) verbs only, therefore leaving no possibility for an index to go anywhere else. Secondly, predicates are more often not part of a presupposition (i.e. the piece of information conveyed by the speaker which is not shared by the hearer) than other constituents in a clause (Lambrecht 1994: 296) and might thus be especially prone to receiving an index as a marker of noteworthiness.

5.6 Conclusion

Our corpus study of index placement in Gutob has revealed that non-verbal indexes in the language are neither a fringe phenomenon nor limited to particularly unusual conditions. The use of preverbal indexes is not exceptional, but is applied actively and frequently by speakers in order to structure their discourse with regard to the piece of information they consider especially noteworthy. Thus, although some elements are more prone to become the host of an index, no lexical item in Gutob is an *a priori* host of an S/A index.

In this respect, Gutob is different from the closely related Kharia, where either the verb or the negative particle preceding the predicate becomes the mandatory host of the person index. It also differs from North Munda languages like Santali, where the index can also attach to various hosts immediately preceding the verb, but is syntactically confined to this position if it does not attach to the verb itself.

Indexing in Gutob is not syntactically obligatory, nor is the placement of the index predetermined. With regard to the question of whether index placement in Gutob can be referred to as *rule governed*, we agree with Givón (2011: 189) who, reflecting on indexes in Ute, states, that "[i]f by 'rule governed' one means the traditional generative statement, with purely syntactic conditioning of the choice of options, the answer is surely no. If, on the other hand, one means that the choices are non-random, and motivated by communicative or cognitive factors, the answer is probably yes." The choices behind

the variablity in Gutob index placement and its contribution to information management are better understood now. The same would be desirable for many other cases of *optional* marking: conditions might be unknown, or the grammaticality of an utterance might not be at stake, but speakers' choices are surely not arbitrary.

Chapter 6

A structural and functional comparison of differential A and P indexing[1]

Abstract

Indexing P arguments on bivalent predicates is often considered more restricted and less often obligatory than A indexing. However, differential A indexing, i.e. the absence vs. the presence of an index referring to the A argument role, is not uncommon either: usually present A indexes can be omitted in particular discourse settings. However, differential A indexing has been a Cinderella subject in the typological study of differential marking, as opposed to differential P indexing or differential A flagging. This paper scrutinizes various cases of both differential A and P indexing and examines structural and functional differences and similarities. It will be shown that exploring differential indexing helps to understand how indexing in general is linked to referential prominence which surfaces as factors such as identifiability, animacy or topicality. Cases where indexing is particularly sensitive to referential prominence, and where it thus is employed only if the referent fulfills certain criteria, bring out the fact that A and P indexing have a common purpose, namely tracking referents through discourse. In this context, the paper also points out that differential A indexing presents an exception from generalizations concerning the amount of material in coding asymmetries.

[1]This chapter is submitted as: Just, Erika. *A structural and functional comparison of differential A and P indexing.*

6.1 Introduction

This paper provides an overview of differential indexing of both A and P referents, suggesting that the investigation of differential indexing (i.e. the variable occurrence of an index for reasons other than the referent's argument role) helps improving our understanding of the functionality of indexing in general, and why it is different from flagging. Whereas the latter can be considered as a strategy of individual role assignment, indexing is a means of reference tracking, which becomes evident through the investigation of differential indexing systems. This has been comprehensively shown for P arguments (e.g. Croft 1988, Iemmolo 2011), but A arguments have been paid less attention to in the cross-linguistic study of differential indexing. Showcasing the parallel behavior of differential A and P indexing, the paper aims at highlighting how indexing, irrespective of argument role, serves to keept track of prominent referents in discourse.

Indexes have been defined as bound markers (commonly on the predicate) expressing argument features (Haspelmath 2013). Indexing, contrary to agreement, does not presuppose any syntactic relationship between the marker and the referential noun phrase, nor whether the latter is obligatorily expressed. In other words, how a language handles overt NP referents on the one hand, and whether it indexes them on the other hand, should be considered as logically independent features (cf. Haig & Forker 2018). Languages show a lot of diversity in how and how often referents are expressed; just as there are languages like German where (S and A) referents are usually expressed by an overt NP and are also obligatorily indexed, there are also a number of languages where there is no indexing, but zero-anaphora is nevertheless an option, i.e. there can be no overt reference of an argument at all. There is no one-to-one relationsip between indexing and the omission of lexical or pronominal referents (see Gilligan 1987). Between those extremes, there are various possibilities how a referent can be indicated. This diversity, coupled with the traditional understanding of agreement in some Indo-European languages (cf. Haspelmath 2013: 209) has led to the agreement vs. pronoun debate: the attempt to classify a bound person marker as either being redundant feature matching and thus coding grammatical relations, or as being the one true instantiation of a referent.

In the context of this debate, often also the morphological status of the index as a clitic or an affix comes up. It is considered irrelevant here, as the latter is often unjustifiably equated with obligatoriness of marking (cf. Haig & Forker 2018: 720), though there are clear examples of syntactically obligatory clitics, as well as of affixal indexes which are syntactically optional. An example for obligatory enclitic person marking comes from Kharia, where the subject index is enclitic to the verb in affirmative clauses, as in (44a), whereas it attaches to the negative particle in negated clauses, as in (44b) (Peterson 2011: 58):

(44) Kharia (Munda, Peterson 2011: 58)

a. *kayom=ta=ɲ*
speak=MID.PRS=1S

'I speak.'

b. *um=iɲ kayom=ta*
NEG=1S speak=MID.PRS

'I do not speak.'

Languages where a syntactically optional index is analyzed as an affix are, for example, Juang (also Munda, Patnaik 2008), or Uralic languages like Hungarian (Coppock & Wechsler 2012), Northern Ostyak (Nikolaeva 1999: 64–76) or Eastern Mansi (Virtanen 2014). In Juang, object indexing by means of a suffix seems to have been very productive, but has been declining in the course of the last decades (Anderson 2007: 83, Patnaik 2008: 529–530). In some Uralic languages, there are two paradigms for the suffixal indexes for an A referent, one of which also indexes the number of the P referent, depending on the information-structural status of the latter. In Eastern Mansi, for instance, the so called subjective paradigm (as used in example (45a) indexes only person and number of the A referent, whereas the objective paradigm (used in example (45b) also indexes the number of the P argument and is used when the P referent is the secondary topic of a proposition (Virtanen 2014: 404).

(45) Eastern Mansi (Uralic, Western Siberia, Virtanen 2015: 28–30)

a. *äj-nø=tee-nø wöär-s-øt*
drink-GER=eat-GER make-PST-3PL

'They made something to eat and drink.'

b. *õõw-øm öät kont-iitø*
door-ACC NEG find-3SG>SG
'He does not find the door.'

That a clitic-affix-distinction is probably not expedient in the discussion of the use of bound person marking is also reflected in cases where even for one and the same language, scholars do not agree on how to categorize bound person forms: Osada (2008) and Anderson (2007) call the same Mundari set of S/A indexes "suffixes" and "clitics" respectively. The same goes, for instance, for Siwi object indexes (Ouali 2011 vs. Souag 2014.) Settling on this morphologically and syntactically unconstrained definition of indexing opens up quite a range of phenomena which can be referred to as *differential* indexing. Differential indexing falls in the category of differential marking as defined by Witzlack-Makarevich & Seržant (2018), which includes both flagging and indexing. It refers to situations where an argument role can be coded in different ways, depending on factors other than the argument role itself. These factors can either be inherent to the referent (like animacy, for instance) or non-inherent (like information-structural status).

The macroroles (or generalized semantic roles) which are of interest in this paper are A and P as the more agent-like and the less agent-like argument of a two-place predicate, giving priority to the semantic relationship between the predicate and its argument over their structural relationship (Bickel & Nichols 2009, Bickel 2011, Witzlack-Makarevich 2019). This study only encompasses transitive predicates, and therefore only A and P arguments, leaving out S arguments of intransitive clauses, as well as T and G in ditransitive predicates. Although there is a strong cross-linguistic tendency towards A and S to align with regard to indexing (Bickel et al. 2013), in many languages S can also align with P, and obligatory S indexing might affect the obligatoriness of P indexing, as brought forward by Haig (2018: 788). Therefore, I consider the investigation of differential indexing and alignment worth to be followed up in its own right.

For the present analysis, I want to exclude clausal properties which lead to differences in indexing, like TAM distinctions, clause type, or polarity, and bring into focus differences in indexing due to referential features of the argument, as well as discourse-structural conditions on the whole clause. Therefore, example (44) from Kharia does not fall under my definition of differential indexing, as the position of the index is absolutely predictable based on po-

larity.[2] Example (45) constitutes my first example for differential indexing, as here the choice of indexing P is controlled by the discourse status of the referent.

In this context, I also have to briefly dwell upon the notions of *topic* and *focus*. Although there is a general consensus about the pragmatic effects associated with each of these categories, the meanings conveyed by different constructions in different languages ascribed to focality or topicality are so manifold that these information-structural categories should actually be seen rather as interpretive effect of certain constructions, and not as being at their core (see Matić & Wedgwood 2013 on focus). Additionally, speakers of languages that have comparable constructions to map information structure, such as clefts or left dislocation, might still use different structures under identical discourse conditions, so there is no one-on-one mapping of information-structural categories and particular constructions (Skopeteas & Fanselow 2010,Ozerov 2018). Differential indexing can also be considered one such construction: as outlined in Section 6.2, the presence of a P index is often connected to the topicality or topicworthiness[3] of the respective referent, whereas the absence of an otherwise present A index has been attributed to the referent being in some kind of focus construction. However, the pragmatic and referential features that indexing is sensitive to have to be considered language specific, even if the differences may seem subtle.

The paper proceeds as follows: In the following Sections 6.2 and 6.3, I will provide examples of differential indexing from various languages, starting with differential indexing of the P argument, as this is probably the more familiar phenomenon in the context of differential marking, and then proceed to the A argument. It will be shown that following the definitions just provided, differential A indexing is not that uncommon and in principle exhibits the same general pattern as differential P indexing. In line with the findings from these two sections, Section 6.4, deals with differential indexing and referential prominence (Haspelmath 2021b), addressing both the role of the referents individually as well as that of co-arguments. Section 6.5 elaborates

[2]Admittedly, demarcating polarity from information structure can be a balancing act. In fact, languages where the position of a syntactically mobile index is determined by discourse effects also very often display an interplay between the placement of an index and polarity (see Cysouw 2003)

[3]Topicworthy referents display semantic features associated with topicality (such as being definite or high on the animacy scale) without necessarily being a topic.

on the functions of indexing in general thus revealed, highlighting the need to consider indexing independently from the overtness of an NP referring to the same referent, and arguing to functionally demarcate indexing more from flagging. I will conclude in Section 6.6, by reflecting on how (or rather if) the thoughts brought forward in the paper can be brought in line with the notion of obligatoriness of marking.

6.2 Differential P indexing

Differential P indexing, in whichever guise it comes along, has been given a lot of attention in the studies on individual languages, language families, and also cross-linguistically. The phenomenon can be encountered, for instance, as "clitic doubling" (see e.g. Aoun 1999 on Arabic dialects), "object reduplication" (see e.g. Friedman 2008 on the languages of the Balkans), or "optional agreement" (see e.g. Muxí 1996 on Catalan).

Many instances of differential indexing have typically involved the presence of an overt lexical NP. It is very often associated with animacy, humanness, specificity, definiteness, or topicality of P. As the P argument is generally not associated with such factors, but with new or contrastive information, a P referent which behaves rather atypically seems to call for a distinct marking pattern like differential flagging or differential indexing (Iemmolo 2011, Dalrymple & Nikolaeva 2011). However, although topicality is often the common ground on which differential P flagging as well as differential P indexing operate, the two have to be distinguished with regard to their function. As the main functions of flagging are to reflect referent features and to discriminate the different actants of an action (Bakker & Siewierska 2009), differential P flagging is often associated with effects of discourse discontinuity, such as topic-shift or topic promotion (Iemmolo 2011: 52). The primary function of indexing, on the other hand, is reference tracking (Lehmann 1982, Givón 1983, Siewierska 1997). Consequently, (differential) P indexing is also associated with tracking an (unexpectedly) topical P referent (Croft 1988, Iemmolo 2011).

However, cross-linguistically, the interpretive effects of differential indexing are very idiosyncratic, which makes it impossible to find a unitary explanation (Kallulli & Tasmowski 2008a: 10). This diversity is connected to

the fact that language internally, there is also often variation, even if a feature has been identified to be the triggering factor for the presence of a P index. For example, in Macedonian, P indexing is clearly associated with definiteness and specificity. If a referent is either inherently definite (like proper nouns) or marked by a definite article, it has to be indexed (Tomić 2008: 70). This is shown in example (46), where P ('the movie director') has to be indexed, as it is marked by a definite article, irrespective of an interpretation as being specific or non-specific.

(46) Macedonian (Balto-Slavic, Tomić 2008: 70)
*Jana *(go)=bara režiser-ot*
Jana 3SG.M.ACC=look.for.3SG movie.director-DEF.M
'Jana is looking for the movie-director (namely for X, who happens to be the movie-director).' or: 'Jana is looking for the movie-director (whoever that may be).'

However, depending on the context, also non-definite referents can be indexed if they are specific. But unlike definiteness, specificity does not force indexing, i.e. there can be specific referents that are not indexed. Also, based on acceptability judgements, humanness plays a role in indexing in Macedonian (Tomić 2008: 71–72). In other languages of the Balkans, P indexing is similarly associated with pragmatic and semantic features of the referent, and has grammaticalized to various extents (or not at all) in the different languages (Ivanov 2012: 350), and this phenomenon was given some attention in an areal context (e.g. Kallulli & Tasmowski 2008a).

There has also been some interest in differential P indexing from a family perspective, for example regarding the Romance languages (e.g. Jaeggli 1981, Miller & Monachesi 2003, De Cat & Demuth 2008, Fischer & Rinke 2013, or Fischer et al. 2019) or Bantu languages (see Downing & Marten 2019: 278–280 for a recent overview of contributions).

Whereas in some languages of the Bantu family, the constraints on P indexing seem to be purely formal in nature, in others, P indexing is licensed by inherent semantic properties of the referent. For instance, in Kagulu, P indexing can be considered as being differential in that it is described as "optional" for animate referents as soon as there is an overt NP referring to the same referent, exemplified in (47a) and (47b); likewise, if there is P indexing

without an NP, as in (47c), this lack of an NP is lead back to the fact that there is already indexing (Petzell 2008: 169).

