Linguistic Typology at the Crossroads 5-2 (2025): 36-65 ISSN 2785-0943

For a *Discourse-Sensitive Typology*: Theoretical and methodological aspects

SIMONE MATTIOLA
UNIVERSITY OF PAVIA

Submitted: 09/04/2025 Revised version: 06/10/2025
Accepted: 06/11/2025 Published: 25/02/2026

Abstract

The aims of this paper are twofold. First, I aim to discuss some theoretical and practical issues concerning the methods employed by linguists when doing typological research. More specifically, I will present the kinds of data typologists usually base their work on, also discussing some related issues that deal with the nature of the data themselves and the possible misinterpretations that they may lead to. Then, I will discuss how typology tends to focus its attention over a limited array of phenomena (i.e., morphosyntactic ones), leaving aside phenomena pertaining to other levels of analysis (e.g., discourse). I thus will exemplify how integrating discourse phenomena might result in more comprehensive analyses and, consequently, in better-suited typological generalizations. Second, I will propose a multi-level method (which I call *discourse-sensitive typology*) based on the *converging evidence* perspective, discussing how it might help us to overcome the abovementioned issues and ultimately make typology truly usage-based.

Keywords: typology; discourse; methods; usage-based.

1. Introduction

This paper aims at discussing some theoretical and practical issues concerning the methods employed by linguists when doing typological research, intended here as cross-linguistic comparison of linguistic phenomena both at the large-scale worldwide

and micro-typological level (i.e., on a specific geographic area or genealogical groupings of languages), proposing a different view and method (that I call *discourse-sensitive typology*). Comparing languages is not an easy task under several respects, and typological literature has, in some way, dealt with some of the issues that doing research in this field has raised, often finding very good solutions or at least creating vivid debates within the scientific community. Since its (recent) foundation (in the 1960s by Joseph Greenberg, at least in its modern sense), typology has incredibly developed its theoretical bases in all the different nuances, reaching a prominent position within linguistics. In fact, on the one hand, we assisted at remarkable theoretical advancements on how to define phenomena cross-linguistically (e.g., comparative concepts and the debate around them, see Dryer 1997; Croft 2001, 2003, 2022; Haspelmath 2007, 2010; Cristofaro 2009; among many others) or on how to explain typological generalizations (see e.g. Schmidtke-Bode et al. 2018), but also in terms of methodological tools (e.g., quantitative typological approaches, see for example the notions of *distributional typology* and *multivariate analysis* in typology as in Bickel 2015 and Bickel 2010, 2011 respectively, but also the methodological apparatus in Guzmán Naranjo & Becker 2022 and in Becker & Guzmán Naranjo 2025, just to mention a few of them); on the other hand, we still have some of these aspects that have basically been neglected or at least for which scarce advancements have been made. Among these, the reflections on the data and the phenomena on which typology bases its analyses remained underexplored. Even though I do not believe in distinguishing between qualitative and quantitative research, at least in typology (since every research has some qualitative and quantitative nuances at the same time), we can say that advances have been made for what concerns the methods in "quantitative" typology, as just mentioned, but "qualitative" typology remains stuck in its position for what concerns the methodological debate.

After introducing some preliminaries in section 2, section 3 aims at discussing the types of data and some related issues for typological research. First, the different possible sources that typologists usually work on are listed and presented (section 3.1) showing how these can be problematic for the typological analysis (section 3.1.1). Second, I briefly discuss the phenomena on which typologists focus their attention (section 3.2) exemplifying how typology could benefit from also looking beyond at less canonical phenomena (section 3.2.1). In section 4, I propose a multi-level method to overcome the issues discussed in previous sections and that would allow typology to become a discourse-sensitive field. Section 5 draws some conclusions.

2. Preliminaries for a *discourse-sensitive typology*: the *why* and the *how*

As already pointed out, the goal of this paper is to propose a discourse-sensitive typology discussing the role that discourse should acquire in typology both in terms of discourse-informed analyses of grammatical phenomena and of linguistic analyses of discourse phenomena.

Linguistic typology is usually defined as "the study of linguistic systems and recurring patterns of linguistic systems" allowing typologists to identify universals which "are typological generalizations based on these recurring patterns" (Velupillai 2012: 15). Croft (2003) identifies at least three different senses that the term typology can have in linguistics: (i) typological classification, (ii) typological generalization, and (iii) typological approach which he defines as follows:

1. ***Typological classification***: "a classification of structural types across languages. […] [A] language is taken to belong to a single type, and a typology of languages is a definition of the types and an enumeration or classification of languages into types" (Croft 2003: 1).
2. ***Typological generalization***: "the study of patterns that occur systematically across languages. […] The patterns found in typological generalization are language **universals**" (Croft 2003: 1, bold in the original).
3. ***Typological approach***: "typology represents an approach or theoretical framework to the study of language that contrasts with prior approaches, such as American structuralism and generative grammar" (Croft 2003: 2).

These three senses represent the three phases of the scientific method: collect and classify the data (*typological classification*); analyze the data, identify possible recurrent patterns and propose generalizations (*typological generalizations*); and account for some patterns which are cross-linguistically recurrent and why generalizations are found (*typological approach*). Croft (2003: 2, bold in the original) himself recognizes typology as "an **empirical scientific** approach to the study of language". As a natural consequence of its strong empirical foundation and its functionalist approach (see Croft 2003: 2), in the last decades, the usage-based approach has acquired more and more relevance in typological(-functionalist) literature. According to this approach, grammar is shaped by (and in some way it adapts to) usage in discourse (cf. Givón 1979a, Du Bois 1985, Bybee & Beckner 2010,

Diessel 2019, among many others). In other words, as Bybee (2006: 730) puts it, "[u]sage feeds into the creation of grammar just as much as grammar determines the shape of usage". This usage-based perspective allowed linguists from different fields to re-consider some grammatical phenomena, like, for example, syntactic ones (see the notion of *spoken syntax* as proposed by Hopper 1987, 1988). However, this perspective has not been fully adopted by all the approaches to the study of language and grammar, including some that share this functionalist view.[1] So far, these approaches have not (at all or not much) paid the required attention to spoken language and to phenomena that mainly pertain with the discourse level, and this is because they have mainly focused on "grammar", leaving aside "discourse". Linguistic typology is one of these disciplines. This is in some way surprising, and I think typologists should address this issue to make typology truly usage-based.

3. What data and what phenomena for typology

The situation described in section 2 seems to originate from two main reasons (at least in "qualitative" typology): (i) the lack of a full awareness of the kind of data typologists usually analyze, and (ii) the array of phenomena that typologists usually investigate. For what concerns the former, typology seems not to have always been fully aware of the nature of the data on which typological investigations are based; while, for the latter, typology, as already noted, tends to give particular attention to some levels of analysis (morphology and syntax) at the expense of others (e.g., phonology and discourse). Needless to say, these two issues are strictly connected with each other since the kind of data typologists analyze in some ways "mirrors" the phenomena investigated, and vice versa (see sections 3.1.1 and 3.2 below). The following sections will focus on these two issues, showing how they can be problematic and why they should be addressed.

