Alignment everywhere all at once: Applying the late aggregation principle to a typological database of argument marking

David Inman[1*], *Alena Witzlack-Makarevich*[2*], *Natalia Chousou-Polydouri*[1], *and Melvin Steiger*[1] *
[1] University of Zurich & Center for the Interdisciplinary Study of Language Evolution
[2] Hebrew University of Jerusalem

ABSTRACT

This article presents the structure of the ATLAs Alignment Module, a typological database designed to exhaustively capture language-internal variation in argument marking (indexing and flagging). The flexible design of our database can be extended to cover further aspects of morphosyntactic alignment. We demonstrate with a small diversity sample how the database can be queried and the data aggregated at different levels of structure (e.g. for a language as a whole or for individual referential types in the form of alignment statements) for the purposes of cross-linguistic comparison. The database is made available in the Cross-Linguistic Data Formats (CLDF), and we provide code that generates an array of aggregations.

Keywords: alignment, typology, database, aggregation, morphology

INTRODUCTION 1

Alignment of argument marking is one of the major morphosyntactic characteristics of languages both in the descriptions of individual languages as well as in comparative studies and typological databases.

*These authors have contributed equally to this work.

All major typological databases, such as WALS (Dryer and Haspelmath 2013) and Grambank (Skirgård *et al.* 2023), include several alignment-related features. Furthermore, a dedicated database tracks the emergence of alignment patterns (Cristofaro *et al.* 2021).[1] Typological work on morphosyntactic alignment (including the aforementioned databases) typically captures only high-level generalizations about alignment at the level of the entire language, e.g. the presence of ergativity in the system of case marking or traces of hierarchical effects on the agreement system. However, many languages have multiple alignments conditioned e.g. by the referential properties of arguments, the tense of the clause, and so on (see e.g. Bickel *et al.* 2015b). Thus, in contrast to some other typological features (e.g. presence of a nominal dual number), alignment is not a typological variable for which there is only one way to make a statement about a language as a whole. Instead, it is a complex and multi-variate component of grammar for which similarities and differences between languages can be established along many different dimensions.

In this article we present the ATLAs Alignment Module (Inman *et al.* in prep), a typological database of morphosyntactic alignment designed to capture existing variation in alignment patterns of a language. By encoding multiple aspects and patterns of alignment within a language all at once, we will show that it is possible to aggregate alignment information at differing levels of structure: for the language as a whole, for individual argument selectors (e.g. nominative case, plural argument marker), for individual referential types (e.g. 1sg, 2pl, masculine nouns), and for argument roles (S, A, and P).

We will begin with an overview of the phenomenon of alignment (Section 2) and discuss how data that describe the phenomenon can be captured in typological databases (Section 3). We will then describe the choices we made for data collection and database design (Section 4) and demonstrate how the data we collected can be used to derive a variety of typological properties (Section 5). Finally, we will offer some concluding remarks and discuss the ways in which our work can

[1] Alignment-related information is also captured in databases dedicated to valency patters (Hartmann *et al.* 2013, Say 2020–). Note, however, that these databases focus on predicate-level details and variations of predicate-specific coding frames.

be extended to answer more questions about alignment (Section 6). Supplementary Materials, including the full database and all code, are available at `https://osf.io/n67mq/`.

2 MORPHOSYNTACTIC ALIGNMENT

The study of morphosyntactic alignment is intimately linked with the broader phenomenon of *grammatical relations*. This label traditionally refers to the relations between a clause or a predicate and its arguments. Some of the most common grammatical relations are subject and object, which are among the basic concepts of many theoretical frameworks. However, starting from the mid-1970s, descriptive linguists and typologists have reported challenges in identifying such traditional grammatical relations in individual languages and in applying them consistently in typological studies (see in particular the collection of papers in Li and Thompson 1976, LaPolla 1993, and Dryer 1996, 1997).

Most typologically informed research adopts a language-specific and construction-specific view of grammatical relations (cf. Comrie 1978; Moravcsik 1978; Van Valin 1981, 1983, 2005; Croft 2001; Bickel 2011; Witzlack-Makarevich 2011, 2019). In this approach, researchers forego assumptions about the universality of grammatical relations, such as subject and object. Instead, they use more robust cross-linguistic concepts as a point of comparison for the relevant morphosyntactic properties of arguments or constructions.[2] In what follows, we first provide an overview of these concepts.

[2] A classic early example of the objectors of this approach is Anderson (1976), who argues that the switch reference construction is the only right way to determine what a subject is in the language Kâte [kate1253] (Nuclear Trans New Guinea; Papua New Guinea), which has ergative flagging and accusative indexing (see Section 2.2). This is a case of prioritizing the identification of a specific grammatical relation (in this case, "subject") over considering all relevant morphosyntactic facts of the language.

2.1 *Arguments*

A common way to capture how arguments of a clause are treated by various morphosyntactic constructions in individual languages is to ask which arguments are marked or behave in the same way. This identity of marking or behavior of certain arguments is what is understood as morphosyntactic alignment. Consider the case marking of the noun 'man' in the Chechen example in (1) and its English translation.

(1) Chechen [chec1245] (Nakh-Daghestanian; Russia; Zarina Molochieva p.c.)

a. **stag** valla.
man die.PRF
'The man died.'

b. **stag-as** xudar de'a-na.
man-ERG porridge eat-PRF
'The man has eaten porridge.'

c. ʒʕala-s **stag** qieri-na.
dog-ERG man frighten-PRF
'The dog frightened the man.'

Whereas the arguments the dead man in (1a), the eating man in (1b) and the frightened man (1c) in the English translation do not have any overt case marking (it is just 'the man'), the Chechen examples have two types of argument marking: the dead man and the frightened man are not marked in any visible way, while the eating man has the dedicated case suffix *-(a)s*, which linguists commonly refer to as *ergative case suffix*. If you translate these sentences into a language which has a special accusative case, the overall picture will be quite different: the frightened man will be marked in a special way and thus differently from the dead man and the eating man, which would be in the (unmarked) nominative case.

This marking is not a special property of the word 'man' and the verbs included in these examples. Instead, it is a pattern found with other nominal and pronominal arguments and other verbs across the language, so we need a way to generalize across arguments of different predicates. As we will outline in this section, we understand arguments

as a composite category made of both *argument role* and *referential properties*. We will first outline how we define argument roles.

The most common argument roles used for the purposes of alignment typology and in descriptive accounts are S, A, and P (or O in some sources).[3] Note, however, that what exactly is understood by these labels varies somewhat between authors (see Haspelmath 2011). We use these terms in the sense of *generalized semantic argument roles* (as opposed to a semantic-syntactic or purely syntactic understanding). A generalized semantic argument role (henceforth *argument role* or just *role*) is an abstraction over *predicate-specific roles* (or *microroles*, as e.g. in Hartmann *et al.* 2013). For example, the verb *hit* has two predicate-specific roles, a HITTER and a HITTEE, the verb *kiss* has a KISSER and a KISSEE, *see* has a SEER and a SEEN, and so on. In the case of the role A, it abstracts over the predicate-specific roles of HITTER, KISSER, and SEER, according to semantic criteria we list below.

Argument roles are first distinguished according to the numerical valency of their predicates: the sole argument of one-argument predicates vs. the two arguments of two-argument predicates. In the case of the sole argument of one-argument (monovalent) predicates, there is no need to distinguish it from anything else; this argument is abbreviated as S, independent of its finer semantic differences. In the case of two-argument (bivalent) predicates, arguments are distinguished on the basis of cross-linguistically viable lexical entailment properties (as in Witzlack-Makarevich 2011, 2019, following Dowty 1991 and Primus 1999, 2006).

Each argument of a bivalent verb accumulates various lexical entailment properties, given in (2). The argument that accumulates more lexical entailments than the other argument of the same verb is the A role, and the other is the P role.

(2) a. causing an event (e.g. A hits P, A kisses P, A goes to P)

[3] The alignment of other argument types, in particular, of the arguments of trivalent or ditransitive verbs, is another common research topic, see e.g. the collections of papers in Malchukov *et al.* 2010b. Due to the project scope, we do not treat any other argument roles apart from S, A, and P. However, the framework presented in Section 4 is equipped and sufficiently flexible to incorporate other domains of alignment, including the alignment of ditransitive verbs.

b. volitional (e.g. A hits P, A kisses P)
c. sentient (e.g. A sees P, A looks at P, A loves P, P pleases A)
d. independently existing (e.g. A bakes P, A makes P)
e. possessing another participant (e.g. A has P, P belongs to A)

For instance, in *Lisa kisses Mario*, *Lisa* is causing the event, she is volitional and sentient, and she exists independently. On the other hand, *Mario* only is sentient and exists independently in this event. Thus, *Lisa* accumulates more of the relevant properties than *Mario* and is classified as A. The remaining argument (*Mario*) is P. Thus, every two-argument predicate will have one argument which can be labelled as A and one which can be labelled as P, following the list of lexical entailments in (2). Note that this labeling process is determined entirely by semantics: there is no reference here to syntactic structure or morphological marking.

With this cross-linguistically applicable set of argument roles, it is possible to calculate alignments by comparing the marking or the behavior of different roles. The five logically possible alignment types are listed in (3). We will refer to them as *basic alignment types*.

(3) Basic alignment types
a. S = A ≠ P corresponds to the (nominative-)accusative alignment pattern (S and A are marked or behave identically but differently from P);
b. S = P ≠ A corresponds to the ergative(-absolutive) alignment pattern;
c. S = A = P corresponds to the neutral alignment pattern;
d. S ≠ A ≠ P corresponds to the tripartite alignment pattern;
e. A = P ≠ S corresponds to the horizontal alignment pattern.

These five basic alignment patterns figure prominently in many typological studies, both dedicated to alignment specifically (e.g. Comrie 2013a,b; Siewierska 2013a) and in large-scale studies of genealogical, geographic, and universal determinants of linguistic patterning (e.g. Nichols 1992). The list in (3) is often expanded with further non-basic alignment types meant to capture specific patterns of

argument marking. For example, Siewierska (2013a) adds active, hierarchical, and split alignment to the list of possible values.

As we have noted at the beginning of this section, arguments have a composite structure in the approach we adopt (see Bickel 2011): In addition to the argument roles, various referential properties of arguments (person, number, definiteness, topicality, specificity, animacy, and also part of speech) can determine the argument's marking by indexing or flagging and thus have an immediate effect on alignment (as demonstrated in Section 2.3).

Argument selectors 2.2

There are two major ways in which some arguments can be treated identically by a language's grammar: via patterns of morphological marking (also called *coding*, or just *marking*) and via patterns of (syntactic) behavior. Coding traditionally encompasses different loci of morphological marking, both case marking on the noun phrase and indexing on the verb (or in clausal inflection), as well as word order (Keenan 1976). We will refer to all ways in which a language groups arguments, either syntactically or morphologically, as *argument selectors*, and will furthermore focus on morphological marking, leaving aside word order. Cross-linguistically, by far the most common argument selectors, as well as the best studied ones, are *flagging* and *indexing*.[4]

We use the term *flagging*, following Malchukov *et al.* (2010a, 8), as a cover term for both morphological case and adposition marking, both of which mark a role within the syntactic domain of a noun phrase. We use the term *indexing* to refer to the marking of verbal agreement or argument cross-referencing on the clause as a whole (again, following Malchukov *et al.*). The present study only concerns the argument selectors of flagging and indexing.

[4] The set of syntactic (or *behavioral*) argument selectors is large and diverse. It includes such syntactic properties as the promotion and demotion of arguments by passivization or antipassivization, the possible relativization site(s) in a relative construction, the possibility to function as either controller or controllee in various control constructions, and conjunction reduction (the interpretation of gapped arguments in coordinated clauses). See Witzlack-Makarevich 2019 for examples and further references.

2.3 *Language-internal variation in argument selection*

The generalized semantic argument roles S, A, and P, and argument selectors (for our purposes, only flagging and indexing) are not sufficient to capture language-internal variation of alignment patterns. Argument selection can vary in two primary ways: by the referential and part-of-speech properties of arguments and by various clause-level conditions.

An example of relatively straightforward variation by referential and part-of-speech information can be seen in English flagging. Some pronouns have a special P form different from the corresponding S and A form (e.g. *me* vs. *I* and *him* vs. *he*), while other pronouns have a single form for all roles (e.g. *you*, *it*). There is no such variation for nominal arguments: they never differentiate between A and P roles (e.g. I_A *kiss* $Lisa_P$ and $Lisa_A$ *kisses* me_P). Capturing this variation requires referencing both the person-number and the part-of-speech properties of arguments.

In addition to argument properties, a number of clausal properties are known to condition language-internal variation in argument selection. The best-known such factors are listed in (4).[5]

(4) a. tense-aspect-mood (TAM) features
b. the nature of the clause (main clause vs. various types of subordinate clauses)
c. polarity
d. scenario (co-presence of particular types of arguments in the clause)

As an example, consider the flagging of P in Aguaruna in (5) (for some generalizations, see Overall 2017). The P argument 'chicken' is

[5] Most of these conditions are long-established in the literature (see Dixon 1994; Bickel 2011) and have been investigated under a variety of labels, including *split alignment* (Silverstein 1976), *differential marking* (Comrie 1989), and *differential object marking* or *DOM* (Bossong 1985, 1991; Witzlack-Makarevich and Seržant 2018). The less-familiar condition is scenario (Zúñiga 2006; Witzlack-Makarevich *et al.* 2016), which represents a more expansive analysis of what has historically been called *hierarchical alignment* (Mallinson and Blake 1981; Nichols 1992; Siewierska 1998).

in the accusative case in (5a) and in the nominative case in (5b). This is a case of *differential object marking* (DOM). However, in contrast to the English pronouns discussed above, it is not the referential nature of the P argument that conditions the accusative case. Rather, it is exclusively the nature of the A argument that determines the marking of the P argument: if the A role references the first person singular, as in (5a), or the third person (not illustrated here), the P argument is flagged accusatively; otherwise it is flagged nominatively.

(5) Aguaruna [agua1253] (Chicham; Peru; Overall 2017, 280)

a. atashu=n yu-a-tata-ha-i
chicken=ACC eat-PFV-FUT-1SG-DECL
'I will eat chicken.'

b. atash yu-a-tata-hi
chicken eat-PFV-FUT-1PL
'We will eat chicken.'

In addition to the cross-linguistically recurrent conditions for variation in argument marking listed in (4), individual language descriptions occasionally include rather idiosyncratic specifications. For instance, when describing the distribution of the overt nominative flagging on S and A arguments in Achumawi [achu1247] (Palaihnihan; USA), de Angulo and Freeland (1930, p. 83) write that "subjectivity need not be indicated either, except as clearness demands it". Such situations are recurrent and there is no principled way to compensate for gaps or vagueness in descriptive accounts.

To account for language-internal variation in argument selection, any database of alignment needs a systematic way to capture such patterns of differential argument marking. In the next section we outline the design principles of such a database, using the existing AUTOTYP alignment database (Bickel *et al.* 2022) as the starting point, and demonstrate how this design captures the multivariate nature of alignment systems.

