Chapter 17

Constituency and convergence in the Americas – Results and discussion

Sandra Auderset[a], Caroline de Becker[b], Gladys Camacho Rios[c], Eric W. Campbell[d], Javier Carol[e], Minella Duzerol[f], Patience Epps[g], Ambrocio Gutiérrez[h], Cristian R. Juárez[i], Magdalena Lemus Serrano[j], Stephen Francis Mann[k], Taylor L. Miller[l], Shun Nakamoto[m], Zoe Poirier Maruenda[i], Andrés Salanova[n], Hiroto Uchihara[o], Natalie Weber[p], Anthony C. Woodbury[g], Dennis Wylie[g] & Adam J. R. Tallman[b]

[a]University of Bern [b]Friedrich-Schiller-Universität Jena [c]State University of New York at Buffalo [d]University of California, Santa Barbara [e]University of Buenos Aires [f]Laboratoire Dynamique du Langage – CNRS [g]University of Texas at Austin [h]University of Colorado, Boulder [i]Max Planck Institute for Evolutionary Anthropology [j]Aix-Marseille Université & Laboratoire Parole et Langage [k]Max Planck Institute for Evolutionary Anthropology, Leipzig [l]State University of New York at Oswego [m]Universidad Nacional Autónoma de México [n]University of Ottawa [o]Tokyo University of Foreign Studies [p]Yale University

This chapter provides a basic conceptual introduction to the planar-fractal method. The method is then contextualized with respect to multivariate typology. The structure of the database based on this method is then described and an illustration of what the database can be used for is also provided. Four issues related to contextualizing constituency in typological context are then assessed in relation to the data gathered in the current volume: (i) the index of synthesis; (ii) the absence of *a priori* wordhood tests; (iii) the relative reliability of wordhood tests; and (iv) the word bisection thesis.

Sandra Auderset and Caroline de Becker and Gladys Camacho Rios and Eric W. Campbell and Javier Carol and Minella Duzerol and Patience Epps and Ambrocio Gutiérrez and Cristian R. Juárez and Magdalena Lemus Serrano and Stephen Francis Mann and Taylor L. Miller and Shun Nakamoto and Zoe Poirier Maruenda and Andrés Salanova and Hiroto Uchihara and Natalie Weber and Anthony C. Woodbury and Dennis Wylie and Adam J. R. Tallman . Constituency and convergence in the Americas – Results and discussion. To appear in: Adam J.R. Tallman, Sandra Auderset & Hiroto Uchihara (eds.). Constituency and convergence in the Americas. Berlin: Language Science Press. [preliminary page numbering]

1 A synopsis of the planar-fractal method

The planar structure is a template over which constituency tests/domains can be coded (Tallman 2021b; Tallman 2024 [this volume]). It constitutes an attempt to apply the ideas of multivariate or distributional typology to the problem of constituency. The planar structure was developed order to assess the degree to which logically distinct constituency tests/domains align and/or nest with each other and explore how much typological variation there is in this regard.

The planar structure can be conceptualized as a template, built out of a "lumping" strategy (Good 2016), which means that the template is designed to describe aspects of linear stipulation over as many constructions as possible, or as a type of phrase structure grammar with constraints imposed on what types of non-terminal nodes are admissible (see Tallman 2024 [this volume]). We should point out that the device is not a "theory of grammar" in the sense of Chomsky (1965). It is a comparative concept used to study a very specific aspect of linguistic structure. In other words, it is a measuring device that could be constructed with different constraints and coding properties for different research questions (for example Good 2016). If we do not use a planar structure or some such measurement technology, we will not have any way of keeping track of when diagnostics align and when they do not.

The "fractal" aspect of the planar-fractal method runs off of the premise that constituency tests, stated in the abstract, can have ambiguous interpretations when applied to actual language data. When a constituency test is applied to a given language we cannot and do not apply the test *as is*. Rather, there is a process of abstraction and then reconcretization in the application of the "test" to a new system. We lift the test from its language specific context, making it abstract, and then add details to apply it to a new language, reconcretizing the test in the process. Every constituency test must be recycled in this fashion if it is to be applied beyond the context for which it was originally developed and used.

We note, for instance, that some span of structure which we call "words" cannot be interrupted by other elements we have already identified as words in some language, let's say English. We abstract away from this property and claim that "non-interruption" is a general diagnostic for the identification of "words." But non-interruption by what? Surely we cannot use words of English to test whether a given span of structure in Hup is a "word" based on non-interruption. So, we tackle the problem by reconcretizing the test, introducing or imputing a Hup–specific interrupting element into the equation. This involves an epistemic leap which might seem so trivial that it passes above conscious awareness.

It is in this reconcretization of recently abstractified "tests" where fracturing comes into play. The problem is that there is often more than one way in which a given constituency test can be reconcretized when it is applied to a new domain. This aspect of linguistic analysis can go unnoticed, especially when linguists are told to find specific categories or structures in novel data, but not told how one could possibly ever justify claiming that the category or structure is not present in a linguistic system (see Tallman's (2024b [this volume]) for a discussion of basic linguistic theory). Therefore, we seek to develop a method that makes the reconcretization explicit and compels us to not discard competing interpretations surreptitiously as a consequence of cognitive biases (Ackermann 1985). If we apply the process repeatedly to more and more languages, we will find that our original "test" has expanded into a number of sub-types. We view this as an application of the autotypologizing method (Bickel & Nichols 2002, Witzlack-Makarevich et al. 2022) to the problem of constituency. The goal of the project is to articulate a taxonomy of domains organized hierarchically from their abstract to their more concrete instantiations. The typology is constructed to discern whether there is statistical order to the patterns we find with these domains in and across languages.

A planar structure can be defined as follows:

(1) Planar structure: a template of consecutively ordered positions from 1 to n. There is a planar structure for each part of speech which is open class. Each planar structure has at least one position for a core element. All other positions are for non-core elements.

Positions can be "fitted out" by core or non-core elements. But for a given planar structure there is at least one position for a given core element. The core element can be defined as follows:

(2) Core: A core is an open class element. Any sentence that is fit out by a planar structure needs to have an overt core element. For instance, a verbal planar structure will have one position for a verb core and all sentences that contain that type of core should be able to be mapped to that planar structure. The core functions as the semantic head (see Croft 2001: 241–280 and Croft 2022: 35–37) of a planar structure and the constructions that it can be fitted out by (see Tallman 2021b and Tallman 2024 [this volume] and Woodbury 2024 [this volume] for discussion).

For instance, a verbal planar structure will have a position for a verb core. A nominal planar structure will have a position for a noun core. If it is necessitated by the facts of the language, we can also add adjectival or adverbial planar

structures. In the present volume, we have limited the scope of the study to verbal planar structures, although two chapters provide preliminary nominal planar structures (Epps 2024, Gutiérrez & Uchihara 2024 [this volume]).[1]

As stated above, a planar structure is composed of a number of POSITIONS. A position has a number, contains elements and is associated with a specific planar structure. Each position is either categorized as a slot or a zone. Slots and zones are defined below.

(3) SLOT: A position which can only be filled by one element at a time.

(4) ZONE: A position which can be filled by more than one element and the elements can occur in any order in the zone.

For expository purposes, we provide a simple planar structure below. We have placed a superscript ^{c} over the core elements of the planar structure. The position with a core is obligatorily filled.[2]

Table 1: Example planar structure

1	slot	a, b, h
2	slot	c
3	slot	d^{c}, e^{c}
4	zone	f, g
5	slot	h

In position 1, there are 3 elements (*a, b, h*). In this position, only one of these elements can occur for a given sentence. This means that *acdfh* is an admissible string according to the planar structure above, but *abcdfgh* or *ahcdfgh* is not. However, in position 4 the elements *f* and *g* can co-occurr and variably order. Thus, *acd**fg**h* and *acd**gf**h* are both admissible strings. Positions can be obligatorily or optionally filled (as with categories in a phrase structure grammar). Positions can be *open* or *closed* contingent on the presence of specific elements or whether a given position is filled. For instance, if we find that element *b* never co-occurs

[1] A given core might be fit out in more than one position. But there can be no positions which can contain the same part of the core in them. For instance, our planar structures are not allowed to have a position 3 and a position 5 both of which could output a core (e.g. a verb root). However, a planar structure could have a core which is composed of two pieces one of which occurs in position 3 and other in position 5. The reason for this restriction, as described in Tallman (2024 [this volume]), is to make the reporting of constituency tests more manageable.

[2] The reverse is not true. We cannot determine that a position is a core position because it must always be filled.

with *h*, we can add a stipulation that position 5 is closed if position 1 is filled with element *a*.

Note that there are two ways of describing the variable ordering of elements in a planar structure. If the variable ordering is *local* in the sense that there are no intervening elements between the elements that variably order, then a zone is posited, as with the elements *f* and *g* above. Zones of this type are useful for defining cases where affixes variably order with one another (Bickel et al. 2007) in a traditional "word" or where adverbs or particles variably order locally (without intervening elements) with one another as well.

If the elements variably order but *around* an element which displays a fixed order, then we simply place the relevant elements in more than one position as with the element *h* above. Allowing *h* to be in position 1 or position 5 means that we can have the order *hd* and *dh*. A typical example of this type of variable ordering is with noun phrases around a complex verb structure in so called non-configurational languages (Austin & Bresnan 1996). A subject NP, for instance, can be given a position on each side of a span of verbal elements.

Finally, we need to define an element.

(5) Element: an element is a morph (Haspelmath 2020), another planar structure or a well-defined subspan of a planar structure.

As a consequence, a nominal planar structure can be an element of a verbal planar structure, or some subspan of a nominal planar structure can be an element of a verbal planar structure, and vice versa. The ability to have elements which are planar structures themselves is necessary to make them practically useful: if this condition was not met, planar structures would not be finite due to recursion. In other words, we do not flatten out phrase structure without limit. While a planar-structure grammar imposes some hierarchical structure by allowing planar structures to embed within each other, notions such as"word" and "phrase" are prohibited.

With the planar structure in hand, we use autotypology as a research method in the application of constituency tests. Constituency tests can now be operationalized as variables which code spans over planar structures of specific languages (see Tallman 2024 [this volume] for more details).

2 Multivariate typology and the constituency variables

Autotypology as a method emerged in the early 2000s as part of the larger AUTOTYP research program, which aims at systematically analyzing variation in the

languages of the world as well as explaining this variation both quantitatively and qualitatively (Bickel & Nichols 2002, Bickel et al. 2017, Witzlack-Makarevich et al. 2022). It has also been referred to as "Multivariate Typology" and "Distributional Typology" (Bickel 2015), although these labels could be seen as more appropriately describing a whole research agenda, rather than only a typological method. However, they share the same approach, so in the remainder of this chapter we will use the label Autotypology as cover term for the methodology and theory behind the AUTOTYP project.

Typological variables always involve a certain degree of abstraction and generalization from language-specific details. In most typological approaches, the variables as well as the possible values they can realize are determined *a priori*, usually based on tradition, theoretical assumptions, and convenience (for more details and examples see Witzlack-Makarevich et al. 2022: 632). Even approaches that try to circumvent the issues with categorization based on tradition and theory by relying on known variation and pilot studies still define the variables top-down. This is also the case for the two largest typological databases currently available, WALS (Dryer & Haspelmath 2013) and Grambank (Skirgård et al. 2023).

Autotypology differs from these more traditional typological approaches in that the variables and their values are developed in a bottom-up fashion and constantly adapted to capture the variation present in the data at hand. The idea behind this methodology is to invest in coding fine-grained variables that adequately account for the diversity of the world's languages and that can be used to investigate a variety of research questions across different theoretical frameworks. While initially more time-consuming than relying on pre-defined, aggregated variables, the methodology ensures that the resulting database can be expanded on and (re-)used by other researchers. In the following, we will describe the methodology and how it was used in developing the diagnostics of the constituency database. As in other frameworks, the starting point for developing variables in Autotypology is usually found in earlier typological studies or theoretical discussions relating to the research question. In the case of constituency, we can draw on a wealth of literature proposing or evaluating diagnostics for constituency and similarly for wordhood and phrasehood (see Tallman 2024 [this volume]). These starting point variables are not seen as static, but rather they are re-evaluated and adjusted with each new language being coded. One type of adjustment frequently encountered with constituency diagnostics is fracturing, that is, the splitting of a diagnostic into multiple diagnostics, driven by details from a language or linguistic system over which one is coding grammatical or phonologigcal properties.

A constituency variable is defined as follows, following Tallman 2021b:

> Constituency variable: ... a generalization within or across constructions that targets or crucially refers to some subspan of a planar structure. A constituency test can only be applied in a given language if it is specific enough such that it refers to a *well-defined* subspan. A subspan is well-defined if it contains a single left-edge (e.g. position 3) and a single right-edge (e.g. position 8).