3.1. Typological data and their issues

Traditionally, typologists retrieve their data from a range of different sources. Among the most commonly adopted, we can list the following: (i) questionnaires specifically

[1] Actually, this view was originally adopted in some typologically-oriented works by exponents of the so-called "West Coast functionalism", like Wallace Chafe, Talmy Givón, Sandra Thompson, and Marianne Mithun (see, e.g., Chafe 1976, 1987, 1994; Givón 1979a, 1979b, 1983, 1984; Mithun 1996, 2015; Thompson 1988), but it did not fully make it into large cross-linguistic investigations.

designed for the purposes of the investigation; (ii) parallel texts; (iii) dedicated scientific papers; and (iv) grammatical descriptions. In what follows, the merits and the imperfections for each of these kinds of data are briefly presented.[2]

Typological questionnaires

The first kind of data are questionnaires. Two sub-types can be identified: questionnaires to be submitted to native speakers and questionnaires to be submitted to linguists. The former is usually composed of a series of sentences to be translated by speakers from a meta-language (e.g., English) into the target language. An example is provided by Dahl's (1985) questionnaire for eliciting data for TAM categories composed of 156 sentences to be translated, as those reported in (1):

(1) Sentences 61-63 of Dahl's (1985) TAM questionnaire (Dahl 1985: 200-201):
   61. [It is cold in the room. The window is closed. Q:] You OPEN the window (and closed it again)?
   62. [Answer to (61):] (Yes,) I OPEN the window
   63. [Answer to (61):] (No,) I not OPEN the window

The sentences reported in this kind of questionnaire usually provide some context (e.g., in (1) the context is between squared brackets), and the uppercase form is the one of interest for the research. This form usually appears in its citation form in order to avoid biased answers from the informant towards the meta-language. So, for example, sentences 61-63 would be the following if our target language is Italian (ita; Indo-European, Italic):[3]

(2) Italian translation of Dahl's (1985) sentences 61-63 (own knowledge):
   61. [It is cold in the room. The window is closed. Q:] ***Apri** la finestra (e la richiudi)?*
   62. [Answer to (61):] *(Sì,)* ***apro** la finestra*
   63. [Answer to (61):] *(No,) non* ***apro** la finestra*

[2] The list is not intended as necessarily exhaustive, but it reports the most common types of sources and briefly discusses them.

[3] Language classification follows the one proposed in Glottolog 5.2 (Hammarström et al. 2025).

Thus, in this way, we would have data for the Italian Present Tense for the verb *aprire* 'open' and similarly for all the languages for which we have informants to answer the questionnaire.

The second type of questionnaire consists of a series of direct questions about the linguistic properties of the target language that the investigator submits to another linguist. For example, see (3):

(3) Questions 1-2 of Corbett's questionnaire on grammatical number (emphasis in the original):[4]

1. Which grammatical numbers are distinguished (singular-plural, singular-dual-plural, etc)?
2. How is number expressed?
   2.1 lexically: are there separate words meaning, say 'plural'? (It would be surprising to find such cases in Europe.)
   2.2 morphologically
      2.2.1 which means are used?
         inflectional: prefixing, suffixing, infixing, ambifixing
         other - suppletion, reduplication
      2.2.2 which lexical categories carry the morphological markers - nouns, verbs, adjectives, pronouns, others?
      2.2.3 within the lexical categories, are all items involved? (i.e. if adjectives mark number, do all or only some adjectives mark number?) It is common for nouns to be defective (singularia tantum, pluralia tantum etc); if so which are involved? Sometimes there are types of noun (e. g. abstracts, mass nouns) which can be predicted to be defective.
   2.3 syntactically
      2.3.1 is there a matching of number marking between different elements (especially in the NP) which could be characterized as agreement?
      2.3.2 if so, are there instances where such matching is violated (e.g. English The committee have decided, Norwegian Pannekaker er godt 'Pancakes is good')?
      2.3.3 are there particular syntactic complications where numerals are involved?

In this way, the researcher would have at her own disposal the required information on the phenomenon she is investigating from linguists who are expert of some languages.

Both these types of questionnaires have some merits but also some shortcomings. First, the questionnaires allow the gathering of data directly on the target phenomenon with specific information on the structures. On the one hand, the first

[4] This questionnaire can be found at the following website: https://www.eva.mpg.de/lingua/tools-at-lingboard/questionnaire/gender-and-number_description.php.

type (questionnaire for speakers) also allows the researcher to see how the specific phenomenon works in the language in controlled situations and specific pre-established contexts. On the other hand, the second type (questionnaire for linguists) allows to have a thorough and informed picture. Coming to the critical issues, the questionnaires can be filled out only by a limited number of people since it is difficult to submit it to large number of speakers/experts in a reasonable time (assuming that she is able to find enough of them) and, thus, the language sample will be relatively small (we can assume about 100 languages). In addition, the first type of questionnaire has a higher possibility to be submitted to informants of languages with a large number of speakers biasing the research towards WEIRD and LOL languages (WEIRD = Western, Educated, Industrialized, Rich, and Democratic, see Henrich et al. 2010 and Majid & Levinson 2010; LOL = Literate, Official, and Lot of users, see Dahl 2015). Then, this kind of data might suffer from influences by the meta-language due to the translation process: no matter how hard we try to avoid influences of this type, it is not possible to eliminate them. Also, in the first type of questionnaire, the context is *de facto* established by the researcher thus eventually biasing the research towards parameters of analysis adopted by the researcher herself. The second type of questionnaire does not provide actual data but rather analyses of data heavily dependent on the linguist's (possibly a non-typologist) own conceptualization of the phenomenon and thus without the possibility to check it through data analysis of any sort.

Parallel texts

The second kind of source typologists usually adopt are parallel texts. Cysouw & Wälchli (2007: 95) provide the following definition: "[p]arallel texts are texts in different languages that can be considered translational equivalent". They also introduce the notion of *massively parallel text* (MPT) "for such texts of which many different translations are available" (Cysouw & Wälchli 2007: 95). In other words, (M)PTs are different translations of the same text in different languages. Several cases of texts are largely translated, just to cite some of them: the Bible, *The Little Prince*, the Harry Potter saga, subtitles of movies, etc. Of course, this kind of data can be easily used to fulfill typological investigations, and they have both very good merits and some relevant shortcomings. First, MPTs allow the typologist to observe exactly the same occurrences of the phenomenon under investigation in the exact same co-text/context for different languages, and this represents a very strong quality. Second,

MPTs are usually digitalized and tagged texts that can be easily used for typological-oriented queries (and beyond). Third, we have at our disposal MPTs for several different languages according to the nature of the original text: from some dozen of languages (e.g., movie subtitles) to a few hundred (*Universal Declaration of Human Rights*, *The little prince* or Andersen's fairy tales), or even thousand (the Bible or part of it) (for a more precise count see Cysouw & Wälchli 2007). However, at the same time, MPTs are heavily biased by the translation since (often) the objective of the translator is not to provide a linguistically equivalent in two (or more) different languages, but rather to provide the more accurate semantic/sense equivalent of the intentions of the authors of the text. This issue can ultimately question if the "same" occurrences in different languages are actually the same, or they are rather "similar" occurrences with almost the same general meaning. This also opens another theoretical problem: are two structures and/or contextual/co-textual environments of two different languages equivalent at all? The answer is not easy to give and, at least for those who advocate for the non-existence of pre-established categories and constructions for typological research (see the "comparative concepts" debate), it usually seems to be "no".[5] However, this is a problem that holds for any kind of source, even though it might have a different impact on the analysis. Another relevant problem is that often parallel texts provide only the original text and its translation without any interlinear gloss making it difficult to pursue more fine-grained linguistic analyses (in particular, morphological and syntactic ones). Finally, MPTs are usually texts written in highly codified language, that is, they are mainly literary texts or, in the best-case scenario, texts that try to mirror spontaneous speech (e.g., subtitles).

Dedicated scientific papers

The third kind of data on which typologists can base their analyses are scientific papers. With this, I refer to papers written by linguists and appeared on scientific (international or not) journals which aim to describe a specific phenomenon in a single language or a group of languages (often from the same language family or geographic area). The main merit of this kind of data is that scientific papers tend to

[5] This position is of course debatable, but since this is not the objective of the present paper to discuss it here, I just refer to the main bibliographic references (already mentioned above in section 1) on it and the references cited therein to which I refer for further information (Dryer 1997; Croft 2001, 2022; Haspelmath 2007, 2010; Cristofaro 2009; among many others). For my purposes, it is only important to note the existence of different views on this that ultimately represent a possible issue.

be very specific about the object of analysis and to report (almost) all the information needed to account for a specific phenomenon in that/those language/s. Sometimes, the information that scientific papers provide is even too specific for typology, whose final aim is to simplify linguistic complexity to identify recurrent patterns and provide them with an explanation. Scientific papers also have a relevant issue for typological research: there will never be enough papers dealing with the object phenomenon for a sufficient number of languages. In other words, the number of languages for which we can find information in dedicated papers is not high enough to be accounted for as a(n actual) language sample. This kind of source can eventually be considered complementary support to the adoption of other sources.