3 DATABASE DESIGN PRINCIPLES

The database presented here is not the first to collect information on alignment. WALS (Dryer and Haspelmath 2013), the first major typological database, has three features/chapters on the topic: Comrie 2013a,b with a sample of 190 languages and Siewierska 2013a with 380 languages. The more recent Grambank database (Skirgård *et al.* 2023) includes information on 2362 languages and has twelve features (GB089–GB094 and GB408–GB410) which capture similar information as WALS in a larger number of binary variables, as well as additional information about the presence of variation in marking (GB095, GB096, and GB098). Finally, Birchall 2014a, a dataset of 95 languages of South America, has a handful of alignment-related features either identical or similar to the ones in Dryer and Haspelmath 2013, as well as several related features focusing on very specific contexts (e.g. ARGEX2-7-1 asks whether verbal person marking for P is variable, obligatory or not realized when the corresponding lexical argument is present in the clause). All these databases essentially classify whole languages or whole language subsystems (e.g. pronouns in Comrie 2013b) as being of a specific alignment type selected from a previously postulated list of possible alignment types.

The database presented here took a design path quite different from the existing databases in several respects. When considered in its entirety, the phenomenon of alignment has many interacting components. We will show that it is advantageous to capture them all at once when collecting data, and to do so in such a way that multiple aggregations can be made over the same database. Our main design principles are an extension of those in AUTOTYP. We now turn to describing those principles and comparing them with those of other alignment databases.[6]

[6] The AUTOTYP database is a large-scale research program with goals in both quantitative and qualitative typology. It was launched in 1996 by Balthasar Bickel and Johanna Nichols and is thus one of the oldest typological databases still in use. AUTOTYP includes a module on grammatical relations and alignment; this has been released as Bickel *et al.* 2022. A variety of follow-up works are based on various aggregations of these data (e.g. Bickel *et al.* 2013, 2014, 2015a,b,c; Witzlack-Makarevich *et al.* 2016).

Perhaps the most common strategy in linguistic typology is to operate with variables which have a closed set of possible values. This set of possible values, either defined entirely beforehand or early on in the coding process, is essentially an etic grid which is used to categorize all individual observations. Such sets can be motivated by tradition (as in the alignment studies by Comrie 2013a,b and Siewierska 2013a), as well as by theoretical considerations or convenience. A major drawback of such pre-defined sets of possible values, especially when they are small, is that they may lack sufficient resolution to capture the full variation present in the data. For instance, the classification of a whole language as showing split alignment of indexing, as in Siewierska 2013a, does not capture what the triggers of such splits are, nor which basic alignment patterns are involved (e.g. Is it neutral and accusative? Ergative and hierarchical? etc.). This philosophy is followed by databases such as WALS (Dryer and Haspelmath 2013) and more recently by Grambank (Skirgård *et al.* 2023). AUTOTYP follows a different set of principles. Among these, the four that are most relevant for this paper are: (1) modularity, (2) autotypology, (3) late aggregation, and (4) use of exemplars (Bickel and Nichols 2002; Witzlack-Makarevich *et al.* 2022).

First, the AUTOTYP database as a whole is built in a *modular* fashion, with each module covering a typological domain. Some modules cover relatively narrow domains with just a few variables (e.g. clusivity), while others include multiple tables and several dozen variables (e.g. clause linkage). The encoding of some linguistic features may be spread across multiple modules (e.g. grammatical relations are spread among the modules on grammatical markers, predicate classes and clause linkage).

The second major design principle of AUTOTYP is *autotypology*. Autotypology means keeping variable values (and even variables themselves) flexible and open during the coding process. That is, there is no closed set of values according to which every language must be categorized. Instead, value sets and even variables can always be adjusted during coding in order to adequately capture the variability of languages. This process characterizes early stages of creation of other typological databases. This represents a radical prioritization of detailed data encoding which transparently maps to statements in reference grammars over encoding variables that match the

researcher's typological questions and previously assumed linguistic types.

The third principle is *late aggregation*. This is the principle of encoding data at a granular (autotypologized) level and only later generalizing over the data to yield cross-linguistically comparable sets of typological properties following a format familiar from conventional typological databases. Since typological categories are in principle not specified at the point of data entry, comparative typological questions are answered by querying an autotypologized database or performing data aggregations (from multiple modules if needed). As a simple example, rather than directly stating that a language has accusative flagging, the database instead lists statements about marking of various nouns and pronouns in S, A, and P roles under various conditions. The presence of accusativity can then be identified algorithmically, that is when nouns that mark S and A roles are marked differently than nouns that mark P roles. One major advantage of late aggregation is that the same data can be used to test different hypotheses and to evaluate the consequences of different operationalizations.

The fourth AUTOTYP principle is the use of *exemplars* for comparative studies, which should be extractable from the underlying data. For methodological or theoretical reasons, in some typological surveys it is desirable to have one data point per language and for these purposes one particular exemplar of a structural domain or a paradigm or a context is selected as representative for the whole domain. In other cases, a particular context or structure may be desirable as a point of comparison, without assumptions about its representativity. The use of exemplars is not unique to AUTOTYP. There are two major differences between AUTOTYP and other databases: the phase at which the exemplar comes into play; and that AUTOTYP allows for multiple exemplars during late aggregation.

4 DATA AND METHODOLOGY

The ATLAs Alignment Module largely follows the design principles of AUTOTYP outlined in Section 3, though these have been modified slightly to accommodate our coding purposes. The dataset used in this

paper is a subsample of the languages that are present in ATLAs (Inman *et al.* in prep), a global database which is focused on North and South American patterns of areality. Modifications to the AUTOTYP principles are presented in Section 4.1, an overview of the database structure is given in Section 4.2, and Section 4.3 discusses the sample and coding procedures.

Database coding 4.1

While we followed the AUTOTYP principles (Section 3) for the most part, we found it practical to depart from them in a few cases. The most significant of these departures has to do with the exhaustivity and scope: for this project, we are interested exclusively in the alignment properties of argument marking in main declarative clauses with verbal predication and with positive polarity. Thus, in a sense, one could argue that due to these limitations of scope there is some collateral violation of the principle of *autotypology*: for any contexts of the phenomenon of argument marking beyond the rather narrow predefined scope we did not expand the set of variables and their values to encode previously unencountered coding patterns. Furthermore, because our sole interest is in the alignment of morphological marking, properties of other grammatical constructions are simply not present in this database.

There are two further cases where for practical reasons we have not followed the principle of autotypology.

First, it is impossible to know in advance all possible variables by which morphological alignment might vary in a sample. The most typical conditions are properties such as TAM and predicate class, and, following the autotypology principle, we have left the possible values of these variables open-ended during coding. However, there are many other possible sites of variation (e.g. word order, the presence or absence of an overt NP, or unknown or insufficiently described conditions). To track these conditions on alignment variation, we have created a single variable called “Miscellaneous conditions” which is used to cover all of these “other” conditions. The set of values that “Miscellaneous conditions” can accommodate is open ended and should in principle be split into separate variables following the

principle of autotypology. However, we have kept this as a single variable since these various conditions are not the primary target of this study.

Second, we included a convenience variable[7] explicitly indicating a highly specific exemplar of flagging and indexing chosen beforehand, instead of computing it after the fact. This adds to rather than detracts from the AUTOTYP way of dealing with exemplars outlined in Section 3, since it only abstracts in a non-algorithmic fashion over information that is already present in other variables. The exemplar we chose in this project is defined by Birchall (2014b, 24–25) (following Lazard 2002, 252). We have adopted and expanded on Birchall's definition and termed it the *exemplar declarative main clause*. This exemplar has the following properties:[8]

(6) Exemplar declarative main clause:
   a. The clause represents a real event (not prospective, not imagined) and is declarative.
   b. The clause is not embedded or a complement of another clause.
   c. The event described in the clause is discrete, perfective or completive, and not ongoing or incomplete.
   d. The clause has positive polarity and is not negated.

Since morphosyntactic alignment is a phenomenon that can vary depending on the characteristics of the arguments, in addition to defining the exemplar clause, the exemplar S, A, and P roles are defined in (7).

[7] By "convenience variable" we mean a variable that is not strictly necessary and does not encode any additional information. As we will outline below, the exemplar variable could in principle be derived algorithmically from the other variables present in the database, although such an algorithm would be cumbersome.

[8] There are consequences to adopting any exemplar. In our case, the definitions in (6) and (7) will preferentially select for accusative alignments, as many languages with split-S marking mark S arguments if they control events the same as A arguments, and thus all these languages will be considered as showing accusative alignment in the exemplar case. We have captured the existence of such systems by making sure to encode monovalent predicate classes where the S lacks control (see the discussion on `Predicate_class` in Appendix A.2).

(7) Exemplar S, A, and P arguments:

a. The S argument is a human that voluntarily performs and controls the event.

b. The A argument is a human that voluntarily performs and controls the event.

c. The P argument is well-individuated, human, and is actually affected by the event.

It is in principle possible to algorithmically derive this exemplar from the TAM, predicate class, and miscellaneous conditions defined for each context. However, because the possible values of these variables are all open-ended, the relevant algorithm would need to include a constantly updated classification of all these conditions (and possibly their interactions) to allow the extraction of only those contexts which represent the exemplar case. Encoding the exemplar in a convenience variable avoids the need to create and continuously update such a list. Though we have encoded this exemplar variable according to the properties defined in Birchall 2014b, this kind of information could be encoded for other exemplars, with each exemplar encoded in a separate convenience variable.

A practical decision was needed as to how and whether to encode the absence of overt marking (or "zero marking"). For nouns and pronouns, we coded contexts for each role S, A, and P, whether they had overt flagging or not. All zero marking in flagging, therefore, is coded explicitly. However, we determined that it was not feasible to do this for indexing. If a language has several slots for indexing, e.g. different slots on the verb for different persons and roles, then there could be many zeros simply indicating that a particular person is absent from a context. In more complex cases, it is unfeasible to code all zeros, or doing so would require making decisions about possibly indeterminate properties (for example, how many slots are present in a certain configuration). There is also a theoretical decision to be made, about whether there is a "true zero" which means something, or if marking is simply absent. This cannot always be determined from available sources.

For the coding of zeros in indexing, we adopted the policy that they need not be explicitly coded, but could be. However, there are some cases where the coding has to be explicit, with a phonologically

zero selector: (1) when a zero is the only reflex of a particular referential type (e.g. 3rd person singular is not marked), or (2) when a zero marker contrasts with an overt marker under certain conditions (e.g. a 3rd person index which is phonologically overt under some conditions and zero under others).[9] However, in other cases, such as the exceptionless absence of indexing for the P role, or the absence of marking in a particular slot in a particular scenario, we allowed for this to be coded explicitly or not, depending on the ease and preference of the coder. This creates a certain level of inconsistency in our database: Sometimes these zeros (both the lack of indexing for a role and the lack of overt marking in a particular case) are present, and sometimes they are not. But in terms of database interpretability, nothing is lost: The absence of explicit information about the indexing of S, A, or P arguments means that there is no overt marking.

4.2 *Database structure*

The ATLAs Alignment Module conforms to the CLDF standard (Forkel *et al.* 2018) and is composed of three basic `csv` files (`contexts.csv`, `selectors.csv`, and `languages.csv`) and the `metadata.json` file that describes how the `csv` files are interrelated. As the CLDF format is customizable and extendable, further information can be added in the form of new columns and even new tables.[10] As Section 5 shows, we add such derived columns and tables as we proceed with querying the database to create data aggregations at different levels (for an overview of the database structure, see Figure 1).

Each of the basic `csv` files are briefly described below in Sections 4.2.1–4.2.3, with an overview of the most important columns

[9] This means that the full list of referential types indexed in a language is always available in `contexts.csv`, unless they behave uniformly in terms of alignment (see Section 4.2.1). In order to perform meaningful aggregations on complex indexing systems (see Section 5), we need a record of all referential types the indexing systems distinguish no matter whether they are overtly marked or not. Thus each referential type must have at least one context indicating its existence. The other possibility would be to have a separate table listing all referential types for all languages.

[10] In the remainder of the paper we use `monospace typeface` for file names and column headers and we enclose variable values in <angle brackets>.

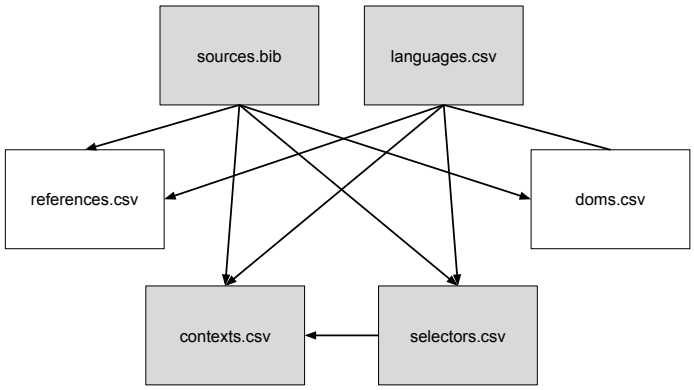

Figure 1: Representation of CLDF database. Basic files with raw data are represented by gray rectangles, additional files populated with scripts by white rectangles. Lines show one-to-one relationships, and arrows one-to-many

and coding decisions for each, along with excerpts from each `csv` file that present the corresponding content for two illustrative languages: Bilua [bilu1245] (isolate; Papua New Guinea and Solomon Islands) and Awa-Cuaiquer [awac1239] (Barbacoan; Colombia and Equador). Bilua is straightforward as far as alignment is concerned: there is no flagging for nouns or pronouns, and S and A roles are indexed with a paradigm of proclitics and P with a paradigm of enclitics (see example (8)). The last three rows in Table 1 represent indexing in Bilua: the same selector <bilua1245-s-a-proclitics-indexing-marker> (corresponding to the proclitic paradigm) is used for both S and A roles and appears in the <clitic -1> slot. The enclitic paradigm for the P role is a separate selector <bilua1245-p-enclitics-indexing-marker> and appears in the <clitic 1> slot.

(8) Bilua [bilu1245] (isolate; Papua New Guinea and Solomon Islands; Obata 2003, 309)[11]

ko=rere=a inio **ko**=pa zuzue=**v**=a
3SG.F=run=PRS SEQ 3SG.F=PROS hug=3SG.M.O=PRS

'She ran and then she hugged him.'

[11] Special glosses for Bilua which extend the Leipzig Glossing Rules (Comrie *et al.* 2008) are: SEQ: sequential coordinator, PROS: prospective marker. The present tense marker in this example is used as historical present tense, and is thus translated using the past tense.