(47) Kagulu (Bantu, Tanzania, Petzell 2008: 169–170)

a. *Awafele ha-wa-koma dijoka*
C2.woman PST-C2.A-kill C5.snake
'The women killed the snake.'

b. *Ka-mu-on-aga imukulu akwe*
PST.3SG.A-C1.P-see-IPFV C1.big 3SG.POSS
'S/he sees his/her older sibling regularly.'

c. *Mheho i-ku-mu-ogoh-es-a*
C9.cold C9.A-PRS-3SG.P-fear-CAUS-FV
'The cold scares him/her.'

I would like to put it a bit differently, namely that P indexing and the overtness of the NP are logically independent, but that they are both sensitive to the discourse pragmatic status of the referent and that in certain cases (depending on the discourse pragmatic effect the speaker wants to achieve) they co-occur. Thus, the information-structural status of some referents allows for the expression by an overt NP, as well as for indexing.

In other Bantu languages, differential indexing is considered on purely syntactic grounds and indexing together with a co-referential NP is limited to cases where the latter appears in a pragmatically marked position. For instance, in Nkore-Kiga, a topicalized P is expressed by an overt NP which appears in a clause initial topic position (the pragmatically neutral position would be after the verb). This P argument then also has to be obligatorily indexed on the verb (Taylor 1985: 78, 91) and the co-occurrence of index and clause-initial NP has thus fully grammaticalized.

That P indexing has become obligatory only in combination with other structural features which can themselves be considered differential is not uncommon. It often co-varies with word order alternations, for instance, in Amuesha (Duff-Tripp 1997) or Burunge (Kießling 1994). Moreover, in some languages, differential indexing has grammaticalized to the extent that it goes hand in hand with differential flagging (see Arkadiev 2013), e.g. in Romanian (Cojocaru 2004) or Lebanese Arabic (Aoun 1999), or in that only a subset of nouns are indexed, e.g. in Makhuwa-Enhara (van der Wal 2009).

Although the co-occurrence of a lexical NP and an index might become obligatory, like in Nkore-Kiga, they should still be considered as logically independent. This becomes especially clear when looking into languages where indexing and the overtness of the referent NP do not totally correlate, but are similarly connected to the information-structural status of a referent. For the Austronesian language Larike (Laidig & Laidig 1990), for instance, the authors state quite clearly how both indexing and the overtness of the NP are independently associated with the prominence[4] of the referent, but that their co-occurrence as well as their joint absence can enhance the discourse structural effect. Although indexes (for both A and P arguments in transitive clauses) generally co-occur with the NPs referring to the same referent, a sentence may consist of only a verb with the appropriate indexes, or there can be an overt NP only, without indexing. In addition, there can be neither, i.e. there can be no overt reference to a P argument at all. The authors state that the choice of how to code a referent is directly linked to the pragmatic status of the referent, and that each way of reference in a clause results in a different interpretation (Laidig & Laidig 1990: 93n11). Despite their insight, however, Laidig & Laidig (1990) unfortunately do not further elaborate on the different discourse effects that differential indexing and its interplay with the lexical argument can have.

In other accounts, the conditions of indexing or not-indexing in the respective language is somewhat clearer, for instance for Teiwa.[5] In Teiwa, P indexing is differential in that firstly, it is strongly associated with animacy: animate referents are obligatorily indexed (48a) whereas inanimate referents are only rarely indexed. Secondly, P indexing is also differential in that its omission is associated with a particular discourse effect: the verbal prefix is omitted, also for animate referents, if they are in contrastive focus; then, a free pronoun is used instead (Klamer & Kratochvíl 2018: 81–82) as in (48b).

(48) Teiwa (Alor-Pantar, Indonesia, Klamer 2010: 407)

a. *Miaag yivar ga-sii*
yesterday dog 3SG.P-bite
'Yesterday a dog bit him.'

[4] In Laidig & Laidig's (1990) terms, which they do not specify.

[5] There are more members of the Alor-Pantar family which are noteworthy with regard to the importance of referential properties as well as lexical restrictions for indexing, see e.g. Fedden et al. (2014).

b. *Miaag yivar ga'an sii*
yesterday dog 3SG bite
'Yesterday a dog bit HIM (not me).'

In this section it was shown how indexing P referents can be associated to various factors: discourse pragmatic effects (as in Macedonian, Nkore-Kiga or Teiwa), or referent semantics (as in Kagulu and Teiwa, where animacy has a strong effect, or Macedonian, where humanness has an impact on indexing). In each language, the relevant factors have different impacts on indexing, and their interplay can be quite complex. Also, the degree of obligatorification ranges from something like "always obligatory if" to "usually not". Probably the only way to do justice, at least to some extent, to the complexity behind differential marking phenomena in many languages is to evaluate the impact of the relevant variables on the basis of corpus annotation as done, for instance, by Schikowski (2013) on object flagging in Nepali, Bresnan et al. (2007) on the English dative alternation, Just & Čéplö (to appear) on P indexing in Maltese, or Goldstein (2021) on the marking of passive agents in Ancient Greek.

Nevertheless, a digest of differential P indexing like the present account demonstrates that, this diversity notwithstanding, the phenomenon often has to do with certain expectations concerning the referent and, depending on whether these expectations are met, the referent is either not indexed or indexed. This also holds for A referents, but often in a reversed manner: the omission of A indexes is often associated with unexpected properties of the A referent. Although is has been observed for various languages that focused A NPs are often in a non-default position with regard to the verb and lack indexing (e.g. Lambrecht & Polinsky 1997, Mereu 1999), differential A indexing has not been addressed as extensively as differential P indexing. Section 6.3, deals with this phenomenon in more depth, ultimately showing that it is driven by the same mechanism as differential P indexing.

6.3 Differential A indexing – the reverse pattern of differential P indexing?

Differential A indexing has been claimed to be globally more restricted than differential P indexing (Haig 2018: 789). As opposed to differential P indexing, where the presence of the index often marks an unusual situation, in the

case of differential A indexing, the absence of a usually present index marks an atypical situation.[6] Like with differential P indexing, lack of A indexing can be induced by semantic features such as non-volitionality or the referent being inanimate (Malchukov & Ogawa 2011: 32–36), but also by discourse pragmatic features. In some languages, differential A indexing is directly related to A displaying certain properties related to focus (such as newness, indefiniteness, contrastiveness), as has been noted by Siewierska (2004: 159–162). On that note, it is in some languages syntactically tied to interrogatives, quantification, or to relative clauses or complement clauses (e.g. Ouhalla 1993, Lambrecht & Polinsky 1997).

In line with this, it has been brought forward that the loss of topicality of the A referent can result in the omission of indexing (e.g. Givón 1976, Lambrecht & Polinsky 1997, Mereu 1999, Malchukov & Ogawa 2011: 29–32). For instance, the suspension of the index for an A argument ("lack of subject-verb agreement") is one of a couple of prosodic and morphosyntactic strategies which languages make use of in order to express sentence focus, i.e. a proposition in which both the predicate as well as the subject are in focus (Lambrecht & Polinsky 1997). This lack of indexing is in some languages realized as impersonal or singular marking also for plural third person referents, like English *There's three women in the room.* Lambrecht & Polinsky (1997) analyze the subject in such a construction as being in focus, together with the predicate, without really stating in detail what pragmatic effects focus entails in this case.

In many Bantu languages, lack of A indexing together with the co-referential noun phrase in a non-default, postverbal position is also often described as reflecting that the referent is in focus. Some Bantu languages feature indexing of a locative noun class instead of the noun class of the A argument,[7] others feature indexing with the P argument, and some seem to have both constructions, like Kirundi (Ndayiragije 1999).

(49) Kirundi (Bantu, Burundi, Ndayiragije 1999)

[6] As for terminology, for differential A indexing one also finds, for instance, "anti-agreement" (Ouhalla 1993, Baier 2018), "impersonal agreement (Lambrecht & Polinsky 1997) or "default agreement" (Borsley et al. 2007).

[7] It has to be added that some Bantu languages which display this locative inversion only allow it with intransitive, some only with unaccusative verbs (see Buell 2005: 48–50).

a. *Abâna ba-ára-nyôye amatá*
C2.children C2-PST-drink:PFV C1.milk
'Children drank milk.'

b. *Amatá y-á-nyôye abâna*
C1.milk C1.PST-drink:PFV C2.children
'Children (not parents) drank milk.'[Lit.: 'Milk drank children.']

c. *Ha-á-nyôye amatá abâna*
C10.LOC-PST-drink:PFV C1.milk C2.children
'Children (not parents) drank milk.'[Lit.: 'There drank milk children.']

In sentence (49a), the constituents are in canonical AVP-order and the A argument is indexed on the verb; in (49b), A and P have swapped their positions and P is indexed. In (49c), neither A nor P are indexed, but there is a prefix for the locative noun class instead. According to Ndayiragije (1999), (49b) and (49c) convey they same meaning and imply a contrastive focus reading of *abâna*. In fact, it has even been argued for some Bantu languages that indexing is forced by topicality rather than the syntactic status of the referent as an argument (e.g. Morimoto 2000), or that an index can be considered an antifocus marker (Zeller 2008). The relation between indexing A and its discourse status as not being the focus of the proposition becomes even more obvious in languages where A referents can only be indexed if they are also overtly marked for topicality (e.g. the Cushitic language Oromo, Malchukov & Ogawa 2011: 31). However, very often indexing is sensitive to the pragmatic status of a referent without this status being morphosyntactically indicated. Some cases reported by Siewierska (2004: 159–162) are Konjo (Austronesian), Chalcatongo Mixtec, and the Arawakan languages Bare, Yagua and Apurinã.

Provided that A indexing – like P indexing– is reserved for referents with a particular discourse status, there can be some other structural features that are involved in differential indexing, which similarly mark or even enhance this discourse status. In Welsh, for instance, there is a restriction with regard to the part of speech of the referential NP: only pronominal referents like in (50a) are indexed, whereas full NPs such as in (50b) are not indexed for number (the verb is in "default form", Borsley et al. 2007). Thus, one could say that only pronominal referents fulfill the information-structural requirements to be indexed.

(50) Welsh (Celtic, Great Britain, Borsley 2009: 3)

a. *Gwel-on nhw ddraig*
see-3PL.PST they dragon
'They saw a dragon.'

b. *Gwel-odd y bechgyn ddraig*
see-3SG.PST DEF boy.PL dragon
'The boy / boys saw a dragon.'

c. **Gwel-on y bechgyn ddraig*
see-3PL.PST DEF boy.PL dragon

Another type of structural dependency can be found in Koorete: A indexing correlates with the presence of the assertive focus marker on the predicate: only if the verb carries the focus marking morpheme *-ko* is A indexed (Mendisu 2010: 166, 172), as in (51a). If the verb does not carry a focus marker, as in (51b) and (51c), there is no indexing. Example (51d) is thus ungrammatical. So the index here is omitted not only if the A argument is in focus, as in (51c), but if there is any deviation from predicate focus, which is considered a universally unmarked discourse configuration (cf. Lambrecht 1994: 296).[8]

(51) Koorete (Omotic, Ethiopia, Mendisu 2010: 172, 180-181)

a. *nun-i doro woon-d-uu-ns'i-ko*
we-NOM sheep buy-PFV-PST-1PL-FOC
'We BOUGHT sheep.'

b. *nun-i doro-ko woon-d-o*
we-NOM sheep-FOC buy-PFV-PST
'We bought SHEEP.'

c. *tamba-ko doro woon-d-a*
me-FOC sheep buy-PFV-REL
'I bought sheep.'

d. **nun-i doro-ko woon-d-uu-ns'i*
we-NOM sheep-FOC buy-PFV-PST-1PL

In another Omotic language, Zargulla, the presence of the focus marker on the verb is a prerequisite for indexing as well. But even if there is focus marking by means of the respective marker, and indexing can therefore occur,

[8] Another example where indexing interacts with the presence as well as the position of overt focus marking, and which has received some attention in the literature, is Somali (e.g. Saeed 1984, Mereu 1999, Tosco 2002): A is only indexed if constituents other than A are focused (Mereu 1999: 231–232).

it is still variable, i.e. not obligatory, but sensitive to identifiability as well as animacy (Amha 2007: 200–201). The association of animacy and A indexing is also known, for instance, from Standard Persian (Sedighi 2010: 35), or Georgian (Harris 2009: 21).

As with differential P indexing, also with A indexing there can be a correlation between indexing and word order. In Anuak, for instance, various constituent orders in transitive clauses are possible, depending on information management. Reh (1996) argues that only a sentence like (52a) with a clause-final verb plus an A index can be considered pragmatically unmarked, (1996: 350–351) and the A referent as topical (Reh 1996: 339, 347-357). In all other cases, i.e. with a topical P in clause (52b), a focalised P in (52c) and a focalised A in (52d) there cannot be an index (Reh 1996: 348–350).[9]

(52) Anuak (Nilotic, Ethiopia, Reh 1996: 348)

a. *ɲìlàal kwǎn ā-cám-ɛ̄*
child porridge PST-eat-3SG.A
'A child ate the porridge.'

b. *kwǎn ā-cám ɲìlàal(-lì)*
porridge PST-eat child(-DEF)
'The child ate the porridge.'

c. *ɲìlàal cám-á kwǎn*
child eat-FOC porridge
'The child eats (the) porridge.'

d. *kwǎn cám-á ɲìlàal(-li)*
porridge eat-FOC child(-DEF)
'A (The) child has eaten the porridge.'

In another language of Ethiopia, Sheko, there is differential A indexing without a correlating differential structure (like word order deviation) or overt pragmatic marking: A referents are usually indexed in main clauses, as in (53a), and only in those clauses where the referent NP is in focus is A indexing omitted, as in (53b):[10] Additionally, similarly to Koorete, indexing in Sheko is not only sensitive to the discourse status of the A referent itself, but also to

[9] A similar case can be found, for instance, in Trentino, where focalized, post-verbal 3rd person referents are not indexed (Mereu 1999: 238).