We start with constituency tests which are frequently found in the literature (displacement, interruption). The diagnostics as they are found in the literature typically require a great deal of refinement to meet the definition provided above. Therefore, much of the intellectual work in developing constituency variables amounts to operationalizing relatively vague heuristics from the morphology, phonology and syntax literature so that they can be applied consistently. This often involves making finer distinctions than what is found in the literature. For instance, non-interruption can be divided into different tests depending on what we choose as the interrupting element. The converse situation also arises. There are cases where the literature attests of apparently distinct diagnostics but, upon closer scrutiny, it is revealed that they are the same; they were just described or conceptualized as different, perhaps by different authors, perhaps in different languages. An example of this concerns the distinction between non-interruption in the morphology/wordhood literature and displacement in the syntax literature. The identity between diagnostics that are often described as if they were distinct becomes apparent when we assess whether convergences between diagnostics might be a spurious consequence of the way such diagnostics are formulated.

The formulation of a constituency test and the operationalization of these tests as variables often elicits protest from certain linguists. It has been claimed that some of the tests used in this study are (or might be) "junk" tests that should be discarded. The basis for such claims often rests on these specific tests not giving a clear result in favor of some or another syntactic model, theory or analysis.

This point is actually partially valid. Many of the constituency tests developed in this book might very well be "junk." However, the protest misses an important point about database construction, measurement, and their relationship to hypothesis testing (Ackermann 1985: 125–149). By coding a constituency test in a database we are not thereby claiming that the test necessarily identifies a constituent in any specific linguistic theory (let alone all theories). A linguist researching within a perspective whereby one of the coded tests is considered useless is free to discard the test and assess what the results show after they have

subset the data so it only contains what they deem relevant. What the database allows, or better yet, compels the researcher to do, is to be consistent and explicit about exactly which data and tests are used. For instance, they cannot discard a test in one language and, at the same time, regard that test as an important piece of evidence in another.

The methodology addresses a concern that linguists might treat a test as reliable only insofar as it confirms a given prejudged analysis and that they discard it otherwise (Croft 2001, 2010, Haspelmath 2011, Tallman 2021a). We argue that constructing a database which samples tests independently of the researcher's analyses attenuates this problem. Another reason we think that the protest against junk tests misses the mark is that it presupposes that we know *a priori* which tests will result in interesting generalizations and which ones not. Further justifying this perspective is the fact that protests about junk tests are not consistent with each other. It turns out that one linguist's trash is another linguist's treasure, a point we return to in §6. Rather, in the perspective adopted in this volume, whether a test turns out to be junk for language description or cross-linguistic generalization is an empirical question. A junk test is just one for which no useful language-internal, nor cross-linguistic generalizations can be made. In order to know which tests are junk, we need to actually code them.

3 The structure of the database and use cases

The constituency tests and the planar structures are collected in an interlinked database designed with AUTOTYP principles in mind. AUTOTYP principles include modularity, autotypology (see §2), separation of definition and data files, and late aggregation (Witzlack-Makarevich et al. 2022). As mentioned above, AUTOTYP is a typological database that has been continuously developed over the past twenty-five years as part of a large-scale research program in order to address problems that have arisen from the creation of more traditional typological databases. One of these issues is the use of fixed, *a priori* categories determined by theoretical considerations, or simply by traditional usage, which often fail to adequately capture a phenomenon across a large and diverse set of languages. The application of the AUTOTYP principles also facilitates the later re-use and expansion of the database. Another design principle concerns the separation of information across multiple files which are linked together via a common, standardized identifier. This flexibility makes it possible to address a larger number of different questions with one data set. As such, these design principles integrate well with the approach taken in this volume. The constituency test results

are coded in a bottom-up fashion and we want to make the data usable for future studies.

The workflow for gathering the data and collecting it in the database is illustrated in Figure 1. It starts with the elaboration of the planar structure by the language expert based on data collected through fieldwork and collaboration with speakers. The planar structure then serves as the basis for applying constituency diagnostics as described in §2. The results are then written up, including discussion of issues with the methodology or application of specific tests that came up during analysis. Finally, the results are entered into the constituency database for cross-linguistic comparison. Given how autotypology works, the structure of the database and the variables are informed by the language-specific analyses and vice versa. In practice, this means that the database and variables are adjusted to accommodate language-specific facts not previously considered, but also that the exact application of a test in a language can be refined or adjusted based on what we learn from other languages.

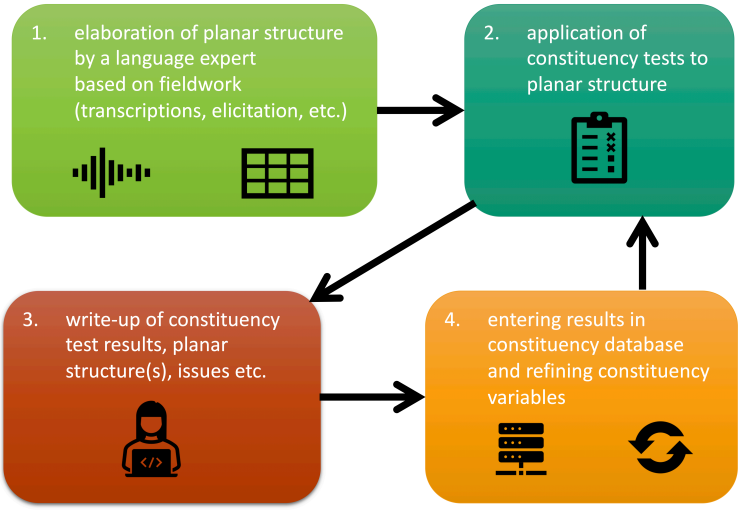

Figure 1: Schematic illustration of the workflow

The structure of the interlinked database is depicted in Figure 2. In the following, we discuss the modules and the variables in more detail, following the outline from left to right and top to bottom. The sources file contains bibliographic information and can be linked to the metadata file with the citekey. The metadata file contains information about languages and contributors, such as commonly used language names, Glottocodes (if available), geographic information, as well

as contributor names and the form of the contribution. The planar structures are collected in the planar file, where each planar structure and each position within it receive a unique identifier. The positions are listed together with the position type (slot vs. zone) and the language-specific elements that can appear in each position. For analyzing convergences and other aspects of test domains, we need to know in which position the base of the planar structure occurs. This information is provided in the overlaps file, which can be linked to the planar file by the planar ID and to the other files by the language ID.

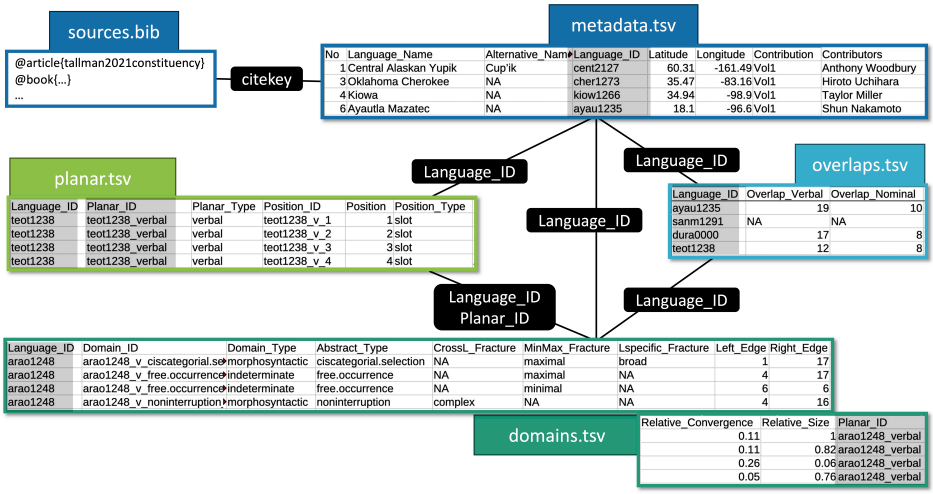

Figure 2: Illustration of the structure of the constituency database with file excerpts. Lines indicate which modules can be connected to each other. Black boxes show the unique identifier(s) that link(s) two modules together and the respective columns in the file excerpts are shaded in grey.

Finally, the test results are recorded in the domains file. This file can be linked to the other files via the language ID and additionally to the planar structure file with the planar ID. For each reported test in a language, we record the position indices that delimit the respective span, as well as information about the type of test applied and measures derived from it, such as span size and the number of other tests the span converges with in this language. Below, we briefly summarize the contents of this file:

(6) Domain Type: the linguistic level that the test applies to. Values:

   a. PHONOLOGICAL: The test makes reference to phonological criteria. An example of this is a domain where vowel elision applies.

b. MORPHOSYNTACTIC: The test makes reference to morphosyntactic criteria. An example of this is a domain delimited by elements that are ciscategorial with the verb.

c. INDETERMINATE: The test can be interpreted as either making reference to phonology or morphosyntax or both. An example of this is free occurrence, as it could be seen as resulting from a phonological constraint or a morphosyntactic one.

(7) ABSTRACT TYPE: standardized classification of constituency tests into abstract classes. Values:

a. CISCATEGORIAL SELECTION: domains where the non-core elements are selectionally restricted to a specific core (e.g. verbal affixes which only combine with the verb);

b. DEVIATIONS: domains where elements display a specific type of deviation from biuniqueness (e.g. extended exponence);

c. FREE OCCURRENCE: domains that identify spans which are free forms;

d. NON-INTERRUPTABILITY: domains that cannot be interrupted by some element;

e. NON-PERMUTABILITY: domains which exhibit fixed ordering of elements;

f. SEGMENTAL: domains that undergo some segmental phonological process;

g. SUPRASEGMENTAL: domains defined by some suprasegmental phonological process;

h. REPAIR: domains that are identified by repair strategies;

i. PAUSING: domains that can be delineated by a pause;

j. PROFORM: domains that can be replaced by a proform;

k. PLAY LANGUAGE: tests that identify spans which are targeted in play language;

l. IDIOM: domains which contain elements that typically form idioms or non-compositional constructions.

(8) Fractures

a. CROSS-LINGUISTIC FRACTURE: a fracture that can be applied across languages with a standardized set of labels or a typological property that helps further subclassify an Abstract type. Such properties can be subtypes of phonological processes, for example, consonant and

vowel deletion as fractures of a segmental domain. Our current data set contains 45 such fractures.

b. Language-specific fracture: a fracture that only applies within a specific language. Those fractures are thus not standardized. Our current data set contains 178 language-specific fractures;

c. Minimal-Maximal fracture: a fracture for the smallest and largest span where a test applies. Minimal-maximal fractures are those that always, by their definition, identify one inner and one outer domain where the former is embedded in the latter. For example, a maximal domain of 2-10, could identify a minimal domain that with a left edge which is the same or smaller than 2 and a right-edge which is the same or larger than 10. The fracture would not be coded in case the minimal and maximal fractures of the test give the same result;

(9) Other coded properties

a. (Right/Left) Edge: The boundary of the span, i.e the first and last positions where the test applies. This is recorded by the position number;

b. Size and Relative size: The size of the span in number of positions and the relative size of the span in number of positions divided the by the largest span identified by a constituency test in the respective language;

c. Convergence and Relative convergence: The number of other spans in the language that this span converges with. The relative convergence is the convergence number divided by the total number of tests applied in the language;

d. Largest: The largest span identified in a language;

e. Position total: The size of the planar structure in number of positions;

f. Tests total: The total number of tests applied in a language.

Due to the modular structure of the database it can be easily expanded upon in the future. The data collected in this volume are available on Zenodo as version 1.0 (Auderset & Tallman 2023), which also includes data from Chacobo (Tallman 2021b) and Siksika (Blackfoot) (Natalie Weber, p.c.) for which we do not yet have an accompanying paper.

The database is designed in such a way that it can be used for investigating a variety of research questions and for providing overviews and summaries regarding constituency. We provide a few examples relevant to the volume here. The

sample languages are plotted on a map in Figure 3, which additionally displays the maximum relative convergence found in each language. The map shows that Cup'ik displays the highest relative convergence, while Chorote has the lowest one. It also shows that even languages spoken in the same geographical area, such as Hup and Yukuna, do not necessarily exhibit the same degree of convergence.

The database also allows one to compare layer sizes and convergences across languages. Figure 4 displays relative convergences versus relative span sizes and shows that there is great cross-linguistic variation in this domain. In terms of relative span size, most of the languages described here have spans of various sizes, ranging from targeting only one position to the whole planar structure, as in South Bolivian Quechua and Chorote. In others, the spans identified by the constituency diagnostics cluster around a few span sizes, as in Martinican, or are skewed to either relatively small spans, as in Kiowa, or relatively large spans, as in Oklahoma Cherokee. In terms of relative convergences, the languages also exhibit vast differences. In a few languages, a clear "winner" emerges, that is, a span that is identified by many diagnostics, while all other spans show no or very little convergences. This is the case for in Cup'ik, for example, where almost half of the diagnostics converge on a span with a relative size of 0.79 (covering 15 out of 19 positions of the planar structure). Martinican and Zenzontepec Chatino both have spans that are targeted by about a third of the diagnostics, but these are much smaller. In Zenzontepec Chatino, the span is has a relative size of 0.19 (covering 4 out of 21 positions) and in Martinican it is even smaller at 0.16 (covering 4 positions out of 25). Furthermore, in some languages, there are no strong convergences at all, as in Chácobo, Hup and Siksika (Blackfoot). These languages approach a situation where each test targets a different span.