Awa-Cuaiquer on the other hand is more complicated: it has coargument sensitivity as well as both a split-S system and a fluid-S system, where fluidity applies only to S arguments of stative verbs and only matters for the markers of the 1st person. The example (9) below is represented with three different contexts in Table 1:

- in the line with ID <awac1239-5>, the reference is a high noun (humans in Awa-Cuaquier) in the P role, which is marked with the accusative case, irrespective of the A coargument being a noun or a pronoun;
- in the line with ID <awac1239-8>, the reference is a pronoun in the A role, which is unmarked, irrespective of the P coargument being a noun or a pronoun;
- in the line with ID <awac1239-15>, the reference is a non-locutor (in example 9, a second person) in the A role with another non-locutor[12] (in the example, a third person) in the P role. The A argument is indexed on the verb with the suffix *-zi*, which is specific to contexts where no locutor (first person) is involved.

(9) Awa-Cuaiquer [awac1239] (Barbacoan; Colombia and Equador; Curnow 1997, 199)

nu=na Juan=**ta** pyan-ti-**zi**
2SG.(NOM)=TOP Juan=ACC hit-PST-NONLOCUT

'You hit Juan.'

For a more detailed description and explanation of all the values for each column, see the Appendices.

4.2.1 contexts.csv

In the `contexts.csv` table (see Table 1),[13] each row represents a context involving either one argument (S in the case of monovalent verbs) or two arguments (A and P in the case of bivalent verbs), and exactly

[12] In Awa-Cuaquier, indexing distinguishes only 1st person (locutor) from 2nd/3rd person (non-locutor).

[13] In Tables 1–3, Table 5, and Table 7, some columns have been omitted for readability.

Table 1: Excerpt from `contexts.csv`

ID	Selector_ID	Slot	Role	Reference	Co-argument_ role	Co-argument_ reference	Exemplar	Predicate_ class	Miscellaneous_ condition
awac1239-1	awac1239-no-flagging		S	Noun-high	NA	NA	any	default	
awac1239-2	awac1239-no-flagging		A	Noun-high	P	any	any	default	
awac1239-3	awac1239-no-flagging		S	Noun-low	NA	NA	any	default	
awac1239-4	awac1239-no-flagging		A	Noun-low	P	any	any	default	
awac1239-5	awac1239-acc-marking-flagging		P	Noun-high	A	any	any	default	
awac1239-6	awac1239-no-flagging		P	Noun-low	A	any	any	default	
awac1239-7	awac1239-no-flagging		S	Pro	NA	NA	any	default	
awac1239-8	awac1239-no-flagging		A	Pro	P	any	any	default	
awac1239-9	awac1239-acc-marking-flagging		P	Pro	A	any	any	default	
awac1239-10	awac1239-suffix-s-1p-indexing-marker	1	P	1	A	2/3	any	default	any
awac1239-11	awac1239-suffix-s-1p-indexing-marker	1	S	1	NA	NA	non-exemplar	stative	unknown condition 1
awac1239-12	awac1239-suffix-w-1s-a-indexing-marker	1	A	1	P	2/3	any	default	any
awac1239-13	awac1239-suffix-w-1s-a-indexing-marker	1	S	1	NA	NA	any	default	control
awac1239-14	awac1239-suffix-w-1s-a-indexing-marker	1	S	1	NA	NA	non-exemplar	stative	unknown condition 2
awac1239-15	awac1239-suffix-zi-2-3s-a-indexing-marker	1	A	2/3	P	2/3	any	default	any
awac1239-16	awac1239-suffix-zi-2-3s-a-indexing-marker	1	S	2/3	NA	NA	any	default	control
awac1239-17	awac1239-suffix-zi-2-3s-a-indexing-marker	1	S	2/3	NA	NA	non-exemplar	stative	any
bilu1245-1	bilu1245-no-flagging		S	any	NA	NA	any	default	
bilu1245-2	bilu1245-no-flagging		A	any	P	any	any	default	
bilu1245-3	bilu1245-no-flagging		P	any	A	any	any	default	
bilu1245-4	bilu1245-p-enclitics-indexing-marker	clitic 1	P	any	A	any	any	default	
bilu1245-5	bilu1245-s-a-proclitics-indexing-marker	clitic -1	S	any	NA	NA	any	default	
bilu1245-6	bilu1245-s-a-proclitics-indexing-marker	clitic -1	A	any	P	any	any	default	

one selector which is associated with this particular context. Each argument present in a particular context is referred to in terms of its role (see Section 2.1). The argument selector (a morpheme or a paradigm of morphemes) associated with a context is identified by a selector ID, which is linked to the `selectors.csv` table, where selector-specific information is collected. Contexts are language-specific, and the language that a context belongs to is specified through a language ID (note that this column has been omitted in Table 1, but the Glottocode is still visible in the ID column).

Because all contexts are associated with exactly one selector, they must minimally be specified for the argument roles and references involved. However, a context may require more information (such as slot, TAM or predicate class) to distinguish it from other contexts in the language which are associated with different selectors.[14]

In most languages, morphological slot can be seen as a property of the selector in question, but this is not always the case. In some languages, such as Puinave [puin1248] (isolate; Colombia and Venezuela), the same paradigm of person indices is used for both the A and P roles but appears in different slots on the verb (Girón Higuita 2008). To have a unified approach, we treat the slot as a property of the context and the same selector can appear in different slots depending on the context.

Another case where more information is needed to identify a context is when a language uses different verbal paradigms for indexing person-number values in different tenses, as is the case for many Indo-European languages. These different paradigms correspond to different selectors, and so the context must be able to distinguish when one paradigm or the other is used. This is accommodated by the dedicated column for TAM. Separate columns for predicate class, co-arguments, and miscellaneous conditions accommodate other cases where contexts may differ. This structure proved sufficient to capture marking variation in the languages we have encountered.

For practical reasons, we do not differentiate between contexts when there is no difference in terms of alignment. For example, we do not list all person and number combinations for person subject indexes

[14] Note that in Table 1, the TAM column has been omitted because it was not relevant for the languages exemplified.

in a language such as Bilua, where there are two paradigms of clitics that behave uniformly (Obata 2003, 49, 303, 309). In such cases, each row represents a bundle of contexts that have in common the same argument role (see Table 1). Thus, in the Bilua example, there are three rows in `contexts.csv`: two that correspond to the subject clitic paradigm (one for indexing S and one for indexing A) and one row for the object clitic paradigm (P indexing). The roles themselves may be combined in one context row in cases of complete absence of verbal indexing for any role.[15]

As a final note, (person indexing) selectors which function in certain contexts as portmanteaus (i.e. they index both A and P arguments)[16] have two entries in the `contexts.csv` table. Since an entry in the `contexts.csv` table represents the marking of both a role and a referential type, such selectors have two entries for the same scenario: one for marking the A role given the appropriate P as its co-argument and another one for marking the P role given the appropriate A as its co-argument. Though this may seem like a kind of double-coding, it is analogous to a single selector used to mark both S and A roles.

selectors.csv 4.2.2

In the `selectors.csv` table, each row corresponds to a morpheme or a paradigm of morphemes (see Table 2). The label of this morpheme or paradigm is given in free form as its `Selector_label`,

[15] We only allow for this collapsing of argument roles in the case of an absence of indexing, and not in the case of an absence of flagging. Unlike verbal indexing, which can be completely absent in a language, flagging is almost always present if we take into account all argument roles. It is very common that other argument roles currently not coded in our database, such as G (goal) or T (theme), have distinct flagging, even if the S, A, and P argument roles do not.

[16] The property of a selector behaving as a portmanteau is commonly seen as inherent to the selector, e.g. an indexing marker is either a simple or a portmanteau marker. However, in some languages the same selector may function as a simple marker in some contexts and as a portmanteau marker in others. As an example, in Huastec [huas1242] (Mayan; Mexico), the marker *tu=* indexes the 1st person plural P role. However, it is also used in all cases where 1st person A acts on 2nd person P (Edmonson 1988, pp. 114–115). In the former case, the morpheme behaves as a simple P marker; but in the latter case, it can only be understood as a portmanteau. We have therefore opted for considering portmanteau behavior as a property of the context rather than the selector.

Table 2: Excerpt from `selectors.csv` corresponding to the contexts given in Table 1

Glottocode	Selector_type	Selector_label	Marker_type	Features
awac1239	flagging	ACC marking	overt	
awac1239	flagging	NO_FLAGGING	zero	
awac1239	indexing marker	suffix -s 1P	overt	person
awac1239	indexing marker	suffix -w 1S/A	overt	person
awac1239	indexing marker	suffix -zi 2/3S/A	overt	person
bilu1245	flagging	NO_FLAGGING	zero	
bilu1245	indexing marker	P enclitics	overt	person + number
bilu1245	indexing marker	S/A proclitics	overt	person + number

which could be an abstract value (like <ergative suffix>) or a more concrete one (such as the phonological shape of a person indexing morpheme, e.g. <mü- 3sgA>). Each selector is given a value for its `Selector_type` which specifies whether this selector is used for flagging or indexing, and a `Marker_type` which specifies if it is phonologically <overt> or <zero>. The `Selector_type` can be <flagging>, <indexing marker>, or <indexing trigger>, the latter of which is a special type indicating a lack of indexing for a role (and thus always has `Marker_type` <zero>). Zero morphemes that encode a specific referential type have `Selector_type` <indexing marker> or <flagging> and `Marker_type` <zero>, while zeros that represent the lack of indexing in general, or the lack of indexing for a particular role, are always `Selector_type` <indexing trigger>. A consistent selector label <NO_FLAGGING> is used for the absence of flagging of a specific argument role.

The `selectors.csv` table includes other information about selectors, such as what features they index (e.g. number, person). Selectors are linked to the language they belong to by the `Glottocode` column.

4.2.3 languages.csv

In the `languages.csv` table, each row is a language characterized by a unique ID and associated information such as family membership, geographical coordinates etc. These data are following Glottolog 4.8

Table 3: Excerpt from `languages.csv`

Glottocode	Name	Macroarea	Latitude	Longitude	Family
awac1239	Awa-Cuaiquer	South America	1.21652	−78.3401	Barbacoan
bilu1245	Bilua	Papunesia	−7.92388	156.663	

(Hammarström *et al.* 2023). There is also a comment field for any unstructured information on the language as a whole, such as the presence or absence of co-referential personal pronouns. An excerpt from the `languages.csv` table is given in Table 3.

Sample and data collection 4.3

We have selected a geographically-balanced diversity sample of 84 phylogenetically unrelated languages (according to Glottolog 4.8, Hammarström *et al.* 2023), equally distributed among each of the world's six macroareas (Hammarström and Donohue 2014). All of our figures and results are based on this sample of languages, the full list of which can be found in the Supplementary Materials in the `languages.csv` file.

The data collected for this dataset were extracted from primary source documents, mostly from reference grammars and linguistic articles. Only occasionally did we consult native speakers and language specialists (via personal communication).

During data collection, in addition to the entries in our database structure, we created a more human-readable summary of each language's flagging and indexing patterns, complete with detailed references and quotes. This summary was used in team discussions, as well as a reference point for necessary adjustments during autotypologizing. Data consistency during the coding procedure was aided by custom scripts, which reported on definitionally impossible entries (e.g. a claim that two morphemes occupy the same slot on the verb in the same context, or that a noun was marked with two different cases in the same context), which were then corrected manually.

5 DATABASE QUERYING AND RESULTS

The database structure described in Section 4.2 does not itself answer any specific typological questions, but the database can be queried to answer a large variety of possible questions. We exemplify a few of the most typical ways of calculating alignment statements. Some of these queries match familiar alignment statements present in other databases, whereas others are impossible to retrieve from statements in other databases. The examples below are by no means exhaustive of the typological properties that can be extracted from our database.

We organize these alignment properties into different levels of linguistic structure. It is possible to specify typological questions about alignment at the level of the language (Section 5.1), at the level of individual argument selectors (Section 5.2), at the level of individual referential types (Section 5.3), and at the level of argument roles (Section 5.4). All queries and aggregations are implemented in individual functions in the accompanying `alignment_aggregation.py` file in the Supplementary Materials.

5.1 *Language-level aggregation*

Several properties of alignment can be established at a language-wide level, without having to calculate per-selector, per-referent, or per-role information. We have defined queries for five of these and implemented them in the `basic_language_level` function in `alignment_aggregation.py`:

(10)
a. the presence of flagging for core arguments
b. the presence of indexing
c. the features which are targeted by indexing, if there is any
d. the presence of an alignment split conditioned by TAM properties
e. the presence of a split-S system

The presence of overt argument flagging (10a) is retrieved from the `selectors.csv` table by querying, for each language, whether

there are any selectors for which the `Selector_type` is <flagging> and the `Marker_type` is <overt>. The presence of indexing (10b) is likewise retrieved from the `selectors.csv` table by querying for rows in which `Selector_type` is <indexing marker> and `Marker_type` is <overt>. The features targeted by indexing (10c) are retrieved by concatenating all unique non-<NA> values in the `Features` column for all indexing selectors of this language.[17]

The presence of an alignment split conditioned by TAM properties (10d) is retrieved from the `contexts.csv` table by querying, for each language, whether there is more than one value present in the `TAM` column. The presence of multiple values indicates that TAM properties are relevant for an alignment split.

Finally, the presence of a split-S system (10e) is also retrieved from the `contexts.csv` table by querying for rows marking the S role which have a `Predicate_Class` value other than `default`.

Some of these properties, such as the presence of a split-S system, occur frequently in studies on alignment, while others, such as the features targeted by indexing, do not. However, answers to both typical and less typical questions can be extracted easily from our database. We can additionally address typological properties at other levels of organization, below the level of the language as a whole, as we will see in the next sections.

The results of these queries for each language are written to `structure-cldf/values.csv`, in accordance with the CLDF format, and another version is optionally written to the non-CLDF compliant `human-readable.csv`, which is organized by language rather than by language and parameter. Statistics can then be calculated on this output.[18] Although the sample size for this dataset is relatively small,

[17] While it is possible to aggregate these values (e.g. a language with a selector which targets <number> and another selector which targets <person> could be aggregated into <person+number>), we chose to keep them separate (e.g. such a language would have a value <person;number> for this query).

[18] All statistics are implemented in the `write_summary_statistics` function of `alignment_aggregations.py`, which reads the CLDF-compliant csv output of each level of aggregation and calculates summary statistics. These statistics are written to file at `summary.csv`, which can be accessed in the Supplementary Materials.

some summary statistics of these language-level aggregations are presented in Table 4.

Table 4: Selected language-level results (N = 84)

Property	Count	Frequency
Presence of argument flagging	47	56%
Presence of argument indexing	59	70%
TAM-based alignment split	3	4%
Split-S system	9	11%
Person + number always indexed together (if indexing present)	42	71%

5.2 *Selector-level aggregation*

In addition to alignment properties at the language level, it is possible to derive alignment statements at the level of individual argument selectors. Selectors mark roles (S, A, or P), either as argument flagging (on the NP) or indexing (on the verb/clause), and an individual selector may mark multiple referential types (e.g. the same verbal index might be used for both 3rd person singular and 3rd person plural A arguments).