[10] Similarly, in Jamsay (Dogon, Heath 2008: 455–456) a focalized preverbal A NP is not indexed, and can be focus-marked or not. In the latter case, the lack of an index alone is an indicator of focus status of the referent.

the pragmatic configuration of the whole proposition in that the position of the index is not fixed: it procliticizes to the verb stem in the case of predicate focus, corresponding to an unmarked topic-comment structure. But the index can also be enclitic to the verb, as in example (53a) in the case of verb polarity focus and thetic sentences. What is more, it can leave the verbal position altogether and can encliticize to any constituent apart from the A NP if this constituent is focused (Hellenthal 2010: 429-432). It is not uncommon cross-linguistically that index placement can be determined by discourse effects, irrespective of the argument role (see Cysouw 2003).

(53) Sheko (Omotic, Ethiopia, Hellenthal 2010: 430–436)

a. *gébèn bây dàdù nyààs=í-k*
Geben female child give.birth=3SG.F.A-REAL
'Geben has given birth to a daughter!'

b. *m-bāyǹ nata gasku-k-ə*
1SG.POSS-wife 1SG insult-REAL-IND
'MY WIFE insulted me.'

This section has shown that differential A indexing – although probably not as common as differential P indexing – is not a rare phenomenon, as its lack of attention from a typological side (compared to other kinds of differential marking) might suggest. In languages which display overt A indexing, the lack of the index can mark deviations of the referent from being high in in some language specific factors usually ascribed to the A role, like topicality or animacy. So structurally, differential A indexing somehow looks like the mirror image of differential P indexing: instead of an additional index showing that something is amiss (such as the P role being occupied by a referent high in topicality), it is the lack of the A index which is often used to indicate something is out of order. However, both phenomena actually simply boil down to the fact that indexing is linked to referential prominence, which will be discussed more in-depth in the following section.

6.4 Differential indexing and referential prominence

6.4.1 Indexing and referential properties

The last two sections served to illustrate that differential indexing should not be considered particularly exceptional, either for P or for A. If there is a P which is high in identifiability (defined by definiteness and specificity), animacy or topicality, it can be indexed, or if an A misses a certain mark with regard to factors such as these, an index can be omitted. In order to account for the connection between various kinds of differential marking and a referent's features (inherent as well as non-inherent), different scales have been proposed, e.g. the *potentiality of agency scale* by Dixon (1979: 85), the *empathy hierarchy* by DeLancey (1981), the *prominence scale* by Aissen (1999), the *D-hierarchy* by Kiparsky et al. (2008) or the *referential hierarchy* (e.g. Bickel 2008). I will not go into these scales or hierarchies (see Witzlack-Makarevich & Seržant 2018: 5-10 for an overview) but use the more convenient term *referential prominence*[11] (Haspelmath 2021b) to refer more generally to a referent's status with regard to identifiability, animacy, or person ranking, etc., whichever factors are relevant in a given language.[12]

Although the precise nature as well as the impact of referential prominence always has to be considered as being language specific, its universal character lies at the heart of what Haspelmath (2021b) refers to as role-reference associations. These role-reference associations imply the ranking of roles with respect to each other (e.g. A ranked higher than P), as well as the role's characteristics regarding prominence. Deviations from these associations, like a scenario in which the role ranking is reversed, or an argument role showing an unexpected degree of prominence, can lead to a number of coding splits: Asymmetries in marking, such as differential object flagging or split ergativity, can ultimately be reduced to deviations from referential features associated with a particular role.

Crucially, Haspelmath also argues that such deviations from role-reference associations tend to be coded by longer forms as they are less predictable and

[11] A term which has similarly been used is *salience* (Croft 1988).

[12] For an extended definition of prominence as a structure-building principle accounting for phenomena on different levels of grammar, and a discussion of its relation to other concepts of referential management see von Heusinger & Schumacher (2019).

less frequent (also see Haspelmath 2021a, as well as Croft 2003 for similar observations). The proposed explanation is that it is more efficient to explicitly mark less frequent meanings, and to not or to a lesser extent mark the more frequent ones. And this generalization works perfectly well for the coding asymmetries he presents, as in differential object (P, but also R and T) flagging, where an NP receives additional or differential case marking if the referent deviates from their role association. It not only works for splits with regard to flagging, but also for the encoding of voice on verbs, like inverse marking, passives, or antipassives, as with these categories, the verbs receive special or longer marking in situations where the arguments deviate from default associations.

As for splits in argument coding, indexing is explicitly exempted from the generalizations concerning the amount of coding material (Haspelmath 2021b: 131n6). The reason for this is that referential prominence is connected to indexing in general, not just with objects, but also with the higher ranked A arguments. Thus, whereas the economical idea to explicitly mark less frequent meanings seemingly fits with differential P indexing (where there is an index for a referent violating their role-reference association), it is not compliant with cases of differential A indexing: the deviating, unexpected, and therefore supposedly less frequent and less predictable construction receives *less* marking in that the index is omitted. Moreover, also with differential P indexing, it is not always the case that the presence of an index is the exceptional pattern, but there are also languages where there is differential P indexing surfacing by the omission of an index. Consider, for instance, the situation in the Austronesian language Makasar, where in transitive clauses both A (proclitic) and P (enclitic) are usually indexed, as in (54a). The exception to this normal transitive pattern occurs if either A or P are in focus (in 54b and 54c), in which case the respective argument NP is fronted[13] and there is no indexing of that argument (Jukes 2015: 55–58):

(54) Makasar (Austronesian, Indonesia, Jukes 2015: 55–58)

a. *Na=cini´=i* *tedong-ku* *i* *Ali*
3SG.A=see=3SG.P buffalo-1SG.POSS PERS Ali
'Ali sees my buffalo.'

[13]without prosodic break, so the construction is different from left dislocation, which has different discourse pragmatic effects and would entail indexing (Jukes 2015: 60)

b. *Kongkong=a a-buno=i miong=a*
dog=DEF AF-kill=3SG.P cat=DEF
'THE DOG killed the cat'

c. *Miong=a na=buno kongkong=a*
cat=DEF 3SG.A=kill dog=DEF
'The dog killed THE CAT.'

This is similar to the Teiwa example shown in (48) above, where P indexing for animate referents has to be omitted if they are in contrastive focus. Thus, the data support this view that indexing is primarily connected to prominence, irrespective of the argument role, and this is where it differs from flagging or other devices which encode the who-does-what (which I will return to in Section 6.5). Indexing indicates a certain level of language specific prominence that a referent has in discourse. If a referent does not have this particular level of prominence, or loses it (e.g. becomes focalized), it is not indexed. Thus, differential indexing should not be considered as a priori marking deviations from role-reference associations (Haspelmath 2021b), but as being only indirectly linked to roles, as an A is typically more prominent (i.e. more index-worthy) than a P, which leads to A indexing becoming grammaticalized more readily. This, in turn, might have distorted our view a bit towards the idea that A arguments (or subjects) are more prone to indexing than other roles (or relations), while this is just a side effect of the prominence level associated to that role.

6.4.2 Properties of the co-argument

Some phenomena which also fall under coding splits are scenario induced, i.e. cases where it is not referential features of the affected argument itself that cause some kind of differential marking, but the nature of the co-argument as well, i.e. "the whole configuration of who is acting on whom" (Witzlack-Makarevich & Seržant 2018: 12). Whereas differential flagging depending on co-argument features is not very common in the world's languages (also see Haspelmath 2021b: 143–151), they play a role in indexing in a number of languages.

Such indexing systems have been referred to as hierarchical alignment (see e.g. Creissels 2009 or Witzlack-Makarevich 2011: 181–194 for discussing whether the combination of the terms 'hierarchical' and 'alignment' is appro-

priate). A distinction has been made between two systems: languages where two roles are considered to compete for a given slot and only the referent outranking the other in terms of a language specific hierarchy is indexed, and languages where indexing a referent is permitted or blocked depending on what kind of referent takes the co-argument role. However, Witzlack-Makarevich et al. (2016) clearly show that both types can be explained in terms of the latter type, namely in terms of co-argument sensitivity: hierarchies are not needed to describe, explain, or compare such systems.

The following examples provided in (55) come from Reyesano, a Tacanan language spoken in Bolivia. Whether A or P are indexed[14] depends on the person as well as on the role of the co-argument: in scenarios involving a locuphoric (i.e. 1st and 2nd person) referent together with a 3rd person, as in (55a) and (55b), the locuphoric referent is indexed, whether it is A or P, while a 3rd person co-argument is indexed only when it is the A argument. In scenarios involving only 3rd person referents in both the A and the P role, exemplified in (55c), only A is indexed (the index for the third person thus encodes the A role). Lastly, in scenarios involving locuphoric referents in both roles (55d) and 55e), A or P is indexed if it is second person (Guillaume 2009: 35–40).[15] For better clarity, Table 6.1 illustrates whether A or P or both are indexed in Reyesano transitive clauses, depending on the co-argument.

(55) Reyesano (Tacanan, Bolivia, Guillaume 2009: 37–40)

a. *K-a-maneme-a awadza*
1PL-PST-kill-PST tapir
'We killed a tapir.' (*'A tapir killed us.')

b. *K-e-dai-ta-da chenu te tue*
1PL-IPFV-cure-3SG.A-IPFV EMP BM 3SG
'She cures us.' (*'We cure her.')

c. *A-kachi-ta-a te iba te awadza*
PST-bite-3SG.A-PST BM jaguar BM tapir
'The tapir bit the jaguar.' OR: 'The jaguar bit the tapir.'

[14]For 1st and 2nd person, indexes do not encode role, i.e. the same set of indexes is used

[15]There is no role distinction in independent pronouns (Guillaume 2009: 31). Therefore, isolated from context, the examples in (55d) and (55e) could be ambiguous in two ways: (55d) could be translated as 'You saw me' with the 2nd person as the A referent, or also be interpreted as having a third person P, i.e. 'You saw him/her'. Similarly, (55e) could also mean 'We won't forget you' in a different context.

d. *Mi-a-b-a te miwe*
2SG-PST-see-PST BM 2SG
'I saw you (crossing the plaza yesterday afternoon).'

e. *Ma te mi-e-deta te ekama*
NEG BM 2SG-FUT-forget BM 1PL
'You won't forget us.' OR: 'We won't forget you.'

Arguments	**Indexing**
A1st with P2nd	P
A1st with P3rd	A
A2nd with P1st	A
A2nd with P3rd	A
A3rd with P1st	A and P
A3rd with P2nd	A and P
A3rd with P3rd	A

Table 6.1: A and P indexing in Reyesano transitive clauses

In a system like this, referential prominence is primarily defined in terms of empathy (cf. DeLancey 1981), and how referents relate to one another with regard to it. That is, the lineup of the referents determines which role is indexed. Thus, indexing based on co-argument sensitivity can be referred to as differential in that it is not the argument role itself which triggers whether a referent is indexed. It is not essentially different from systems like those found in Koorete (examples in 51) or Anuak (examples in 52) or other languages where it is not or not only some inherent referential features playing into prominence and thus trigger indexing, but where the whole configuration of a clause is relevant.

6.4.3 Prominence and lexical NPs

What can also be considered in the light of prominence-level and is thus ultimately linked to role associations, is whether an argument is expressed by an overt NP. Lexical NPs are used for new information, contrastive information, topic shifts or for referents at a long lexical distance; non-lexical forms, on

the other hand, are used for more accessible information (Givón 1983, Ariel 1990), i.e. for referents higher in prominence.

As the A role is usually occupied by referents which are high in identifiability, animacy and topicality, A arguments are less commonly occupied by lexical NPs than P arguments (see DuBois 1987, as well as Haig et al. 2020 for confirming the observation in corpus data from a sample of typologically diverse languages). P, on the other hand, is more commonly used to introduce new or non-prominent referents than A (e.g. Givón 1976, DuBois 1987, Comrie 1988, Schnell et al. 2020). Therefore, to be expressed by a lexical NP is more of an exception for the A role than for the P role (also see Lambrecht 1994: 189-190).

And this is the point where we come back to the issue of indexing and the co-occurrence of a full lexical NP, one of the parameters often used in the agreement-vs-pronoun-debate. Although I do not want to overgeneralize too much here – prominence obviously has a common denominator cross-linguistically, but it also has a specific character in every language – indexing and the overtness and also the position of a lexical NP in the clause are subject to referential prominence.

6.5 Some more thoughts on the function of indexing

Regarding indexing and overt NPs, their co-occurrence can in some cases grammaticalize, either with an additional prerequisite (word order, differential flagging etc., like in the case of Nkore-Kiga preverbal P NPs), or unconditionally, in languages with grammatical agreement (Bresnan & Mchombo 1987, Siewierska 1999). Although grammatical agreement, where both the index and the overt NP (aka the controller) are confined to the same clause, is very rare cross-linguistically (Siewierska 1999), it has been considered as the logical endpoint on the basis of which deviating scenarios can be described (Corbett 2006).

It seems to be hard to give up on the assumption that indexing has to be investigated with regard to the co-occurrence of a referential NP. Even Haspelmath (2013) who argues for treating indexes as “phenomena sui generis” (2013: 213) in order to detach them from the futile agreement-vs-pronoun

discussion, eventually classifies them on the basis of whether they can, may, or must co-occur with a co-nominal in the same clause (cross-indexes, gramm-indexes and pro-indexes, respectively, Haspelmath 2013: 218–221).

In order to uncouple indexing from overt NPs and to truly consider them as logically independent (as also advocated by Haig & Forker 2018), one has to bear in mind that indexes, in the course of their emergence, become less referential than they used to be in their time as an anaphoric pronoun (Kibrik 2011). Anaphoric pronouns, which are by definition used for prominent referents, can be considered as the source material for indexes (e.g. Givón 1976, Lehmann 1982). When grammaticalizing into indexes, they gradually lose referential potential (Siewierska 1999: 225) and it is the transmission of the prominence level of the referent which remains. This would explain how indexing is employed to facilitate tracking of prominent referents, irrespective of the argument role (e.g. Lehmann 1982, Givón 1983, Siewierska 1997) which can best be demonstrated by looking into cases of differential indexing as presented in Sections 6.2 and 6.3. An index is omitted if the referent is not prominent enough (however prominence might be spelled out in the respective language) even though this index might have grammaticalized in other contexts. Conversely, there can be an index which is syntactically optional in order to assign the appropriate level of prominence to the respective referent.