Grammatical descriptions (or descriptive grammars)

The last type of data that typologists have at their disposal are grammatical descriptions (henceforth simply grammars). Grammars are probably the most used and the most traditional source for typological investigations. They consist of thorough descriptions of the grammatical structures of a language made by field linguists who collect texts (of different genres)[6] by speakers of the object language and provide an analysis of the structures they are able to identify in such a sample of texts.

As with all the other kinds of data, grammars have merits and shortcomings. The first merit is undoubtedly the fact that we have at our disposal grammatical descriptions of some kind for a large number of languages. According to Glottolog 5.2 (Hammarström et al. 2025), we have at least a full grammar or a grammar sketch for about 4.900 languages of the world, out of which at least ca. 2.900 languages have a full grammar. If we take for granted the overall number of about 7.800 languages of the world, we are talking about 63% of the total amount of languages.[7] Thus, this source would allow us to have incredibly extensive language samples (i.e., thousands of languages), a coverage that other kinds of data can hardly achieve. Another merit of grammars is that they provide a(n) (presumptive) overall description of the grammatical structures of a language, thus allowing to check for possible correlations among different constructions. Finally, one last value is that, quite often (at least in

[6] With the terms *text* and *genre*, I intend (here and elsewhere) them in their widest senses, that is, any possible type of linguistic text: written or oral, formal or informal, narrative or conversational, and so on.

[7] Of course, the figures are approximated for convenience, for the current figures see https://glottolog.org/langdoc/status.

grammars published in the last couple of decades), they also have a sample of (glossed) texts (at the end of the grammar) that allow us to look at the phenomenon/a we are investigating directly in texts. However, it is worth mentioning that, unfortunately, these sample texts often tend to be quite small.

Grammars also have some shortcomings. We can mention at least two of them that are relevant to the present discussion. First, the information we find within grammars are (at least) second-hand data, that is, they are not raw data to be analyzed by typologists, but they rather are data already analyzed by the grammarian (this is true also for other kinds of data, such as the one from questionnaires for linguists or dedicated scientific papers). This means that typologists must trust grammarians and their descriptions without debating much on what they find since they are not language experts. Of course, the grammarian might be contacted for further information, but this is not always possible or not for all the doubts that a typologist might have. Then, a follow-up of this shortcoming is that field linguists cannot have in mind all the questions that typologists would like to answer. This means that we cannot necessarily presume to find what we are looking for in a grammar or enough information on it and this does not necessarily mean that that specific language lacks the phenomenon we are investigating, but it simply means that the grammarian did not write anything about it because of a series of possible reasons such as the phenomenon is not relevant for the language or is categorized differently in the linguistic tradition of the language or any other possible reasons. All these reasons are valid and do not underestimate the immense work that fieldworkers do: typology would not exist without fieldwork. In the end, grammars cannot be considered by definition fully exhaustive for typological purposes, and this is because of the distinction between descriptive categories of individual languages and comparative concepts, two notions that are "in a many-to-many relationship" (Haspelmath 2010: 665) with each other.

The "perfect" source for typology

As already noted in the previous sections, none of the kinds of data presented above can be considered completely satisfactory. All of them show some shortcomings: some are more relevant or difficult to be faced, others are less. In any case, the fact that none of the sources is fully appropriate for the purposes of a thorough typological investigation clearly emerges from what has been shown above. This does not mean that typological research cannot be carried out, but rather that the data typologists analyze are not necessarily exhaustive. The most important consequence is that we

(typologists) must be aware of this and bear it in mind, in particular when proposing generalizations. Otherwise, typological data might bias our generalizations and, more in general, our research.

Picking one of these sources instead of the others is just a matter of personal choice and preference or of better adaptation of the source to the aims of our research. It is beyond the purposes of the current paper to discuss which is the best source of data, but I think that a personal positioning is in order in a paper like the current. In my opinion, descriptive grammars remain the best choice for several kinds of typological investigations and this is because they tend to be free from translational biases (much more than other sources) and pay the right tribute to language experts and field linguists who, by describing the languages, disclose their access to typologists thus allowing us to do cross-linguistic investigations.

The "perfect"[8] source would be to have at our disposal a significant corpus of (glossed) texts for a very large number of languages of the world, alongside a grammatical description. In this way, we would be able to investigate our object of analysis directly in the texts of the given language (preferably different textual genres) also having its context of use and co-text. In addition, the grammatical description would serve as a complementary tool to decipher and better understand the structures of a language in which we are not experts. Unfortunately, to date, this source is not available. Yet, in the last decade or so, a few projects have been carried out aiming to fill this gap. I am referring to projects like: CorpAfroAs (see Mettouchi et al. 2010, Mettouchi et al. 2015) which is specifically dedicated to Afroasiatic (AA) languages and contains texts and grammatical information for 16 AA languages; DoReCo (Seifart et al. 2024) which is a collection of spoken language corpora for 53 typologically different languages; and MultiCAST (Haig & Schnell 2023) which collects comparable corpora of 20 languages. Even though these projects are extremely important and praiseworthy, the final goal is still far, and probably a lot of time will pass before having this kind of source for a large number of languages.

3.1.1 How typological data may affect typological analyses

Let us now turn to how the kind of data adopted might impact the typological analysis. In the great majority of the cases, typologists use data from grammatical descriptions (as already noted above): grammars, being the best available source or

[8] Since talking about "perfectness" in science is always rather utopistic, I decided to use this term only between double quotation marks.

not, are still the most widely adopted source for typological research. In the best-case scenario, in descriptive grammars, we find data that are out of their narrative and/or pragmatic context and generally used by the grammarian to explain and exemplify some kind of phenomena of the object language in the specific paragraph/chapter of the grammar itself.

The lack of context (and co-text) can give rise to a possible 'partial' reading of the data that can ultimately lead to a possible misinterpretation of the data themselves. To interpret at best examples of phenomena found in grammars, the researcher should infer what in cognitive linguistics is called *construal*, that is, how "an experience is framed" in the sense of "how the speaker conceptualizes the experience to be communicated, for the understanding of the hearer" (Croft & Cruse 2004: 19). The narrative context/co-text[9] is indeed part of the construal since it would determine the conceptualization and the understanding of the linguistic structure (in its widest sense). From this, it follows that having an example of a specific phenomenon without its narrative context/co-text is like having incomplete information that ultimately can affect the interpretation. This, in turn, can lead to a possible biased typological generalization (if not even wrong): if typologists analyze a single (or a few) example(s) for the phenomenon they investigate in a specific language and this example might be wrongly interpreted given the lack of narrative context/co-text, then it follows that the overall understanding of the phenomenon might eventually be biased. Needless to say, this is a key issue for typology.

In the remainder of this section, I will illustrate through a couple of examples, based on the analysis on the pluractional marker *-pödï* of Akawaio (ake; Cariban, Venezuelan Cariban) in Mattiola & Gildea (2023), how this issue might impact typological analyses. With pluractional marker (PM), I intend any morphological strategy encoding a plurality of the event expressed by the verb (cf. Mattiola 2019: 164). Cross-linguistically, PMs can express a wide range of different functions, going from iterative to habitual and from durative to reciprocals (see Mattiola 2019: 21-42). As for any typological work, identifying the correct function for a formal strategy is fundamental. As noted above, to interpret the function of a strategy, we need to understand the construal of the construction and, specifically for PMs, the *event construal*: how the event is conceptualized and then expressed from a linguistic point

[9] With *narrative context*, I refer to the general context in and through which information (narration, in its widest sense) flows to be communicated. With *narrative co-text*, I intend the particular linguistic context in which a specific construction is inserted.

of view. This means we need to analyze the semantic-functional value that the PM adds to the lexical value of the verb and how it impacts the overall semantics of the example.