The first question that can be answered about an argument selector is: "Which role(s) are marked by this specific marker?" For example, an argument selector may mark S and A roles, but not P; or S and P, but not A. Once it has been determined which argument roles a selector marks, an alignment statement can be calculated for that selector. This selector-based "alignment" is not quite the same as reference-based alignment, which is what is prototypically referred to by the term (see Sections 2.1 and 5.3). At the level of an individual selector (disregarding for the time being what it is referencing), it either marks a particular role, or it does not (e.g. a specific case suffix either marks S arguments or it does not). There is in this sense no such thing as a tripartite alignment for selectors: since its presence is a binary value, it is impossible to have the state $S \neq A \neq P$. For the same reason, selectors which function exclusively as portmanteaus (such as a morpheme marking 2>1) will always have a horizontal alignment

at the selector level (the marking of A and P but not S). This differs from reference-based alignment, where a horizontal alignment means that for a given reference (e.g. 2*sg*) the A and P (but not S) roles are marked by the same morpheme.

Note that zero selectors (the absence of marking) can also have an alignment. A zero-marked nominative case (contrasting with an overtly-marked accusative case) still has a selector-based (nominative-)accusative alignment. The only case in which a non-overt selector does not have an alignment (Alignment is <NA>) is when a role is not marked at all. As we discussed in Section 4.1, this is possible with indexing (the selector type <indexing trigger>), but not with flagging.[19] Selector-based alignment is closely related to the concept of trigger potential (Siewierska 2003; Bickel *et al.* 2013), because it describes which roles *can* trigger the appearance of a particular morpheme. As such, selector-based alignment can only have the values neutral, accusative, ergative, and horizontal.

For the selector-level aggregation, we wrote queries to add four columns to the `selectors.csv` table (see Table 5). The first three columns, `S_references`, `A_references`, and `P_references`, keep track of which references a selector marks. The values of these columns are generated by looking in the `contexts.csv` table for all instances of a given `Selector_label`, and entering into the appropriate column in the `selectors.csv` table which referential types that selector can reference. If a referential type is conditioned by a co-argument, they are concatenated, e.g. if a selector only marks 1st person A when P is 2nd person, the value entered is <1_2>. If no reference is marked for that role, the value <NONE> is entered in the references list.

Finally, the fourth column, `Alignment`, is added. The value of this column is calculated based on the presence or absence of referential types in the `S_references`, `A_references`, and `P_references` columns, regardless of what values are present. For example, if a

[19] As explained in Section 4.2.1, we consider slot a property of the context, rather than of the selector. Therefore, for each language there is at most one zero selector for flagging and one for indexing and they can appear in different slots. These zero selectors are of the type <flagging> or <indexing marker> respectively and are treated identically to other selectors of the same type.

Table 5: Excerpt from `selectors.csv` with added columns from the selector-level queries

Glottocode	Selector_type	Selector_label	S_references	A_references	P_references	Alignment
awac1239	flagging	ACC marking	NONE	NONE	Noun-high; Pro	accusative
awac1239	flagging	NO_FLAGGING	Noun-high; Noun-low; Pro	Noun-high; Noun-low; Pro	Noun-low	neutral
awac1239	indexing marker	suffix -s 1P	1	NONE	1_2/3	ergative
awac1239	indexing marker	suffix -w 1S/A	1	1_2/3	NONE	accusative
awac1239	indexing marker	suffix -zi 2/3S/A	2/3	2/3_2/3	NONE	accusative
bilu1245	flagging	NO_FLAGGING	any	any	any	neutral
bilu1245	indexing marker	P enclitics	NONE	NONE	any	accusative
bilu1245	indexing marker	S/A proclitics	any	any	NONE	accusative

Variable	Value	Count	Frequency
Selector flagging of S	True	16	26%
Selector flagging of A	True	30	49%
Selector flagging of P	True	35	57%
Selector flagging alignment	Accusative	40	66%
	Ergative	17	28%
	Neutral	3	5%
	Horizontal	1	2%
Selector indexing of S	True	222	58%
Selector indexing of A	True	247	65%
Selector indexing of P	True	204	54%
Selector indexing alignment	Accusative	238	63%
	Neutral	53	14%
	Ergative	45	12%
	Horizontal	44	12%

Table 6: Aggregations of flagging selectors (N = 61) and indexing selectors (N = 380)

particular selector has a non-<None> entry in the `S_references` and `A_references` columns, but <None> in the `P_references` column, then its value for `Alignment` is <accusative>, even if the values present in the `S_references` and `A_references` columns are different.

As we did with language-level aggregation, we present summary statistics at the level of selectors. Here, we have only calculated these statistics for overt markers. These statistics could also be balanced per language, so that languages with many selectors are weighted evenly with languages that have fewer. We present the unbalanced, selector-level statistics for some of these properties in Table 6.

Reference-level aggregation 5.3

Another possible level of aggregation is at the level of referential types. For pronouns and verbal indexing, the relevant referential types are the various person-number combinations attested in the language, while the relevant referential types for nouns are the different groups of nouns (if any) that behave uniformly as far as argument flagging is

concerned (e.g. masculine, feminine, singular, plural, etc.). Thus, it is possible for a language to have e.g. a tripartite alignment for first person singular indexing (different selectors are used for each of the S, A, and P roles), but an accusative alignment for second person singular indexing (S and A roles are indexed with the same selector, while P has a distinct one). Similarly, nouns in the singular may exhibit accusative flagging (nominal S and A arguments are in the nominative case, whereas nominal P arguments are in the accusative case), while nouns in plural may have neutral flagging (the same nominative form is used for all three roles). The reference-level alignment can also be different under different conditions, such as TAM or different predicate classes. In such cases a reference-level alignment is calculated for each of those different conditions.

The reference-level aggregation is implemented in the `reference_alignment.py` script, available in the Supplementary Materials. This script extracts, per language, how each combination of role and referential type is marked. If further conditions are relevant, such as TAM, then the marking of each role and reference combination is calculated per condition. In cases of co-argument sensitive marking, there is no single marking strategy for a role and reference combination, but several, dependent on the co-argument. In such cases, the script extracts a series of marking strategies depending on the co-argument. A detailed example of the script functionality and code flow can be found in Appendix C.

The results of the aggregation at the reference level are written to a separate `references.csv` file (see Table 7). Each row in this table represents a particular referential type of a particular language under specific conditions. Each row is identified with a unique `ID` and is linked with the corresponding language through the `Glottocode` column, while the relevant referential type is listed in the `Referential_type` column. The `references.csv` table also includes several additional columns that specify the conditions (`Monovalent_predicate_class`, `Bivalent_predicate_class`, `TAM`, `Condition`), one column per role (`S`, `A`, and `P`), and two alignment columns (`Alignment` and `Alignment_not_local`), which are calculated based on the role columns. For a more complete description of the `references.csv` table, its columns and possible values, see the Appendices.

Table 7: Excerpt from `references.csv`

Glottocode	Domain	Referential_type	Exemplar	Monovalent Predicate_class	Condition	S	A	P	Alignment	Alignment_not_local
awac1239	Verb	1	exemplar	default	control	suffix -w 1S/A_overt	suffix -w 1S/A_overt	suffix -s 1P_overt	accusative	accusative
awac1239	Verb	2/3	exemplar	default	control	suffix -zi 2/3S/A_overt	suffix -zi 2/3S/A_overt_coarg:2/3 ; INFERRED_NULL_zero_coarg:1	INFERRED_NULL_zero	sensitive	sensitive
awac1239	Verb	1	all	stative	unknown condition 1	suffix -s 1P_overt	suffix -w 1S/A_overt	suffix -s 1P_overt	ergative	ergative
awac1239	Verb	2/3	all	stative	unknown condition 1	suffix -zi 2/3S/A_overt	suffix -zi 2/3S/A_overt_coarg:2/3 ; INFERRED_NULL_zero_coarg:1	INFERRED_NULL_zero_coarg	sensitive	sensitive
awac1239	Verb	1	all	stative	unknown condition 2	suffix -w 1S/A_overt	suffix -w 1S/A_overt	suffix -s 1P_overt	accusative	accusative
awac1239	Verb	2/3	all	stative	unknown condition 2	suffix -zi 2/3S/A_overt	suffix -zi 2/3S/A_overt_coarg:2/3 ; INFERRED_NULL_zero_coarg:1	INFERRED_NULL_zero	sensitive	sensitive
awac1239	Verb	1	all	default	control	suffix -w 1S/A_overt	suffix -w 1S/A_overt	suffix -s 1P_overt	accusative	accusative
awac1239	Verb	2/3	all	default	control	suffix -zi 2/3S/A_overt	suffix -zi 2/3S/A_overt_coarg:2/3 ; INFERRED_NULL_zero_coarg:1	INFERRED_NULL_zero	sensitive	sensitive
awac1239	Noun	Noun-high	exemplar	default		NO_FLAGGING_zero	NO_FLAGGING_zero	ACC marking_overt	accusative	accusative
awac1239	Noun	Noun-low	exemplar	default		NO_FLAGGING_zero	NO_FLAGGING_zero	NO_FLAGGING_zero	no marking	no marking
awac1239	Noun	Noun-high	all	default		NO_FLAGGING_zero	NO_FLAGGING_zero	ACC marking_overt	accusative	accusative
awac1239	Noun	Noun-low	all	default		NO_FLAGGING_zero	NO_FLAGGING_zero	NO_FLAGGING_zero	no marking	no marking
awac1239	Pro	Pro	exemplar	default		NO_FLAGGING_zero	NO_FLAGGING_zero	ACC marking_overt	accusative	accusative
awac1239	Pro	Pro	all	default		NO_FLAGGING_zero	NO_FLAGGING_zero	ACC marking_overt	accusative	accusative
bilu1245	Verb	any	exemplar	default		S/A proclitics_overt	S/A proclitics_overt	P enclitics_overt	accusative	accusative
bilu1245	Verb	any	all	default		S/A proclitics_overt	S/A proclitics_overt	P enclitics_overt	accusative	accusative
bilu1245	Noun	any	exemplar	default		NO_FLAGGING_zero	NO_FLAGGING_zero	NO_FLAGGING_zero	no marking	no marking
bilu1245	Noun	any	all	default		NO_FLAGGING_zero	NO_FLAGGING_zero	NO_FLAGGING_zero	no marking	no marking
bilu1245	Pro	any	exemplar	default		NO_FLAGGING_zero	NO_FLAGGING_zero	NO_FLAGGING_zero	no marking	no marking
bilu1245	Pro	any	all	default		NO_FLAGGING_zero	NO_FLAGGING_zero	NO_FLAGGING_zero	no marking	no marking

The `Alignment` column takes into account all contexts, while the `Alignment_not_local` column excludes from the calculation local scenarios (1st person acting on 2nd and vice versa), since in many languages politeness may have affected the alignments of such scenarios (see e.g. Heath 1998; DeLancey 2021). For example, if the `S` and `A` column for the referential type <noun> are both <NOM_zero> and the `P` column is <ACC_overt>, then the `Alignment` would be <accusative>, as would be the `Alignment_not_local`. If there is co-argument sensitivity for any role, then the `Alignment` is given the value <sensitive>.[20] If this sensitivity is due only to local scenarios, then the `Alignment_not_local` column would have a non-sensitive value.[21] Finally, if all markers involved in the flagging of a particular referent are non-overt, the resulting `Alignment` is <no marking>, a special type of neutral alignment. Another case of neutral alignment but with overt markers is attested more often in indexing than flagging, namely, in cases where the same set of markers is used for all roles. We call this alignment pattern <overt neutral>.

With the `references.csv` table, we can once again perform summary statistics, in this case over all referential types. First, we can calculate, per language, what the most common reference-based alignment pattern is (using the `Alignment` column rather than the `Alignment_not_local` one), weighting all referential types equally. The results are summarized in Table 8, which reports for each selector subtype (Flagging on nouns, Flagging on pronouns, Indexing) all the most frequent reference-based alignments per language that occur at least 5% of the time in our sample.

Another kind of aggregation that can be done at the reference level (and without further aggregation per language) is the presence of paradigmatic zeros in indexing, i.e. which referential types are not

[20] The value <sensitive> is not a proper alignment the way that accusative, ergative, etc. are. It is a bundle of different alignments that are dependent on co-argument references. It is in principle possible to further decompose this state into individual alignment statements not per reference but per combination of reference and co-argument reference, as in Witzlack-Makarevich *et al.* 2016.

[21] In Awa-Cuaiquer it is not possible to assign a scenario involving a 1st person as mixed or local, since second and third persons are not distinguished. We have opted conservatively to keep all potentially non-local scenarios, and the <sensitive> alignment is retained; see Table 7.

Variable	Value	Count	Frequency
Flagging on nouns	no marking	44	52%
	accusative	14	17%
	ergative	9	11%
	no marking/accusative	7	8%
Flagging on pronouns	no marking	41	49%
	accusative	27	32%
	ergative	8	10%
Indexing	accusative	33	39%
	no marking	27	32%
	co-argument sensitive	14	17%

Table 8: Most frequent reference-based alignments per language (>5%)

Person	Role	Count	Frequency
Zero indexing for 1	S	6	3%
	A	7	4%
	P	14	9%
Zero indexing for 2	S	0	0%
	A	17	10%
	P	17	13%
Zero indexing for 3	S	55	22%
	A	65	26%
	P	51	23%

Table 9: Zeros in indexing by person reference

indexed for S, A, and P roles, while other referential types are marked under the same conditions. Table 9 presents cases of zero indexing broken down by person (without distinguishing number, i.e. 2sg and 2pl each count as independent examples of indexing of 2). As Table 9 shows, in our sample the P role more frequently lacks indexing than S and A roles, as is expected from previous research (Siewierska 2013b). Furthermore, the 3rd person more frequently lacks indexing than 1st and 2nd (see e.g. Bickel *et al.* 2015c).

5.4 Role-level aggregation

Information about argument marking can also be aggregated at the level of the role (S, A, P). Such aggregations are not alignment, as typically conceived, since they concern exclusively the manner of marking of individual argument roles, i.e. the manner of S marking (on its own), the manner of A marking, and the manner of P marking. This aggregation allows one to capture the various patterns of differential argument marking (see Witzlack-Makarevich and Seržant 2018 for a recent overview). The best studied type of differential argument marking which corresponds to this level of aggregation is differential object marking (or DOM, see Bossong 1985, 1991). In addition to DOM, there are differential A marking (or DAM) and differential S marking (including split-S or active-stative systems, see Section 5.1).

For this paper, we have only aggregated information about DOM. For our present purposes, we are considering DOM "in the broad sense" (see Witzlack-Makarevich and Seržant 2018), that is, we treat as DOM any case of variation in the marking of the P argument irrespective of the condition triggering it, such as different referential types (e.g. definite vs. indefinite), different TAM of the clause, etc, so long as this change is also accompanied by a change in alignment.[22] It is possible for a language to have complex systems of DOM with more than one factor conditioning the split, e.g. person and TAM. In these cases, we present the combined conditioning factors causing the split.