Therefore, when considering indexes as a role identifier like case marking (e.g. Dixon 2010, Haspelmath 2019), one probably runs into danger of confusing cause and effect. What I mean by this is the following: as indexes serve to keep track of a prominent referent, indexing of a particular role depends on which kind of referent (semantically and/or pragmatically speaking) occupies this role. Following this assumption, an index indexing referential features of a particular argument role and thus marking this role on the predicate should be considered as a side effect of following this particular referent through the discourse.[16]

So I think that what has been stated by (Iemmolo 2011: 47–60), that differential P flagging and differential P indexing, albeit often sensitive to similar referential features, do not serve the same purpose, can be expanded

[16]Arguably, there are languages (like German) where over the course of time indexing has become more of an automatism and dissociated from prominence, which might have led to statements that grammatical agreement is a purely redundant expression of features (Bresnan & Mchombo 1987: 741).

to flagging and indexing in general. True role assignment can be achieved through flagging which serves to distinguish the referents involved in an action (Bakker & Siewierska 2009), whereas indexing a referent, occupying a particular role, marks this one as prominent and as to be tracked. That flagging is immediately tied to roles in contrast to indexing would also explain why flagging splits can be explained by deviations from role-reference associations which ultimately lead to longer coding for unexpected and/or less frequent combinations (Haspelmath 2021b), while this does not really work for indexing: in cases of differential A indexing, the deviating and supposedly less frequent and less predictable construction receives less marking instead of longer coding (see Section 6.4.1).

6.6 Conclusion

In a nutshell, investigating differential indexing opens a window to looking at the core of indexing in general, and to understanding its link to referential prominence. Prominent arguments, be it A or P, tend to be indexed more readily than arguments which are low in identifiability, animacy or topicality. That both A and P indexing have a common purpose, namely tracking referents through discourse, becomes especially evident in languages where indexing is particularly sensitive to referential prominence, and where it thus is employed only if the referent fulfills certain prominence criteria.

The A role is more commonly associated with prominent referents, so indexing it is in many languages more prevalent than P indexing, and has more often grammaticalized. But indexing can also be associated with the P argument, if the referent fulfills certain requirements.[17] (Haig 2018) discusses whether the grammaticalization paths of A and P indexes are the same. He argues that differential P indexing constitutes an attractor state, i.e. a pattern constituting the endpoint towards which languages as complex dynamic systems tend to settle during the course of change. This would mean that differential indexing is more preferable for P referents than "fully obligatory" (Haig 2018: 788) indexing.

[17] Categorical splits due to information structural effects can be considered rare (Schultze-Berndt 2018), and phenomena which are not hard-coded (i.e. not syntactically required) might escape grammatical description (Fauconnier & Verstraete 2014: 10n1). So not only differential P indexing, but also differential A indexing is might be more wide spread than it appears.

On that note, I would like to point out again that even with languages where indexing is described as obligatory (like in Bantu languages), it can be differential to satisfy certain discourse pragmatic effects. And even if such a pragmatic effect is seemingly identified as focus or emphasis, it still remains to be discerned what this actually means for the language in question (cf. Matić & Wedgwood 2013, Ozerov 2018). In addition, in cases where there is an interplay of various factors involved, the question remains whether these factors can be translated into obligatorily choosing one pattern over another. It is probably more fruitful to conceive of many differential marking patterns as tendencies rather than rules (cf. Witzlack-Makarevich & Seržant 2018: 28), and the often cited optionality can be well motivated, although the motivation for the choice of a particular marking pattern might not be fully understood.

Chapter 7

General discussion and conclusion

7.1 Recalling the goals

This research studied the contribution of the variables which can underlie the intra-linguistic variation in indexing, with a focus on referential and discourse-structural factors. The first objective of this thesis (1) addressed the fact that differential indexing characterized through the absence vs. the presence of marking can very often not be attributed to one single factor. Possible interplays of various factors were visualized on the basis of quantitatively analyzed corpus data in the case studies in Chapters 3 and 4. Both studies, like the majority of the literature on differential indexing, were concerned with the P role. However, variability in A indexing, which has been studied less extensively, is not a rare phenomenon either. This led to the question in (3) whether differential indexing is different for different argument roles across languages. The formal and functional comparison of differential indexing for the A and P argument roles was dealt with in Chapter 6, where it was shown on the basis of cross-linguistic data that the referent tracking function of indexing holds irrespective of the argument role. This explains the often parallel behaviour of differential A and P indexing. Finally, following the lines of objective (2), Chapter 5 dealt with the observation that variablity in indexing is not always about the presence of the respective marker, but can also involve its placement

in the syntactic environment. The findings from this study also point towards the fact that there is more to indexing than indicating reference.

7.2 The multivariate character of differential indexing: the corpus-linguistic perspective

The literature on differential indexing shows that the same underlying factors (such as animacy, identifiability or topicality) can be encountered again and again across languages. However, the exact manifestation of these factors has to be viewed language-specifically. Not only can languages differ with regard to the relevant factors themselves, but also with regard to where a line is drawn on the respective hierarchies, or whether there is a precise line to be drawn at all. For instance, in Sambaa (Chapter 3), P indexing is determined by the animacy hierarchy and is thus obligatory for proper names, titles and first and second person referents. However, although the chance of P indexing continually decreases as we follow the animacy hierarchy further down, there is no cut-off point at which indexing becomes ungrammatical: albeit rare, it is still possible with inanimate referents (Riedel 2009: 46). Thus, other factors must be involved as well.

And once there is more than one factor identified as potential variable for indexing (such as animacy + discourse status), it remains to determine the extent to which these factors impact or depend on one another. This complex kind of language-internal interplay was showcased in two corpus-driven quantitative studies on P indexing in the two unrelated languages Ruuli and Maltese. Differential P indexing is similar in both languages in that firstly, there is a strong correlation of P indexing together with a particular constituent order, and secondly, that it could be attributed to the topicality of the referent (as has been suggested, in fact, for Maltese, as well as for other Bantu languages similar to Ruuli). However, the notion of topicality can evidently be broken down further, and its components are weighted differently in different languages.

In the analyses, differential P indexing in both the Ruuli and the Maltese data is strongly associated with constituent order, but there is no absolute correlation in either of the two languages. Apart from constituent order, for both languages discourse givenness, identifiability, and part of speech of the

head are significant variables, but their interrelations differ, which can be retraced through the visualization in the conditional inference trees. It should be pointed out, however, that for both languages, the reality of indexing is most definitely more complex than conveyed by the tree models. Some of the variable values are rather coarse-grained (e.g. there is unquestionably more to referent accessibility than identifiability and textual givenness) and, at least for Maltese, there is reason to believe that differential indexing is also a matter of style or text type. Finally, the various discourse effects often subsumed under the notion of topicality (as well as focality, cf. Ozerov 2018) cannot be covered by the choice of the discourse-structural proxies chosen for annotation. However, the studies do clearly show that ascribing differential indexing in languages like Ruuli and Maltese to topicality, or any of the other variables involved alone would not do justice to the reality of the phenomenon, and to the motivated choices the speakers make.

Similarly, the study on variable index placement in Gutob based on systematic corpus annotation reveals more about its true nature. Previous accounts had ascribed index placement in Gutob to expressive discourse and exceptional rhetoric conditions, without these being defined any further; the verb was considered as the default host for an index. However, it was shown that in clauses where there are alternatives to verbal placement, this claim does not hold, as in the majority of cases, indexes in fact seek different hosts. Additionally, the annotation made it possible to easily browse the various examples with non-verbal and verbal indexes and compare them in their respective contexts, in order to determine the effect that index placement has on the host constituent: speakers can mark constituents as particularly noteworthy by attaching the index to them, to ensure the information is met with the adequate level of attention.

Although the case studies treat different instantiations of differential indexing, they have in common that they show quite clearly that indexing in these three languages (and probably many more) is not (just) role assignment, and that differential indexing is not (just) induced by a referent's deviations from assumed associations with a particular argument role.

7.3 Parallels between differential A and P indexing: the cross-linguistic perspective

Chapter 6 compared the more systematically studied differential P indexing with differential A indexing, pulling together cross-linguistic data for both phenomena. Is was shown that indexing for any role is rooted in the referent reaching a particular level of prominence (in Haspelmath's 2021b terms), and indexing facilitates following this prominent referent through the discourse (cf. Iemmolo 2011) by signaling that this referent continues to hold this prominence status. As every role has its associations with prominence as well, differential indexing through the absence vs. the presence of an index, in many cases involves opposite levels of the prominence related factors. Thus, not to index an unexpectedly un-topical or inanimate A referent, and to index a surprisingly topical or animate referent, are both determined by the same underlying main driver, although this main driver's exact characteristics are, of course, language-specific. And even though this can result in the unexpected structural variant also being less frequent (e.g. both in Ruuli and Maltese, indexing P in the presence of a lexical NP is less frequent than not indexing it), indexing a referent or not can probably not be attributed to efficiency. This would also explain why indexing is different from other coding asymmetries like differential case marking, where this line of reasoning actually works (cf. Haspelmath 2021a, Haspelmath 2021b; also see Fauconnier & Verstraete 2014 who show that differential case marking for A and P are clearly triggered by distinct motivations).

Two notions often associated to indexing in reference grammars and other studies are obligatoriness and optionality. As has been argued in Section 2.1, obligatoriness should not be defined as depending on presence or absence of a co-referential NP. Also, for some languages where A indexing is described as being obligatory – i.e. present in what is considered the specific language's most frequent or default configuration – there can still be differential indexing to satisfy certain discourse-pragmatic effects. The use of a default configuration also has a discourse-pragmatic effect, as there is no such thing as a "pragmatically neutral" clause (Lambrecht 1994: 16). Some clauses or structures can probably fit more discourse effects than other, more specialized configurations, and thus evoke the impression of being "most normal" and, in fact, become the most frequent due to their versatility (Lambrecht

1994: 16-17, 126). However, for many languages it is difficult to translate the probability of indexing which is sensitive to discourse and/or semantics into obligatoriness. Concerning optionality in indexing, one should keep in mind that although a speaker might be free to choose, the decision made is not random (cf. Givón 2011: 189) but well motivated, though not necessarily by syntax.

7.4 Final reflections and prospects for future research

Due to the definition of differential indexing adopted in this thesis, various instantiations of differential indexing could be considered, opening up a path towards a better understanding of what lies at the core of indexing, and of what can make it differential. This thesis is to my knowledge the first work to deal with both differential A and P indexing, demonstrating that both omitting and adding an index can be explained on the basis of the same motivation, namely the function of the index signaling a referent's particular status in terms of semantic and discourse-structural factors. Additionally, it addresses the issue of variable index placement and how it can equally respond to pragmatic realities of the proposition.

This work highlights that supplementing grammar-based investigations with corpus-based studies turns out to be very meaningful and can substantiate the observations previously made on the basis of qualitative data: although judgments based on intuition can identify the basic semantic and discourse-structural factors which may underlie a split, they more often than not cannot do justice to the intricacies of those factors.

The evaluation of systematic corpus annotation is a very promising way to investigate the tendencies in marking splits. And although working on phenomena such as differential indexing on the basis of reference grammars often leaves open several questions, I do not argue that grammatical description needs statistical support; on the contrary. Reference grammars are the primary tools for establishing what linguistic structures there are, their characteristics, and how they are distributed. As of now, corpora cannot replace reference grammars as the main data source in linguistic typology (cf. Levshina 2021: 2). However, more corpus-driven accounts, especially for lesser-

described languages, backing up descriptive work and highlighting the probabilistic character of some features, are certainly a worthwhile investment for linguistic typology. As for further research on differential indexing, it should go beyond focusing on object roles, as initiated in this thesis. Furthermore, variable index placement should be paid more attention to from a typological as well as functional perspective. Both objectives would lead to a deeper understanding of the functions of indexing, and could also provide implications for the understanding of other coding splits and syntactic optionalities.

Appendices

Appendix A

Ruuli coding scheme

A.1 Annotated texts

All annotated texts are part of the corpus of spoken Ruuli (Witzlack-Makarevich et al. 2019); they are all transcripts from converstions recorded in February 2017, in the villages of Kayunga, Nakasongola, and Kibbale. The following table lists the text IDs with some meta information on the speakers (sex and year of birth) and the topics of the conversations:

The texts and their translations are copied into a spreadsheet for annotation, each variable is coded in a devoted column.

A.2 Step 1: Identifying P

Manually tag transitive predicates. The following instances are excluded:

a) clauses where P is a headless relative clause or a complement clause
b) cases of light verb constructions/fixed expressions without individuated participants
c) passives, reflexives, reciprocals
d) unclear ditransitives (e.g. 'to note')

With these predicates, we identify the less agent-like argument as P.

ID	Speakers	Topics
II-R-NAKASONGOLA-170225-FS-1	M 1967 F 1980	expectations towards family members; gender roles; traditional medicines; working abroad
II-N-BBALE-170220-FS-4	M 1950 F 1955 F unknown	childhood; taboos; marriage
II-N-BBALE-170220-FS-2	M 1968 F 1958	Banyala historical accounts; expectations for the future
II-N-WSKAYUNGA-170218-FS-1A	M 1964 M 1939	culture in general; traditional religion among the Banyala; political history; names and naming strategies
II-M-KIBBALE-170221-FS-4B	M 1974 M 1960	types and uses of trees; wild and domestic animals
II-R-NAKASONGOLA-170224-FS-1A	M 1962 M 1962 M 1951	political structure, blacksmithing, pottery, beer making

Table A.1: Metadata on annotated Ruuli texts

A.3 Step 2: Determine the head of P

If P is represented by a complex expression, such as the ones below, we annotate the head of the phrase. The following scheme will be used:

- NP1 + conjunction + NP2 → NP1
- determiner/adjective/numeral + noun → noun
- noun1 + preposition + noun2 → noun1 (e.g. *abasaiza ba irai* 'men of the past')
- pronoun + NP → pronoun (e.g. *owa Kangulumira* 'one of the Kangulumira')
- noun + possessive pronoun → noun (e.g. *engeso zaalyo* 'its norms')

There can also be a zero head e.g. *owamu aliayi* 'where is yours [i.e. your child]?'. In that case, code only the properties that can be inferred from the context. Code "NA" for the rest. In this example, there will be a "NA" for part of speech and subcategory of the head, whereas 3SG, noun class 1, human, definite, and given can be inferred.