The PM *-pödï* of Akawaio can express several pluractional functions that we found in other languages of the world (Mattiola & Gildea 2023: 463-470). Let's now try to understand the functional reading of this marker in a couple of examples. Consider (4).

(4) Akawaio (Cariban, Venezuelan Cariban; Caesar-Fox 2003: 336)
*yöi naga'pï po y-enggurumï-**bödï**-Ø-ng*
tree stump on 3-wait-**PLAC**-PRES-STYLE
'He **would just rest** there on top of a piece of tree stump'

In (4), we may consider *-pödï* as expressing a continuative function ("a single situation that is prolonged during a period of time" Mattiola 2019: 34): the subject ("he") seems to be in the state of resting on the tree stump for a long time. Another possible reading might be to understand *-pödï* to express a wish, a sort of optative marker. However, this interpretation contrasts with the cross-linguistic understanding of PMs and thus appears to be less likely. But if we consider the narrative context the picture changes. The sentence in (4) is taken from a narrative story in which the main characters are a duck and an owl. Every day, both of them leave their own houses, but while the duck goes outside to go to work, the owl goes outside to rest on a tree stump waiting for the end of the day. Thus, a different interpretation seems to consider the state of resting as repeated over a long past time frame ("every day the owl rests on a tree stump"), that is, a habitual function ("a situation that is repeated customarily, i.e. that is typical of a period of time" Mattiola 2019: 31-32).

A similar situation is found when interpreting the sentence in (5).

(5) Akawaio (Cariban, Venezuelan Cariban; Caesar-Fox 2003: 522)
*ö'rö pe y-eji-Ø y-aburö-**bödï**-Ø*
what ATTR 3-be-PRES 3-praise-**PLAC**-PRES
'Why **is** she **being praised**?'

In this case, the PM might be interpreted as giving again a continuative reading or, eventually, an iterative reading to the situation ("the situation occurs multiple times,

but the repetitions are limited to a single and the same occasion" Mattiola 2019: 23). The sentence is a question in which the speaker asks the reasons why a girl/woman is being praised and we might interpret it as the action occurs several times (iterative) and/or prolonged through time (continuative). However, the example is found in a personal narrative in which the informant is reporting a party for the retirement of a nurse. Again, given the full contextual environment, the first-glance interpretation might be erroneous. Also in this case, it is more likely a different construal of the situation. In fact, a more well-suited interpretation seems to be participant plurality ("the plurality of situations will be distributed over different participants" Mattiola 2019: 26) or, but less likely, a frequentative reading ("the repetitions of a specific situation are performed over multiple and different occasions" Mattiola 2019: 24): several single events of praising in which several people congratulate the nurse for the retirement or a single participant (but also several) that congratulate with the nurse on different occasions during the party (frequentative). In any case, the first interpretation turned out to be wrong.

From these two very short examples, it clearly comes out the problem to which I was referring at the beginning of the current section: how data typologists collect and analyze might lead to possible misinterpretation if we do not take into consideration their narrative context. Again, this is a highly problematic issue because an incorrect (or wrong) interpretation of the data can result in a biased analysis and thus lead to incorrect (or wrong) generalizations.

In section 4, I will be back on how to try to avoid this issue with the data we have at our disposal.

3.2. *Investigate 'em all*: discourse phenomena in typological perspective

The second issue that deals with typology as a truly usage-based field concerns the range of phenomena typically examined by typologists. Often, typological investigations focus their attention on mainly "structural" and "grammatical" phenomena, that is, on morphology and syntax with phonetics/phonology and lexicon being much less represented in typological studies. Just as an example, suffice it to say that in the online version of the pioneering WALS project (Dryer & Haspelmath 2013), one of the most important sources and enterprise in the history of typology, out of the 144 chapters (to date), we can count 54 and 57 chapters respectively for morphosyntax (10 morphology, 28 nominal categories, 16 verbal categories) and

syntax (7 nominal syntax, 19 word order, 24 simple clauses, 7 complex sentences), 19 for phonology and 10 for lexicon (plus 2 for sign languages and 2 classified as "others", i.e., clicks and writing systems). The picture is thus quite straightforward: typology heavily prefers morphology and syntax over the other levels of analysis. On the one hand, we can, of course, account for this imbalance making reference to the historical relevance that morphology and syntax have for typology since its foundation. On the other, however, the imbalance is impressive and cannot be easily dismissed by referring to historical reasons and tradition. The picture becomes even harsher if we consider what is generally called "discourse phenomena", which, following Barotto & Mattiola (2023a), are defined as:

> linguistic elements and constructions that help to manage the organization, flow and outcome of communication (cf. Schiffrin 1987; Du Bois 2003). They have to do with what can be called information packaging, that is, the ways speakers organize their discourse and turns, link and give cohesion to utterances while clarifying the relationship occurring between them or foreshadow the status of information in what will come after. Moreover, discourse phenomena can also facilitate the intersubjectivity between speakers, as in the case of politeness or hedging. (Barotto & Mattiola 2023a: 1)

In fact, except for a few notable and recent cases (e.g., several papers on different phenomena in Barotto & Mattiola 2023b; Dingemanse 2012 and Lahaussois et al. 2024 on ideophones; Lahaussois & Treis 2019 and Ponsonnet et al. 2023 on interjections; Pakendorf & Rose 2025 on fillers), typology has almost completely ignored discourse phenomena. Studies focusing on discourse phenomena are pretty rare both at the cross-linguistic level and at the language-specific level for non-WEIRD and non-LOL languages. This is something that is not in line with the view that I gave in section 2, for which typology understands itself as a usage-based field (grammar is shaped by discourse use). In my opinion, we (typologists) must realize that giving the right importance to all the levels of analysis and in particular to discourse phenomena (which are by no means less relevant than morpho-syntactic ones) and their cross-linguistic variability can provide important information allowing us to better understand how languages shape their own grammar and why they are organized in such a way. Of course, this is not simple at all, and I will be back on how to do this in section 4. In the next section, I am going to show how typology (and discourse studies) can benefit from studying discourse phenomena in cross-linguistic

perspective and/or in non-WEIRD and non-LOL languages. To do so, I will briefly analyze a discourse construction found in Akawaio (Cariban, Venezuelan Cariban).

3.2.1 A discourse marker in a non-WEIRD/non-LOL language: ti'tuik prarö *in Akawaio*

Akawaio is a variety of Kapóng,[10] a Cariban language belonging to the Venezuelan branch of the family (Pemón sub-branch), spoken by the Akawaio tribe (circa 10.000 people) in Guyana, South America. The corpus I consulted (found as an appendix to Caesar-Fox 2003) is composed of about 10.800 words and was collected, analyzed, and glossed by Desrey Caesar-Fox (with Spike Gildea). The corpus is composed of twenty-seven texts belonging to different genres: traditional stories (12), personal narratives (5), Tareng healing chants (6), and traditional praising rhymes for children (4). In this corpus, I came across a construction, *ti'tuik prarö*, whose formal and functional properties revealed to be peculiar. For this reason, I decided to thoroughly investigate it, but only 13 occurrences were identified in the corpus. So, what follows represents just a tentative analysis that cannot be corroborated without referring to additional evidence. Despite this, I think it deserves to be examined exactly because of the reasons discussed above and, also, because it clearly shows the points made in the previous sections on the relationship between typology and discourse phenomena.

From the formal side, *ti'tuik prarö* construction is composed of two elements: a participial form of the verb 'know' (which is circumfixal, *t-V-ik*, like in proto-Cariban **te-V-ce* as reconstructed by Gildea 1998: 140-151) and a negative emphatic particle (6). The overall construction is literally translated as 'not knowing(ly)'.

(6) Morphological structure of *ti'tuik prarö*:

ti'tuik	***prarö***
t-i'tu-ze	*bra-rö*
ADV-know-PTCP	NEG-EM

'lit. not knowing(ly)'

[10] According to Glottolog 5.2, Akawaio is now considered a language and no longer a variety (the language is currently named Akawaio-Ingariko). However, since this contrast with the general view on Akawaio by experts and since this is not the objective of this paper to discuss the linguistic status of Akawaio, I still refer to it as a variety of Kapóng.

