Research Article

Adam J. R. Tallman* and Sandra Auderset

Measuring and assessing indeterminacy and variation in the morphology-syntax distinction

https://doi.org/10.1515/lingty-2021-0041
Received July 5, 2021; accepted February 5, 2022; published online March 24, 2022

Abstract: We provide a discussion of some of the challenges in using statistical methods to investigate the morphology-syntax distinction cross-linguistically. The paper is structured around three problems related to the morphology-syntax distinction: (i) the boundary strength problem; (ii) the composition problem; (iii) the architectural problem. The boundary strength problem refers to the possibility that languages vary in terms of how distinct morphology and syntax are or the degree to which morphology is autonomous. The composition problem refers to the possibility that languages vary in terms of how they distinguish morphology and syntax: what types of properties distinguish the two systems. The architecture problem refers to the possibility that languages vary in terms of whether a global distinction between morphology and syntax is motivated at all and the possibility that languages might partition phenomena in different ways. This paper is concerned with providing an overarching review of the methodological problems involved in addressing these three issues. We illustrate the problems using three statistical methods: correlation matrices, random forests with different choices for the dependent variable, and hierarchical clustering with validation techniques.

Keywords: clustering; language variation; morphological autonomy; morphology-syntax distinction; random forest; typology

***Corresponding author: Adam J. R. Tallman** [ˈærm̩ ɖʒejmz ɹas ˈtalmn̩], Department of English and American Studies, Friedrich-Schiller-Universität Jena, Jena, Germany,
E-mail: adam.james.ross.tallman@uni-jena.de. https://orcid.org/0000-0003-3524-9300
Sandra Auderset [ˈsanˌdɾɑ ˈɔuˌdərsɛt], Department of Linguistic and Cultural Evolution, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany; and Department of Linguistics, University of California Santa Barbara, Santa Barbara, USA,
E-mail: sandra_auderset@eva.mpg.de. https://orcid.org/0000-0002-4673-4814

1 Introduction

A global distinction between morphology and syntax is presupposed across much linguistic work, whether it professes to be theoretical or descriptive (Stewart 2016). Many theoretical approaches assume that the internal arrangement of words is distinct from that of syntactic structures such as phrases and clauses. For instance, many morphologists argue that morphology is realizational rather than combinatorial like syntax (Anderson 1992; Blevins 2016; Stump 2016; Zwicky 1985). Theories differ with respect to how these two systems are integrated or interact and what their essential properties are. Theories vary in the extent to which words play an important part in the overall distinction between morphology and syntax (Stewart 2016) and in terms of how they deal with purported 'boundary cases' such as clitics, compounds, incorporation, and periphrasis. Clitic position, for example, can be treated morphologically or syntactically, depending on the type of clitic (e.g. 'simple' vs. 'special') and the theoretical orientation of the author (Spencer and Luís 2012b). Similarly, incorporation can be treated strictly morphologically (Rosen 1989; Spencer 1995), involve a syntactic movement operation (Baker 1988), or overlapping morphological and syntactic constituent structures (Sadock 1980).

A global morphology-syntax distinction is an important part of the organization of many descriptive grammars or the description of individual languages. Descriptive grammars vary in the extent to which they imply the distinction is arbitrary, motivated, or necessary for describing the language or the languages in question. For instance, in his description of Central Alaskan Yupik (CAY, cent2127) Miyaoka (2012: 18) states that "The 'word' weighs heavily in the grammar of CAY (…) perhaps much more than those of a majority of the world's languages". In contrast, Epps (2004: 125) refers to the suffix label in her grammar of Hup as "to some degree a language-specific convenience". Grammars typically discuss numerous intermediate or indeterminate cases (affixes that behave as clitics, clitics that behave as words, etc.) without necessarily questioning whether the ubiquity of such cases undermines the basic distinction between morphology and syntax or whether it suggests that a different partition is necessary. While some linguists have assumed that words (and therefore the morphology-syntax distinction) are given (Sapir 1921: 33–34), others have discussed recalcitrant problems in word-segmentation (Peterson 2006) and have even suggested that for some languages the distinction is not motivated or important (Pike 1972; Tallman 2018a). Finally, even if these issues regarding the description of specific languages are ignored, the issue of comparability across descriptions remains: the notion of word or what phenomena are considered morphological is not the same across descriptions (Haspelmath 2011; Russell 1999). One then wonders whether the

distinction between morphology and syntax in descriptive grammars is a matter of expositional convenience, rather than one that is motivated linguistically or instantiated in the minds of speakers.