A.4 Step 3: Code for semantic and formal variables of P, the overtness of the A referent and word order

1. Index
   - 0 (no index)
   - 1 (index)
2. Presence of referential NP
   - 0 (no NP)
   - 1 (NP present)
3. POS of the head
   - noun
   - pronoun
   - NA (in case there's no head)
4. Subcategory of the head
   a) For nouns: "proper" vs. "common", the following are instances of "proper"
      - personal names
      - place names
      - institutions

      The rest are coded as common nouns.
   b) Pronouns: personal, possessive, demonstrative, interrogative, other
   c) Other parts of speech or zero head: NA
5. Noun class of the head (for Ruuli noun classes see Namyalo et al. 2021: 43-49)
6. Modification of head noun
   - NA (if there is no head)
   - none
   - possessive pronoun
   - adjective
   - relative clause
   - prepositional phrase

- quantifier
- numeral
- demonstrative
- interrogative
- other
- multiple (if there is more than one modifier)

7. Referent person and number
   - 1SG
   - 1PL
   - 2SG
   - 2PL
   - 3SG
   - 3PL
8. Semantic class of the referent: we always code the referent, not the noun (e.g. *kanisa* 'church' could be a physical object or an institution, depending on context)
   - human
   - kinship term
   - environment (field, river etc.)
   - animal
   - physical object (can be touched)
   - abstract entity (cannot be touched), but not an event
   - event (involves the time dimension)
   - organization
   - anthropomorphic (god, angels, demons, etc.)
9. Identifiability: Covers both definiteness and specificity.
   - definite: the referent can be identified by both the speaker and the hearer, e.g. *My father has bought the car I told you about.*
   - specific: identifiable by the speaker only, e.g. *I've just bought a car.*
   - non-specific: not identifiable by neither the speaker nor the hearer, e.g. *I want to buy a car [any car]*. Also used in impersonal contexts (e.g. *They are going to build a plant*) and in generic statements when

the focus is on any arbitrary member of the class (e.g. *A computer is a useful device*).

10. Discourse accessibility of the referent (givenness):
    - given: textually given (i.e. mentioned previously + inferable from previous mentioning).
    - new: first mention of a referent.
    - NA: impersonal uses of pronouns, interrogative and relative pronouns.

    We do not include a status such as "accessible" (cf. Chafe 1976), meaning not previously mentioned, but inferable using background knowledge. Speakers' background knowledge is very hard to know, so we end up with textual givenness in terms of previously mentioned vs. not mentioned.
11. Presence of the A noun phrase
    - 0 (no NP)
    - 1 (NP present)
12. Word order (linear order of A NP, P NP and verb)
    - AVP
    - APV
    - PAV
    - PVA
    - PV
    - AV
    - VP
    - V

Appendix B

Maltese coding scheme

B.1 Texts

From the transcripts of parliamentary debates from the *bulbulistan* corpus[1], all (orthographic) sentences (cf. Čéplö 2018a: 63-64) containing the keyword *nagħmlu* (without index) and *nagħmluha* (with an index for 3SG.F) are extracted; the preceding and the following 1000 characters are extracted as well, in order to account for context. We conceive of these transcripts as "coming close to naturalistic speech" and not "naturalistic speech", as a comparison of randomly selected transcripts with their audio recordings has made it clear that some editing was executed (Čéplö 2018a: 58).

B.2 Annotation

In a spreadsheet, the clauses containing the keywords are annotated, taking into consideration the left and right context. Instances with as well as without co-referential object NP are considered. If this NP is a complex one, we consider the features of the head of the object NP. EJ annotates all instances of *nagħmluha*, SČ takes on instances of *nagħmlu*.

The following formal and referential features are coded, in one dedicated column each:

1. Index

[1] see Čéplö 2018a and http://www.bulbul.sk/bonito2/

- 0 (no index)
- 1 (index)

2. Presence of referential NP
   - 0 (np NP)
   - 1 (NP present)
3. POS of the head
   - noun
   - pronoun
   - NA (in case there's no head)
4. Subcategory of the head
   - proper noun
   - common noun
   - personal pronoun
   - impersonal use of personal pronoun
   - possessive pronoun
   - demonstrative pronoun
   - interrogative pronoun
5. Modification of head NP
   - NA (if there is no head)
   - none
   - adjective
   - relative clause
   - determiner
   - possessive
   - numerl
   - demonstrative
   - multiple
6. Semantic class of the referent
   - physical object (can be touched)
   - abstract entity (cannot be touched), but not an event
   - event (involves the time dimension)
   - organization

7. Identifiability: Covers both definiteness and specificity
    - definite: the referent can be identified by both the speaker and the hearer, e.g. *my father has bought the car I told you about*; can be overtly marked for definiteness, but does not have to be
    - specific: identifiable by the speaker only
    - non-specific: not identifiable by neither the speaker nor the hearer
8. Givenness
    - given: textually given, i.e. explicitly mentioned, within the previous 1000 characters
    - new: not mentioned within the previous 1000 characters
9. Clause type
    - main clause
    - relative clause
    - adverbial clause
    - complement clause
10. Polarity
    - positive
    - negative
11. Sentence type
    - declarative
    - imperative
    - interrogative
    - exhortative
12. Order of subject, object and verb
    - SVO
    - VO
    - OV
    - V

Appendix C

Gutob coding scheme

C.1 Annotated texts

ID	Speakers	Topics
Gutob-0444-20161125_3	Gurbari, ~70y	interview on traditions
Gutob-0444-20161205_9	Tulsa, Komu, Rotika, ~20-35y	fairy tale
Gutob-0444-20161220	Komla, ~45y	life experiences
Gutob-0444-20170105_1	Sukri, ~45y	life experiences
gutob-0444-20170116_3	Donnai, ~20y	fairy tale
Gutob-0444-20170119_4	Komla, ~45y	daily life experiences
Gutob-0444-20170130	Rotika, ~20y	fairy tale
Gutob-0444-20170131	Komla, ~45y	interview on traditions
Gutob-0444-20170209_1	Komla, ~45y	interview on traditions
Gutob-0444-20170210	Komla, ~45y	interview on traditions
Gutob-0444-20170215	Rotika, ~20y	fairy tale
Gutob-0444-20170327_2	Sukri, ~45y	life experiences

Table C.1: Metadata on annotated Gutob texts

Our study is based on a corpus collected during a recent language documentation project (Voß 2018) between 2016 and 2018. Our subcorpus for annotation contains 32669 words and is comprised of 12 narratives and stories from everyday life (approx. 360 min) by 7 speakers (see Table C.1). As only very few women in the parent generation, usually the eldest daughters in the family, can speak Gutob, and the youngest male speakers are in the grandparent generation or up, all of the speakers here are female.

C.2 Clauses coded

We only annotate finite clauses. Clauses with non-verbal predication are excluded, as well as conditional and sequential clauses. Conditional clauses entail then conditional verbal suffix *-na*, glossed as COND. Sequential clauses are recognized by the final clitic *=su*, glossed as =and.

C.3 Variables

In a spreadsheet, we code for every clause

- person and number of the S/A argument
  - 1SG
  - 1PL
  - 2SG
  - 2PL
  - 3SG
  - 3PL
- whether the referent is expressed by an NP (values 1 or 0)
- whether there is a non-verbal index (1 or 0)
- whether there is a verbal index (1 or 0)
- the host of the non-verbal index
  - adverbial phrase
  - adverb
  - demonstrative
  - interrogative pronoun

  - numeral
  - object NP
  - other
  - NA (in case of no non-verbal index)
- whether a preverbal index would have been sytactically possible (values 1 or 0)

Bibliography

Aikhenvald, Alexandra Y. 2003. Typological parameters for the study of clitics, with special reference to Tariana. In Robert M. W. Dixon & Alexandra Y Aikhenvald (eds.), *Word: A Cross-linguistic Typology*, 42–78. Cambridge University Press.

Aissen, Judith. 1999. Markedness and subject choice in Optimality Theory. *Natural Language and Linguistic Theory* 17(4). 673–711.

Aissen, Judith. 2003. Differential Object Marking: Iconicity vs. Economy. *Natural Language and Linguistic Theory* 21. 435–483.

Amberber, Mengistu. 2009. Differential case marking of arguments in Amharic. In Andrej Malchukov & Andrew Spencer (eds.), *The Oxford Handbook of Case*, 742–755. Oxford: Oxford University Press.

Amha, Azeb. 2007. Questioning forms in Zargulla. In Rainer Voigt (ed.), *From Beyond the Mediterranean: Akten des 7. Internationalen Semitohamitistenkongresses*, 197–210. Düren: Shaker Verlag.

Amha, Azeb. 2009. The morphosyntax of negation in Zargulla. In W. Leo Wetzels (ed.), *The Linguistics of Endangered Languages*, 197–220. Utrecht: LOT.

Anderson, Gregory. 2001. A new classification of South Munda: Evidence from comparative verb morphology. *Indian linguistics* 62(1-4). 21–36.

Anderson, Gregory D. S. 2007. *The Munda Verb: Typological Perspectives*. Berlin: De Gruyter Mouton.

Anderson, Gregory D. S. 2008. *The Munda languages*. London: Routledge.

Anderson, Gregory D.S. & K. David Harrison. 2008. Remo (Bonda). In Gregory D.S. Anderson (ed.), *The Munda languages*, 557–632. London: Routledge.

Anderson, Gregory D.S. & Norman H Zide. 2001. Recent advances in the reconstruction of the Proto-Munda verb. In Laurel J. Brinton (ed.), *Historical Linguistics 1999: Selected papers from the 14th International Conference on Historical Linguistics, Vancouver, 9–13 August 1999*, 13–30. Amsterdam: John Benjamins.

Anderson, Stephen R. 1992. *A-Morphous Morphology*. Cambridge: Cambridge University Press.

Anderson, Stephen R. 1993. Wackernagel's revenge: clitics, morphology, and the syntax of second position. *Language* 69(1).

Anderson, Stephen R. 2005. *Aspects of the theory of clitics*. Oxford: Oxford University Press.

Aoun, Joseph. 1999. Clitic-doubled arguments. In Kyle Johnson & Ian Roberts (eds.), *Beyond principles and parameters: Essays in Memory of Osvaldo Jaeggli*, 13–42. Dordrecht: Springer.

Aquilina, Joseph. 1940. *The structure of Maltese: A study in mixed grammar and vocabulary*. London: School of Oriental and African Studies.

Aquilina, Joseph. 1958. Maltese as a mixed language. *Journal of Semitic Studies* 3(1). 58–79.

Aquilina, Joseph. 1987. *Maltese-English Dictionary*. Malta: Midsea Books.

Ariel, Mira. 1990. *Accessing noun-phrase antecedents*. London/New York: Routledge.

Arkadiev, Peter. 2010. Clitic doubling: Towards a typology. Paper presented at the Workshop on Clitics and Syntactic Typology. `https://www.academia.edu/1695634/Clitic_doubling_Towards_a_typology`.

Arkadiev, Peter. 2013. Double-marking of prominent objects: a cross-linguistic typology. Handout of talk presented at the 10th Conference of the Association of Linguistic Typology in Leipzig, 2013.

Baier, Nicholas Benson. 2018. *Anti-agreement*. Berkeley: University of California PhD dissertation.

Baker, Brett J. 2002. How referential is agreement? In Nicholas D. Evans & Hans-Jürgen Sasse (eds.), *Problems of polysynthesis*, Berlin: Akademieverlag.

Bakker, Dik & Anna Siewierska. 2009. Case and alternative strategies: Word order and agreement marking. In Andrej Malchukov & Andrew Spencer (eds.), *The Oxford handBook of case*, 290–303. Oxford: Oxford University

Press.

Barbosa, Pilar. 1996. Clitic placement in European Portuguese and the position of subjects. In Aaron L. Halpern & Arnold M. Zwicky (eds.), *Approaching second: Second position clitics and related phenomena*, 1–40. Stanford, CA: CSLI Publications.

Baumann, Stefan. 2012. *The Intonation of Givenness: Evidence from German*. Tübingen: Max Niemeyer Verlag.

Béjar, Susana. 1999. Agreement alternations and functional licensing in Selayarese. *Toronto Working Papers in Linguistics* 16(2). 51–61.

Benveniste, Émile. 1971. *Problems in General Linguistics*. Coral Gables, FL: University of Miami Press.

Berger, Peter. 2015. *Feeding, Sharing, and Devouring: Ritual and Society in Highland Odisha, India*. Berlin: De Gruyter Mouton.

Bickel, Balthasar. 2003. Referential density in discourse and syntactic typology. *Language* 79.

Bickel, Balthasar. 2008. On the scope of the referential hierarchy in the typology of grammatical relations. In Greville G. Corbett & Michael Noonan (eds.), *Case and Grammatical Relations: Studies in Honor of Bernard Comrie*, 191–210. Amsterdam: John Benjamins.

Bickel, Balthasar. 2011. Grammatical relations typology. In Jae Jung Song (ed.), *The Oxford HandBook of Language Typology*, 399–444. Oxford: Oxford University Press.

Bickel, Balthasar, Giorgio Iemmolo, Taras Zakharko & Alena Witzlack-Makarevich. 2013. Patterns of alignment in verb agreement. In Dik Bakker & Martin Haspelmath (eds.), *Languages across boundaries: Studies in memory of Anna Siewierska*, 15–36. Berlin: De Gruyter Mouton.

Bickel, Balthasar & Johanna Nichols. 2007. Inflectional morphology. In Timothy Shopen (ed.), *Language Typology and Syntactic Description*, vol. 3, 169–240. Cambridge: Cambridge University Press.

Bickel, Balthasar & Johanna Nichols. 2009. Case-marking and alignment. In Andrej Malchukov & Andrew Spencer (eds.), *The Oxford HandBook of Case*, 304–321. Oxford: Oxford University Press.

Blois, Kornelis Frans de. 1970. The augment in the Bantu languages. *Africana linguistica* 4(1). 85–165.

Bolotin, Naomi. 1995. Arabic and parametric VSO agreement. In Mushira Eid (ed.), *Perspectives on Arabic Linguistics : Papers from the Annual Symposium on Arabic Linguistics Volume VII*, 7–28. Amsterdam: John Benjamins.

Borg, Albert & Marie Azzopardi-Alexander. 1997. *Maltese*. London: Routledge.

Borg, Albert & Marie Azzopardi-Alexander. 2009. Topicalisation in Maltese. In Bernard Comrie, Ray Fabri, Elizabeth Hume, Manwel Mifsud, Thomas Stolz & Martine Vanhove (eds.), *Introducing Maltese Linguistics: Selected papers from the 1st International Conference on Maltese Linguistics*, 71–81. Amsterdam: John Benjamins.