The presence of DOM is extracted from the `contexts.csv`, `selectors.csv`, and `references.csv` tables. First, we select, per language, all the rows in the `contexts.csv` table which have their `Role` marked as <P>, and which encode flagging information (the associated selector in the `selectors.csv` table has the `Selector_type` <flagging>). If these rows contain different selectors and at least one has `Marker_type` <overt> (i.e. not all are <zero>), then the `references.csv` table is checked for whether these P selectors are as-

[22] We consider <overt neutral> and <no marking> as the same, since they are both subtypes of neutral alignment.

sociated with different alignments. If this is the case, then a language with DOM has been found.[23]

Once a language is established as having DOM, then the conditions which cause the differential P marking are calculated. The potential set of DOM-triggering conditions in the `contexts.csv` table can be found in the columns `TAM`, `Reference`, `Co-argument_reference`, `Miscellaneous_condition`, and `Predicate_class`. If one of these columns has values which are each associated with unique P markers and different alignment statements, then that column is the conditioning environment for the DOM. However, as we mentioned above, it is possible that the DOM is conditioned by two (or even more) conditions; if single columns fail to distinguish between different P markings, then each possible combination is tested. The full details of this extraction are given in the `alignment_aggregation.py` script.

Once calculated, the different DOMs are output to `doms.csv`. Each row represents a single DOM and indicates the language (`Glottocode`) and the conditioning factor which causes it (the `Conditioning` column), e.g. a different reference, TAM, and so forth. In addition, there is a series of columns for each marking, the set of alignments it is associated with, and the corresponding conditions (see the Appendices for more details).

Table 10 shows a simplified example taken from the table generated by our DOM aggregation. Central Kanuri [cent2050] (Saharan; Cameroon, Niger, Nigeria, and Chad) has a DOM in which the P marker *-ga* appears under specific word orders (categorized under <Miscellaneous_condition>), while in the standard word order P is not marked. Brahui [brah1256] (Dravidian; Pakistan, Iran, and Afghanistan) has a DOM based on definiteness (categorized under <Reference>): indefinite nominal P arguments are not marked for

[23] Because the `references.csv` table does not calculate alignment according to fixed coarguments but generalizes across them (see Section 5.3), P selectors that occur for the same reference with different coarguments cooccur in a single cell. In such cases, the relevant row receives the label <sensitive>, indicating coargument-sensitive alignment. The code for calculating DOMs made available in the Supplementary Materials makes the assumption that all such coargument-sensitive differential P flagging necessarily implies the presence of DOM, without calculating all fixed coargument alignments. A manual check confirms that this assumption is correct, at least for the data present in our database.

Table 10: Excerpt from `doms.csv`

Glotto-code	Condi-tioning	Marking_1	Alignment_1	Condition_1	Marking_2	Alignment_2	Condition_2
cent2050	Misc_cond	NO_FLAGGING	no marking	default	P marker -ga	accusative	non-standard order
brah1256	Reference	ACC -e	accusative	Noun-def; Pro	NO_FLAGGING	no marking	Noun-indef

case, while definite nominal and pronominal P arguments have accusative P marking.

With this role-level aggregation, we can derive yet another language-level property, namely, whether the language has DOM at all and what the triggering condition is for the DOM. This is added to the `values.csv` table, using the `doms.csv` table to derive this information. We found that in our sample, DOM is fairly common (20% of languages), with the majority (71%) having a reference-based split.

6 CONCLUSION

When doing typological comparison on complex and multi-layered parts of grammar, such as morphosyntactic alignment, there are many possible points of comparison for the analyst to choose from. One valid method of comparison is to select a well-defined exemplar and compare languages based strictly on the exemplar case. Another possibility is to enumerate each possible pattern and ask whether each occurs in the language above a certain frequency (or whether it occurs at all). With a carefully constructed database, it is possible to encode linguistic data in a way that allows for "late aggregation" (Witzlack-Makarevich *et al.* 2022) for multiple points of comparison based on the same data structure.

We have presented such a database for alignment and shown how it can be used to answer many types of typological questions concerning core argument flagging and indexing. This includes many traditional concepts of alignment (such as alignment per referential type), broader alignment-related phenomena (such as differential object marking), and more expansive questions about argument flagging

and indexing (such as the presence of indexing at all, and which persons and roles lack indexing or are indexed by phonologically null elements). Our database is extensible and there are several additional phenomena that could be added: other roles (such as Theme and Goal); other predicate classes (beyond the major class of bivalent verbs); other types of argument selectors beyond indexing and flagging (e.g. various syntactic properties); and so on. Further aggregations of the data are also possible, besides the ones we have demonstrated. Differential agent marking, differences in alignment based on targeted features (person, person + number, or number only), and a finer distinction among zero-indexing for 3rd persons (separating by number and even gender) are some of the most obvious extensions. Beyond adding more data and more aggregations, another direction for future research could include a more streamlined user interface for data entry and quantitative comparisons with other databases of a different design philosophy. The work presented here, both in database design and ways to query data for typological properties, represents a step forward in the direction of creating generalized, multi-purpose typological databases which can be used to answer many typological questions all at once.

AUTHOR CONTRIBUTIONS

The database structure and data aggregations were conceived by A.W.M., D.I, and N.C.P. Data were collected by A.W.M. and D.I. All computer code was written by M.S. (for reference-based alignment) and D.I. (for other aggregations). The paper was written by A.W.M., D.I., and N.C.P.

ACKNOWLEDGEMENTS

We would like to thank our colleagues and research assistants who contributed to data entry: Marine Vuillermet, Anna Graff, Tai Hong, and Alexandra Nogina, as well as Balthasar Bickel for conceptual discussions. We also would like to thank our reviewers for their helpful feedback on our manuscript. D.I., N.C.P., and M.S. were supported by the Swiss National Science Foundation (SNSF) Sinergia Project "Out of Asia" CRSII5_183578.

APPENDICES

A GENERIC CLDF DATASET DESCRIPTION

The generic CLDF dataset includes a `metadata.json` file, a `sources.bib` file and five tables: `languages.csv`, `contexts.csv`, `selectors.csv`, `references.csv` and `doms.csv`. Of these, the first three tables are basic and correspond to raw data collected from grammars, while the other two are populated algorithmically through scripts. The `metadata.json` file describes the whole dataset and how the different tables are interrelated. The `sources.bib` file contains the bibliographic references. The tables are described in detail below.

A.1 *languages.csv*

Each doculect is identified through its Glottocode and its Glottolog name. This table also contains information about family membership (`Family_Name` column), macroarea, and geographic coordinates (Hammarström *et al.* 2023).

Additionally, there is a `Comment` column for any further unstructured information.

A.2 *contexts.csv*

Each context has a unique `ID`, and is linked to the doculect it belongs to through the `Glottocode` column and to a selector (the morpheme or paradigm of morphemes used in this context) through the `Selector_ID` column. Bibliographic references are given in the `Source` column and the responsible person in the `Coder` column. Finally, any additional remarks are kept in the `Comment` column.

The `Role` and `Reference` columns refer to the argument, while the `Co-argument_role` and `Co-argument_reference` columns to the co-argument. Note that as explained in Section 4.2.1, all contexts involving two arguments are written as two separate contexts where each argument is considered as the primary argument and the other

as the co-argument. The `Role` column can only take one of three values: <S> (for Sole argument of monovalent verb), <A> (for Agent-like argument of bivalent verb), or <P> (for Patient-like argument of bivalent verb). The `Co-argument_role` column can take only one of the following three values: <P> (when the argument is A), <A> (when the argument is P), <NA> (for Not Applicable when the argument is S). In the present form of the ATLAs Alignment Module, the `Co-argument_role` column is redundant since it can be predicted by the `Role` column. However, in an extended form of the database, where e.g. arguments of trivalent verbs are included, more combinations of argument and co-argument roles would be possible, since A could be combined with Theme or Goal.

The `Reference` column can take a variety of values depending on the doculect in question. For indexing, it can take any relevant person-number combination, such as <1sg>, <1pl.incl>, <3pl>, as well as any relevant gender distinction, e.g. <3sg.M>, <3sg.F>. For pronouns, the possible values are the same as for indexing for most languages, but they are always followed by the string "Pro" (e.g. <2sgPro>, <1duPro>, <3pl.F.Pro>). For nouns, the relevant categories are noun classes or other kinds of noun groups that behave uniformly in terms of alignment, always including the string "Noun" (e.g. <Noun-M>, <Noun-sg>, <Noun-pl-indef>, <Noun-high>). The `Co-argument_reference` column can take the same values as the `Reference` column, as appropriate for the co-argument restrictions of each context. The `Selector_ID` column always refers to the marking of the argument (rather than the co-argument) in each context. This is true even for portmanteau morphemes that mark both the A and P roles, since such morphemes appear in two different context rows, one for marking the A argument and one for marking the P argument. The `Portmanteau` column has also been filled out only for indexing, and indicates whether the selector involved in the context functions as a portmanteau which indexes both A and P roles; it takes three possible values: <NA>, <simple>, and <portmanteau>.

The `Slot` column is optional and contains information about the relative orders in which the argument markers appear. This column does not capture slot in the strict sense of a fully articulated morphological template, as determining this for every language in our large typological study proved impractical (for example, there may be

Table 11: Possible values for `Slot` in `contexts.csv`

Value	Interpretation
1, 2, etc.	suffix at the 1st, 2nd etc. slot
−1, −2, etc.	prefix at the 1st, 2nd etc. slot
0	infix or stem change (tone, ablaut, etc.)
1/−1	mixed paradigm that contains both prefixes and suffixes and their corresponding slots
1&2	the suffix slot could be 1 or 2 depending on the analysis
clitic 1	enclitic
clitic -1	proclitic
multiple	for ∅ morphemes on verbs with multiple slots for indexing; the number of posited ∅ morphemes in such cases can vary depending on the analysis
AUX −1, AUX 1, etc.	affixes at the corresponding slot on an auxiliary verb
NA	not applicable; for languages with no argument indexing

many optional slots for grammatical voice markers and TAM information that are not fully listed in the description). Instead, the value in our `Slot` column is only guaranteed to be correct in a relative sense: a <2> indicates a suffix further to the right of the stem than a <1>, for example. For languages with complex templatic structures, we used the slot values given in the grammar. Otherwise, we generated our own slot information based on what was present in the description of the indexing paradigm. The possible values for slot and their interpretation can be seen in Table 11.

The `Exemplar` column is a convenience column for the extraction of alignment patterns per referential type and contains information about our exemplar monovalent and bivalent context as explained in Section 4.1. It can take the values: <exemplar>, <non-exemplar>, and <any>. The `Exemplar` column value <exemplar> corresponds to cases where the context or context bundle in question fits the properties of our exemplar exactly. This value is not attested in our data, due to our exemplar being highly specified. The value <non-exemplar> indicates that the context or context bundle in question does not fit the properties of our exemplar in some regard (e.g. the A may be non-human; or it may not be in control of the action; the

action may have not happened yet; etc.). Finally, the value <any> indicates that this bundle of contexts can contain both exemplar and non-exemplar situations.

Non-exemplar contexts are entered as separate rows in the `contexts.csv` only if they are marked in a way that produces a different alignment pattern. Otherwise, they are bundled appropriately in corresponding <any> contexts.

Common conditions that cause splits in alignment and yield non-exemplar alignment patterns, such as TAM, predicate class and co-argument reference, are marked in the homonymous columns, while all other conditions are listed in the `Miscellaneous_condition` column. The `TAM` column can take any value following the language description, such as <progressive>, <perfective>, <future>, etc. The `Predicate_class` column has at least one <default> monovalent predicate class and one <default> bivalent predicate class. Beyond the default classes, languages may have any number of other predicate classes, such as <stative verbs>. For the present study, we have coded additional bivalent predicate classes only if they contain meanings that at least some of the time meet our exemplar conditions, as well as additional monovalent predicate classes where the S argument lacks control. This restriction is motivated by reasons of practicality (it is often difficult to find details about all predicate classes in a language and/or it takes longer to code) and because our broader study was specifically interested in the "split S" phenomenon.

Finally, several of these columns have a special value type, <any>, which is used as a "wildcard": an <any> in the `TAM`, `Miscellaneous_condition`, and `Exemplar` columns signifies that the context bundle contains contexts that have all possible values of the relevant variable for this doculect. This is a way to avoid duplicate encoding of contexts which are not sensitive to conditions that may be operative in other parts of the language. For example, the language Lavukaleve [lavu1241] (isolate; Solomon Islands) sometimes drops S and A indexing on the verb in unknown discourse contexts, but always indexes P on the verb. In this case, there are contexts for S and A indexing, conditioned on `Miscellaneous_condition` <default>, and contexts for a lack of S and A indexing, conditioned on another `Miscellaneous_condition` (descriptively, <unknown conditions, may be discourse-based>). The indexing for P, however, occurs in

both conditions. So the P context has the `Miscellaneous_condition` <any>, which means that it occurs for all possible values of `Miscellaneous_condition`. The wildcard <any> can also be used for `Reference` and `Co-argument_Reference`, where it refers to any possible referential value. In the case where a bivalent context is not influenced by its co-argument, the value for that variable is <any>. In cases where all indexation or all flagging uses the same paradigm, regardless of referential properties, the `Reference` is set to <any>.

A.3 *selectors.csv*

Each selector has a unique `ID` and is linked to the corresponding doculect through the `Glottocode` column. Analogous to the `contexts.csv`, there are independent columns for primary reference, `Source`, and the coder who entered the data, `Coder`. Selectors have a name (either a high-level description or their phonological form, e.g. 'ACC case' or '-ú 3plS/A'), which is entered in the `Selector_label` column. The `Selector_type` column can take three values in our database: <flagging> for case marking or adpositions, <indexing marker> for selectors involved in verbal indexing, and <indexing trigger> for roles that lack verbal indexing. The `Marker_type` column is a boolean type column involving two values: <overt> and <zero>, for overt and null markers respectively. The `Features` column encodes which features a selector indexes; this column has only been filled out for indexing. It can take one of six values for our data: <NA>, <person>, <number>, <person+number>, <gender>, and <other>. The value <other> covers a variety of features that are more rarely attested, such as proximate/obviative, specificity, honorificity, etc.

The table also includes four columns whose values are not entered by hand, but are derived algorithmically, as described in Section 5.2: `S_references`, `A_references`, `P_references`, and `Alignment`.