The database can also be used to explore tendencies associated with certain test types across languages. Figure 5 displays the distribution of relative span size according to the type of constituency test. Many of the test types have similar bimodal distributions, with a larger peak targeting a smaller span and a smaller peak targeting a larger span. This reflects the minimum and maximum fractures of said tests. Deviations from biuniqueness, however, exhibit a different distribution: they overwhelmingly target small spans (with a peak around 0.15), with very few tests resulting in larger spans above 0.5. This could explain why deviations from biuniqueness are often seen as good wordhood tests – they capture almost exclusively small spans that can felicitously be interpreted as "words."

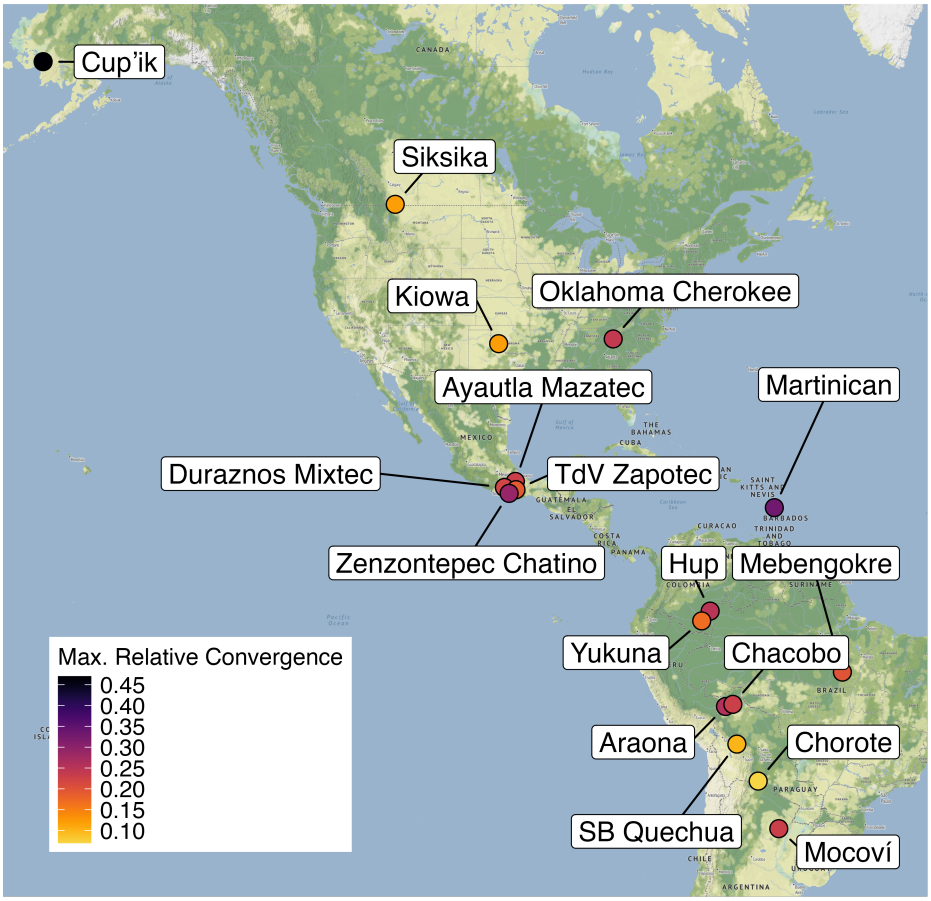

Figure 3: Location of the sample languages with maximum relative convergence (= the maximum number of test convergences per language divided by the total number of tests) represented as a color gradient.

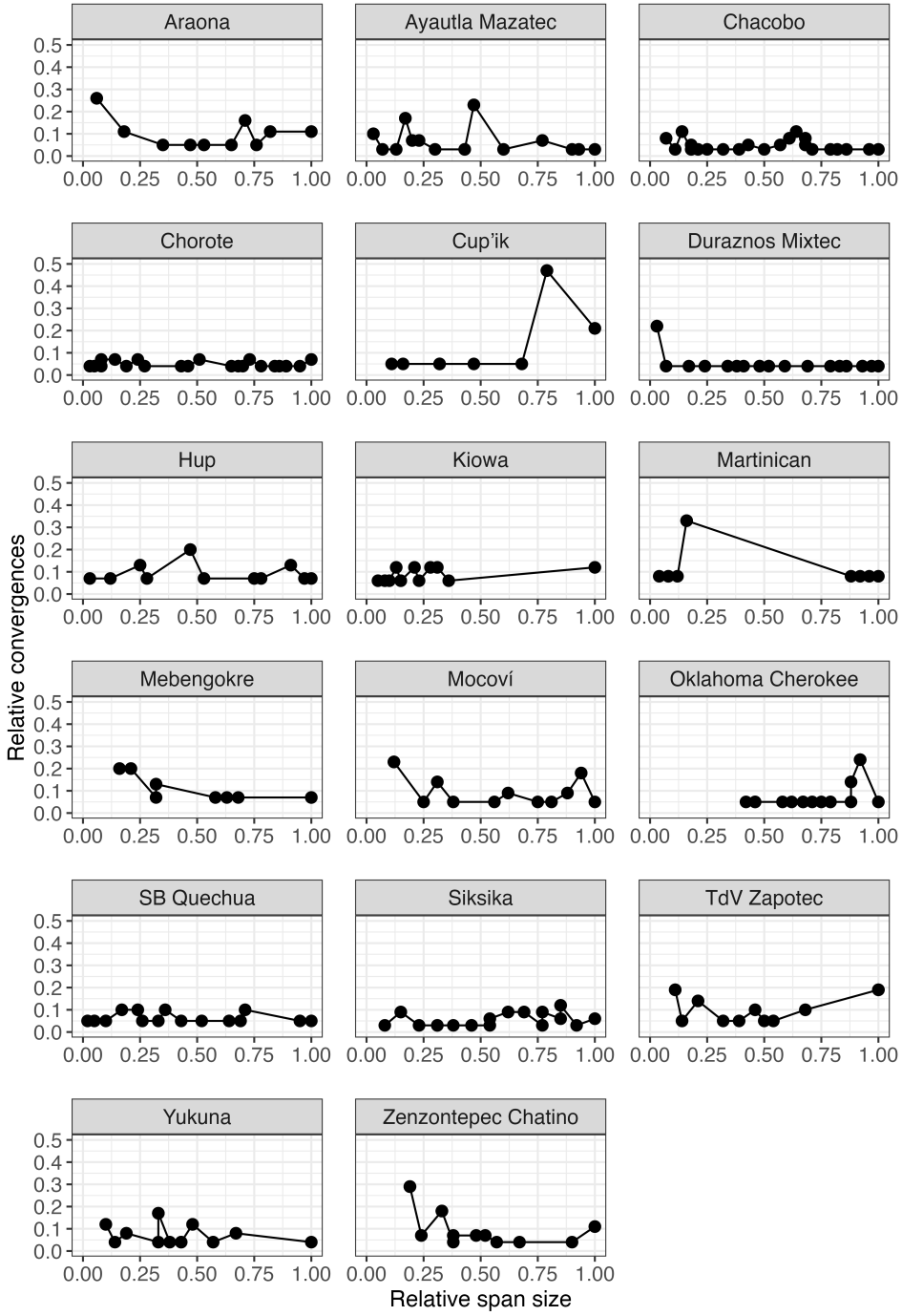

Figure 4: Visualization of relative convergences per relative span size across the languages of the sample in the verbal domain.

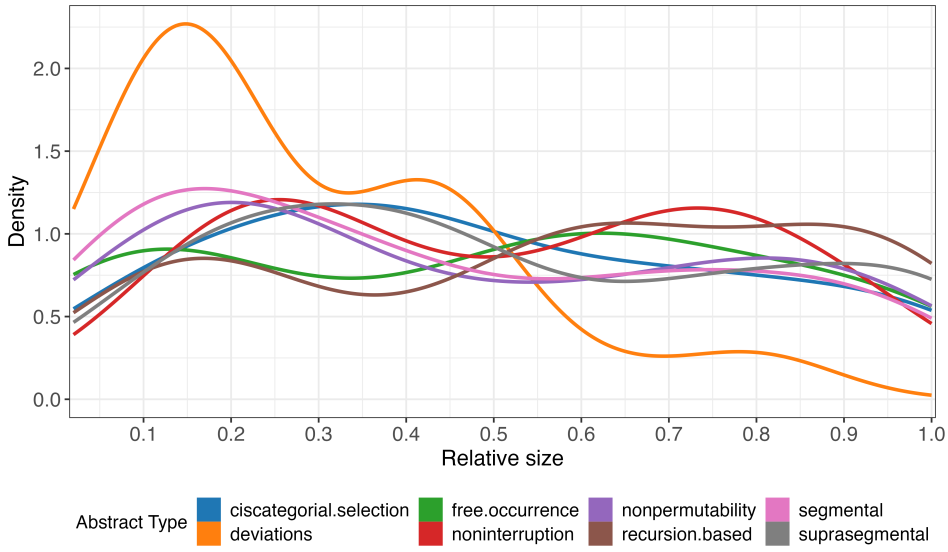

Figure 5: Density of relative size of spans by abstract type across sample languages

4 The index of synthesis reconsidered

Traditionally languages are described as varying in terms of their degree of synthesis. The degree of synthesis that a language displays refers roughly to its tendency to pack more or less concepts into a single word (Sapir 1921). For typological comparison the notion has been operationalized by counting the number of segmentable morphs that occur within each orthographically spaced out word on average over some text (Greenberg 1954, Easterday et al. 2021).[3] Such studies rely on orthographic words and they rest on the assumption that either orthographic words are legitimate units of comparison or are approximations to some unit of

[3]It is important to realize that the "number of segmentable morphs", once some criterion for morph segmentation is provided, is not the same as how "easily such morphs can be segmented". The former is most relevant for the analytic-synthetic distinction, while the latter speaks to the traditional distinction between agglutination and fusion. In an obvious sense, both clines are destabilized by the current study because they both make reference to the internal structure of words. We do not treat the agglutination-fusion cline in this chapter as it has already been shown to rely on empirically incorrect assumptions independent of its reliance on the notion of word Haspelmath (2009). We would also suggest that a metric of "exponence complexity" (Tallman & Auderset 2023) that measures deviations from biuniqueness across the grammar is more useful because it does not conflate distinct properties, as the traditional metric of fusion does.

comparison across languages. The analytic-synthetic continuum also forms an important aspect of describing variation and change in certain language families (e.g. Schwegler 1990, Ledgeway 2017 for Romance; Arcodia & Basciano 2020 for Sino-Tibetan).

The results of the language-specific studies in this volume highlight the fact that the synthetic status ascribed to a language can be contingent on which constituency tests are deemed to be appropriate wordhood diagnostics. There is no unified notion of synthesis, but a spectrum of different notions or candidates that may or may not align on their left and/or right edges in a given language.

Even in languages with a high degree of convergence it should be noted that not all diagnostics target what is traditionally considered a word. Clear examples come from Cup'ik and South Bolivian Quechua, which have both been described as polysynthetic languages in the literature. However, if we take the criterion of "conventionalized coherence" as definitional (Dixon & Aikhenvald 2002, Dixon 2010, Aikhenvald et al. 2020), both of the languages are much closer to being analytic. These languages contain pockets of word-like chunks or clusters within their traditionally defined words. This is not because linguists working with these languages have simply ignored wordhood criteria. On the contrary: free occurrence, non-interruption, phonological criteria such as stress, or syllabification all hit a domain of structure attached to the notion of"word" used by Inuit–Yupik–Unanganists and Quechuanists, respectively.

The observation that so-called polysynthetic languages have word-like pockets inside their grammatical words is not a theoretically innocuous observation: many morphologists and researchers in corpus linguistics propose that morphological structure emerges as a distinct component from syntax via "chunking." (Bybee 2001, 2010, Lorenz & Tizón-Couto 2019). This refers to a process whereby multiple pieces of structure are gradually reinterpreted as a single unit for processing and production, presumably on the basis of "conventionalized coherence."

If such a theory of morphological development and maintenance is maintained, then it follows that the traditional "word" in these languages is a phrasal (or post-word) constituent. But the convergences of wordhood tests around this domain still provide evidence for dichotomous structuring of some sort. Insisting that morphology is defined through conventionalized coherence does not result in the purported "morphological complexity" of polysynthetic languages disappearing but simply displaces it to a different terminological realm: we would now claim that many polysynthetic languages display dichotomous patterning in their "syntax."