A common response to the latter concerns is to state that the data suggest a continuum approach. Morphology and syntax should not be thought of as discrete components, rather formatives and constructions are distributed on a cline from more morphology-like to more syntax-like phenomena. This idea introduces the question as to how specific languages distribute constructions and elements on such a cline. If elements are distributed uniformly, then it is unclear how the morphology-syntax distinction is motivated. If elements cluster into groups along this cline, then perhaps a fuzzy distinction between the two systems can be motivated. For typologists, these considerations mean moving away from categorizing constructions and elements with respect to presupposed components or sub-components (e.g. Is this clitic positioned with respect to the normal syntax or phrasal morphology of the language? Is this compound a word or a phrase?) to investigating language-internal and cross-linguistic variation with respect to patterns related to the variables that could make up the morphology-syntax distinction cross-linguistically. We refer to this approach to the issue as the 'variationist' approach to the morphology-syntax distinction, since it is takes as its starting point the goal of measuring cross-linguistic variation in the domain under question (Bickel 2010a).This paper explores some of the methodological problems involved in such a perspective.

We identify three problems that a variationist approach to the morphology-syntax distinction could address.

a. **Boundary strength problem**: Languages may vary in the degree to which morphology and syntax are distinct. Some languages might display more indeterminacy than others.
b. **Composition problem**: Languages may differ with respect to how the distinction between morphology and syntax is made. Certain properties (e.g. high degree of allomorphy) may distinguish morphology and syntax in one language and not another. Languages may also vary in the degree to which certain properties help distinguish morphology and syntax.
c. **Architecture problem**: Languages may vary with respect to whether morphology and syntax are distinct at all. Languages may have more than two systems for organizing form-meaning relations or just one.

Section 2 provides further background on the morphology-syntax distinction focusing on the concept of morphological autonomy. The concept of morphological autonomy posits that morphology is distinct from syntax because morphology displays deviations from biunique mappings between form and meaning, while

syntax (typically) does not. This idea is crucial for some of the statistical methods used throughout the paper because it allows us to theoretically ground a distinction between independent and dependent variables. Section 3 provides a description of the languages used in this study and some discussion concerning why they are interesting with respect to the issues laid out above. Section 4 addresses the boundary strength problem, providing different ways of measuring morphological autonomy in individual languages as a typological index (Tallman and Epps 2020). We argue that languages can be distinguished according to how much indeterminacy there is between morphology and syntax. Section 5 describes some aspects of the data and provides a basic methodology that approaches the **boundary strength** problem. Section 6 discusses the issues of addressing the **composition problem** using Random Forest models for illustration. We argue that the problem is difficult to address because it requires selection of a dependent variable, but it is not clear what this should be. Section 7 discusses the **architecture problem**. Addressing the architecture problem involves leveraging statistical methods to assess whether the morphology-syntax distinction is motivated at all in specific languages and across them (Haspelmath 2011). To a certain extent, exploratory methods combined with validation techniques can determine whether a language makes a basic partition into two systems or not, but it is much less clear how to determine what the optimal number of systems actually is. Section 8 provides some concluding remarks and suggestions for future research.

2 Exponence complexity inside and outside morphology

In this section, we focus on the issue of morphological autonomy and its relationship to categories which seem to be indeterminate between morphological and syntactic analyses. Morphological autonomy refers to the general idea that there are properly morphological phenomena which *cannot* be accounted for by reference to syntax or general phonological rules (Anderson 1992, 2015; Aronoff 1994; Booij 1997a, 1997b; Matthews 1991). A popular argument in favor of autonomy with relation to syntax refers to differences in the way form and meaning relate to each other in each of these domains.