A.4 *references.csv*

The `references.csv` is entirely derived by the `reference_alignment.py` script, the logic of which is detailed further below in Appendix C. The table lists references for every doculect

and every relevant condition and gives their alignments. Each reference has a unique `ID` and is linked to the corresponding doculect through the `Glottocode` column, and the language name is given in human-readable format in the `Language` column. The domain to which the reference applies is given in `Domain`, and has three possible values: <Noun>, <Pro> (i.e. pronoun), and <Verb>. The referential type itself is given in `Referential_type` and takes an open-ended set of values, which correspond to the referential types present in that language. In the `Exemplar` column it is indicated if the referential type and associated conditions are among the exemplar ones (value <exemplar>) or if it includes non-exemplar conditions and referential types as well (value <all>). Note that for a language that has no non-exemplar contexts (that are behaving differently from exemplar contexts as far as alignment is concerned) these sets of rows will be identical. By construction, every referent for every language will have at least one row with `Exemplar` marked as <all>. The value of other conditions relevant to the alignment statement is given in the columns `Monovalent_predicate_class`, `Bivalent_predicate_class`, `TAM`, and `Condition`. For languages that have multiple monovalent predicate class and/or multiple bivalent predicate classes, each monovalent predicate class is combined with each bivalent predicate class to produce alignment statements. The `S`, `A`, and `P` columns give the selector(s) which encode that role for each reference, and the `Alignment` and `Alignment_not_local` columns abstract over `S`, `A`, and `P`, generating an alignment per referent (per condition). As explained in Section 5.3, the `Alignment` column takes into account all scenarios in the alignment calculation, while for the `Alignment_not_local` columns, local scenarios are excluded. Finally, the `Source` column amalgamates the sources from the `contexts.csv` and `selectors.csv` tables that were used to generate this alignment, and the `Coder` column likewise amalgamates the coders.

doms.csv A.5

The `doms.csv` table is entirely derived by the `dom_aggregation` function in the `alignment_aggregation.py` script, as described in Section 5.4. The table lists all DOMs (Differential Object Marking) present

in the sample, each of which has a unique `ID` and is linked to the corresponding doculect through the `Glottocode` column. The condition(s) that generate the DOM are given in the `Conditioning` column, which can take the values <Reference>, <Miscellaneous_condition>, <TAM>, and <Co-argument_reference>, or in the case of complex conditions, two or more of these joined by a <+>. The `doms.csv` table also includes an open-ended series of columns, `Marking_X`, `Alignment_X`, and `Condition_X`, for X = 1, 2, ..., for as many conditions as there are encountered in the data for the same doculect. Each `Marking_X` column gives one of the possible markings of P, each `Alignments_X` column gives the set of alignments associated with the marking, and each `Condition_X` column gives the condition in which that marking appears. DOMs definitionally have at least two different markings under two different conditions, but in our data we have one language with three different markings following three different conditions. Finally, there is a `Source` column which amalgamates the sources in `contexts.csv` and `selectors.csv` from which this DOM was derived, and a `Coder` column which concatenates the coders.

B STRUCTURE CLDF DATASET

The structure CLDF dataset has a `metadata.json` and three tables: `languages.csv` (which is an identical copy of the one in the generic CLDF dataset), `parameters.csv`, and `values.csv`.

The `parameters.csv` table contains all language-level aggregations (including ones which are derived from selector, reference, and role-level aggregations), in the form of a unique `Parameter_ID` and a `Question`, which describes the typological property that is derived in the form of a question.

The value of a particular doculect for a particular parameter corresponds to a row in `values.csv`, and is associated with the `languages.csv` and the `parameters.csv` tables via the `Glottocode` and `Parameter_ID` columns respectively. The value itself is stored in the `Value` column. Finally, values have a `Coder`, which is the concatenation of all coders responsible for the raw data which generated this value, and a `Source`, which is the amalgamation of all sources in the

raw data which generated this value. As mentioned in Appendix 5.1, an alternative view of this information – as a matrix with one row per doculect and a column for each parameter – can be generated from our scripts and output by default to `human-readable.csv`. This file is not part of the structure CLDF dataset.

C REFERENCE-LEVEL AGGREGATION CODE FLOW EXAMPLE

In this section, we present two examples of the code flow of the `reference_alignment.py` script, which calculates reference-based alignment, first for Kamu [kamu1258] (Kamu; Australia) and then for Marind [nucl1622] (Anim; Indonesia and Papua New Guinea). Note that the tables we present are slightly simplified with invariant or non-relevant columns removed for the sake of readability.

Kamu exemplifies a moderately complex system of both flagging and indexing, each of which can change under different conditions. Kamu has 26 rows in the `contexts.csv` table (four for argument flagging and 22 for verbal indexing, see Table 12), and five selectors in the `selectors.csv` table (two for flagging and three for indexing, see Table 13).

First, we will exemplify the calculation of alignment for referential types that receive flagging. By filtering the `contexts.csv` table for selectors which are used in flagging (whose corresponding `Selector_type` in the `selectors.csv` is <flagging>), we see that there is only one referential type, the special type <any>, indicating that all pronouns and nouns behave identically with respect to alignment, and two miscellaneous conditions (<default> and <non-default>). For each referential type (in this case, only <any>), we filter the table for every possible miscellaneous condition (here, <default> and <non-default>), always matching <any> with all other values, as explained in A.2. As an example, filtering for the referential type <any> and the miscellaneous condition <default> yields Table 14.

This resulting table is used to fill in the corresponding row for referential type <any> and miscellaneous condition <default> in the

Table 12: Kamu contexts

ID	Selector_ID	Slot	Role	Reference	Co-argument_ reference	Exemplar	Miscellaneous_ condition
kamu1258-1	kamu1258-erg-marking-flagging		A	any	any	any	default
kamu1258-2	kamu1258-no-flagging		A	any	any	any	non-default
kamu1258-3	kamu1258-no-flagging		S	any	NA	any	any
kamu1258-4	kamu1258-no-flagging		P	any	any	any	any
kamu1258-5	kamu1258-null-marker	1	P	3sg.nonhum	any	non-exemplar	any
kamu1258-6	kamu1258-null-marker	1	P	3sg.hum	any	any	unknown condition
kamu1258-7	kamu1258-p-enclitics-indexing-marker	1	P	1sg	any	any	any
kamu1258-8	kamu1258-p-enclitics-indexing-marker	1	P	1pl	any	any	any
kamu1258-9	kamu1258-p-enclitics-indexing-marker	1	P	2sg	any	any	any
kamu1258-10	kamu1258-p-enclitics-indexing-marker	1	P	2pl	any	any	any
kamu1258-11	kamu1258-p-enclitics-indexing-marker	1	P	3pl	any	any	any
kamu1258-12	kamu1258-p-enclitics-indexing-marker	1	P	3sg.hum	any	any	default
kamu1258-13	kamu1258-s-a-prefixes-indexing-marker	AUX -1	A	1sg	any	any	any
kamu1258-14	kamu1258-s-a-prefixes-indexing-marker	AUX -1	A	2sg	any	any	any
kamu1258-15	kamu1258-s-a-prefixes-indexing-marker	AUX -1	A	2pl	any	any	any
kamu1258-16	kamu1258-s-a-prefixes-indexing-marker	AUX -1	A	1pl	any	any	any
kamu1258-17	kamu1258-s-a-prefixes-indexing-marker	AUX -1	A	3pl	any	any	any
kamu1258-18	kamu1258-s-a-prefixes-indexing-marker	AUX -1	A	3sg.nonhum	any	non-exemplar	any
kamu1258-19	kamu1258-s-a-prefixes-indexing-marker	AUX -1	A	3sg.hum	any	any	any
kamu1258-20	kamu1258-s-a-prefixes-indexing-marker	AUX -1	S	1sg	NA	any	any
kamu1258-21	kamu1258-s-a-prefixes-indexing-marker	AUX -1	S	1pl	NA	any	any
kamu1258-22	kamu1258-s-a-prefixes-indexing-marker	AUX -1	S	2sg	NA	any	any
kamu1258-23	kamu1258-s-a-prefixes-indexing-marker	AUX -1	S	2pl	NA	any	any
kamu1258-24	kamu1258-s-a-prefixes-indexing-marker	AUX -1	S	3pl	NA	any	any
kamu1258-25	kamu1258-s-a-prefixes-indexing-marker	AUX -1	S	3sg.nonhum	NA	non-exemplar	any
kamu1258-26	kamu1258-s-a-prefixes-indexing-marker	AUX -1	S	3sg.hum	NA	any	any

Table 13: Kamu selectors

ID	Selector_type	Selector_label	Marker_type	Features
kamu1258-erg-marking-flagging	flagging	ERG marking	overt	
kamu1258-no-flagging	flagging	NO_FLAGGING	zero	
kamu1258-null-marker	indexing marker	NULL_MARKER	zero	NA
kamu1258-p-enclitics-indexing-marker	indexing marker	P enclitics	overt	person + number
kamu1258-s-a-prefixes-indexing-marker	indexing marker	S/A prefixes	overt	person + number

Table 14: Kamu contexts: filtering for selector type <flagging> and <default> condition

ID	Selector_ID	Role	Reference	Co-argument_reference	Exemplar	Miscellaneous_condition
kamu1258-1	kamu1258-erg-marking-flagging	A	any	any	any	default
kamu1258-3	kamu1258-no-flagging	S	any	NA	any	any
kamu1258-4	kamu1258-no-flagging	P	any	any	any	any

references.csv table as follows: The column S contains the selector (and if it is overt or not) for referential type <any> when in the S role, and the same for columns A and P. The result of this process is shown in the first row of Table 15. In the same way, now filtering for <any> and <non-default> condition, we can fill in the second row of Table 15. Note that there are two sets of rows in the references.csv table: one set of rows is marked <exemplar> in the Exemplar column and includes only exemplar contexts, and the other is marked <all> and includes all contexts (exemplar and non-exemplar). In a subsequent step, the S, A, P columns of references.csv are used to calculate the alignment for each referential type. Here, all nouns and pronouns (referential type <any>) have an ergative alignment in the <default> condition, since only the A argument is marked with an overt marker. In the <non-default> condition all nouns and pronouns receive no marking since all selectors are of Marker_type <zero>.

Indexing in Kamu changes depending on referential properties that are included in our exemplar, with some referential types conditionally marked by a null morpheme. We again filter for each combination of referential type, exemplar case, and any relevant miscellaneous condition (i.e. a condition within a certain exemplar case). In this case, Exemplar <non-exemplar> always has Miscellaneous_condition <any>, and Exemplar <any> has <unknown condition>, <default> or <any>, so the following combinations are filtered for in different iterations of the script: each person-number combination for Exemplar <any> and Miscellaneous_condition <unknown condition> or <any>, and each person-number combination for Exemplar <any> and Miscellaneous_condition <default> or <any>. During each iteration, a corresponding line is filled in references.csv, following the same process as for flagging. In the case of Kamu indexing, all referential types have an accusative alignment, even though the P marking changes under certain conditions. The full reference-based alignment table for both flagging and indexing in Kamu, after all processing is done, is presented in Table 15.

Our other example, Marind, has no flagging at all, but a different kind of complexity in its indexing system, including both a split in S marking according to predicate class and co-argument sensitivity for 3pl. Contexts for Marind are given in Table 16 (22 rows total) and selectors in Table 17 (nine rows total).

Table 15: Kamu reference-based alignments

Domain	Referen-tial_ type	Exemplar	Condition	S	A	P	Alignment	Alignment_ not_local
Noun	any	exemplar	default	NO_FLAGGING_zero	ERG marking_overt	NO_FLAGGING_zero	ergative	ergative
Noun	any	exemplar	non-default	NO_FLAGGING_zero	NO_FLAGGING_zero	NO_FLAGGING_zero	no marking	no marking
Noun	any	all	default	NO_FLAGGING_zero	ERG marking_overt	NO_FLAGGING_zero	ergative	ergative
Noun	any	all	non-default	NO_FLAGGING_zero	NO_FLAGGING_zero	NO_FLAGGING_zero	no marking	no marking
Pro	any	exemplar	non-default	NO_FLAGGING_zero	NO_FLAGGING_zero	NO_FLAGGING_zero	no marking	no marking
Pro	any	exemplar	default	NO_FLAGGING_zero	ERG marking_overt	NO_FLAGGING_zero	ergative	ergative
Pro	any	all	non-default	NO_FLAGGING_zero	NO_FLAGGING_zero	NO_FLAGGING_zero	no marking	no marking
Pro	any	all	default	NO_FLAGGING_zero	ERG marking_overt	NO_FLAGGING_zero	ergative	ergative
Verb	2pl	exemplar	unknown condition	S/A prefixes_overt	S/A prefixes_overt	P enclitics_overt	accusative	accusative
Verb	2sg	exemplar	unknown condition	S/A prefixes_overt	S/A prefixes_overt	P enclitics_overt	accusative	accusative
Verb	1pl	exemplar	unknown condition	S/A prefixes_overt	S/A prefixes_overt	P enclitics_overt	accusative	accusative
Verb	3sg.hum	exemplar	unknown condition	S/A prefixes_overt	S/A prefixes_overt	NULL_MARKER_zero	accusative	accusative
Verb	3pl	exemplar	unknown condition	S/A prefixes_overt	S/A prefixes_overt	P enclitics_overt	accusative	accusative
Verb	1sg	exemplar	default	S/A prefixes_overt	S/A prefixes_overt	P enclitics_overt	accusative	accusative
Verb	2pl	exemplar	default	S/A prefixes_overt	S/A prefixes_overt	P enclitics_overt	accusative	accusative
Verb	2sg	exemplar	default	S/A prefixes_overt	S/A prefixes_overt	P enclitics_overt	accusative	accusative
Verb	1pl	exemplar	default	S/A prefixes_overt	S/A prefixes_overt	P enclitics_overt	accusative	accusative
Verb	3sg.hum	exemplar	default	S/A prefixes_overt	S/A prefixes_overt	P enclitics_overt	accusative	accusative
Verb	3pl	exemplar	default	S/A prefixes_overt	S/A prefixes_overt	P enclitics_overt	accusative	accusative
Verb	1sg	all	unknown condition	S/A prefixes_overt	S/A prefixes_overt	P enclitics_overt	accusative	accusative
Verb	2pl	all	unknown condition	S/A prefixes_overt	S/A prefixes_overt	P enclitics_overt	accusative	accusative
Verb	2sg	all	unknown condition	S/A prefixes_overt	S/A prefixes_overt	P enclitics_overt	accusative	accusative
Verb	1pl	all	unknown condition	S/A prefixes_overt	S/A prefixes_overt	P enclitics_overt	accusative	accusative
Verb	3sg.nonhum	all	unknown condition	S/A prefixes_overt	S/A prefixes_overt	NULL_MARKER_zero	accusative	accusative
Verb	3sg.hum	all	unknown condition	S/A prefixes_overt	S/A prefixes_overt	NULL_MARKER_zero	accusative	accusative
Verb	3pl	all	unknown condition	S/A prefixes_overt	S/A prefixes_overt	P enclitics_overt	accusative	accusative
Verb	1sg	all	default	S/A prefixes_overt	S/A prefixes_overt	P enclitics_overt	accusative	accusative
Verb	2pl	all	default	S/A prefixes_overt	S/A prefixes_overt	P enclitics_overt	accusative	accusative
Verb	2sg	all	default	S/A prefixes_overt	S/A prefixes_overt	P enclitics_overt	accusative	accusative
Verb	1pl	all	default	S/A prefixes_overt	S/A prefixes_overt	P enclitics_overt	accusative	accusative
Verb	3sg.nonhum	all	default	S/A prefixes_overt	S/A prefixes_overt	NULL_MARKER_zero	accusative	accusative
Verb	3sg.hum	all	default	S/A prefixes_overt	S/A prefixes_overt	P enclitics_overt	accusative	accusative
Verb	3pl	all	default	S/A prefixes_overt	S/A prefixes_overt	P enclitics_overt	accusative	accusative