At the syntactic level, *ti'tuik prarö* can be found in four different positions in texts: (i) as a clause modifier (adverbial adjunct) (7a), (ii) as a V(P) modifier (7b), (iii) as a N(P) modifier (7c), and (iv) in autonomous position at the end of the clause (7d).

(7) Syntactic position of *ti'tuik prarö* (Caesar-Fox 2003: 317, 485, 310, 430):

a. ***t-i'tu-ze*** ***bra-rö*** *mïgï* *agïdï-bödï-bök* *y-eji*
ADV-know-PTCP NEG-EM this cut-PLAC-PROG 3-be
'**Unknowingly and without care**, he would cut up whatever he wants to'

b. *meguru yek* *pömï-u-ya* *chigaru* *yek* *pömï-u-ya*
banana plant plant-1-ERG sugar.cane plant plant-1-ERG
t-i'tu-ze ***bra-rö*** *Ø-e'-pömï* *ibïra* *rö*
ADV-know-PTCP NEG-EM 1-DETR-plant with.no.doubt EM
'I plant banana plants, I plant sugar cane plants, (or) I plant **anything else(/without knowing)**, no problem'

c. ***t-i'tu-ze*** ***bra-rö*** *murang* *eji* *mörö kwaro'nai* *kubi'ta*
ADV-know-PTCP NEG-EM charm be AI(?) ginger.charm herb.charm
kwak *ta* *tok* *ya*
charm(sp.) say 3PL ERG
'There are **numerous (types of)** charms, 'kwaro'nai, kubi'ta, kwak, so they say'

d. *ane* *ji* *an-egama-gö* *pandong* *wayamori* ***t-i'tu-ze*** ***bra-rö***
wait.IMP EM 3O.IMP-tell-IMP story turtle **ADV-know-PTCP NEG-EM**
'Please tell a story then, about the turtle **or anything**'

At the functional level, the situation is quite complex, mainly because of the scarcity of occurrences in the corpus. However, I was able to identify at least three different functions: (i) when *ti'tuik prarö* is used as a general extender[11] (see (7d) above); (ii) when *ti'tuik prarö* is used to express an heterogeneous set of entities (similar to an hypernym or a category label, see (7c) above); (iii) when *ti'tuik prarö* is used to express manner (see (7a) above). To these, I also identified a case in which the function is ambiguous, this is exemplified in (7b) where *ti'tuik prarö* can both have a GE or a manner reading.

[11] A general extender (GE) is generally defined as "a form that indicates additional members of a list, set, or category [and that combines] with a named exemplar (or exemplars)" (Overstreet 1999: 11), e.g., *and the like, and things like that, and so on, etcetera*, etc.

Despite the scarcity of occurrences, we can try to go further and propose a tentative path of evolution for the *ti'tuik prarö* construction (8) mainly based on the formal and functional properties just shown and on their frequency in the corpus, also in comparison with other constructions that are similar to *ti'tuik prarö*. Of course, this proposal must be conceived just as an attempt rather than an actual empirically-based diachronic scenario.

(8) Tentative path of evolution of *ti'tuik prarö*:[12]
Mod C (adverbial adjunct - manner) > Mod V(P) (adverb - manner, but also GE) > Mod N(P) (after a list of elements, heterogeneity or GE) > Autonomous (after a list, GE)

The first stage would probably be *ti'tuik prarö* as a clausal modifier encoding manner. This is, indeed, the original position and function of the **ti-V-ce* construction as reconstructed in Proto-Cariban by Gildea (1998: Ch. 8). I found only one occurrence (out of 13) of *ti'tuik prarö* used in such a way. I also checked other *ti-V-ze* constructions in the Akawaio corpus, and I found 12 occurrences (out of 27) of other participles with the same structure showing these function and syntactic position. The second stage would consist of *ti'tuik prarö* used as V(P) modifier (another original position of **ti-V-ce* construction in Proto-Cariban) expressing again manner, but eventually also employed as a GE. In this case, I found 2 occurrences (out of 13) in the corpus, and I also found 9 occurrences (out of 27) of other participles with *ti-V-ze* structure displaying this function/position. The third stage would predict *ti'tuik prarö* as N(P) modifier after a list of elements conveying what I called heterogeneity (i.e., a heterogeneous set of entities, hypernym/category label). The occurrences of *ti'tuik prarö* with this function/position are 4 in the corpus (out of 13). I found 5 occurrences (out of 27) of other participles with *ti-V-ze* structure that have this function, but these cases do not modify a N(P), strictly speaking, but a nominalized verb mainly with a predicative function (thus a much more verb-like entity). The fourth and final stage would suggest *ti'tuik prarö* in an autonomous syntactic position at the end of a clause (after a list) used as a GE. I found 6 occurrences (out of 13) of this case and no occurrences (out of 27) of other participles with the same structure of *ti'tuik prarö*

[12] This path of evolution was discussed by the author with Spike Gildea that I would like to thank for the generous support. However, all possible mistakes, misunderstandings, and misinterpretations (if any) must be considered solely mine.

showing these functional-syntactic properties. To sum up, it might be probable that *ti'tuik prarö* started being a truly participial form modifying clauses or V(P)s with an adverbial function (with some VP also like a sort of GE). These two situations are the less frequently found for *ti'tuik prarö* (3 out of 13) but the most common for other *ti-V-ce* constructions (19 out of 27). Then, it started being used as a modifier for N(P)s, maybe with situations employing nominalized verbs or nouns with a predicative function as a bridging context and then extended to "true" nominals (through reanalysis) expressing heterogeneity (because of its negative element). This situation is quite frequent for *ti'tuik prarö* (4 out of 13) but much less for other participial constructions (5 out of 27). Finally, *ti'tuik prarö* might have been reanalyzed as a GE, which is the most common situation for *ti'tuik prarö* (6 out of 13) and is not attested with other *ti-V-ce* constructions (none out of 27).

Of course, the evidence for such an evolution path is scanty and relies a lot on frequency in the corpus, which is ultimately based on very low figures, not allowing me to draw strong conclusions. However, the increasing frequency figures of *ti'tuik prarö* when going on the right along the path and the concomitant decreasing of occurrences of similar constructions are compelling evidence and, in my opinion, makes the path in (8) at least plausible.

In conclusion, we can say that Akawaio *ti'tuik prarö* is presumably developing as a particular discourse marker, i.e., a general extender. At the typological level, GEs are defined as elements with (Mauri & Sansò 2017: 65, my free translation):

> associative referential function […]. We can describe this function referring to three kinds of entities to which GEs make reference to:
>
> i. one or more explicit exemplars,
> ii. additional non-explicit elements X, that are associated to exemplars according to a shared property that is relevant for the context,
> iii. a wider category that includes both explicit exemplars and implicit additional elements X

We find all these properties in several of the situations in which *ti'tuik prarö* is found.

The *ti'tuik prarö* construction seems to be peculiar to Akawaio solely within the language family: I quickly consulted some corpora and grammars for other Cariban languages looking for similar constructions and I did not find any of them (also confirmed by Gildea p.c.). However, Trió (tri; Cariban, Guianan) displays a quite similar and interesting construction: *ookinenpen* (that.INAN-PST-CONT) 'all kinds of things (lit. those things that used to be)'. Even though this construction seems to have a

different function (rather heterogeneity – 'all kinds of X' – than a true GE), I only found 2 occurrences in the Trió texts and, in addition, this probably has a verbal origin, too, but not a cognate of *ti'tuik prarö*, leaving open possible speculations over its evolution path.