Other languages pose even starker problems for the traditional analytic-synthetic distinction. It can be observed that in Chácobo, Duraznos Mixtec and Hup, a shift in perspective regarding which criterion we regard as defining the word can result in these languages being recategorized as isolating or (poly)synthetic. Put another way, these language could be classified as isolating or polysynthetic depending on which test is regarded as word-identifying. In both Chácobo and Hup, a focus on non-permutability (that is, contiguity or fixedness of order) and certain interpretations of non-interruption would result in the classification of these languages as isolating or at least highly analytic. If we shift our focus to free occurrence domains, the languages become (poly)synthetic, and the facts that were rallied to argue that they were isolating now become indicative of the languages displaying a "syntax-like" morphology (Payne 1990, Tallman & Epps 2020). Moreover it is too simplistic to claim that this is a difference between only two types of criteria: free occurrence vs. non-permutability or non-interruption. Mixtec is isolating or polysynthetic depending on how free occurrence is treated as a constituency test, the minimal fracture providing an isolating result and the maximal fracture providing a highly synthetic result. We are reminded of Boas' observation that in some languages (Tsimshian was his example) the division or combination of forms into separate or single words can be fairly arbitrary (Boas 1911: 28), but importantly languages may vary in terms of how arbitrary this division is (Boas 1911: 26; see Bazell 1953: 68 as well).

Claims about synthetic status usually make reference to morphological complexity (e.g. Easterday et al. 2021). But synthesis could also be discussed in terms of phonological domains – in terms of segmentable morphs per phonological word. This approach would run into the same problem, however, as there are competing definitions of the phonological word for many of the languages of the study. The notion of a phonological word is not unified in a single criterion either and so couching synthesis in terms of phonological integration does not necessarily simplify this notion.

These considerations do not mean that the analytic-synthetic notion should be abandoned for typological research, but rather that it should at least be refined. As a language has less and less converging wordhood criteria, the notion of synthesis becomes more complex and graded in that language. In this way, we could understand the index of synthesis as not only multidimensional (as it can be decomposed into a number of logically distinct variables) but as an index that interacts with other architectural properties of a language, as in how strongly the languages displays dichotomous patterning or how fuzzy the boundary between morphology and syntax is in the language (e.g. Tallman & Epps 2020 for this perspective).

5 No *a priori* wordhood tests

In a sense, the notion that there are wordhood tests presupposes that there are words to begin with (Lara 2004). If we claim that wordhood tests are not always picking out a unified notion of word, then what are these wordhood tests picking out? The apparent paradox is resolved once we recognize that words are a species of constituent which we assign special status because it represents a cut-off point between two different realms of structural organization. From this perspective it is somewhat misleading to even refer to "wordhood" tests as such. Rather, if the whole idea of a word is interesting because it indexes our belief that languages display some sort of modular[4] structure (with word-formation being distinct from phrase and sentence-level formation), then words emerge from patterns of structural groupings over utterances reoccurring over the domain, not from singular diagnostics applied in the abstract. From this perspective, there are reasons to think there should be no coherent notion of "wordhood test", as distinct from phrasehood test, at least not *a priori*.

The fact that there is no clear distinction between wordhood and phrasehood tests can be discerned in two ways. First, when we put formulations of wordhood and phrasehood tests side by side, we find that they are difficult to distinguish. Tallman (2024 [this volume]) gives the examples of non-interruption as a wordhood test versus displacement as a phrasehood test.

Another indication that constituency tests cannot be clearly grouped into wordhood and phrasehood tests arises when one considers that in numerous cases a diagnostic that hits a "word" according to its definition in one language (or linguistic tradition), hits a subword unit in the second language, and an apparently phrasal unit in a third. For instance, non-interruptability by a free form lines up with the traditional word in Cherokee (the orthographic word and what is considered to be a word by Iroquianists) (Uchihara 2024 [this volume]). The same is true of non-interruptability in Martinican (Duzerol 2024 [this volume]). The derivational prefix, the verb root and two pronominal indexes make up the orthographic word in Martinican as long as the pronominal indices are second or third person. However, if we take the way the word is described in Araona (and the Takanan tradition generally) the same interruption test identifies a subword unit, in fact, just the verb root, rather than the large polysynthetic structure described as

[4]Note that claiming that languages display modular structure does not entail that the modular structure is innate, nor that there are some fuzzy boundaries between domains. In cognitive science and biology generally it is well recognized that modularity is a matter of degree (Rasskin-Gutman 2005; Carruthers 2006: 14) and that it can be emergent (Coltheart 1999, Zerilli 2020).

a word by some linguists who have described the language (Pitman 1980, Emkow 2019). The converse problem is also attested. Non-interruption by a single free form identifies a span of structure *higher* (e.g. a phrase) than what Gutierrez & Uchihara argue is the best candidate for phonological word in Teotitlán del Valle Zapotec. Therefore, non-interruption by a single free form identifies a word, a subword or phrasal domain depending on the language. Should we still consider non-interruption a "wordhood test"? Another example is extended exponence. In Araona, extended exponence lines up with the traditional word, but in Central Alaskan Yup'ik, the same diagnostic identifies a subword constituent with respect to the traditional word of this language. Again, should extended exponence be identified with a word or a subword?

In the phonological domain these issues are so endemic that it is difficult to know where to start. Bickel et al. (2009) show that there is no overall tendency for phonological domains to cluster around a universal "prosodic word". Furthermore, once prosodic words are classified for the type of phonological generalization that defines them (e.g. rhythm, epenthesis etc.), there is no overall tendency for any specific phonological process to identify higher or lower domains, except for "stress", which shows a tendency to identify relatively higher domains (Schiering et al. 2012).

"Words" refer to boundaries between domains of different structural organization. But it is doubtful that a "wordhood test", abstracted from the rest of the structure of a language, is a useful starting point for typological investigation. Constituents, domains or groupings are a better starting point since they do not presuppose that we know *a priori* the properties of the modules we are interested in investigating, which may be subject to cross-linguistic variation.

6 Reliable and unreliable tests

6.1 Introduction

The literature on wordhood and constituency often implies that certain tests are better or more reliable than others. For instance, Dixon & Aikhenvald (2002) distinguish certain "main criteria" (cohesiveness, fixed order, conventionalized coherence). But the test of "isolatability", for example, only identifies words as a "tendency" (Dixon & Aikhenvald 2002: 25). Similarly, Payne (2006: 162) claims regarding coordination that it "can't be the major way of determining constituent structure", compared to the other constituency tests he discusses (Adger 2003: 125 and Carnie 2010: 21 for related claims).

Writers on these topics apparently do not agree with each other. Dixon & Aikhenvald (2002: 25) state that "the principle of uninterruptability ... is only a tendency – which may apply more to phonological than to grammatical words – but can be a useful support for the other criteria." Bauer (2017: 17) has a discussion concerning "criteria involving structural integrity", which appears to be similar if not the same as non-interruption. He makes nearly the opposite claim regarding the reliability of this wordhood test: "The uninterruptability of the word is, in general terms, a much stronger criterion". Martinet (1962: 92) states "[a]s a matter of fact, inseparability is one of the most useful criteria for distinguishing what is formally one word from what is a succession of different words" (see Brown & Miller 1980: 164–165 as well). Booij (2005: 185-187) describes non-interruption as definitional of word constituents. Some of the apparent disagreement could be a result of authors interpreting the criteria in different ways[5], but the point remains that there is a re-occurring tendency to regard some tests as better or more reliable than others in some sense, yet it is unclear from the literature which ones should be regarded as more reliable.

It is worth asking on what basis such claims about the relative reliability of tests could be made. In the literature, the relative superiority of some tests over others is generally asserted without any justification. In some cases it is pointed out that a test is unreliable because it does not converge with a predefined or established constituency analysis (e.g. Payne 2006: 162, Carnie 2010: 21), which appears to be a circular argument. More charitably, what some of these researchers might mean is that unreliable tests are just those tests that are prone to not be applied correctly (presumably by linguists who are not as skilled at syntactic analysis as they are). Yet an articulation of the proper interpretation of a potentially unreliable test is never given, except insofar as it means "in line with my own theoretical expectations."[6]

[5]For instance, Dixon & Aikhenvald make a distinction between cohesiveness and non-interruption that the other authors do not make, to our knowledge. Non-interruption seems to also involve a pause, whereas "cohesiveness" is the more general term for any non-interruptable piece of structure. The ambiguity regarding how to interpret the diagnostics as they are formulated in the literature is perhaps one of the reasons why it appears so difficult to refute them. If one finds that a diagnostic is not working, one can be accused of misinterpreting it. Indeed as we have shown throughout the chapters of this volume, the diagnostics have multiple interpretations.

[6]In the context of coordination tests, Phillips states: "Traditionally, the results of movement tests have tended to be taken more seriously, and the results of other tests have been made to fit with these." (Phillips 1996: 27). As Phillips shows, one ends up with a quite distinct view of constituency structure if coordination is put on par with the other tests (see Osborne 2018 as well for relevant discussion).

In the context of the literature on word identification, we might speculate that the widespread sense that there are some tests which are better than others is based on how well a given test lines up with prescriptive orthographic conventions within some speech community. Given that prescriptive orthographic practices are socially constructed (not all languages/speech communities have them), it is not clear that they would correlate to the *same* degree with the *same* diagnostics cross-linguistically. Disagreements between linguists with respect to the reliability of some diagnostics together with the widespread feeling that some tests are better than others might be a reflection of the languages (or perhaps even constructions in languages) that these linguists are most familiar with and the degree to which the orthographic conventions of these languages line up with this or that diagnostic stated in literature.

We might, however, consider "convergence" to be a more empirically grounded, and perhaps theoretically grounded, way of assessing the relative reliability of tests. The convergence of logically distinct diagnostics has been used to justify categories such as "word" and "phrase" as valid linguistic units, as the quotations from Matthews (2002) and Levine (2017) below illustrate respectively.

For words:

> No criterion is either necessary or sufficient, as Bazell ... made clear long ago. But they are relevant insofar as, in particular languages, they do tend to coincide. A form which is cohesive need not logically consist of elements whose order is fixed. (Matthews 2002: 276)

For phrasal constituents:

> The two phenomena which appeal to unithood must, in other words, be fundamentally independent. Normal methodological considerations then make it highly unlikely that the joint appeal to syntactic unithood from two independent sources envisioned here arose from coincidence. (Levine 2017: 13)

If we work our way backwards from such statements, then tests are reliable insofar as they tend to converge, because insofar as they tend to converge they are identifying (abstract?) constituents.

In what follows, we attempt to assess the relative reliability of certain tests by assessing the degree to which they converge with other tests in general. We report two findings: (i) there are some clear correlations between certain specific tests (e.g. free occurrence and segmental phonological processes); (ii) there is no overall tendency for any constituency test to be more reliable than another as

judged by convergence. What this means is that for some tests, one can predict with some degree of accuracy what other tests they are more or less likely to converge with. However, for a given test one cannot say whether it is more likely to converge than any other test in general. Where possible, we point to some fairly straightforward functional motivations which have been pointed out in the literature. Overall the results suggest that edges ("junctures", "boundaries") might be a source of more meaningful generalizations as opposed to span-defined units such as "word" or "phrase".

6.2 Correlations between domains

Before presenting the results, some remarks regarding comparison are in order. The comparison of word/constituency tests cross-linguistically is complicated by a number of factors, two of which should be mentioned. First, we can compare constituency convergence in terms of convergence at individual edges of structure (e.g. left or right edge) or at both edges simultaneously. We will refer to the former as EDGE CONVERGENCE and the latter as SPAN CONVERGENCE. Secondly, constituency domains can be compared on different levels of abstraction. For instance, we could ask how well non-interruption, regardless of whether and how it is fractured into subtests, converges with domains related to accent/stress marking. If we wanted to get more granular we could ask how well non-interruption by a single free form converges with the minimal fracture of an accent-based domain. We will, therefore, be presenting results at different levels of abstraction corresponding to different levels in the taxonomic hierarchy of constituency tests that emerges from fracturing.

We exclude discussion of some test types that only have one example in our data set (e.g. "play language" in Zenzontepec Chatino).[7] We note that our results are preliminary as they only contain 463 test results from 17 languages. Furthermore, future research might involve applying and or further operationalizing more constituency domains which could change the results. We will also ignore fractures of recursion-based diagnostics such as those based on whether the marking is syndetic or asyndetic, or same or different subject clauses etc. This is done in order to simplify the discussion.

In what follows, we assess the relationships between individual domains using correlation matrices. Correlation matrices present the correlations between different tests. In order to present these correlations all variables are coded as binary variables. We use the Kendall rank correlation coefficient, referred to as

[7] This does not mean that we think this test is irrelevant. Rather, it means that future research is needed in order to compare the relevant domain cross-linguistically.