Table 16: Marind contexts

ID	Selector_ID	Slot	Role	Reference	Co-argument_reference	Exemplar	Predicate_class
nucl1622-1	nucl1622-no-flagging		A	any	any	any	default
nucl1622-2	nucl1622-no-flagging		P	any	any	any	default
nucl1622-3	nucl1622-no-flagging		S	any	NA	any	default
nucl1622-4	nucl1622-p-affix-indexing-marker	−1/1	P	any	any	any	default
nucl1622-5	nucl1622-s-a-prefix-1pl-nak-e-indexing-marker	−1/1	S	any	NA	non-exemplar	Sp class
nucl1622-6	nucl1622-s-a-prefix-1pl-nak-e-indexing-marker	−1	S	1pl	NA	any	default
nucl1622-7	nucl1622-s-a-prefix-1sg-no-nak-indexing-marker	−1	A	1pl	any	any	default
nucl1622-8	nucl1622-s-a-prefix-1sg-no-nak-indexing-marker	−1	S	1sg	NA	any	default
nucl1622-9	nucl1622-s-a-prefix-2pl-e-indexing-marker	−1	A	1sg	any	any	default
nucl1622-10	nucl1622-s-a-prefix-2pl-e-indexing-marker	−1	S	2pl	NA	any	default
nucl1622-11	nucl1622-s-a-prefix-2sg-o-indexing-marker	−1	A	2pl	any	any	default
nucl1622-12	nucl1622-s-a-prefix-2sg-o-indexing-marker	−1	S	2sg	NA	any	default
nucl1622-13	nucl1622-p-affix-indexing-marker	−1	A	2sg	any	any	default
nucl1622-14	nucl1622-s-a-prefix-3pl-e-3pl-1-indexing-marker	−1	A	3pl	1sg	any	default
nucl1622-15	nucl1622-s-a-prefix-3pl-e-3pl-1-indexing-marker	−1	A	3pl	1pl	any	default
nucl1622-16	nucl1622-s-a-prefix-3pl-n-not-in-3pl-1-indexing-marker	−1	S	3pl	NA	any	default
nucl1622-17	nucl1622-s-a-prefix-3pl-n-not-in-3pl-1-indexing-marker	−1	A	3pl	2sg	any	default
nucl1622-18	nucl1622-s-a-prefix-3pl-n-not-in-3pl-1-indexing-marker	−1	A	3pl	2pl	any	default
nucl1622-19	nucl1622-s-a-prefix-3pl-n-not-in-3pl-1-indexing-marker	−1	A	3pl	3sg	any	default
nucl1622-20	nucl1622-s-a-prefix-3pl-n-not-in-3pl-1-indexing-marker	−1	A	3pl	3pl	any	default
nucl1622-21	nucl1622-s-a-prefix-3sg-a-0-indexing-marker	−1	S	3sg	NA	any	default
nucl1622-22	nucl1622-s-a-prefix-3sg-a-0-indexing-marker	−1	A	3sg	any	any	default

Table 17: Marind selectors

ID	Selector_type	Selector_label	Marker_type	Features
nucl1622-no-flagging	flagging	NO_FLAGGING	zero	
nucl1622-p-affix-indexing-marker	indexing marker	P affix	overt	person + number
nucl1622-s-a-prefix-1pl-nak-e-indexing-marker	indexing marker	S/A prefix 1pl nak-…(e-)	overt	person + number
nucl1622-s-a-prefix-1sg-no-nak-indexing-marker	indexing marker	S/A prefix 1sg no-/nak-	overt	person + number
nucl1622-s-a-prefix-2pl-e-indexing-marker	indexing marker	S/A prefix 2pl e-	overt	person + number
nucl1622-s-a-prefix-2sg-o-indexing-marker	indexing marker	S/A prefix 2sg o-	overt	person + number
nucl1622-s-a-prefix-3pl-e-3pl-1-indexing-marker	indexing marker	S/A prefix 3pl e- (3pl > 1)	overt	person + number
nucl1622-s-a-prefix-3pl-n-not-in-3pl-1-indexing-marker	indexing marker	S/A prefix 3pl n- (not in 3pl > 1)	overt	person + number
nucl1622-s-a-prefix-3sg-a-0-indexing-marker	indexing marker	S/A prefix 3sg a-/0-	overt	person + number

Table 18: Marind contexts filtered for < default > monovalent predicate class and < 1pl >

ID	Selector_ID	Slot	Role	Reference	Co-argument_reference	Predicate class
nucl1622-4	nucl1622-p-affix-indexing-marker	−1/1	P	any	any	default
nucl1622-6	nucl1622-s-a-prefix-1pl-nak-e-indexing-marker	−1	S	1pl	NA	default
nucl1622-7	nucl1622-s-a-prefix-1pl-nak-e-indexing-marker	−1	A	1pl	any	default

Marind has two monovalent predicate classes (<default> and <Sp class>), so there are two different alignment calculations: one comparing monovalent default S with bivalent default A and P and one comparing monovalent Sp class S with bivalent default A and P. An alignment is calculated for each referential type and for each condition (in this case, default and Sp class). Filtering for the <1pl> referential type and the <default> monovalent predicate class results in Table 18. Note that referential type <any> matches all specific referential types, including <1pl>. The alignment for <1pl> and the <default> condition is accusative as can be seen in the fifth row of Table 20 for the set of exemplar alignments and in the 17th row for the set of all alignments. Note that in the set of all alignments, an additional alignment statement for <1pl> is attested (in the eleventh row of the table), this time with a different monovalent predicate class (Sp) and its alignment value is <ergative>. Predicates of the Sp class indicate actions where the S has no control, and therefore they are not included in the set of exemplar alignments, since our chosen exemplar requires that the S has control over the event (see Section 3).

When co-argument sensitivity is involved, a referential type will participate in multiple contexts with the same role but with different co-arguments, as is the case for 3pl in Marind. This can be seen in Table 19, which filters Marind contexts for referential type <3pl> and the <default> monovalent predicate class. An alignment for this referential type cannot be calculated because there is no single marker for the A role (although one could calculate an alignment if the co-arguments were fixed, i.e. the marking of 3pl when its co-argument, if any, is 1sg – e.g. an alignment of 3pl S vs. A (with 1sg P) vs. P (with 1sg A) – but this is not something we have done here). Instead, in `references.csv` all the different ways that 3pl A is marked depending on the co-argument are concatenated within the same cell of the `A` column (see the first, seventh and thirteenth rows in Table 20). When we calculate reference-based alignments, cases such as 3pl in Marind get the pseudo-alignment <sensitive>, indicating that there is no single alignment statement that can be made without the co-argument role being fixed.

Once all of these calculations are done for every reference and every condition, the output for reference-based alignment of indexing in Marind is as in Table 20.

Table 19: Marind contexts filtered for <default> monovalent predicate class and <3pl>

ID	Selector_ID	Slot	Role	Reference	Co-argument_ reference	Predicate class
nucl1622-4	nucl1622-p-affix-indexing-marker	−1/1	P	any	any	default
nucl1622-14	nucl1622-s-a-prefix-3pl-e-3pl-1-indexing-marker	−1	A	3pl	1sg	default
nucl1622-15	nucl1622-s-a-prefix-3pl-e-3pl-1-indexing-marker	−1	A	3pl	1pl	default
nucl1622-16	nucl1622-s-a-prefix-3pl-n-not-in-3pl-1-indexing-marker	−1	S	3pl	NA	default
nucl1622-17	nucl1622-s-a-prefix-3pl-n-not-in-3pl-1-indexing-marker	−1	A	3pl	2sg	default
nucl1622-18	nucl1622-s-a-prefix-3pl-n-not-in-3pl-1-indexing-marker	−1	A	3pl	2pl	default
nucl1622-19	nucl1622-s-a-prefix-3pl-n-not-in-3pl-1-indexing-marker	−1	A	3pl	3sg	default
nucl1622-20	nucl1622-s-a-prefix-3pl-n-not-in-3pl-1-indexing-marker	−1	A	3pl	3pl	default

Table 20: Marind reference-based alignments

Referential_type	Exemplar	Monovalent_predicate_class	S	A	P	Alignment
3pl	exemplar	default	S/A prefix 3pl n-(not in 3pl>1)_overt	S/A prefix 3pl e- (3pl>1)_overt_coarg:1sg ; S/A prefix 3pl e- (3pl>1)_overt_coarg:1pl ; S/A prefix 3pl n- (not in 3pl>1)_overt_coarg:2sg ; S/A prefix 3pl n- (not in 3pl>1)_overt_coarg:2pl ; S/A prefix 3pl n- (not in 3pl>1)_overt_coarg:3sg ; S/A prefix 3pl n- (not in 3pl>1)_overt_coarg:3pl ; INFERRED_NULL_zero_coarg:else	P affix_overt	sensitive
2pl	exemplar	default	S/A prefix 2pl e-_overt	S/A prefix 2pl e-_overt	P affix_overt	accusative
1sg	exemplar	default	S/A prefix 1sg no-/nak-_overt	S/A prefix 1sg no-/nak-_overt	P affix_overt	accusative
2sg	exemplar	default	S/A prefix 2sg o-_overt	S/A prefix 2sg o-_overt	P affix_overt	accusative
1pl	exemplar	default	S/A prefix 1pl nak-…(e-)_overt	S/A prefix 1pl nak-…(e-)_overt	P affix_overt	accusative
3sg	exemplar	default	S/A prefix 3sg a-/0-_overt	S/A prefix 3sg a-/0-_overt	P affix_overt	accusative
3pl	all	Sp class	P affix_overt	S/A prefix 3pl e- (3pl>1)_overt_coarg:1sg ; S/A prefix 3pl e- (3pl>1)_overt_coarg:1pl ; S/A prefix 3pl n- (not in 3pl>1)_overt_coarg:2sg ; S/A prefix 3pl n- (not in 3pl>1)_overt_coarg:2pl ; S/A prefix 3pl n- (not in 3pl>1)_overt_coarg:3sg ; S/A prefix 3pl n- (not in 3pl>1)_overt_coarg:3pl ; INFERRED_NULL_zero_coarg:else	P affix_overt	sensitive
2pl	all	Sp class	P affix_overt	S/A prefix 2pl e-_overt	P affix_overt	ergative
1sg	all	Sp class	P affix_overt	S/A prefix 1sg no-/nak-_overt	P affix_overt	ergative
2sg	all	Sp class	P affix_overt	S/A prefix 2sg o-_overt	P affix_overt	ergative
1pl	all	Sp class	P affix_overt	S/A prefix 1pl nak-…(e-)_overt	P affix_overt	ergative
3sg	all	Sp class	P affix_overt	S/A prefix 3sg a-/0-_overt	P affix_overt	ergative
3pl	all	default	S/A prefix 3pl n-(not in 3pl>1)_overt	S/A prefix 3pl e- (3pl>1)_overt_coarg:1sg ; S/A prefix 3pl e- (3pl>1)_overt_coarg:1pl ; S/A prefix 3pl n- (not in 3pl>1)_overt_coarg:2sg ; S/A prefix 3pl n- (not in 3pl>1)_overt_coarg:2pl ; S/A prefix 3pl n- (not in 3pl>1)_overt_coarg:3sg ; S/A prefix 3pl n- (not in 3pl>1)_overt_coarg:3pl ; INFERRED_NULL_zero_coarg:else	P affix_overt	sensitive
2pl	all	default	S/A prefix 2pl e-_overt	S/A prefix 2pl e-_overt	P affix_overt	accusative
1sg	all	default	S/A prefix 1sg no-/nak-_overt	S/A prefix 1sg no-/nak-_overt	P affix_overt	accusative
2sg	all	default	S/A prefix 2sg o-_overt	S/A prefix 2sg o-_overt	P affix_overt	accusative
1pl	all	default	S/A prefix 1pl nak-…(e-)_overt	S/A prefix 1pl nak-…(e-)_overt	P affix_overt	accusative
3sg	all	default	S/A prefix 3sg a-/0-_overt	S/A prefix 3sg a-/0-_overt	P affix_overt	accusative
any	exemplar	default	NO_FLAGGING_zero	NO_FLAGGING_zero	NO_FLAGGING_zero	no marking
any	all	default	NO_FLAGGING_zero	NO_FLAGGING_zero	NO_FLAGGING_zero	no marking
any	exemplar	default	NO_FLAGGING_zero	NO_FLAGGING_zero	NO_FLAGGING_zero	no marking
any	all	default	NO_FLAGGING_zero	NO_FLAGGING_zero	NO_FLAGGING_zero	no marking

REFERENCES

Stephen R. Anderson (1976), On the notion of subject in ergative languages, in Charles N. Li, editor, *Subject and topic*, pp. 1–23, Academic Press, New York.

Balthasar Bickel (2011), Grammatical relations typology, in Jae Jung Song, editor, *The Oxford handbook of linguistic typology*, pp. 399–444, Oxford University Press, Oxford.

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

Balthasar Bickel and Johanna Nichols (2002), Autotypologizing databases and their use in fieldwork, in Peter Austin, Helen Dry, and Peter Wittenburg, editors, *Proceedings of the International LREC Workshop on Resources and Tools in Field Linguistics, Las Palmas, 26–27 May 2002*, MPI for Psycholinguistics, Nijmegen.

Balthasar Bickel, Johanna Nichols, Taras Zakharko, Alena Witzlack-Makarevich, Kristine Hildebrandt, Michael Riessler, Lennart Bierkandt, Fernando Zúñiga, and John B. Lowe (2022), The AUTOTYP database (v1.1.0), doi:10.5281/zenodo.6793367.

Balthasar Bickel, Alena Witzlack-Makarevich, Kamal K. Choudhary, Matthias Schlesewsky, and Ina Bornkessel-Schlesewsky (2015a), The neurophysiology of language processing shapes the evolution of grammar: Evidence from case marking, *PloS One*, 10(8):e0132819.

Balthasar Bickel, Alena Witzlack-Makarevich, and Taras Zakharko (2015b), Typological evidence against universal effects of referential scales on case alignment, in Ina Bornkessel-Schlesewsky, Andrej L. Malchukov, and Marc Richards, editors, *Scales and hierarchies: A cross-disciplinary perspective*, pp. 7–43, de Gruyter Mouton, Berlin.

Balthasar Bickel, Alena Witzlack-Makarevich, Taras Zakharko, and Giorgio Iemmolo (2015c), Exploring diachronic universals of agreement: Alignment patterns and zero marking across person categories, in Jürg Fleischer, Elisabeth Rieken, and Paul Widmer, editors, *Agreement from a diachronic perspective*, pp. 29–52, de Gruyter Mouton, Berlin.

Balthasar Bickel, Taras Zakharko, Lennart Bierkandt, and Alena Witzlack-Makarevich (2014), Semantic role clustering: An empirical assessment of semantic role types in non-default case assignment, *Studies in Language*, 38(3):485–511.