Borsley, Robert D. 2009. On the superficiality of Welsh agreement. *Natural Language & Linguistic Theory* 27(2). 225–265.

Borsley, Robert D., Maggie Tallerman & David Willis. 2007. *The Syntax of Welsh*. New York: Cambridge University Press.

Bossong, Georg. 1982. Der präpositionale Akkusativ im Sardischen. In Otto Winkelmann & Maria Braisch (eds.), *Festschrift für Johannes Hubschmid zum 65. Geburtstag: Beiträge zur allgemeinen, indogermanischen und romanischen Sprachwissenschaft*, 579–599. Bern: Francke.

Bresnan, Joan, Anna Cueni, Tatiana Nikitina & Harald Baayen. 2007. Predicting the dative alternation. In Gerlof Boume, Irene Krämer & Joost Zwarts (eds.), *Cognitive Foundations of Interpretation*, 69–94. Amsterdam: Koniklijke Nederlandse Akademie van Wetenschapen.

Bresnan, Joan & Sam A. Mchombo. 1987. Topic, pronoun, and agreement in Chichewa. *Language* 63. 741–782.

Brustad, Kristen. 2000. *The syntax of spoken Arabic: A comparative study of Moroccan, Egyptian, Syrian, and Kuwaiti dialects*. Washington: Georgetown University Press.

Buell, Leston Chandler. 2005. *Issues in Zulu verbal morphosyntax*: University of California at Los Angeles PhD dissertation.

Butt, Miriam & Wilhelm Geuder. 2003. Light verbs in Urdu and grammaticalization. In Regine Eckardt, Klaus von Heusinger & Christoph Schwarze (eds.), *Words in time: diachronic semantics from different points of view*, 295–350. Berlin: Mouton de Gruyter.

Capell, Arthur. 1972. The affix-transferring languages of Australia. *Linguistics* 10(87). 5–36.

de Cat, Cécile. 2004. On the impact of French subject clitics on the information structure of the sentence. In Bart Hollebrandse, Brigitte Kampers-Manhe, Petra Sleeman & Reineke Bok-Bennema (eds.), *Romance Languages and Linguistic Theory: Selected Papers from "Going Romance", Groningen, 28-30 November 2002*, 33–46. Amsterdam: John Benjamins.

Čéplö, Slavomír. 2014. An overview of object reduplication in Maltese. In Alexandra Vella Albert Borg, Sandro Caruana (ed.), *Perspectives on Maltese Linguistics*, 201–222. Berlin: De Gruyter Mouton.

Čéplö, Slavomír. 2018a. *Constituent order in Maltese: A quantitative analysis*. Prague: Charles University in Prague Doctoral dissertation.

Čéplö, Slavomír. 2018b. Maltese Universal Dependencies Treebank. In Joakim Nivre et al. (ed.), *Universal Dependencies 2.3, LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL)*, Prague: Charles University.

Chafe, Wallace L. 1976. Givenness, contrastiveness, definiteness, subjects, topics, and point of view. In Charles N. Li (ed.), *Subject and topic*, 27–55. New York: Academic Press.

Cojocaru, Dana. 2004. *Romanian Grammar*. Durham: Slavic and East European Language Research Center.

Comrie, Bernard. 1988. Topics, grammaticalized topics, and subjects. In *Proceedings of the Fourteenth Annual Meeting of the Berkeley Linguistics Society*, 265–279.

Comrie, Bernard & Michael Spagnol. 2016. Maltese loanword typology. In Gilbert Puech & Benjamin Saade (eds.), *Shifts and Patterns in Maltese*, 315–330. De Gruyter Mouton.

Coppock, Elizabeth & Stephen Wechsler. 2012. The objective conjugation in Hungarian: Agreement without phi-features. *Natural Language & Linguistic Theory* 30(3). 699–740.

Corbett, Greville G. 2006. *Agreement*. Cambridge: Cambridge University Press.

Cowell, Mark W. 1964. *A Reference Grammar of Syrian Arabic: Based on the dialect of Damascus*. Washington: Georgetown University Press.

Creissels, Denis. 2005. A typology of subject marker and object marker systems. In Erhard F.K. Voeltz (ed.), *Studies in African linguistic typology*, 445–459. Amsterdam: John Benjamins.

Creissels, Denis. 2009. Ergativity/Accusativity Revisited. Presented at ALT VIII, Berkeley (www.deniscreissels.fr/public/Creissels-ergativity.pdf).

Croft, William. 1988. Agreement vs. Case Marking and Direct Objects. In Michael Barlow & Charles Ferguson (eds.), *Agreement in Natural Language: Approaches, Theories, Descriptions*, Stanford, CA: CSLI Publications.

Croft, William. 2001. *Radical Construction Grammar: Syntactic Theory in Typological Perspective*. Oxford: Oxford University Press.

Croft, William. 2003. *Typology and Universals*. Cambridge: Cambridge University Press.

Culbertson, Jennifer. 2010. Convergent evidence for categorial change in French: from subject clitic to agreement marker. *Language* 86(1). 85–132.

Cysouw, Michael. 2003. Towards a typology of pronominal cliticization. Handout presented at the 5th International Conference of the Association for Linguistic Typology. `http://cysouw.de/home/presentations_files/cysouwCLITICS_handout.pdf`.

Cysouw, Michael. 2004. The rise of person inflection with special reference to the Munda languages. Paper presented at the 11th International Morphology Meeting 2004.

Dahl, Östen. 2000. Egophoricity in discourse and syntax. *Functions of Language* 7(11). 37–77.

Dalrymple, Mary & Irina Nikolaeva. 2011. *Objects and information structure*. Cambridge: Cambridge University Press.

De Cat, Cécile & Katherine Demuth. 2008. *The Bantu-Romance Connection: A comparative investigation of verbal agreement, DPs, and information structure*. Amsterdam: John Benjamins.

DeLancey, Scott. 1981. An interpretation of split ergativity and related patterns. *Language* 57. 626–657.

Diercks, Michael, Rodrigo Ranero & Mary Paster. 2015. Evidence for a clitic analysis of object markers in Kuria. In Elizabeth C. Zsiga & One Tlale Boyer (eds.), *Selected Proceedings of the 44th Annual Conference on African Linguistics*, Somerville, MA: Cascadilla Proceedings Project.

Dixon, Robert M. W. 1979. Ergativity. *Language* 55. 59–138.

Dixon, Robert M. W. 2002. *Australian Languages*. Cambridge: Cambridge University Press.

Dixon, Robert M. W. 2010. *Basic Linguistic Theory*, vol. 2: Grammatical Topics. Oxford: Oxford University Press.

Dixon, Robert M. W. & Alexandra Y. Aikhenvald. 2002. *Word: A Cross-linguistic Typology*. Cambridge: Cambridge University Press.

Donegan, Patricia & David Stampe. 2004. Rhythm and the synthetic drift of Munda. *The yearbook of South Asian languages and linguistics* 7. 3–36.

Downing, Laura & Lutz Marten. 2019. Clausal morphosyntax and information structure. In Mark Van de Velde, Koen Bostoen, Derek Nurse & Gérard Philippson (eds.), *The Bantu languages*, 270–307. London: Routledge.

Downing, Laura J. 2011. The prosody of 'dislocation' in selected Bantu languages. *Lingua* 121(5). 772 – 786.

Downing, Laura J. 2018. Differential object marking in Chichewa. In Alena Witzlack-Makarevich & Ilja A. Seržant (eds.), *Diachrony of differential argument marking*, 41–67. Berlin: Language Science Press.

Dryer, Matthew. 1986. Primary Objects, Secondary Objects, and Antidative. *Language* 62. 808–845.

Dryer, Matthew S. 1997. Are grammatical relations universal? In Joan Bybee, John Haiman & Sandra A. Thompson (eds.), *Essays on Language Function and Language Type Dedicated to T. Givón*, 115–143. Amsterdam: John Benjamins.

DuBois, John W. 1987. The discourse basis of ergativity. *Language* 63. 805–855.

Duff-Tripp, Martha. 1997. *Gramática del idioma yanesha' (amuesha)* Serie Lingüística Peruana. Lima: Ministerio de Educación and Instituto Lingüístico de Verano.

Duranti, Alessandro. 1979. Object clitic pronouns in Bantu and the topicality hierarchy. *Studies in African Linguistics* 10(1). 31–45.

Erika & Alena Witzlack-Makarevich. accepted. A corpus-based analysis of P indexing in Ruuli (Bantu, JE103). Manuscript accepted for publication.

Fabri, Ray. 1993. *Kongruenz und die Grammatik des Maltesischen*. Tübingen: Max Niemeyer Verlag.

Fabri, Ray & Albert Borg. 2002. Topic, focus and word order in Maltese. In Abderrahim Youssi, Fouzia Benjelloun, Mohamed Dahbi & Zakia Iraqui Sinaceur (eds.), *Aspects of Dialects of Arabic Today*, 354–363. Rabat: Amapatril.

Facundes, Sidney da Silva. 2000. *The Language of the Apurinã People of Brazil (Maipure/Arawak)*. Buffalo, NY: SUNY PhD dissertation.

Falk, Yehuda N. 2006. *Subjects and Universal Grammar: An Explanatory Theory*. Cambridge: Cambridge University Press.

Fauconnier, Stefanie & Jean-Christophe Verstraete. 2014. A and O as each other's mirror image? Problems with markedness reversal. *Linguistic Typology* 18(1). 3–49.

Fedden, Sebastian, Dunstan Brown, František Kratochvíl, Laura C Robinson & Antoinette Schapper. 2014. Variation in pronominal indexing: lexical stipulation vs. referential properties in Alor-Pantar languages. *Studies in Language* 38(1). 44–79.

Fernandez, Frank. 1983. The morphology of the Remo (Bonda) verbs. *International Journal of Dravidian Linguistics* 12. 15–45.

Fischer, Susann, Mario Navarro & Jorge Vega Vilanova. 2019. The clitic doubling parameter: Development and distribution of a cyclic change. In *Cycles in Language Change*, 52–70. Oxford University Press.

Fischer, Susann & Esther Rinke. 2013. Explaining the variability of clitic doubling across Romance: a diachronic account. *Linguistische Berichte* 2013(236). 455–472.

Forker, Diana. 2016. Floating agreement and information structure: The case of Sanzhi Dargwa. *Studies in Language* 40(1). 1–25.

Franks, Steven & Tracy Holloway King. 2000. *A handbook of Slavic clitics*. Oxford: Oxford University Press.

Friedman, Victor A. 2008. Balkan object reduplication in areal and dialectological perspective. In Dalina Kallulli & Liliane Tasmowski (eds.), *Clitic doubling in the Balkan languages*, 35–63. Amsterdam: John Benjamins.

Gaby, Alice Rose. 2006. *A Grammar of Kuuk Thaayorre*. Melbourne: University of Melbourne PhD dissertation.

García-Miguel, José M. 2015. Variable coding and object alignment in Spanish: A corpus-based approach. *Folia Linguistica* 49(1). 205–256.

Gilligan, Gary Martin. 1987. *A cross-linguistic approach to the pro-drop parameter*. Los Angeles: University of Southern California Doctoral dissertation.

Givón, T. 2011. *Ute Reference Grammar*. Amsterdam: John Benjamins.

Givón, Talmy. 1976. Topic, pronoun, and grammatical agreement. In Charles N. Li (ed.), *Subject and Topic*, New York: Academic Press.

Givón, Talmy (ed.). 1983. *Topic continuity in discourse: a quantitative cross-language study*. Amsterdam: John Benjamins.

Goldstein, David M. 2021. A multifactorial analysis of differential agent marking in Herodotus. *Journal of Greek Linguistics* 21(1). 3–57.

Gómez González, María de los Ángeles. 1997. On Theme, Topic and Givenness: The state of the art. *Moenia* 135–155.

Gries, Stefan Th. 2020. On classification trees and random forests in corpus linguistics: Some words of caution and suggestions for improvement. *Corpus Linguistics and Linguistic Theory* 16(3). 617–647.

Griffiths, Arlo. 2008. Gutob. In Gregory D. S. Anderson (ed.), *The Munda languages*, 633–681. London: Routledge.

Guillaume, Antoine. 2009. Hierarchical agreement and split intransitivity in Reyesano. *International journal of American linguistics* 75(1). 29–48.

Gunlogson, Christine. 2001. Third-Person Object Prefixes in Babine-Witsuwit'en. *International Journal of American Linguistics* 67(4). 365–395.

Haig, Geoffrey. 2018. The grammaticalization of object pronouns: Why differential object indexing is an attractor state. *Linguistics* 56(4). 781–818.

Haig, Geoffrey & Diana Forker. 2018. Agreement in grammar and discourse: A research overview. *Linguistics* 56(4). 715–734.

Haig, Geoffrey, Nils N. Schiborr & Stefan Schnell. 2020. On potential statistical universals of grammar in discourse: Evidence from Multi-CAST. Talk presented at the 42. Jahrestagung der Deutschen Gesellschaft für Sprachwissenschaft, Hamburg.

Harris, Alice C. 2000. Where in the word is the Udi clitic? *Language* 76(3). 593–616.

Harris, Alice C. 2009. *Georgian syntax: a study in relational grammar* Cambridge studies in linguistics. Cambridge [u. a.]: Cambridge: Cambridge University Press. Includes index. Bibliography: p. [309]-318.

Haspelmath, Martin. 2010. Comparative concepts and descriptive categories in crosslinguistic studies. *Language* 86(3).

Haspelmath, Martin. 2011. On S, A, P, T, and R as comparative concepts for alignment typology. *Linguistic Typology* 15. 535–567.

Haspelmath, Martin. 2013. Argument indexing: A conceptual framework for the syntactic status of bound person forms. In Dik Bakker & Martin Haspelmath (eds.), *Languages across boundaries: Studies in memory of Anna Siewierska*, 197–226. Berlin: De Gruyter Mouton.

Haspelmath, Martin. 2019. Indexing and flagging, and head and dependent marking. *Te Reo* 62(1). 93–115.

Haspelmath, Martin. 2021a. Explaining grammatical coding asymmetries: Form–frequency correspondences and predictability. *Journal of Linguistics* 1—29.

Haspelmath, Martin. 2021b. Role-reference associations and the explanation of argument coding splits. *Linguistics* 59(1). 123–174.

Heath, Jeffrey. 2008. *A grammar of Jamsay*. Berlin, New York: De Gruyter Mouton. `http://www.degruyter.com/view/books/9783110207224/9783110207224/9783110207224.xml`.