This very brief case study is particularly important for several reasons. The *ti'tuik prarö* construction represents an analytical strategy in which we cannot identify any connective, which is a typical property of GEs as described in the cross-linguistic literature[13] as proposed by Mauri & Sansò (2017: 66) for GE structure: connective + indefinite/generic element + similative element. In addition, *ti'tuik prarö* has a verbal origin that represents a totally new source for GEs cross-linguistically (cf. Mauri & Sansò 2017). At the same time, the pattern NEG + 'know' is not fully unknown for other types of discourse markers (e.g., *non so* 'don't know' in Italian, cf. Lo Baido 2020, or *I don't know* in English). Thus, this widens our understanding of GE in the languages of the world, showing how they might be more varied (both synchronically and diachronically) than attested so far, but also shows how there might be some commonalities between different discourse markers in different languages.

This case study might represent an important starting point. On the one hand, typologists must realize that giving the right importance to discourse phenomena and their cross-linguistic variety can provide important information, allowing us to account for possible evolution path involving different level of analysis (e.g., morphosyntax and discourse, like in the case of *ti'tuik prarö*) and thus better understand how languages shape their own grammar and why they are organized in such a way. Still, it would also allow typology to adopt a truly usage-based approach. On the other hand, discourse studies should be aware of and investigate the cross-linguistic variety (thus in non-WEIRD and non-LOL languages, too) of discourse phenomena since this can help in better describing, understanding, and assessing them also in single and specific WEIRD and LOL languages.

4. How can we make data speak to us: A multi-level methodology

The aim of this paper is not solely to discuss some issues and criticize the methods and the data on which typology has been founded so far. I would also like to propose

[13] However, we must bear in mind that in Cariban languages connectives are not very frequent strategies, these languages tend to employ more frequently juxtapositions.

a possible method that would help solve, at least partially, some of the issues considered in the previous sections.

We saw how the best kind of data for typology would be having "primary" data along with grammatical descriptions (corpora on which the latter are based on) since these could allow us to look for/at linguistic phenomena (from morphosyntax to discourse) in their own context/co-text. This is exactly what the abovementioned projects (e.g., MultiCAST, DoReCo, CorpAfroAs, etc.) have tried/are trying to do. However, as already noted, the path is very long, and this is not fully viable right now. So, what can we do? In my opinion, the only possibility we have is to follow a multi-level *converging evidence* method. The converging evidence method consists of adopting a multifaceted perspective to investigate the object of analysis, that is, taking data and pieces of evidence from different kinds of sources and perspectives. This is what Mauri & Masini (2022) call "the 3D methodology", which "combin[es] Discourse analysis with cross-linguistic Diversity and/or Diachrony" (Mauri & Masini 2022: 101) to which I would add sociolinguistic and areal/contact information (if available). In other words, for what concerns typology, this means to analyze linguistic phenomena from a cross-linguistic perspective, also looking at their diachrony (possible sources and evolution paths) and their discourse properties mainly through corpus-based language-specific analyses. All these perspectives are mutually connected, and comprising all of them is fundamental to maximizing the possibility to catch and account for the whole complexity of what we are investigating.

Typology, by definition, simplifies language complexity (found in single languages) in order to identify recurrent patterns and propose generalizations that can ultimately provide hints on the cognitive organization of information. However, "simplification" should not correspond to "oversimplification",[14] which is the most dangerous risk for typology. This is because oversimplified data would end up in possible not well-suited (or even wrong) cross-linguistic generalizations and predictions. In order to avoid oversimplifications, typologists need to base their investigations over the wider range of evidence as possible.

[14] I am aware that identifying a boundary between "simplification" and "oversimplification" is indeed difficult. However, discussing such an issue is not an objective of the present paper, and, in addition, it is not fundamental to my purposes. This is because the boundary is not discrete and may vary depending on the phenomenon and the research objectives. Typology should, by definition, simplify complexity to the bare minimum just to identify generalizations (regardless of where we put the boundary), and to do so, interpreting the data as best as possible is fundamental.

The method I am proposing here goes exactly in this direction and is based on the *converging evidence* perspective. This method (which can be simply called *discourse-sensitive typological method*) consists of a cross-linguistic 'multi-level method' that allows one to maximize the possibility of finding data and, thus, let emerge the widest cross-linguistic variety while looking at phenomena also in their discourse environment.[15] This multi-level typological method consists of three different levels of investigation: (i) the horizontal level (large-scale typological investigation), (ii) the intermediate level (a more 'qualitative' typological investigation), and (iii) the vertical level (intralinguistic investigations, i.e., case studies).

Horizontal level

The first level consists of a horizontal investigation, that is, the traditional way of doing typological research. The researcher will investigate the object phenomenon through a balanced sample of languages, generally a variety (and convenience) sample, composed of about 250 or more languages. Large variety samples are preferred over the other types because they are specifically designed to maximize the degree of cross-linguistic variety giving more relevance to internal complexity at the genealogical level and less importance to the actual statistical balancing (e.g., the Diversity Value – cf. Rijkhoff et al. 1993 and Rijkhoff & Bakker 1998 – and the Genus-Macroarea – cf. Dryer 1989, Miestamo 2005, Miestamo et al. 2016 – techniques). However, the sample size and type may vary according to the objectives of the research.

This level is based on grammar mining, as it is generally called within typology, which consists of looking for data through the analysis of grammatical descriptions. However, some potentially useful and practical techniques may be adopted to identify all possible patterns. For example, picking the "best" grammar for a language is very important. In general, the best choice would be the grammar that is most recent, the most exhaustive,[16] and the easiest to find. Another useful tool is to create a list of terms and glosses to which the object phenomenon can be referred to in the grammatical description in order to look for them (in the table of contents, the

[15] This method has already been tested in some works by the author of this paper. A preliminary description and an application of this method to a specific phenomenon can be found in Masini & Mattiola (2019).

[16] It is not easy to identify the most exhaustive grammar *a priori* (i.e., without reading it), so, in this case, the number of pages should be considered as an indicative clue.

analytical index, etc.) and thus detect all the possible information available. For example, for reduplication, we might use the following terms: *reduplication*, *repetition*, *duplication*, *multiplication*, *serialization*, *doubling*, *iteration*, *RED/RDP/REDUP*, and so on. This is particularly useful in digital versions of grammars that usually are automatically searchable (alongside other techniques of data mining).

Through this level, the researcher will have a first survey on the presence of linguistic phenomena and on how they work in a large sample of the world's languages. However, the general 'imperfections' of large-scale typology are still there, such as not having much information on the discourse usage and properties of the phenomenon. Despite this, the traditional typological method (which allows us only to scratch the surface of linguistic complexity) is still fundamental to have a general picture of what languages of the world display. However, as pointed out above, this method alone cannot suffice for a discourse-sensitive approach, and other levels of investigation are needed.

Intermediate level

The second level involves more detailed typological investigations. In this case, the researcher will design a smaller sample of languages (not necessarily balanced) of approximately 20-30 languages. Again, the numbers must be conceived as approximations and rather should adapt to the objectives of the research itself.

This level is based on a more fine-grained analysis of grammatical descriptions (e.g., taking in consideration different grammars or sources, like dictionaries, if relevant) and, alongside this, it also requires the analysis of texts of the languages, such as those found at the end of descriptive grammars or made available by linguists specifically working on that language (freely available or directly asking for them to experts). This phase is mandatory since it allows the researcher to look for and observe the phenomena directly in the texts of the language (also with the narrative context and linguistic co-text at one own disposal), and thus find also patterns that are not described within the grammatical descriptions (e.g., discourse phenomena) or not identified by the grammarian (e.g., because of the non-overlapping between descriptive categories – described within grammars – and comparative concepts – comparative definitions).

This level allows us to get over the problems of 'traditional' typology, pointed out in the previous sections, by verifying directly in a corpus (even if a small one) what the horizontal level might have not brought out. In addition, through the intermediate level, discourse starts playing a crucial role both in terms of description and in terms of explanation.

Vertical level

Finally, the vertical level consists of a much more detailed linguistic analysis of a very small language sample (from 2 up to 5 languages) comprising languages as typologically different as possible. This level relies on analyses of (large) corpora made available by linguists who are experts in a language (possibly first-hand data and glossed) or are freely available (e.g., on Sketch Engine). In this case, the researcher must be an expert in the language or should count on the help of experts since these are truly corpus-based analyses. Through these analyses, the full linguistic complexity of the phenomenon emerges (even if only for a few languages), making (virtually) all the existent patterns able to be identified and analyzed in detail. At this level, possible information on the discourse, pragmatic, and sociolinguistic levels might emerge more clearly, allowing for a thorough account of the object phenomenon.