Kendall's tau, as our correlation metric. This metric measures the ordinal association between two variables. The meaning of a correlation metric in relation to constituency test convergence requires some commentary. Imagine that we have two tests, *x* and *y*. If *x* always converges with *y*, the correlation coefficient will be 1. If these tests never converge with one another, the correlation coefficient will be -1, which could be conceptualized as predictable divergence. If two tests have no tendency to either converge or diverge, the correlation coefficient will be 0. Constituency domains which tend to converge with one another will, therefore, show positive correlations. Note that two constituency tests can be non-convergent on their spans, but convergent on one of their edges.

The correlation plot in Figure 6 shows the correlations between tests in terms of span convergences. The correlation plot in Figure 7 provides correlations for left and right edges, respectively. These figures provide overviews of the tests coded by "Abstract Type." This means that the results pool fractures of constituency tests (e.g. minimal and maximal domains of free occurrence are coded together).[8]

Looking at spans as a whole, there are positive correlations and most of them are under 0.2, i.e. very weak. In fact, tests at an abstract level are more likely to be misaligned than not, since most correlations are negative. The strongest negative correlation, which is still considered moderate, is between recursion-based tests and suprasegmental domains (-0.23). When we consider span convergence, therefore, tests in the abstract are less likely to converge than not. When we look at edge convergences separately, cf. Figure 7, we see a different pattern.

The correlations become positive in the aggregate and statistically stronger when we consider edges by themselves. For left-edge convergence, there is a relatively strong correlation between non-interruption and non-permutability (0.54), followed by moderate correlations between free occurrence and non-interruption (0.41), and non-interruption and ciscategorial selection (0.36).

In general, the majority of test domains exhibit moderate or weak positive correlations with each other, especially those involving non-permutability. Deviations from biuniqueness, however, tend not to converge on left edges.

For right-edge convergence, the strongest positive correlations are found between free occurrence with segmental and suprasegmental processes (0.38, 0.28).

Domains defined by free occurrence tend to align more strongly than other domains on the right edge in general: we also see moderate correlations with ciscat-

[8] Abbreviations used in the figures: Deviations = "Deviations from biuniqueness"; Non-interrupt. = "Non-interruptability"; Non-permut. = "Non-permutability"; Free_occur. = "Free occurrence"; Selection = "Ciscategorial selection"; Segmental. = "Segmental phonological processes/domains"; Supraseg. = "Suprasegmental phonological processes/domains".

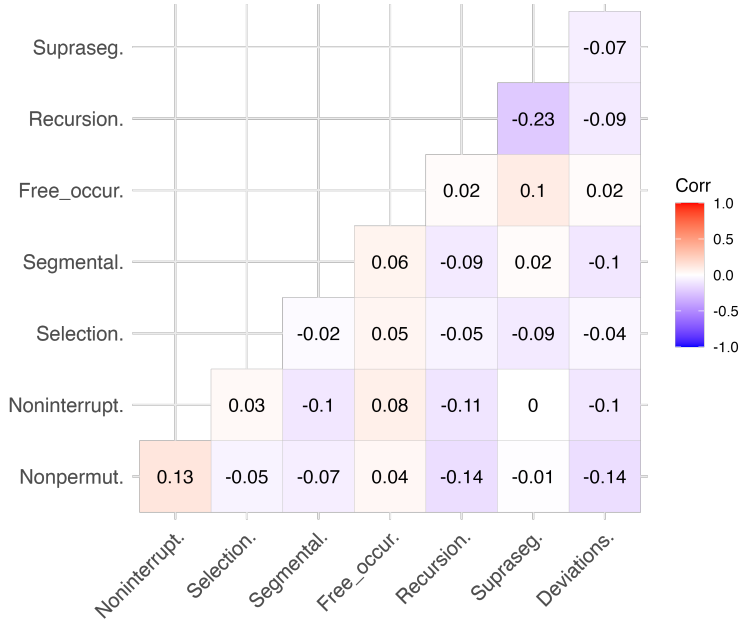

Figure 6: Correlations between test domains over the whole span by abstract type

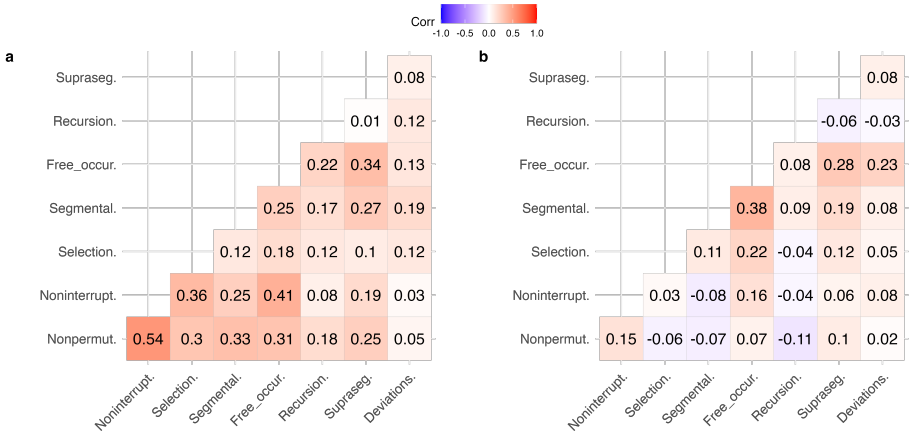

Figure 7: Correlations between test domains on the left (a) and right (b) edge by abstract type

egorial selection and deviations from biuniqueness. Segmental and suprasegmental processes are also weakly correlated with each other. On diachronic grounds, it is not surprising that suprasegmental and segmental processes should line up on an edge. The presence of a prominent syllable can result in segmental changes over time (Bybee et al. 1998), for instance, but prominent syllables are almost always attached to the edge of their domains.

The domains found in Figures 6–7 are perhaps too abstract to develop specific explanations. We consider more fine-grained domains next, taking into account domains fractured according to whether they are minimal or maximal, where this fraction is available. If not, we break apart tests by highly frequent cross-language fractures. In the case of non-interruption, the test is broken up into distinctions between simplex, complex, and multipositional interrupting elements. In the case of non-permutability, we break them apart according to whether the tests have scopal or non-scopal interpretations. Deviations of biuniqueness are not fractured at all, because there are no recurrent cross-language fractures nor minimal/maximal domains. All other tests are fractured across minimal and maximal domains.

Overall, correlations across spans are weak also when taking into account more specific fractures. There are a few moderate positive correlations (>0.2), all but one with minimal domains, as illustrated in Table 2. We can see that all but one of the test pairs involves a free occurrence test. The minimal free occurrence spans have a weak tendency to converge with minimal spans of segmental processes and recursion-based tests. Maximal free occurrence tests have a weak tendency to converge with non-interruption by a simplex form. Spans defined by non-interruption by a form that can variably order are weakly correlated with minimal spans defined by ciscategorial selection. The full table is found on Zenodo (Auderset & Tallman 2023).

Table 2: Pairwise correlations (Kendall's τ) between test domains over cross-language and minimal-maximal fractures across spans. Rows with weak or no correlations ($-0.2 \geq x \leq 0.2$) were excluded.

Test1	Test2	Correlation
Noninterrupt.simpl	FreeOccur.max	0.20
Noninterrupt.multipos	Selection.min	0.24
Recursion.min	FreeOccur.min	0.26
FreeOccur.min	Segmental.min	0.30

Once again, when we consider edge convergences, stronger relationships appear, as can be seen in Table 3. First, we observe that there are more meaningful

and stronger convergences on the left edges than on the right edges, as was already the case when considering only abstract types as a whole. We also see that there are more convergences in minimal domains than maximal ones. Many of the minimal test domains target only the verb core and thus have a higher probability to converge than maximal spans, which mostly target spans larger than the verb core. There are no negative correlations below -0.2, that is, there is no general tendency to be misaligned when considering only edges.

Many of the stronger correlations involve the minimal domain of free occurrence, segmental and suprasegmental processes, often combined with non-permutability and non-interruption. Maximal domains overall tend to have lower convergences than minimal ones. A few chapters of this volume suggest that maximal domains might be more likely to indicate phrase-level structures (Gutiérrez & Uchihara 2024, Tallman 2024a). If convergences are more likely to hit edges of structural shift from morph to utterance (i.e. words), this difference between minimal and maximal domain convergence is potentially understandable.

Table 3: Pairwise correlations (Kendall's τ) between test domains over cross-language and minimal/maximal fractures on the left and right edges. Minimal domains are listed first, followed by maximal domains. Rows with weak or no correlations ($-0.2 \geq x \leq 0.2$) were excluded.

Test1	Test2	Corr.Left	Corr.Right
Noninterrupt.simpl	Noninterrupt.compl	0.06	0.20
Noninterrupt.simpl	Nonpermut.rigid	0.38	0.11
Noninterrupt.simpl	Selection.min	0.26	-0.08
Noninterrupt.simpl	Recursion.min	0.09	0.22
Noninterrupt.simpl	Deviations	0.23	0.04
Noninterrupt.simpl	Segmental.min	0.29	0.05
Noninterrupt.compl	Nonpermut.scopal	0.21	-0.06
Noninterrupt.compl	Recursion.min	0.46	0.05
Noninterrupt.compl	Segmental.min	0.21	0.06
Noninterrupt.compl	Supraseg.min	0.23	0.13
Noninterrupt.multipos	Selection.min	0.39	0.23
Nonpermut.rigid	Nonpermut.scopal	0.26	0.24
Nonpermut.rigid	Selection.min	0.20	0.11
Nonpermut.rigid	Supraseg.min	0.36	0.02
Selection.min	FreeOccur.min	0.26	0.36
Selection.min	Supraseg.min	0.24	0.11
Recursion.min	FreeOccur.min	0.31	0.29

Recursion.min	Segmental.min	0.30	0.12
Recursion.min	Supraseg.min	0.24	0
FreeOccur.min	Deviations	0.22	0.29
FreeOccur.min	Segmental.min	0.43	0.41
FreeOccur.min	Supraseg.min	0.37	0.23
Deviations	Segmental.min	0.38	0.13
Deviations	Supraseg.min	0.23	0.15
Segmental.min	Supraseg.min	0.36	0.23
Noninterrupt.simpl	Noninterrupt.compl	0.06	0.20
Noninterrupt.simpl	Nonpermut.rigid	0.38	0.11
Noninterrupt.simpl	FreeOccur.max	0.26	0.28
Noninterrupt.simpl	Deviations	0.23	0.04
Noninterrupt.simpl	Segmental.max	0.27	-0.01
Noninterrupt.compl	Nonpermut.scopal	0.21	-0.06
Noninterrupt.compl	FreeOccur.max	0.23	0.09
Noninterrupt.compl	Supraseg.max	0.23	0.13
Nonpermut.rigid	Nonpermut.scopal	0.26	0.24
Nonpermut.rigid	FreeOccur.max	0.52	0.01
Nonpermut.rigid	Supraseg.max	0.36	0.02
Nonpermut.scopal	FreeOccur.max	0.26	0.11
Selection.max	FreeOccur.max	0.25	0.13
FreeOccur.max	Supraseg.max	0.28	-0.05
Deviations	Supraseg.max	0.23	0.15
Segmental.max	Supraseg.max	0.30	-0.02

6.3 Predicting convergence

In this section we attempt to discern whether there is an overall tendency for some domains to converge more than others. First we need to discuss some metrics of convergence. One can discern the relative importance of domains based on how often they converge with other diagnostics. Each coded domain or test result can be coded with a an ABSOLUTE CONVERGENCE number. If a domain converges with no other tests in a language, its ABSOLUTE CONVERGENCE is 1. We assign each domain a RELATIVE CONVERGENCE METRIC by language. This takes the absolute convergence level and divides it by the total number of tests applied in a language. Thus a domain which converges with no other domains in a language for which 10 tests were applied has a relative convergence of 0.1. In a given language the relative convergence level is perhaps a more accurate metric

of the convergence strength of a given test. The reason is that we expect overall convergence to increase as a matter of chance as the number of tests increases in a given language (Tallman 2021b).

Note that there are three types of absolute and relative convergence: span convergence, left-edge convergence and right-edge convergence.

Figure 8 provides density distributions showing span convergence (blue), right edge convergence (green) and left edge convergence (orange). The density distribution of span convergence is heavily skewed leftwards towards lower numbers. Most domains do not span-converge. Right edge convergence is less skewed to lower relative convergence values, and left edge convergence presents something approaching a uniform distribution (or else shows weakly distinguished bimodality).

