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

Just, E.C.

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 nilh'ën.*
       man moose look.at.it.3SG

       'The man is looking at a moose.'

b. *Hida   dinï yi-niɬh'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.