This level is extremely important for typological analyses mainly because of two reasons: (i) it allows us to verify the typological generalizations identified in the first two levels and (ii) it allows us to investigate phenomena in much greater detail, letting possible characteristics and patterns (and eventually also unexpected parallelisms among different structures) that (traditional) typology cannot detect emerge.

The three levels must be considered as strictly intertwined and mutually dependent with each other. This means that if a pattern and/or a parameter of analysis, not previously considered, emerges from one of the levels, then the researcher should go back and forth through the levels and implement the analysis of each of them by integrating such a pattern and/or parameter accordingly.

This three-level method should help us in the difficult process of collecting consistent data in a typological sample of languages on linguistic phenomena (both “grammatical” and “discourse” phenomena) and analyzing them thoroughly. In this way, we retrieve the more varied data we can find in the sources we have at our disposal. This would help us a lot in proposing consistent analyses and strong generalizations. In fact, each level requires the adoption of the converging evidence perspective *per se* comprising cross-linguistic, diachronic, sociolinguistic, and areal data. All these data would “speak” to us and with each other, making our investigation and findings as solid and informative as possible.

In my opinion, a consistent methodology, like the one proposed here, is in order in (qualitative) typology. Even though some typological works have already introduced some of the phases of this multi-level method, it is still extremely difficult to find

investigations that fully comprise and adopt in a methodologically consistent way the suggestions made here.

5. Conclusion

This paper focused on some aspects of the scientific research (the data and the methods on which the discipline bases its research) on which typological community and literature have hardly discussed in detail. I first presented some theoretical preliminaries based on functionalist views that understand linguistic typology as a usage-based approach. I then focused on the aspects of such a perspective that typologists often tend not to fully integrate into their research. More specifically, I discussed two issues. First, I presented the different types of sources on which typological works are usually based, showing how they can generate possible issues for the analysis. Second, I discussed how typology tends to focus its attention on phenomena of specific levels (morphosyntax), almost completely disregarding phenomena of other levels of analysis (e.g., discourse). This represents an important shortcoming for typology since its final aim is to provide a comprehensive account and explanation to cross-linguistic variety of linguistic phenomena. Also in this case, I supported this view by exemplifying through a brief case study. Finally, I concluded the paper by proposing a multi-level discourse-sensitive method based on the converging evidence perspective that would allow (qualitative) typology and typologists to reach their objectives and finally become a truly usage-based discipline.

Acknowledgements

I would like to thank two anonymous reviewers for their valuable comments and suggestions, which helped me improve the paper. The ideas in this paper stem from personal reflections and challenges I faced during my research, as well as from discussions with people who shared their perspectives and expertise over the years. I would like to thank (in alphabetical order): Alessandra Barotto, Francesca Masini, Caterina Mauri, and Marianne Mithun. I also thank Silvia Ballarè for reading and commenting on a previous version of this work. Finally, I owe a special thanks to Spike Gildea, who generously shared the Akawaio texts, his expertise on Cariban languages, and several hours of his time to discuss the *-pödi* and *ti'tuik prarö* constructions with me. All shortcomings are my own.

Abbreviations

1 = 1st person
3 = 3rd person
ADV = adverb(ializer)
AI = addressee involvement
ATTR = attributive
CONT = continuative
DETR = detransitivizer
EM = emphatic
ERG = ergative
IMP = imperative
INAN = inanimate
NEG = negative
O = object
PL = plural
PLAC = pluractional
PRES = present
PROG = progressive
PST = past
PTCP = participial
STYLE = stylistic element

References

Barotto, Alessandra & Simone Mattiola. 2023a. Discourse phenomena in typological perspective: An overview. In Alessandra Barotto & Simone Mattiola (eds.), *Discourse phenomena in typological perspective*, 1-9. Amsterdam: John Benjamins.

Barotto, Alessandra & Simone Mattiola (eds.). 2023b. *Discourse phenomena in typological perspective*. Amsterdam: John Benjamins.

Becker, Laura & Matías Guzmán Naranjo. 2025. Replication & methodological robustness in typology. *Linguistic Typology* 29(3). 463–505.

Bickel, Balthasar. 2010. Capturing particulars and universals in clause linkage: A multivariate analysis. In Isabelle Bril (ed.), *Clause-hierarchy and clause-linking: The syntax and pragmatics interface*, 51–101. Amsterdam: Benjamins.

Bickel, Balthasar. 2011. Multivariate typology and field linguistics: A case study on detransitivization in Kiranti (Sino-Tibetan). In Peter K. Austin, Oliver Bond, David Nathan & Lutz Marten (eds.), *Proceedings of Conference on Language Documentation and Linguistic Theory 3*, 3–13. London: SOAS.

Bickel, Balthasar. 2015. Distributional Typology: Statistical inquiries into the dynamics of linguistic diversity. In Bernd Heine & Heiko Narrog (eds.), *The Oxford Handbook of Linguistic Analysis*, 901-923. Oxford: Oxford University Press.

Bybee, Joan L. 2006. From Usage to Grammar: The Mind's Response to Repetition. *Language* 82(4). 711-733.

Bybee, Joan L. & Clay Beckner. 2010. Usage-based theory. In Bernd Heine & Heiko Narrog (eds.), *The Oxford Handbook of Linguistic Analysis*, 827–855. Oxford: Oxford University Press.

Caesar-Fox, Desrey Clementine. 2003. *Zauro'nödok Agawayo Yau: variants of Akawaio spoken at Waramadong*. Doctoral dissertation, Rice University, Houston.

Chafe, Wallace. 1976. Givenness, contrastiveness, definiteness, subjects, topics, and point of view. In Charles N. Li (ed.), *Subject and Topic*, 25-55. New York, NY: Academic Press.

Chafe, Wallace. 1980. The deployment of consciousness in the production of a narrative. In Wallace Chafe (ed.), *The Pear Stories*, 9-50. Norwood, NJ: Ablex.

Chafe, Wallace. 1987. Cognitive constraints on information flow. In Russell Tomlin (ed.), *Coherence and grounding in discourse*, 21-51. Amsterdam: John Benjamins.

Chafe, Wallace. 1994. *Discourse, consciousness, and time: The flow and displacement of conscious experience in speaking and writing*. Chicago, IL: The University of Chicago Press.

Corbett, Greville. No date. Agreement: Gender and Number. (Typological tools for field linguistics, MPI-EVA, Leipzig, available online at: https://www.eva.mpg.de/lingua/tools-at-lingboard/questionnaire/gender-and-number_description.php)

Cristofaro, Sonia. 2009. Grammatical categories and relations: Universality vs. language-specificity and construction-specificity. *Language and Linguistics Compass* 3(1). 441–479.

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

Croft, William. 2003. *Typology and Universals, 2nd edn*. Cambridge: Cambridge University Press.

Croft, William. 2022. *Morphosyntax*. Cambridge: Cambridge University Press.

Croft, William & Alan D. Cruse. 2004. *Cognitive Linguistics*. Cambridge: Cambridge University Press.

Cysouw, Michael & Bernhard Wälchli. 2007. Parallel texts: using translational equivalents in linguistic typology. *STUF – Language Typology and Universals* 60(2). 95-99.

Dahl, Östen. 1985. *Tense and Aspect Systems*. Oxford: Blackwell.

Dahl, Östen. 2015. How WEIRD are WALS languages? Presentation at the conference *Diversity Linguistics: Retrospect and Prospect*, MPI for Evolutionary Anthropology, Leipzig, 1 May 2015.

Diessel, Holger. 2019. *The Grammar Network. How Linguistic Structure is Shaped by Language Use*. Cambridge: Cambridge University Press.