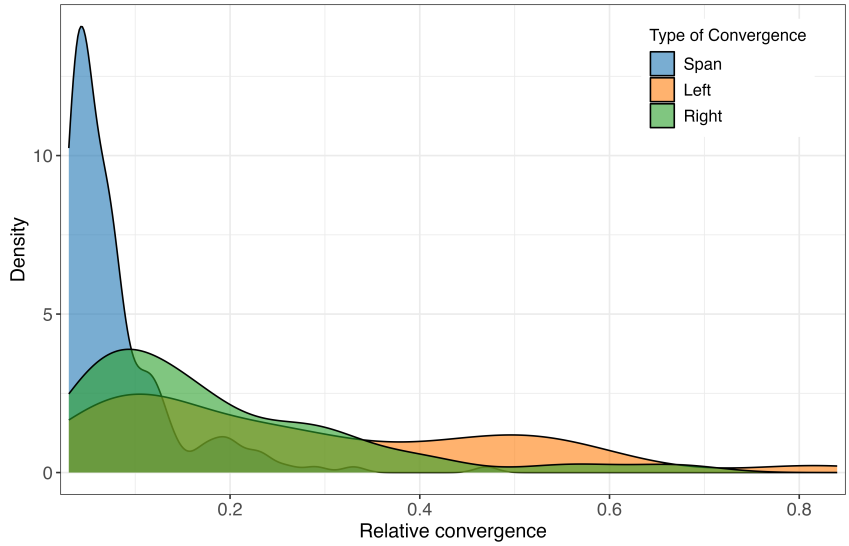

Figure 8: Density distributions of relative convergence at the right and left edge and the whole span across the sample languages.

One way we can discern whether certain domains are more convergent than others would be through comparing their distribution along relative convergence compared to the distribution of all the domains pooled. A more convergent test would exhibit a distribution more skewed to the right compared to the distributions of the domains as a whole. Figures 9 through 11 suggest that none of the tests are obviously more convergent than any others, as they all display distributions which are similarly left-skewed.

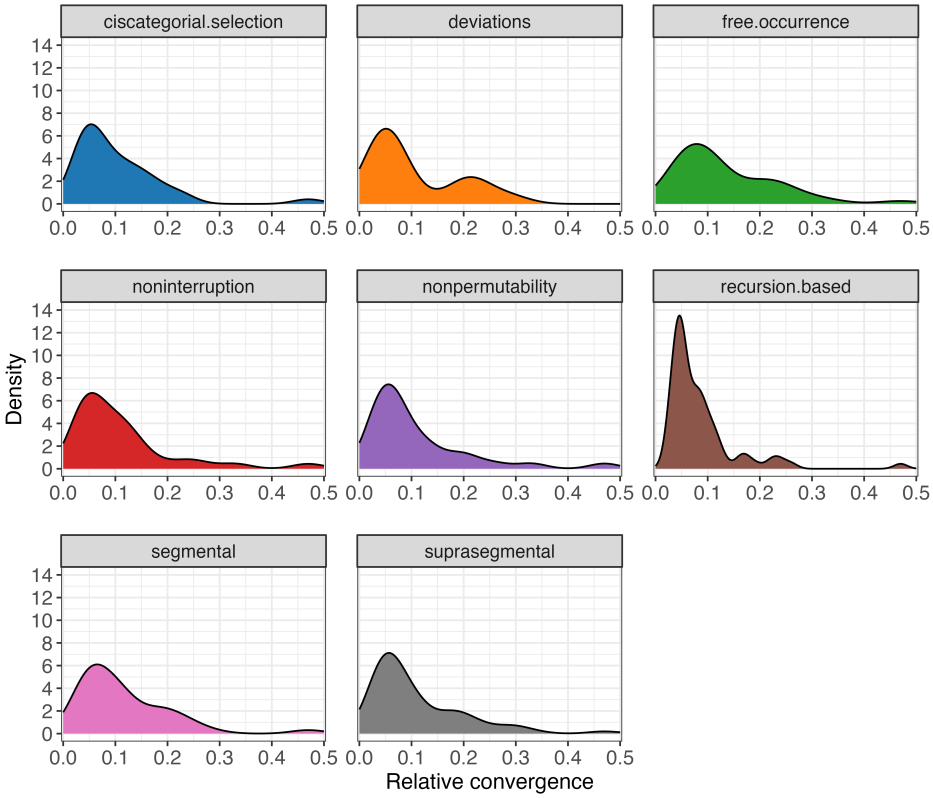

Figure 9: Density distributions of abstract types on relative span convergence in the verbal domain across sample languages. Types with fewer than 5 data points are excluded.

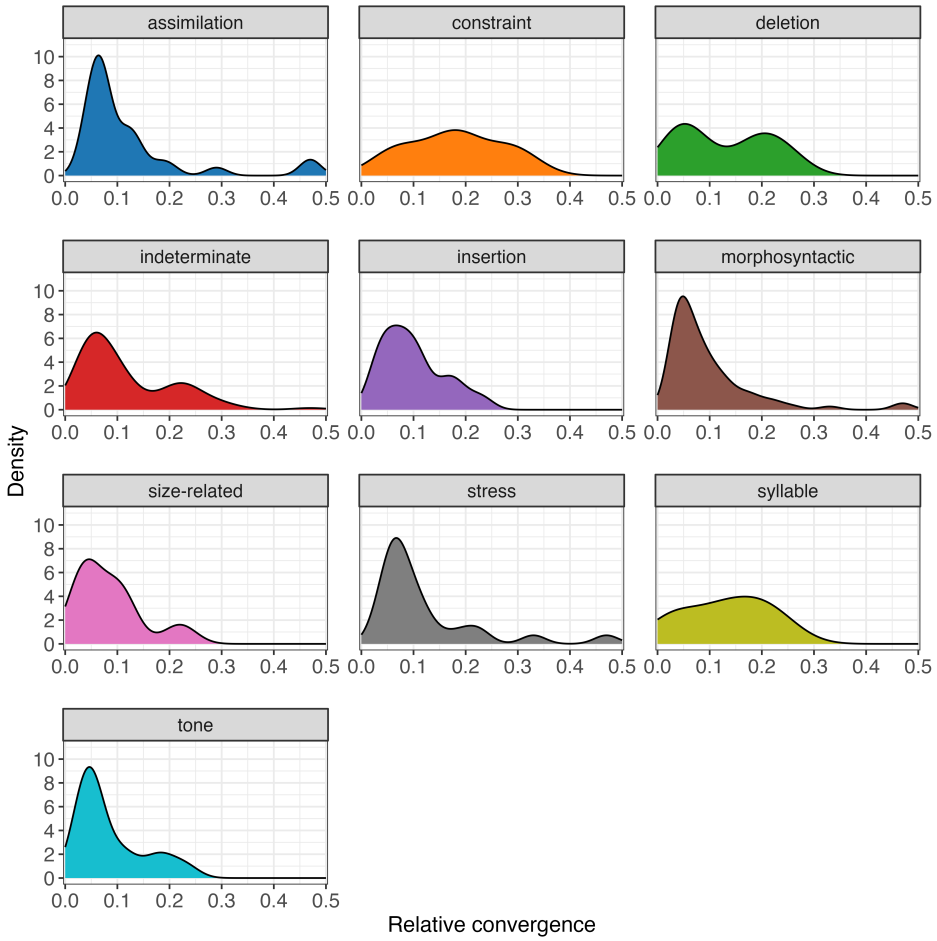

Figure 10: Density distributions of prosodic word domains inspired by the presentation in Bickel et al. (2009) for relative span convergence in the verbal domain across sample languages. Domains with fewer than 5 data points are excluded.

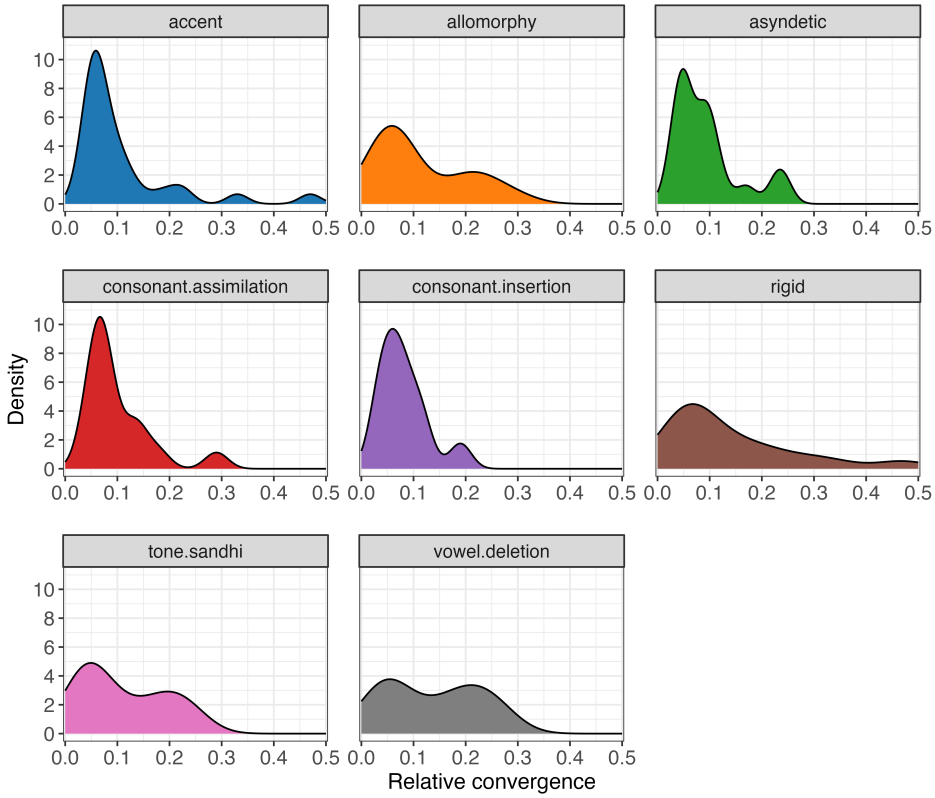

Figure 11: Density distributions of relative span convergence across cross-linguistic fractures with 10 or more tokens in the verbal domain across sample languages.

A statistical test of reliability might attempt to predict convergence level from the domain type. A more reliable domain should predict a higher convergence level than a less reliable one.

For span convergence a random forest was constructed in order to assess whether any of the domains might be good predictors of convergence level. A random forest model is a classification algorithm that aggregates over a multitude of decision trees. It is often used for variable selection and other classification tasks. It requires a dependent variable, which is the variable to be predicted, and predictor variables. We use absolute convergence as the dependent variable and the classification of different domains at all levels of abstraction as predictors. This includes abstract types, cross-language fractures, minimal and maximal domains, and the prosodic-word domain classifications. The model outputs error rates for each level of the dependent variables in a confusion matrix and an overall error rate for the model. However, the accuracy of the model should not be interpreted by itself. Rather it has to be interpreted against the baseline value in order to adjust for the skewness of the data. The baseline value can be understood as the accuracy value an RF would have it it simply chose the most frequent value for the dependent variable every time.

The random forest always predicts level 1 convergence for all domains. The baseline classification rate for the random forest is 0.409 and the accuracy is 0.411. This means that if all data points were classified as the most frequent category, the accuracy is roughly 40.9%. The random forest model outperforms the baseline by a negligible amount; its accuracy is at 41.1%. We do not interpret the model as significantly better than chance. As such constituency test classification does not appear to be an obvious predictor of convergence. If we can use convergence as a metric to rank constituency tests in terms of their reliability, then we currently do not have any good reason to think that any constituency tests are better than any others. Future research with a larger dataset, with new or differently defined constituency tests might provide evidence that some tests are superior to others, but we currently do not have strong empirical reasons to make such judgements.

7 The word bisection thesis

Another hypothesis that the data structures developed in this volume can test is the (empirical) word bisection thesis. Tallman (2024 [this volume]) notes that there are two versions of the word bisection thesis. The fiat-based word bisection thesis assumes that a universal distinction between morphosyntactic and phonological words can be maintained because diagnostics for the relevant constituents

can be concocted. There is no sense in arguing against this claim because it has the status of a tautology. The empirical word bisection thesis is more interesting because it maintains that the relevant diagnostics tend to converge with one another to support the morphosyntactic versus phonological word dichotomy in language after language.

Tallman (2021b) attempts to test the empirical word bisection thesis with data from Chácobo. He shows that there are few convergences within morphosyntactic domains and within phonological domains. The paper attempts to articulate the word bisection thesis as a falsifiable hypothesis concerning the (mis)alignment of wordhood tests. Phonological and morphosyntactic tests may misalign with others, but morphosyntactic tests should tend to align with other morphosyntactic tests and phonological tests should tend to align with other phonological tests. Based on this methodology, convergence between tests is not meaningful by itself, however. As the number of tests increases, the probability that two or more tests align by chance increases. Some notion of "chance convergence" has to be constructed in order to assess an empirically contentful notion of the word bisection thesis. Hypotheses which are falsifiable in principle are not necessarily falsifiable in practice if methods cannot be designed to test them. Literature in the philosophy of science has emphasizes that scientific activity is not only narrowly concerned with theory construction, but also with designing experimental ideas, analytic techniques and new kinds of technologies that can be used to test (falsify) hypotheses (Hacking 1983: 214; Mayo 2018). Tallman (2021b) develops a methodology for calculating chance convergence between wordhood tests that relies on a simulated null distribution. The results suggest no support for the version of the word bisection hypothesis he constructs.