Joshua Birchall (2014a), Argument marking (argex), in Harald Hammarström, Olga Krasnoukhova, Neele Müller, Joshua Birchall, Simon van de Kerke, Loretta O'Connor, Swintha Danielsen, Rik van Gijn, and George Saad, editors, *South American Indian language structures (SAILS) online*, Max Planck Institute for the Science of Human History, http://sails.clld.org.

Joshua Thomas Rigo Birchall (2014b), *Argument marking patterns in South American languages*, Utrecht: LOT.

Georg Bossong (1985), *Empirische Universalienforschung: Differentielle Objektmarkierung in neuiranischen Sprachen [Empirical research on universals: Differential object marking in New Iranian languages]*, Narr, Tübingen.

Georg Bossong (1991), Differential object marking in Romance and beyond, in Dieter Wanner and Douglas A. Kibbee, editors, *New analyses in Romance linguistics. Selected papers from the XVIII Linguistic Symposium on Romance Languages Urbana-Champaign, April 7–9, 1988*, pp. 143–170, John Benjamins, Amsterdam.

Bernard Comrie (1978), Ergativity, in Winfred Philipp Lehmann, editor, *Syntactic typology: Studies in the phenomenology of language*, pp. 329–394, University of Texas Press, Austin.

Bernard Comrie (1989), *Language universals and linguistic typology*, Blackwell, Oxford.

Bernard Comrie (2013a), Alignment of case marking of full noun phrases, in Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*, Max Planck Institute for Evolutionary Anthropology, Leipzig, https://wals.info/chapter/98.

Bernard Comrie (2013b), Alignment of case marking of pronouns, in Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*, Max Planck Institute for Evolutionary Anthropology, Leipzig, https://wals.info/chapter/99.

Bernard Comrie, Martin Haspelmath, and Balthasar Bickel (2008), The Leipzig Glossing Rules: Conventions for interlinear morpheme-by-morpheme glosses, *Department of Linguistics of the Max Planck Institute for Evolutionary Anthropology & the Department of Linguistics of the University of Leipzig*.

Sonia Cristofaro, Vittorio Ganfi, and Guglielmo Inglese, editors (2021), *The Pavia DEmA (Diachronic Emergence of Alignment) database*, Università di Pavia, Pavia, https://su-lab.unipv.it/tasf/index.php/dema/.

William Croft (2001), *Radical Construction Grammar: Syntactic theory in typological perspective*, Oxford University Press, Oxford.

Timothy Curnow (1997), *A grammar of Awa Pit (Cuaiquer): An indigenous language of south-western Colombia*, Ph.D. thesis, Australian National University, http://monolith.eva.mpg.de/~haspelmt/AwaPit.pdf.

Jaime DE ANGULO and Lucy S. FREELAND (1930), The Achumawi language, *International Journal of American Linguistics*, 6(2):77–120.

Scott DELANCEY (2021), Differential innovation in 2nd person pronouns and agreement indexation in Trans-Himalayan languages, *Folia Linguistica*, 55(s42-s1):155–174, doi:10.1515/flin-2021-2017.

Robert M. W. DIXON (1994), *Ergativity*, Cambridge University Press, Cambridge.

David R. DOWTY (1991), Thematic proto-roles and argument selection, *Language*, 67(3):547–619.

Matthew S. DRYER (1996), *Grammatical relations in Kutenai*, Voices of Rupert's Land, Winnipeg.

Matthew S. DRYER (1997), Are grammatical relations universal?, in Joan BYBEE, John HAIMAN, and Sandra A. THOMPSON, editors, *Essays on language function and language type: Dedicated to T. Givón*, pp. 115–143, Benjamins, Amsterdam.

Matthew S. DRYER and Martin HASPELMATH, editors (2013), *The World Atlas of Language Structures Online (v2020.3)*, Max Planck Institute for Evolutionary Anthropology, doi:10.5281/zenodo.7385533.

Barbara Wedemeyer EDMONSON (1988), *A descriptive grammar of Huastec (Potosino dialect)*, Ph.D. thesis, Tulane University, Ann Arbor.

Robert FORKEL, Johann-Mattis LIST, Simon J. GREENHILL, Christoph RZYMSKI, Sebastian BANK, Michael CYSOUW, Harald HAMMARSTRÖM, Martin HASPELMATH, Gereon A. KAIPING, and Russell D. GRAY (2018), Cross-Linguistic Data Formats, advancing data sharing and re-use in comparative linguistics, *Scientific Data*, 5(1):180–205, doi:10.1038/sdata.2018.205.

Jesús Mario GIRÓN HIGUITA (2008), *Una gramatica del Wansöjöt (Puinave)*, Ph.D. thesis, University of Amsterdam.

Harald HAMMARSTRÖM and Mark DONOHUE (2014), Some principles on the use of macro-areas in typological comparison, *Language Dynamics and Change*, 4(1):167–187, doi:10.1163/22105832-00401001.

Harald HAMMARSTRÖM, Robert FORKEL, Martin HASPELMATH, and Sebastian BANK (2023), glottolog/glottolog-cldf: Glottolog database 4.8 as CLDF, doi:10.5281/zenodo.8131091.

Iren HARTMANN, Martin HASPELMATH, and Bradley TAYLOR, editors (2013), *The Valency Patterns Leipzig online database*, Max Planck Institute for Evolutionary Anthropology, Leipzig, `https://valpal.info/`.

Martin HASPELMATH (2011), On S, A, P, T, and R as comparative concepts for alignment typology, *Linguistic Typology*, 15(3):535–567.

Jeffrey Heath (1998), Pragmatic skewing in 1 > 2 pronominal combinations in Native American languages, *International Journal of American Linguistics*, 64:83–104.

David Inman, Natalia Chousou-Polydouri, Marine Vuillermet, Kellen Parker van Dam, Shelece Easterday, Françoise Rose, Alena Witzlack-Makarevich, Kevin M Bätscher, Oscar Cocaud-Degrève, Anna Graff, Selma Hardegger, Tai Hong, Thomas C Huber, Diana Krasovskaya, Raphaël Luffroy, Nora Muheim, André Müller, Alexandra Nogina, David Timothy Perrot, and Balthasar Bickel (in prep), The ATLAs database: Areal typology of the languages of the Americas.

Edward L. Keenan (1976), Remarkable subjects in Malagasy, in Charles Li, editor, *Subject and Topic*, Academic Press, New York.

Randy J. LaPolla (1993), Arguments against 'subject' and 'object' as viable concepts in Chinese, *Bulletin of the Institute of History and Philology, Academia Sinica*, 63:759–813.

Gilbert Lazard (2002), Transitivity revisited as an example of a more strict approach in typological research, *Folia Linguistica*, 36(3–4):141–190, doi:10.1515/flin.2002.36.3-4.141.

Charles N. Li and Sandra A. Thompson (1976), Subject and topic: a new typology of language, in Charles N. Li, editor, *Subject and topic*, New York.

Andrej Malchukov, Martin Haspelmath, and Bernard Comrie (2010a), Ditransitive constructions: A typological overview, in Andrej Malchukov, Martin Haspelmath, and Bernard Comrie, editors, *Studies in ditransitive constructions: A comparative handbook*, pp. 1–35, De Gruyter Mouton, Berlin.

Andrej Malchukov, Martin Haspelmath, and Bernard Comrie, editors (2010b), *Studies in ditransitive constructions: A comparative handbook*, De Gruyter Mouton, Berlin.

Graham Mallinson and Barry Blake (1981), *Language typology: Cross-linguistic studies in syntax*, North-Holland, Amsterdam.

Edith Moravcsik (1978), On the distribution of ergative and accusative patterns, *Lingua*, 45(3–4):233–279.

Johanna Nichols (1992), *Linguistic diversity in space and time*, University of Chicago Press, Chicago.

Kazuko Obata (2003), *A grammar of Bilua: A Papuan language of the Solomon Islands*, Research School of Pacific and Asian Studies, Australian National University, Canberra.

Simon Overall (2017), *A grammar of Aguaruna (Iiniá Chicham)*, De Gruyter Mouton, Berlin.

Beatrice Primus (1999), *Cases and thematic roles*, Niemeyer, Tübingen.

Beatrice Primus (2006), Mismatches in semantic role hierarchies and the dimensions of role semantics, in Ina Bornkessel, Matthias Schlesewsky, Bernard Comrie, and Angela D. Friederici, editors, *Semantic role universals and argument linking: Theoretical, typological and psycholinguistic perspectives*, pp. 53–88, Mouton de Gruyter, Berlin.

Sergey Say, editor (2020–), *BivalTyp: Typological database of bivalent verbs and their encoding frames*, `https://www.bivaltyp.info`.

Anna Siewierska (1998), On nominal and verbal person marking, *Linguistic Typology*, 2:1–55.

Anna Siewierska (2003), Person agreement and the determination of alignment, *Transactions of the Philological Society*, 101:339–370.

Anna Siewierska (2013a), Alignment of verbal person marking, in Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*, Max Planck Institute for Evolutionary Anthropology, Leipzig, `http://wals.info/chapter/100`.

Anna Siewierska (2013b), Verbal person marking (v2020.3), in Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*, Max Planck Institute for Evolutionary Anthropology, Leipzig, `http://wals.info/chapter/102`.

Michael Silverstein (1976), Hierarchy of features and ergativity, in Robert M. W. Dixon, editor, *Grammatical categories in Australian languages*, pp. 112–171, Humanities Press, New Jersey.

Hedvig Skirgård, Hannah J. Haynie, Damián E. Blasi, Harald Hammarström, Jeremy Collins, Jay J. Latarche, Jakob Lesage, Tobias Weber, Alena Witzlack-Makarevich, Sam Passmore, Angela Chira, Luke Maurits, Russell Dinnage, Michael Dunn, Ger Reesink, Ruth Singer, Claire Bowern, Patience Epps, Jane Hill, Outi Vesakoski, Martine Robbeets, Noor Karolin Abbas, Daniel Auer, Nancy A. Bakker, Giulia Barbos, Robert D. Borges, Swintha Danielsen, Luise Dorenbusch, Ella Dorn, John Elliott, Giada Falcone, Jana Fischer, Yustinus Ghanggo Ate, Hannah Gibson, Hans-Philipp Göbel, Jemima A. Goodall, Victoria Gruner, Andrew Harvey, Rebekah Hayes, Leonard Heer, Roberto E. Herrera Miranda, Nataliia Hübler, Biu Huntington-Rainey, Jessica K. Ivani, Marilen Johns, Erika Just, Eri Kashima, Carolina Kipf, Janina V. Klingenberg, Nikita König, Aikaterina Koti, Richard G. A. Kowalik, Olga Krasnoukhova, Nora L.M. Lindvall, Mandy Lorenzen, Hannah Lutzenberger, Tônia R.A. Martins, Celia Mata German, Suzanne van der Meer, Jaime Montoya Samamé, Michael Müller, Saliha Muradoglu, Kelsey Neely, Johanna Nickel, Miina Norvik, Cheryl Akinyi Oluoch, Jesse Peacock, India O.C. Pearey, Naomi Peck, Stephanie Petit, Sören Pieper, Mariana Poblete, Daniel Prestipino, Linda Raabe, Amna

Raja, Janis Reimringer, Sydney C. Rey, Julia Rizaew, Eloisa Ruppert, Kim K. Salmon, Jill Sammet, Rhiannon Schembri, Lars Schlabbach, Frederick W.P. Schmidt, Amalia Skilton, Wikaliler Daniel Smith, Hilário de Sousa, Kristin Sverredal, Daniel Valle, Javier Vera, Judith Voss, Tim Witte, Henry Wu, Stephanie Yam, Jingting Ye, Maisie Yong, Tessa Yuditha, Roberto Zariquiey, Robert Forkel, Nicholas Evans, Stephen C. Levinson, Martin Haspelmath, Simon J. Greenhill, Quentin D. Atkinson, and Russell D. Gray (2023), Grambank reveals global patterns in the structural diversity of the world's languages, *Science Advances*, 9, doi:10.1126/sciadv.adg6175.

Robert D. Van Valin, Jr. (1981), Grammatical relations in ergative languages, *Studies in Language*, 5(3):361–394.

Robert D. Van Valin, Jr. (1983), Pragmatics, ergativity and grammatical relations, *Journal of Pragmatics*, 7(1):63–88.

Robert D. Van Valin, Jr. (2005), *Exploring the syntax-semantics interface*, Cambridge University Press, Cambridge.

Alena Witzlack-Makarevich (2011), *Typological variation in grammatical relations*, Ph.D. thesis, Unversity of Leipzig, Leipzig.

Alena Witzlack-Makarevich (2019), Argument selectors: A new perspective on grammatical relations. An introduction, in Alena Witzlack-Makarevich and Balthasar Bickel, editors, *Argument Selectors: A new perspective on grammatical relations*, pp. 1–38, John Benjamins, Amsterdam.

Alena Witzlack-Makarevich, Johanna Nichols, Kristine A. Hildebrandt, Taras Zakharko, and Balthasar Bickel (2022), Managing AUTOTYP data: Design principles and implementation, in Andrea L. Berez-Kroeker, Bradley McDonnell, Eve Koller, and Lauren B. Collister, editors, *The Open Handbook of Linguistic Data Management*, pp. 631–642, The MIT Press.

Alena Witzlack-Makarevich and Ilja A. Seržant (2018), Differential argument marking: Patterns of variation, in Ilja A. Seržant and Alena Witzlack-Makarevich, editors, *Diachrony of differential argument marking*, pp. 1–40, Language Science Press, Berlin.

Alena Witzlack-Makarevich, Taras Zakharko, Lennart Bierkandt, Fernando Zúñiga, and Balthasar Bickel (2016), Decomposing hierarchical alignment: co-arguments as conditions on alignment and the limits of referential hierarchies as explanations in verb agreement, *Linguistics*, 54(3):531–561.

Fernando Zúñiga (2006), *Deixis and alignment: inverse systems in indigenous languages of the Americas*, John Benjamins, Amsterdam.

David Inman
0000-0003-1892-591X
david.inman@uzh.ch

Department of Comparative Language Science & Center for the Interdisciplinary Study of Language Evolution
University of Zurich

Alena Witzlack-Makarevich
0000-0003-0138-4635
awitzlack@maoil.huji.ac.il

Hebrew University of Jerusalem

Natalia Chousou-Polydouri
0000-0002-5693-975X
nchousoupolydouri@gmail.com

Department of Comparative Language Science & Center for the Interdisciplinary Study of Language Evolution
University of Zurich

Melvin Steiger
steigermelvin@gmail.com
0000-0001-7300-0704

Department of Informatics
University of Zurich

David Inman, Alena Witzlack-Makarevich, Natalia Chousou-Polydouri, and Melvin Steiger These authors have contributed equally to this work. (2024), *Alignment everywhere all at once: Applying the late aggregation principle to a typological database of argument marking*, Journal of Language Modelling, 12(2):287–347
https://dx.doi.org/10.15398/jlm.v12i2.360