Hellenthal, Anneke Christine. 2010. *A grammar of Sheko*: University of Leiden PhD dissertation.

Hoffmann, Johann Baptist. 1903. *Mundari grammar*. Calcutta: Bengal Secretariat Press.

Hoop, Helen De & Andrej L. Malchukov. 2008. Case-Marking Strategies. *Linguistic Inquiry* 39(4). 565–587.

Hothorn, Torsten, Kurt Hornik & Achim Zeileis. 2006. Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical statistics* 15(3). 651–674.

Hundt, Marianne. 2018. It is time that this (should) be studied across a broader range of Englishes: A global trip around mandative subjunctives. In Sandra C. Deshors (ed.), *Modeling World Englishes: Assessing the interplay of emancipation and globalization of ESL varieties*, 217–244. Amsterdam: John Benjamins.

Iemmolo, Giorgio. 2010. Topicality and differential object marking: Evidence from Romance and beyond. *Studies in Language* 34(2). 239–272.

Iemmolo, Giorgio. 2011. *Towards a typological study of Differential Object Marking and Differential Object Indexation*: Università degli Studi di Pavia PhD dissertation.

Iemmolo, Giorgio & Alena Witzlack-Makarevich. 2013. When is there agreement? Typologizing suspension restrictions on agreement. Talk at the 10th Biennial Conference of the Association for Linguistic Typology, Leipzig.

Ivanov, Ivan P. 2012. L2 acquisition of Bulgarian clitic doubling: A test Case for the Interface Hypothesis. *Second Language Research* 28(3). 345–368. http://www.jstor.org/stable/43103900.

Jaeggli, Osvaldo. 1981. *Topics in Romance syntax*. Dordrecht: Foris Publications.

Jukes, Anthony. 2015. Focus and argument indexing in Makasar. In *Proceedings of the Second International Workshop on Information Structure of Austronesian Languages, Research Institute for Languages and Cultures of Asia and Africa, Tokyo University of Foreign Studies*, 53–63.

Just, Erika. submitted. A structural and functional comparison of differential A and P indexing. Manuscript submitted for publication.

Just, Erika & Slavomír Čéplö. to appear. Differential object indexing in Maltese - a corpus based pilot study. In Przemysław Turek & Julia Nintemann (eds.), *Maltese: Contemporary changes and historical innovations*, Berlin: De Gruyter Mouton.

Kallulli, Dalina. 2000. Direct object clitic doubling in Albanian and Greek. In Frits H. Beukema & Marcel Den Dikken (eds.), *Clitic phenomena in European languages*, 209–248. John Benjamins.

Kallulli, Dalina & Liliane Tasmowski. 2008a. Clitic doubling, core syntax and the interfaces. In Dalina Kallulli & Liliane Tasmowski (eds.), *Clitic doubling in the Balkan languages*, 1–34. Amsterdam: John Benjamins.

Kallulli, Dalina & Liliane Tasmowski. 2008b. *Clitic doubling in the Balkan languages*. Amsterdam: John Benjamins.

Kawalya, Deo, Koen Bostoen & Gilles-Maurice De Schryver. 2014. Diachronic semantics of the modal verb-sóból-in Luganda: A corpus-driven approach. *International Journal of Corpus Linguistics* 9(1). 60–93.

Kibrik, Aleksandr E. 1997. Beyond subject and object: towards a comprehensive relational typology. *Linguistic Typology* 1. 279–346.

Kibrik, Andrej A. 2011. *Reference in discourse*. Oxford: Oxford University Press.

Kießling, Roland. 1994. *Eine Grammatik des Burunge*, vol. 13 Afrikanistische Forschungen. Hamburg: Research and Progress Verlag.

Kiparsky, Paul et al. 2008. Universals constrain change; change results in typological generalizations. *Linguistic universals and language change* 23–53.

Klamer, Marian. 2010. *A grammar of Teiwa*. Berlin: De Gruyter Mouton.

Klamer, Marian & František Kratochvíl. 2018. The evolution of differential object marking in Alor-Pantar languages. In Alena Witzlack-Makarevich & Ilja A. Seržant (eds.), *Diachrony of differential argument marking*, 69–95. Berlin: Language Science Press.

Klavan, Jane, Maarja-Liisa Pilvik & Kristel Uiboaed. 2015. The Use of Multivariate Statistical Classification Models for Predicting Constructional Choice in Spoken, Non-Standard Varieties of Estonian. *SKY Journal of Linguistics* 28. 187–224.

Laidig, Wyn D. & Carol J. Laidig. 1990. Larike Pronouns: Duals and Trials in a Central Moluccan Language. *Oceanic Linguistics* 29(2). 87–109.

Lambrecht, Knud. 1994. *Information Structure and Sentence Form*. Cambridge: Cambridge University Press.

Lambrecht, Knud & Maria Polinsky. 1997. Typological variation in sentence-focus constructions. *Papers from the Thirty-Third Regional Meeting of the Chicago Linguistic Society: Panels on Linguistic Ideologies in Contact*. 141–165.

Lehmann, Christian. 1982. Universal and typological aspects of agreement. In Hansjakob Seiler & Franz J. Stachowiak (eds.), *Apprehension: Das sprachliche Erfassen von Gegenständen, Teil II: Die Techniken und ihr Zusammenhang in Einzelsprachen*, 201–267. Tübingen: Narr.

Levin, Aryeh. 1987. The Particle la as an Object Marker in some Arabic Dialects of the Galilee. *Zeitschrift für arabische Linguistik* 17. 31–40.

Levshina, Natalia. 2015. *How to do linguistics with R*. Amsterdam: John Benjamins.

Levshina, Natalia. 2021. Corpus-based typology: applications, challenges and some solutions. *Linguistic Typology* 25. 1–32.

Li, Charles N. & Sandra A. Thompson. 1976. Subject and topic: a new typology of language. In Charles N. Li (ed.), *Subject and topic*, 466–489. New York: Academic Press.

Lyons, Christopher. 1999. *Definiteness*. Cambridge: Cambridge University Press.

Macaulay, Monica. 1996. *A Grammar of Chalcatongo Mixtec*. Berkeley: University of California Press.

Malchukov, Andrej & Akio Ogawa. 2011. Towards a typology of impersonal constructions: A semantic map approach. In Andrej Malchukov & Anna Siewierska (eds.), *Impersonal constructions: A cross-linguistic perspective*, 19–56. Amsterdam: John Benjamins.

Marten, Lutz & Nancy C. Kula. 2012. Object marking and morphosyntactic variation in Bantu. *Southern African Linguistics and Applied Language Studies* 30(2). 237–253.

Matić, Dejan & Daniel Wedgwood. 2013. The meanings of focus: The significance of an interpretation-based category in cross-linguistic analysis. *Journal of Linguistics* 49(1). 127–163.

McConvell, Patrick. 1996. The functions of split-wackernagel clitic systems: pronominal clitics in the Ngumpin languages. In Aaron L. Halpern & Arnold M. Zwicky (eds.), *Approaching Second: second position clitics and related phenomena*, 299–332. Stanford, CA: CSLI Publications.

Mendisu, Binyam Sisay. 2010. *Aspects of Koorete verb morphology*. Cologne: Rüdiger Köppe.

Mereu, Lunella. 1999. Agreement, Pronominalization and Word Order in Pragmatically-Oriented Languages. In *Boundaries of Morphology and Syntax*, 231. Amsterdam: John Benjamins.

Miller, Philip & Paola Monachesi. 2003. Les pronoms clitiques dans les langues romanes. In D. Godard (ed.), *Langues romanes, problèmes de la phrase simple*, 67–123. Editions du CNRS.

Mithun, Marianne. 1992. Is Basic Word Order Universal? In Doris L. Payne (ed.), *Pragmatics of Word Order Flexibility*, 15–61. Amsterdam: Benjamins.

Morimoto, Yukiko. 2000. *Discourse configurationality in Bantu morphosyntax*: Stanford University PhD dissertation.

Morimoto, Yukiko. 2002. Prominence Mismatches and Differential Object Marking in Bantu. In Miriam Butt & Tracy Holloway King (eds.), *The Proceedings of the LFG '02 Conference*, Stanford, CA: CSLI Publications.

Morimoto, Yukiko. 2006. Agreement properties and word order in comparative Bantu. *ZAS Papers in Linguistics* 43. 161–187.

Mushin, Ilana. 2006. Motivations for second position: evidence from North-Central Australia. *Linguistic Typology* 10(3). 287–326.

Mushin, Ilana & Jane Simpson. 2008. Free to Bound to Free? Interactions between Pragmatics and Syntax in the Development of Australian Pronominal Systems. *Language* 84(3). 566–596.

Muxí, Isabel. 1996. Optional participial agreement with direct object clitics in Catalan. *Catalan working papers in linguistics* 5(1). 127–145.

Nabirye, Minah. 2016. *A corpus-based grammar of Lusoga*: Ghent University PhD dissertation.

Næss, Åshild. 2007. *Prototypical Transitivity*. Amsterdam: John Benjamins.

Namyalo, Saudah, Alena Witzlack-Makarevich, Anatole Kiriggwajjo, Amos Atuhairwe, Zarina Molochieva, Ruth Mukama & Margaret Zellers. 2021. *A Ruruuli-Lunyala–English dictionary and grammar sketch*. Berlin: Language Science Press.

Ndayiragije, Juvénal. 1999. Checking Economy. *Linguistic Inquiry* 30(3). 399–444.

Neukom, Lukas. 2001. *Santali*. Munich: Lincom Europa.

Ngoboka, Jean Paul & Jochen Zeller. 2017. The conjoint/disjoint alternation in Kinyarwanda. In Jenneke Wal & Larry M. Hyman (eds.), *The Conjoint/Disjoint Alternation in Bantu*, 350–389. Berlin: De Gruyter Mouton.

Ngonyani, Deo & Peter Githinji. 2006. The asymmetric nature of Bantu applicative constructions. *Lingua* 116(1). 31–63.

Nikolaeva, Irina. 1999. *Ostyak*. Munich: Lincom Europa.

Okrand, Marc. 1977. *Mutsun Grammar*. Berkeley: University of California at Berkeley PhD dissertation.

Osada, Toshiki. 2008. Mundari. In Gregory D. S. Anderson & Norman H. Zide (eds.), *The Munda languages*, 99–164. Routledge London & New York.

Ouali, Hamid. 2011. *Agreement, pronominal clitics and negation in Tamazight Berber: A unified analysis*. London: A&C Black.

Ouhalla, Jamal. 1993. Subject-extraction, negation and the anti-agreement effect. *Natural Language and Linguistic Theory* 11(3). 477–518.

Ozerov, Pavel. 2018. Tracing the sources of Information Structure: Towards the study of interactional management of information. *Journal of Pragmatics* 138. 77–97.

Ozerov, Pavel. 2021. Multifactorial Information Management (MIM): summing up the emerging alternative to Information Structure. *Linguistics Vanguard* 7(1). 1–17.

Patnaik, Manideepa. 2008. Juang. In Gregory D. S. Anderson & Norman H. Zide (eds.), *The Munda languages* Routledge language family series, 508–556. London: Routledge.

Peterson, John. 2011. *A grammar of Kharia*. Leiden: Brill.

Petzell, M. 2008. *The Kagulu Language of Tanzania: Grammar, Texts and Vocabulary*. Cologne: Rüdiger Köppe.

Pinnow, Heinz-Jürgen. 1966. A comparative study of the verb in the Munda languages. *Studies in comparative Austroasiatic linguistics* 5. 96–193.

Preminger, Omer. 2009. Breaking agreements: Distinguishing agreement and clitic doubling by their failures. *Linguistic Inquiry* 40(4). 619–666.

Press, Margaret L. 1979. *Chemehuevi: A Grammar and Lexicon*. Berkeley: University of California Press.

Priestley, Carol. 2008. *A grammar of Koromu (Kesawai), a Trans New Guinea language of Papua New Guinea*. Canberra: Australian National University PhD dissertation.

R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing Vienna, Austria. `https://www.R-project.org/`.

Rajan, Jamuna & Herold Rajan. 2001. *Grammar write-up of Gutob-Gadaba*. Lamptaput: Asha Kiran Society.

Ranero, Rodrigo. 2019. Deriving an object dislocation asymmetry in Luganda. In Emily Clem, Peter Jenks & Hannah Sande (eds.), *Theory and description in African Linguistics: Selected papers from the 47th Annual Conference on African Linguistic*, 595–622. Berlin: Language Science Press.

Reh, Mechthild. 1996. *Anywa Language: Description and Internal Reconstructions*. Cologne: Rüdiger Köppe.

Rezaee, Abbas Ali & Seyyed Ehsan Golparvar. 2017. Conditional inference tree modelling of competing motivators of the positioning of concessive clauses:

The case of a non-native corpus. *Journal of Quantitative Linguistics* 24. 89–106.

Riedel, Kristina. 2009. *The syntax of object marking in Sambaa: A comparative Bantu perspective*: Universiteit Leiden PhD dissertation.

Saeed, John I. 1984. *The syntax of focus and topic in Somali* Kuschitische Sprachstudien. Hamburg: Helmut Buske.

Schikowski, Robert. 2013. *Object-conditioned differential marking in Chintang and Nepali*: University of Zürich PhD dissertation.

Schneider-Blum, Gertrud. 2007. *A grammar of Alaaba, a Highland East Cushitic language of Ethiopia*. Cologne: Rüdiger Köppe.

Schnell, Stefan. 2018. Whence subject-verb agreement? Investigating the role of topicality, accessibility, and frequency in Vera'a texts. *Linguistics* 56(4). 735–780.

Schnell, Stefan, Geoffrey Haig, Nils N. Schiborr & Maria Vollmer. 2020. Introducing new referents: A corpus-based cross-linguistic perspective. *Paper presented at the 53rd Annual Meeting of the Societas Linguistica Europaea, Bucharest* .

Schultze-Berndt, Eva. 2018. Universal vs. language-specific influences on agent prominence and differential agent marking: a view from Down Under. Paper presented at Second International Conference "Prominence in Language", Cologne.

Sedighi, Anousha. 2010. *Agreement restrictions in Persian*. Leiden: Leiden University Press.

Seidl, Amanda & Alexis Dimitriadis. 1997. The discourse function of object marking in Swahili. *CLS* 33. 373–389.