The constituency database allows us to give a first pass assessment of the word bisection hypothesis with more languages. Ideally, a method would also be used to construct chance probability, but we will leave that for future research. Here, we will present simpler metrics that can be derived from basic arithmetic. There are two main results from the current study that we wish to emphasize: (i) There is interesting language variation with respect to how strongly convergent word constituents are supported (see Figure 4 above). (ii) While there are some constituents that are strong word candidates for "word" in terms of convergence, cases where morphosyntactic *and* phonological words appear to be motivated based on convergence are less common and/or less obvious.

Figure 12 displays the relative convergence levels for phonological tests. Each panel displays a nominal or verbal planar structure in a given language of the sample. The y-axis shows the absolute number of convergences per relative span size, which is represented on the x-axis. We can give a preliminary assessment of

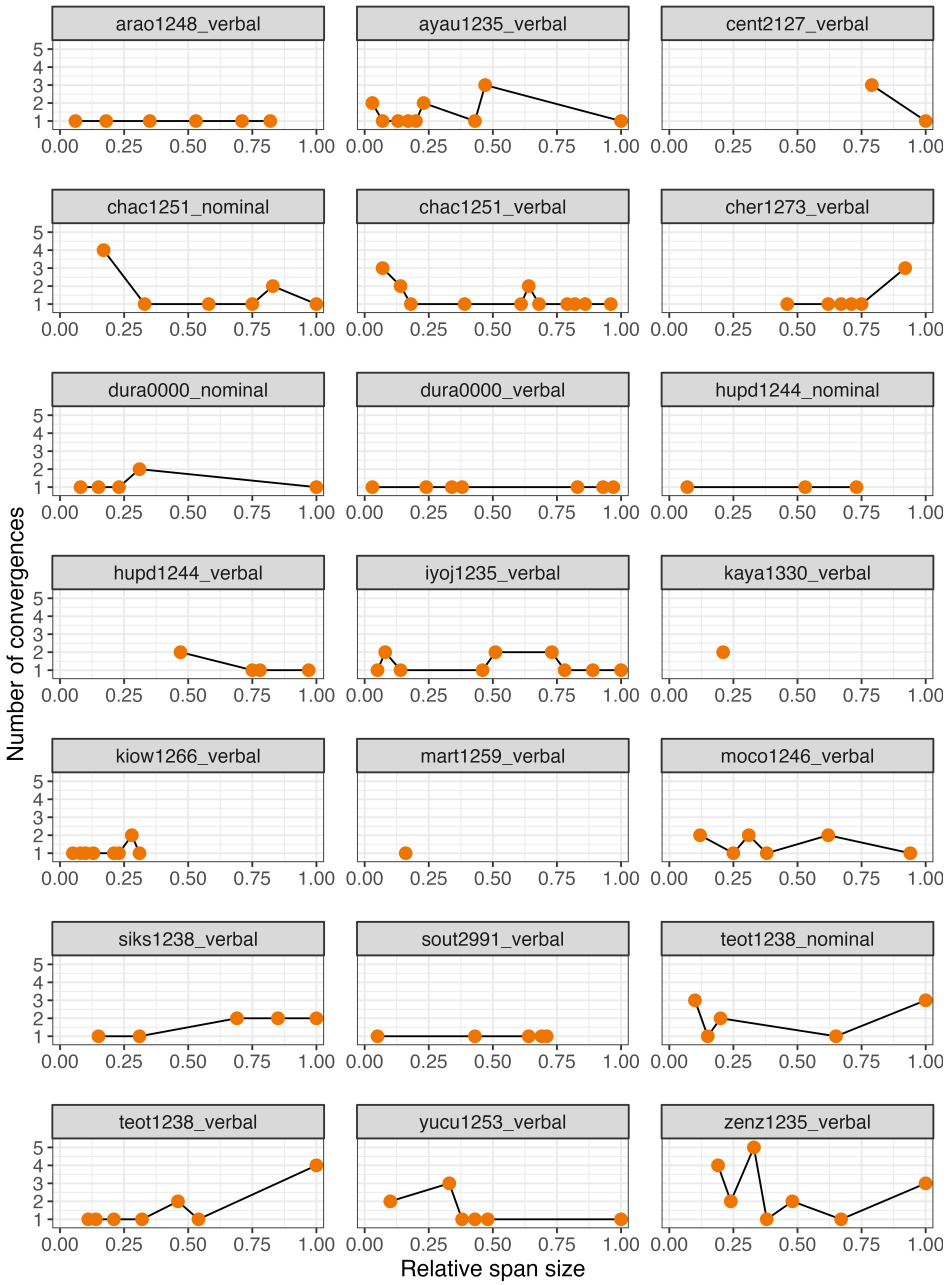

Figure 12: Distribution of relative convergence versus span size in phonological domains by planar structure.

the strength of a wordhood proposal based on a combination of absolute convergence and the number of tests that were applied in each language. An ideal case where phonological wordhood is supported would show a spike upward (high convergence) in relation to a relatively low number of tests applied. To the extent that convergence supports phonological wordhood, the strongest case appears in the verbal domain of Zenzontepec Chatino (with 5 convergences). The Chácobo nominal domain also displays some evidence for phonological wordhood (contrast this with the verbal domain, Tallman 2021b). The case of Teotitlán del Valle Zapotec is somewhat difficult, because although there are a relatively large number of convergences, these appear in a domain that most authors would consider to be an utterance/sentence level grouping (see Gutiérrez & Uchihara (2024) for discussion). We would also say that the phonological word in the Central Alaskan Yupik verbal domain is relatively well supported. While the convergence level is only 3, only 4 phonological tests were applicable in this case.

In the morphosyntactic domain (Figure 13), no layer of structure goes beyond a convergence level of 4. Central Alaskan Yupik, Zenzontepec Chatino, and Duraznos Mixtec seem to display the strongest candidates for morphosyntactic wordhood. Note that the latter is somewhat weaker because in Duraznos a larger repertoire of morphosyntactic tests could be applied. Slightly weaker domains appear for Oklahoma Cherokee, Siksika, Mocovi, and Mẽbêngôkre verbal domains.

There are only two languages that provide some type of support for the word-bisection thesis: Central Alaskan Yupik and Zenzontepec Chatino. In both cases, there are domains with relatively high convergences in both morphosyntax and phonology. While some degree of convergence appears to be the norm, the more typical pattern thus far is that either there is a highly convergent phonological domain or a highly convergent morphosyntactic one, but not both.

We emphasize again that the meaningfulness of the (non)convergences across languages is an open question both on methodological and theoretical grounds. On methodological grounds, more realistic simulation methods might find that the apparently highly convergent patterns are not surprising given factors such as the number of tests applied, the number of languages considered, the tendency for tests to nest, and the hypothesis space for test alignment (e.g. the planar structure). On theoretical grounds, researchers could challenge the idea that convergence is the right notion for the assessment of the word bisection thesis. We might also find independent reasons to consider some tests as more theoretically relevant than others. There are other tests that have not yet been included in the database (e.g. proform replacement), but whose inclusion might change the picture as well.

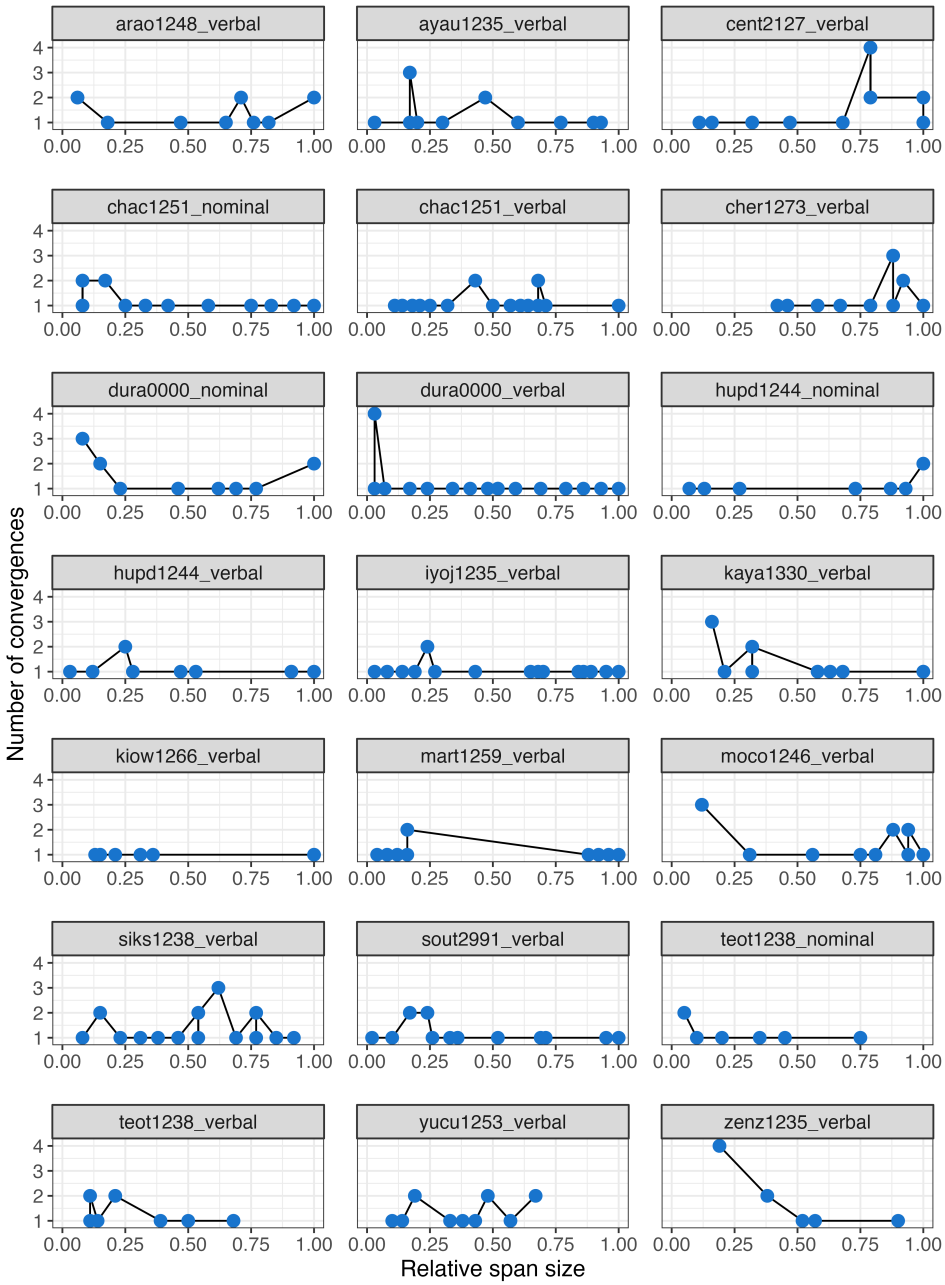

Figure 13: Distribution of relative convergence versus span size in morphosyntactic domains by planar structure

8 Summary and conclusion

The goal of this chapter is to summarize the construction, conceptualization and structure of the constituency and convergence data set. We also, in a general sense, show how the data set can be used to investigate typological questions in linguistics.

Apart from developing simulation methods as described in the previous section, future research can be concerned with developing more constituency tests, attempting to tease out an operationalizable distinction between wordhood and phrasehood level tests (or levels in general). A fuller account of convergences in nominal domains also needs to be provided. In this book we focused mostly on the verb, because we viewed this category as more consistently associated with problems of wordhood, probably because of its relatively high syntagmatic complexity compared to the noun. If both verbal and nominal domains are considered, an actual assessment of the degree to which verbal and nominal constituency structures are homologous could be given (e.g. some version of X' theory could be tested empirically rather than assumed).

A number of phonological domains are also likely missing across the languages. For instance, there is a relative absence of claims or information concerning utterance level phenomena in the studies of this volume. This is a natural consequence of the project starting with a focus on wordhood, but now that it has been revealed that a focus uniquely on wordhood is at best methodologically problematic and, at worst, incoherent, higher-level prosodic domains ought to be included.

Deviations from biuniqueness are also relatively superficially considered in the current approach. This is because in the current approach, deviation domains are fractured according to the type of deviation from biuniqueness (e.g. extended exponence, suppletion etc.). A great deal of complexity and variation is hidden behind such designations. Future research might be concerned with finding some way of syncing current studies on paradigmatic complexity and morphomic structure (e.g. Herce 2023) with a broader study of constituency.

References

Ackermann, Robert John. 1985. *Data, instruments, and theory: A dialectical approach to understanding science*. New Jersey: Princeton University Press.

Adger, David. 2003. *Core syntax: A minimalist approach*. Cambridge: Oxford University Press.

Aikhenvald, Alexandra Y., R. M. W. Dixon & Nathan M. White. 2020. The essence of 'word'. In Alexandra Y. Aikhenvald, R. M. W. Dixon & Nathan M. White (eds.), *Phonological word and grammatical word: A cross-linguistic typology*, 1–24. Oxford: Oxford University Press. DOI: 10.1093/oso/9780198865681.003.0001.