Dingemanse, Mark. 2012. Advances in the cross-linguistic study of ideophones. *Language and Linguistics Compass* 6(10). 654–672.

Dryer, Matthew S. 1989. Large linguistic areas and language sampling. *Studies in Language* 13(2). 257-292.

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

Dryer, Matthew S. & Martin Haspelmath (eds). 2013. *The World Atlas of Language Structures Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. Available online at: http://wals.info (Accessed 2025.03.26).

Du Bois, John W. 1985. Competing motivations. In John Haiman (ed.), *Iconicity in syntax*, 343–365. Amsterdam: John Benjamins.

Du Bois, John W. 2003. Discourse and grammar. In Michael Tomasello (ed.), *The new psychology of language: Cognitive and functional approaches to language structure, Vol. 2*, 47–87. Mahwah, NJ: Lawrence Erlbaum Associates.

Gildea, Spike. 1998. *On reconstructing grammar: Comparative Cariban morphosyntax*. Oxford: Oxford University Press.

Givón, Talmy. 1979a. From discourse to syntax: Grammar as a processing strategy. In T. Givón (ed.), *Discourse and syntax*, 81–113. New York, NY: Academic Press.

Givón, Talmy. 1979b. *On understanding grammar*. New York, NY: Academic Press.

Givón, Talmy (ed.). 1983. *Topic continuity in discourse*. Amsterdam: John Benjamins.

Givón, Talmy. 1984. *Syntax*. Amsterdam: John Benjamins.

Guzmán Naranjo, Matías & Laura Becker. 2022. Statistical bias control in typology. *Linguistic Typology* 26(3). 605-670.

Haig, Geoffrey & Schnell, Stefan (eds.). 2023. *Multi-CAST: Multilingual corpus of annotated spoken texts. Version 2311*. Bamberg: University of Bamberg. Available online at: multicast.aspra.uni-bamberg.de (Accessed 2025.03.26)

Hammarström, Harald, Robert Forkel, Martin Haspelmath, Sebastian Bank. 2025. *Glottolog 5.2*. Leipzig: Max Planck Institute for Evolutionary Anthropology. Available online at: http://glottolog.org (Accessed 2025.12.09)

Haspelmath, Martin. 2007. Pre-established categories don't exist: Consequences for language description and typology. *Linguistic Typology* 11(1). 119–132.

Haspelmath, Martin. 2010. Comparative concepts and descriptive categories in cross-linguistic studies. *Language* 86(3). 663–687.

Henrich, Joseph, Steven J. Heine & Ara Norenzayan. 2010. The weirdest people in the world? *The Behavioral and brain sciences* 33(2-3). 61-83.

Hopper, Paul J. 1987. Emergent grammar. In Jon Aske, Natasha Beery, Laura Michaelis & Hana Filip (eds.), *Proceedings of the Thirteenth Annual Meeting of the Berkeley Linguistics Society: General Session and Parasession on Grammar and Cognition*, 139–157. Berkeley, CA: Berkeley Linguistics Society.

Hopper, Paul J. 1988. Emergent grammar and the a priori grammar postulate. In Deborah Tannen (ed.), *Linguistics in Context*, 117–134. Norwood, NJ: Ablex.

Lahaussois, Aimée & Yvonne Treis. 2019. Ideophones and interjections. Workshop at SLE 2019 Conference, University of Leipzig.

Lahaussois, Aimée & Julie Marsault and Yvonne Treis (eds.). 2024. Ideophones: honing in on a descriptive and typological concept. Special Issue of *Linguisdtic typology at the crossroads* 4(1). 1-444.

Majid, Asifa & Stephen C. Levinson. 2010. WEIRD languages have misled us, too. *The Behavioral and brain sciences* 33(2-3). 103.

Masini, Francesca & Simone Mattiola. 2019. Come fare tipologia con categorie non tradizionali? In Chiara Gianollo & Caterina Mauri (eds.), *CLUB Working papers in linguistics 3*, 282-294. Bologna: AMS Acta – Alma Mater Studiorum – Università di Bologna.

Mattiola, Simone. 2019. *Typology of pluractional constructions in the languages of the world*. Amsterdam: John Benjamins.

Mattiola, Simone & Spike Gildea. 2023. The pluractional marker *-pödï* of Akawaio (Cariban) and beyond. *International Journal of American Linguistics* 89(4). 457-491.

Mauri, Caterina & Francesca Masini 2022. Diversity, discourse, diachrony: A converging evidence methodology for grammar emergence. In Miriam Voghera (ed.), *From Speaking to Grammar*, 101-150. Berlin: Peter Lang.

Mauri, Caterina & Andrea Sansò. 2017. Un approccio tipologico ai general extenders. In Marina Chini & Pierluigi Cuzzolin (eds.), *Tipologia, Acquisizione, Grammaticalizzazione. Typology, Acquisition, Grammaticalization Studies*, 63-72. Milano: Franco Angeli.

Mettouchi, Amina, Dominique Caubet, Martine Vanhove, Mauro Tosco, Bernard Comrie, Shlomo Izre'el. 2010. CORPAFROAS, A Corpus for Spoken Afroasiatic Languages: Morphosyntactic and Prosodic analysis. In Frederick Mario Fales & Giulia Francesca Grassi (eds.), *CAMSEMUD 2007, Proceedings of the 13th Italian Meeting of Afro-Asiatic Linguistics*, 177-180. SARGON: Padova.

Mettouchi, Amina, Martine Vanhove & Dominique Caubet (eds). 2015. *Corpus-based Studies of Lesser-described Languages: The CorpAfroAs corpus of spoken AfroAsiatic languages.* John Benjamins: Amsterdam.

Miestamo, Matti. 2005. *Standard negation: The negation of declarative verbal main clauses in a typological perspective*. Berlin: Mouton de Gruyter.

Miestamo, Matti, Dik Bakker & Antti Arppe. 2016. Sampling for variety. *Linguistic Typology* 20(2). 233–296.

Mithun, Marianne (ed.). 1996. *Prosody, grammar, and discourse in Central Alaskan Yup'ik*. Santa Barbara, CA: Linguistics Department, University of California Santa Barbara.

Mithun, Marianne. 2015. Discourse and grammar. In Heidi Hamilton, Deborah Schiffrin & Deborah Tannen (eds.), *Handbook of Discourse Analysis. 2nd ed.*, 9-41. Oxford: Blackwell.

Overstreet, Maryann. 1999. *Whales, candlelight, and stuff like that: General extenders in English discourse*. Oxford: Oxford University Press.

Pakendorf, Brigitte & Françoise Rose (eds.). 2025. *Fillers: Hesitatives and placeholders*. Berlin: Language Science Press.

Ponsonnet, Maïa, Aimée Lahaussois & Yvonne Treis. 2023. Typologizing Interjections. Workshop organized at Dynamique Du Langage, Lyon, 21 november 2023.

Rijkhoff, Jan, Dik Bakker, Kees Hengeveld & Peter Kahrel. 1993. A method of language sampling. *Studies in Language* 17(1). 169–203.

Rijkhoff, Jan & Dik Bakker. 1998. Language sampling. *Linguistic Typology* 2(3). 263–314.

Schiffrin, Deborah. 1987. *Discourse Markers*. Cambridge: Cambridge University Press.

Schmidtke-Bode, Karsten, Natalia Levshina, Susanne Maria Michaelis & Seržant Ilja (eds.). 2018. *Explanation in typology: Diachronic sources, functional motivations and the nature of the evidence*. Berlin: Language Science Press.

Seifart, Frank, Ludger Paschen & Matthew Stave (eds.). 2024. Language Documentation Reference Corpus (DoReCo) 2.0. Lyon: Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2).

Thompson, Sandra A. 1988. A discourse approach to the cross-linguistic category 'adjective'. In John Hawkins (ed.), *Explanations for language universals*, 167-185. London: Basil Blackwell.

Velupillai, Viveka. 2012. *Introduction to linguistic typology*. Amsterdam: John Benjamins.

CONTACT

simone.mattiola@unipv.it