Sidwell, Paul. 2015. *The Palaungic languages: Classification, reconstruction and comparative lexicon*. Lincom.

Siewierska, Anna. 1997. The formal realization of case and agreement marking: a functional perspective. In Anne-Marie Simon-Vandenbergen, Kristin Davidse & Dirk Noël (eds.), *Reconnecting Language: Morphology and Syntax in Functonal Perspectives*, 181–212. Amsterdam: John Benjamins.

Siewierska, Anna. 1999. From anaphoric pronoun to grammatical agreement marker: why objects don't make it. *Folia linguistica* 33(1-2). 225–251.

Siewierska, Anna. 2004. *Person*. Cambridge: Cambridge University Press.

Siewierska, Anna. 2013. Alignment of Verbal Person Marking. In Matthew S. Dryer & Martin Haspelmath (eds.), *The World Atlas of Language Structures Online*, Leipzig: Max Planck Institute for Evolutionary Anthropology. https://wals.info/chapter/100.

Siewierska, Anna & Dik Bakker. 2008. Case and alternative strategies. In Andrej L. Malchukov & Andrew Spencer (eds.), *The Oxford handbook of case*, 290–303. Oxford: Oxford University Press.

Sikuku, Justine M., Michael Diercks & Michael R. Marlo. 2018. Pragmatic effects of clitic doubling: Two kinds of object markers in Lubukusu. *Linguistic Variation* 18(2). 359–429.

Skopeteas, Stavros & Gisbert Fanselow. 2010. Focus types and argument asymmetries: a cross-linguistic study in language production. In Carsten Breul & Edward Göbbel (eds.), *Comparative and contrastive studies of information structure*, 169–197. Amsterdam: John Benjamins.

Smith, Ian & Steve Johnson. 1985. The syntax of clitic cross-referencing pronouns in Kugu Nganhcara. *Anthropological linguistics* 27. 102–111.

Souag, Lameen. 2014. The development of dative agreement in Berber: Beyond nominal hierarchies. *Transactions of the Philological Society* 113(2). 232–248.

Souag, Lameen. 2017. Clitic doubling and language contact in Arabic. *Zeitschrift für Arabische Linguistik* 66. 45–70.

Spagnol, Michael. 2011. *A tale of two morphologies: Verb structure and argument alternations in Maltese*: University of Konstanz PhD dissertation.

Stolz, Thomas. 2011. Maltese. In Bernd Kortmann & Johan van der Auwera (eds.), *The Languages and Linguistics of Europe: A Comprehensive Guide*, 241–256. Berlin: De Gruyter Mouton.

Ström, Eva-Marie. 2013. *The Ndengeleko language of Tanzania*: Göteborgs Universitet PhD dissertation.

Sutcliffe, Edmund Felix. 1936. *A grammar of the Maltese language: With chrestomathy and vocabulary*. Oxford University Press.

Swain, Rajashree. 1997. *A Grammar of Bonda Language*: Poona: Deccan College PhD dissertation.

Tagliamonte, Sali A & R Harald Baayen. 2012. Models, forests, and trees of York English: Was/were variation as a case study for statistical practice.

Language variation and change 24(2). 135–178.

Taylor, Charles. 1985. *Nkore-Kiga*. London: Croom Helm.

Tognini-Bonelli, Elena. 2001. *Corpus linguistics at work*. Amsterdam: John Benjamins.

Tomić, Olga Mišeska. 2008. Towards grammaticalization of clitic doubling: Clitic doubling in Macedonian and neighbouring languages. In Dalina Kallulli & Liliane Tasmowski (eds.), *Clitic doubling in the Balkan languages*, 65–88. Amsterdam: John Benjamins.

Tosco, Mauro. 2002. A whole lotta focusin'goin'on information packaging in Somali texts. *Studies in African Linguistics* 31(1 & 2). 28–53.

Uganda Bureau of Statistics. 2016. *The national population and housing census 2014 – main report*. Kampala: Bureau of Statistics.

Van Valin, Robert D. Jr. 2005. *Exploring the Syntax-Semantics Interface*. Cambridge: Cambridge University Press.

Van de Velde, Mark. 2019. Nominal morphology and syntax. In Mark Van de Velde, Koen Bostoen, Derek Nurse & Gérard Philippson (eds.), *The Bantu languages*, 237–269. London and New York: Routledge.

Vella, Alexandra. 2009. On Maltese prosody. In Bernard Comrie, Ray Fabri, Elizabeth Hume, Manwel Mifsud, Thomas Stolz & Martine Vanhove (eds.), *Introducing Maltese Linguistics*, 47–68. Amsterdam: John Benjamins.

Virtanen, Susanna. 2014. Pragmatic direct object marking in Eastern Mansi. *Linguistics* 52(2). 391–413.

Virtanen, Susanna. 2015. *Transitivity in Eastern Mansi: An information structural approach*: University of Helsinki PhD dissertation.

von Heusinger, Klaus & Petra B. Schumacher. 2019. Discourse prominence: Definition and application. *Journal of Pragmatics* 154. 117–127. https://www.sciencedirect.com/science/article/pii/S0378216619305776.

Voß, Judith. 2015. Person markers in Gutob. *Journal of South Asian Languages and Linguistics* 2(2). 215–240.

Voß, Judith. 2018. Documentation and Grammar of Gutob (Munda). Endangered Languages Archive. http://hdl.handle.net/2196/00-0000-0000-000F-CB59-B.

van der Wal, Jenneke. 2009. *Word order and information structure in Makhuwa-Enahara*: University of Leiden PhD dissertation.

Wistrand Robinson, Lila & James Armagost. 1990. *Comanche dictionary and grammar*. Arlington: Summer Institute of Linguistics and The University of Texas at Arlington.

Witzlack-Makarevich, Alena. 2011. *Typological variations in grammatical relations*: University of Leipzig PhD dissertation.

Witzlack-Makarevich, Alena. 2019. Argument selectors: a new perspective on grammatical relations: an introduction. In Alena Witzlack-Makarevich & Balthasar Bickel (eds.), *Argument selectors: a new perspective on grammatical relations*, 1–38. Amsterdam: John Benjamins.

Witzlack-Makarevich, Alena, Saudah Namyalo, Anatol Kiriggwajjo, Zarina Molochieva & Amos Atuhairwe. 2019. *A corpus of spoken Ruuli*. Kampala and Jerusalem: Makerere University and Hebrew University of Jerusalem.

Witzlack-Makarevich, Alena & Ilja A. Seržant. 2018. Differential argument marking: Patterns of variation. In Ilja A. Seržant & Alena Witzlack-Makarevich (eds.), *Diachronie of Differential Argument Marking*, 1–40. Language Science Press.

Witzlack-Makarevich, Alena, Taras Zakharko, Lennart Bierkandt, Fernando Zúñiga & Balthasar Bickel. 2016. Decomposing hierarchical alignment: co-arguments as conditions on alignment. *Linguistics 54(3)* 531–561.

Zeller, Jochen. 2008. The subject marker in Bantu as an antifocus marker. *Stellenbosch Papers in Linguistics* 38(0). 221–254.

Zeller, Jochen. 2015. Argument prominence and agreement: explaining an unexpected object asymmetry in Zulu. *Lingua* 156. 17–39.

Zellers, Margaret, Saudah Namyalo & Alena Witzlack-Makarevich. 2020. Investigating relationships between intonational and syntactic phrasing in Ruuli/Lunyala. *ISCA* 394–398.

Zerbian, Sabine. 2006. *Expression of information structure in the Bantu language Northern Sotho*: Humboldt-Universität zu Berlin PhD dissertation.

Zide, Norman H. 1997. Gutob pronominal clitics and related phenomena elsewhere in Gutob-Remo-Gta. *Languages of tribal and indigenous peoples of India: The ethnic space* 307–334.

Zwicky, Arnold M & Geoffrey K Pullum. 1983. Phonology in syntax: The Somali optional agreement rule. *Natural Language & Linguistic Theory* 1(3). 385–402.

Nederlandse samenvatting

Deze dissertatie richt zich op differentiële indexering (dat wil zeggen: variatie in gebonden persoonsmarkering in het werkwoord binnen één taal) en de referentiële en discourse-structurele factoren die dit veroorzaken. Ze bestaat uit vier artikelen: drie gedetailleerde casestudy's over het Ruuli (Bantoe), het Maltees (Semitisch) en het Gutob (Moenda), gevolgd door een typologische discussie over het fenomeen zelf. De term indexering wordt hier zonder enige theoretische lading gebruikt: zij vooronderstelt geen enkele syntactische relatie tussen de marker en de referentiële NP. Ook doet zij geen uitspraken over de morfologische status van de index als een cliticum of een affix. Dit maakt het mogelijk om een aantal gevallen van differentiële indexering te onderzoeken die in de wetenschappelijke literatuur als ongelijksoortig bestempeld zijn, zoals 'clitic doubling' of 'optional agreement'.

In het verleden zijn er zowel taalfamiliespecifieke als typologische studies aan differentiële indexering van P (het minder agentieve argument van een tweeplaatsig predicaat) gewijd. Differentiële indexering van A (het agentievere argument) heeft daarentegen minder aandacht genoten, en de indexeringen van beide argumenten worden over het algemeen als verschillende fenomenen beschouwd, met name wat betreft de mate waarin ze verplicht zijn. Differentiële P-indexering behelst vaak de aanwezigheid van een index voor een referent die in onverwacht sterk mate bezield of identificeerbaar is of als topic beschikbaar is. Differentiële A-indexering omvat juist vaak de omissie van een normaliter aanwezige index voor referenten die deze eigenschappen ontberen of in focuspositie staan.

Hoewel de onderliggende factoren voor differentiële indexering verschillende talen veelal dezelfde zijn, moet de precieze manifestatie van deze factoren voor elke taal afzonderlijk onderzocht worden. Talen verschillen namelijk niet alleen wat betreft de factoren zelf, maar ook wat betreft de vraag of er in de daaraan verbonden hiërarchieën grenzen getrokken worden, en zo ja: waar precies. Ook de mate waarin meerdere factoren van elkaar afhankelijk zijn of elkaar beïnvloeden moet per taal worden vastgesteld.

Deze complexiteit binnen één gegeven taal komt in deze dissertatie goed naar voren in twee corpusgedreven kwantitatieve studies over P-indexering in het Ruuli en het Maltees. In de analyses komt naar voren dat differentiële P-indexering in zowel het Ruuli als het Maltees sterk verbonden is met de constituentvolgorde, hoewel geen van beide talen een absolute correlatie laat zien. Naast constituentvolgorde zijn in beide talen ook givenness, identificeerbaarheid en de woordsoort van het syntactische hoofd significante predictoren, maar de relatie tussen deze variabelen verschilt. Deze vondsten bekrachtigen eerdere oordelen die op basis van intuïtie gemaakt zijn en onderstrepen de hogere-orde-intergerelateerdheid van verschillende factoren.

In Gutob gaat differentiële indexering niet zozeer om dier aan- of afwezigheid als wel om de wisselende plaatsing: indexen kunnen aan het predicaat en enig ander constituent in de (deel)zin bevestigd worden zonder een vaste syntactische positie in te nemen. Eerdere behandelingen schreven indexplaatsing in het Gutob toe aan buitengewone retorische voorwaarden en namen het werkwoord als de plek waar een index standaard aan wordt bevestigd. De casestudy hier toont op basis van systematische corpusannotatie aan dat deze claim geen water houdt in gevallen waarin er syntactische alternatieven voor plaatsing bij het werkwoord zijn. In de meeste gevallen zoeken de indexen namelijk een andere gastheer. Systematische zoekopdrachten in het geannoteerde corpus heeft het bovendien mogelijk gemaakt om diverse voorbeelden van niet-verbale en verbale indexen in hun eigen context met elkaar te vergelijken. Hiermee is het effect van indexplaatsing op de gastheerconstituent bepaald en aangetoond hoe indexplaatsing een actieve rol speelt in aandachtsmanagement.

Het vierde paper biedt een overzicht van differentiële A- en P-indexering (specifiek de aan- of afwezigheid van een index) in diverse talen, waarbij met name gelet wordt wordt op structurele en functionele overeenkomsten en verschillen. Dit paper toont aan dat aandacht voor differentiële index-

ering ons kan helpen begrijpen hoe indexering in het algemeen verbonden is met de prominentie van een referent (namelijk: zijn of haar mate van givenness, identificeerbaarheid of bezieldheid) onafhankelijk van de rol die de referent toegewezen krijgt. Referenten kunnen geïndexeerd worden als zij een bepaald prominentieniveau (blijven) bezitten. Als een referent geen taalspecifieke prominentielimiet overschrijdt of deze kwijtraakt, wordt hij of zij niet geïndexeerd.

Deze casestudy's en de taaltypologische beschrijving tonen samen aan dat indexering in veel talen niet zomaar gelijk staat aan roltoekenning, maar dat communicatieve behoeften sprekers ertoe brengen om te indexeren en hierin keuzes maken op basis van semantische en discourse-structurele overwegingen. Deze dissertatie onderstreept ook dat functioneel-georiënteerde, corpusgedreven beschrijvingen een waardevolle toevoeging vormen op studies die op basis van grammatica's geschreven worden, met name voor talen die slechts oppervlakkig geanalyseerd zijn. Ze onderbouwen beschrijvend onderzoek en bieden toegang tot de probabilistiek achter taalkundige structuren.

Curriculum Vitae

Erika Just (née Weinberger) was born in Cham (Eastern Bavaria, Germany) on October 15th 1990. She was raised in the same district and attended the Benedikt-Stattler-Gymnasium in Bad Kötzting, finishing in 2010 with main exams in English and French.

She began her studies in General Linguistics and English Studies at Kiel University in 2011, obtaining her BA with a thesis on a typological topic in 2014. She continued her studies at the University of Bremen and graduated with an MA in Language Sciences with a thesis on Maltese L1-acquisition in 2016. In the same year, she started working as a research associate back at the Department of Linguistics at Kiel University and began her dissertation project in October 2017. During the first year of her Ph.D. she received a Federal State Funding at Kiel University. In 2018 she did some field work in Uganda, supporting the compilation of the corpus and the dictionary during the Ruruuli-Lunyala documentation project funded by the Volkswagen Foundation. Parts of this corpus were used as the data basis for the case study in Chapter 3 of this dissertation. She joined the doctorate program at Leiden University as an external Ph.D. researcher in January 2020.