Arcodia, Giorgio Francesco & Bianca Basciano. 2020. Morphology in Sino-Tibetan languages. In *Oxford Research Encyclopedias, Linguistics*. Oxford: Oxford University Press. DOI: 10.1093/acrefore/9780199384655.013.530.

Auderset, Sandra & Adam J. R. Tallman. 2023. *Constituency and convergence/constituency database: 1.0.0 [data set]*. DOI: 10.5281/zenodo.10076550.

Austin, Peter & Joan Bresnan. 1996. Non-configurationality in Australian Aboriginal languages. *Natural Language & Linguistic Theory* 14(2). 215–268.

Bauer, Laurie. 2017. *Compounds and compounding*. Cambridge: Cambridge University Press.

Bazell, Charles Ernest. 1953. *Linguistic form*. Istanbul: Istanbul Press.

Bickel, Balthasar. 2015. Distributional typology: Statistical inquiries into the dynamics of linguistic diversity. In Bernd Heine & Heiko Narrog (eds.), *Oxford Handbook of Linguistic Analysis*, 2nd edn., chap. 37, 901–924. Oxford: Oxford University Press.

Bickel, Balthasar, Goma Banjade, Martin Gaenszle, Elena Lieven, Netra Prasad Paudyal, Ichchha Purna Rai, Manoj Rai, Novel Kishore Rai & Sabine Stoll. 2007. Free prefix ordering in Chintang. *Language* 83(1). 43–73.

Bickel, Balthasar, Kristine A. Hildebrandt. & René Schiering. 2009. The distribution of phonological word domains: A probabilistic typology. In Janet Grijzenhout & Kabak Baris (eds.), *Phonological Domains: Universals and Deviations*, 47–78. Berlin: De Gruyter Mouton. DOI: 10.1515/9783110217100.1.47.

Bickel, Balthasar & Johanna Nichols. 2002. Autotypologizing databases and their use in fieldwork. In *Proceedings of the international LREC workshop on resources and tools in field linguistics, Las Palmas*, vol. 2627.

Bickel, Balthasar, Johanna Nichols, Taras Zakharko, Alena Witzlack-Makarevich, Kristine A. Hildebrandt, Michael Riessler, Lennart Bierkandt, Fernando Zúñiga & John B. Lowe. 2017. *The AUTOTYP typological databases. Version 0.1.0*. https://github.com/autotyp/autotyp-data/tree/0.1.0.

Boas, Franz. 1911. Introduction. In *Handbook of American Indian languages, bulletin 40, part 1*, 1–83. Washington D.C.: Bureau of American Ethnology.

Booij, Geert. 2005. *The grammar of words*. Oxford & New York: Oxford University Press.

Brown, E. Keith & Jim E. Miller. 1980. *Syntax: A linguistic introduction to sentence structure*. London: Hutchinson University Library.

Bybee, Joan. 2001. *Phonology and language use*. Cambridge: Cambridge University Press.

Bybee, Joan. 2010. *Language, usage and cognition*. Cambridge: Cambridge University Press.

Bybee, Joan, Parmotima Chakraborti, Dagmar Jung & Joanne Scheibman. 1998. Prosody and segmental effect: Some paths of the evolution of word stress. *Studies in Language* 2(22). 267–314.

Carnie, Andrew. 2010. *Constituent structure*. New York: Oxford University Press.

Carruthers, Peter. 2006. *The architecture of mind: Massive modularity and the flexibility of thought*. Oxford: Clarendon Press. DOI: 10.1093/acprof:oso/9780199207077.001.0001.

Chomsky, Noam. 1965. *Aspects of the theory of syntax*. Cambridge: MIT Press.

Coltheart, Max. 1999. Modularity and cognition. *Trends in Cognitive Science* 3(3). DOI: 10.1016/s1364-6613(99)01289-9.

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

Croft, William. 2010. Ten unwarranted assumptions in syntactic argumentation. In Kasper Boye & Elisabeth Engberg (eds.), *Language usage and language structure*, 313–350. Berlin: De Gruyter Mouton.

Croft, William. 2022. *Morphosyntax: Constructions of the world's languages*. Cambridge: Cambridge University Press.

Dixon, R. M. W. 2010. *Basic linguistic theory, Vol. 2: Grammatical Topics*. Oxford: Oxford University Press.

Dixon, R. M. W. & Alexandra Y. Aikhenvald. 2002. Word: A typological framework. In R. M. W. Dixon & Alexandra Y. Aikhenvald (eds.), *Word: A Cross-linguistic Typology*, 1–41. Cambridge: Cambridge University Press. DOI: 10.1017/CBO9780511486241.002.

Dryer, Matthew S. & Martin Haspelmath. 2013. *WALS online*. Matthew S. Dryer & Martin Haspelmath (eds.). http://wals.info/. Leipzig.

Duzerol, Minella. 2024. Constituency in Martinican (creole, Martinique). In Adam J.R. Tallman, Sandra Auderset & Hiroto Uchihara (eds.), *Constituency and convergence in the Americas*, 419–446. Berlin: Language Science Press. DOI: ??.

Easterday, Shelece, Matthew Stave, Marc Allassonnière-Tang & Frank Seifart. 2021. Syllable complexity and morphological synthesis: A well-motivated positive complexity corerlation across subdomains. *Frontiers in Psychology* 12. DOI: 10.3389/fpsyg.2021.638659.

Emkow, Carola. 2019. *A grammar of Araona* (Outstanding grammars from Australia 19). Munich: LINCOM.

Epps, Patience. 2024. Constituency in Hup: Synchronic and diachronic perspectives. In Adam J.R. Tallman, Sandra Auderset & Hiroto Uchihara (eds.), *Constituency and convergence in the Americas*, 447–482. Berlin: Language Science Press. DOI: ??.

Good, Jeff. 2016. *The linguistic typology of templates*. Cambridge: Cambridge University Press.

Greenberg, Joseph. 1954. A quantitative approach to the morphological typology of language. In Robert F. Spencer (ed.), *Method and perspective in anthropology*, 192–220. Minneapolis: University of Minnesota Press.

Gutiérrez, Ambrocio & Hiroto Uchihara. 2024. Words as emergent constituents in Teotitlán del Valle Zapotec. In Adam J.R. Tallman, Sandra Auderset & Hiroto Uchihara (eds.), *Constituency and convergence in the Americas*, 305–365. Berlin: Language Science Press. DOI: ??.

Hacking, Ian. 1983. *Representing and intervening: Introductory topics in the philosophy of natural science*. Cambridge: Cambridge University Press.

Haspelmath, Martin. 2009. An empirical test of the agglutination hypothesis. In Sergio Scalise, Elisabetta Magni & Antonietta Bisetto (eds.), *Universals of language today*, 13–29. Cham: Springer Science. DOI: 10.1007/978-1-4020-8825-4_2.

Haspelmath, Martin. 2011. The indeterminacy of word segmentation and the nature of morphology and syntax. *Folia Linguistica* 45(1). 31–80. DOI: 10.1515/flin.2011.002.

Haspelmath, Martin. 2020. The morph as a minimal linguistic form. *Morphology* 30. 117–134.

Herce, Borja. 2023. *The typological diversity of morphomes: A cross-linguistic study of unnatural morphology*. Oxford: Oxford University Press.

Lara, Luis Fernando. 2004. ¿Es posible una teoría de la palabra? *Lexis* 1-2(XXVII). 401–427.

Ledgeway, Adam. 2017. Syntheticity and analyticity. In Andres Dufter & Elisabeth Stark (eds.), *Manual of Romance Morphosyntax and syntax (MRL 17)*, Günter Holtus and Fernando Sánchez Miret, chap. 23, 839–886. Berlin: Walter de Gruyter.

Levine, Robert D. 2017. *Syntactic analysis: An HPSG-based approach*. Cambridge: Cambridge University Press.

Lorenz, David & David Tizón-Couto. 2019. Chunking or predicting - frequency information and reduction in the perception of multi-word sequences. *Cognitive Linguisics* 4(30). 751–784.

Martinet, André. 1962. *A functional view of language*. Oxford: Clarendon Press.

Matthews, Peter H. 2002. What can we conclude? In R. M. W. Dixon & Alexandra Y. Aikhenvald (eds.), *Word: A cross-linguistic typology*, 266–281. Oxford: Oxford University Press.

Mayo, Deborah. 2018. *Statistical inference as severe testing: How to get beyond the statistics wars*. Cambridge: Cambridge University Press.

Osborne, Timothy J. 2018. Tests for constituents: What they really reveal about the nature of syntactic structure. *Language Under Discussion* 1(5). 1–41.

Payne, Doris L. 1990. Morphological characteristics of lowland South American languages. In Doris L. Payne (ed.), *Amazonian linguistics: Studies in lowland South American languages*, 213–241. Austin: University of Texas Press.

Payne, Thomas E. 2006. *Exploring language structure: A student's guide*. New York: Cambridge University Press.

Phillips, Colin. 1996. *Order and structure*. Cambridge Massachusetts: MIT. (Doctoral dissertation).

Pitman, Donald. 1980. *Bosquejo de la gramática araona* (Notas Lingüísticas de Bolivia 9). Cochabamba: Summer Institute of Linguistics.

Rasskin-Gutman, Diego. 2005. Modularity: Jumping forms within morphospace. In Wener Callebaut, Diego Rasskin-Gutman & Herbet A. Simon (eds.), *Modularity. Uderstanding natural complex systems*, 207–219. Cambridge: MIT Press.

Sapir, Edward. 1921. *Language*. New York: Harcourt, Brace & World.

Schiering, René, Balthasar Bickel & Kristine A. Hildebrandt. 2012. Stress-time = word-based? Testing a hypothesis in prosodic typology. *STUF-Language Typology and Universals* 65(2). 157–168.

Schwegler, Armin. 1990. *Analyticity and syntheticity: A diachronic perspective with special reference to Romance languages*. Berlin: Mouton de Gruyter.

Skirgård, Hedvig, Hannah J Haynie, Damián E Blasi, Harald Hammarström, Jeremy Collins, Jay J Latarche, Jakob Lesage, Tobias Weber, Alena Witzlack-Makarevich, Sam Passmore, et al. 2023. Grambank reveals the importance of genealogical constraints on linguistic diversity and highlights the impact of language loss. *Science Advances* 9(16). eadg6175.

Tallman, Adam J. R. 2021a. Analysis and falsifiability in practice. *Theoretical Linguistics* 47(1-2). 95–112.

Tallman, Adam J. R. 2021b. Constituency and coincidence in Chácobo (Pano). *Studies in Language* 45(2). 321–383. DOI: 10.1075/sl.19025.tal.

Tallman, Adam J. R. 2024a. Graded constituency in the Araona (Takana) verb complex. In Adam J.R. Tallman, Sandra Auderset & Hiroto Uchihara (eds.), *Constituency and convergence in the Americas*, 545–602. Berlin: Language Science Press. DOI: ??.

Tallman, Adam J. R. 2024b. Introduction: Phonological and morphosyntactic constituency in cross-linguistic perspective. In Adam J.R. Tallman, Sandra Auderset & Hiroto Uchihara (eds.), *Constituency and convergence in the Americas*, 1–84. Berlin: Language Science Press. DOI: ??.

Tallman, Adam J. R. & Sandra Auderset. 2023. Measuring and assessing indeterminacyand variation in the morphology-syntax distinction. *Linguistic Typology* 27(1). 113–156.

Tallman, Adam J. R. & Patience Epps. 2020. Morphological complexity, autonomy, and areality in western Amazonia. In Francesco Gardani & Peter Arkadiev (eds.), *The complexities of morphology*, 230–264. Oxford: Oxford University Press. DOI: 10.1093/oso/9780198861287.003.0009.

Uchihara, Hiroto. 2024. Constituency in Oklahoma Cherokee. In Adam J.R. Tallman, Sandra Auderset & Hiroto Uchihara (eds.), *Constituency and convergence in the Americas*, 139–177. Berlin: Language Science Press. DOI: ??.

Witzlack-Makarevich, Alena, Johanna Nichols, Kristine A. Hildebrandt, Taras Zakharko & Balthasar Bickel. 2022. Managing AUTOTYP data: Design principles and implementation. In Eve Koller Lauren B. Collister Andrea L. Berez-Kroeker Bradley McDonnell (ed.), *The open handbook of linguistic data management*. Cambridge: MIT Press.

Woodbury, Anthony C. 2024. Constituency in Cup'ik and the problem of holophrasis. In Adam J.R. Tallman, Sandra Auderset & Hiroto Uchihara (eds.), *Constituency and convergence in the Americas*, 85–138. Berlin: Language Science Press. DOI: ??.

Zerilli, John. 2020. *The adaptable mind: What neuroplasticity and neural reuse tell us about language and cognition*. Oxford: Oxford University Press.