Since Matthews (1991) it has been common to emphasize that morphology differs from syntax in that morphological units and constructions display deviations from biunique correspondences between form and meaning. A biunique relation between form and meaning is one where form is predictable from

meaning and meaning is predictable from form. Words are biunique units of syntax because one can predict their meaning from their form. Phrases are biunique insofar as their meaning is predictable from the composition of biunique words. These conditions are not supposed to hold of the units of words or morphological constructions. Below we describe different types of deviations from biuniqueness.

Multiple forms can correspond to a single meaning. There are two ways this can occur: (i) extended (or multiple) exponence and (ii) allomorphy. Extended exponence refers to cases where a specific meaning appears to be realized by more than one formative. For instance, negation in Araona (Bolivia, Takana, arao1248) is expressed by a circumfix (two formatives) *pi-… -ma* as in *pi-mimi-ma* 'not speak'. Allomorphy refers to cases where a single meaning maps to multiple forms, but the forms are mutually exclusive such that they do not occur in the same context. For instance, the inflectional suffixes of Romance verb conjugations display (suppletive) allomorphy in this sense (e.g. *-aba/-ía* both mark imperfective in Spanish (Indo-European: Italic, stan1288) depending on the inflectional class). To the extent that variations in form occur for elements at the syntactic level, that variation should be predictable from general phonological rules or constraints of the language.

A single form can correspond to multiple meanings. Again, there are two types: (i) syncretism and (ii) cumulative exponence. Syncretic exponence typically refers to cases where a single form fills out two or more forms in a paradigm. An example is the widespread pattern in Indo-European languages whereby the nominative and accusative are expressed by the same form in neuter paradigms (e.g. Sanskrit (Indo-European: Indo-Aryan, sans1269) *man-as* 'mind-NOM/ACC.SG.N' and *man-āṃsi* 'mind-NOM/ACC.PL.N'). Cumulative exponence refers to cases where a single form fuses together multiple semantic concepts – typically understood as abstract or grammatical concepts rather than lexical ones. For instance, Tariana (Arawak, tari1256) contains a set of morphemes that express tense and evidentiality together (Aikhenvald 2003: 289) (*=naka* 'visual, present'; *mha* 'non-visual present').

The absence of form can correspond to a single meaning. Morphological analyses sometimes posit that the absence of a certain morph in some position indexes a specific meaning: zero or null morphs (Jakobson 1939; Mel'čuk 2006: 308). For instance, in Algonquian languages (algo1257), independent verb forms call for positing person number affixes and a theme affix that indicates which participant is acting on which (*ni-wa:pam-a-w* 1SG-see-SAP-3-3 'I see him'). In the conjunct mode (roughly a type of subordinate clause), the third person and the theme marker could be analyzed as zero because there is only one analyzable

suffix and the prefix slot is taken up by the conjunct marker (*e-wa:pam-ø-ak-ø* [CONJ-see-SAP>3-1-3] '(that) I saw him') (Wolfart 1973: 51–56).

A form can correspond to no meaning. A frequently cited example is the third stem *-t* of Latin (lati1261) (e.g. *laudā-t* 'have been praised') which occurs in the perfect participle form. The future participle form seems to be built from the perfect participle in that it always contain *-t* (*laudā-t-ūr 'will be praising'*). According to (Aronoff 1994: 32) the perfect participle is 'usually passive' and, thus, the *-t* appears to lose its meaning in the context of the perfect participle (see Mel'čuk 1993: 46–50 for discussion).

There are two types of arguments against the claim that deviations from biuniqueness distinguish morphology and syntax. The first is that not all units described as 'morphological' necessarily display such deviations. The sheer volume of languages described as 'agglutinating' suggests that languages the world over are full of counterexamples to the purported tendency for morphological relations to display deviations from biuniqueness. For instance, the suffixes of South Bolivian Quechua (sout2991), illustrated in example (1), do not vary in form nor substantially in meaning across different verb forms and constructions.

(1) South Bolivian Quechua
suya-raya-chi-sha-lla-wa-nku=puni
wait-without.moving-CAUS-PROG-keep-1OBJ-3PL=certainly
'They keep making me wait' (Gladys Camacho-Rios p.c.)

The second type of counterexample comes from cases where deviations from biuniqueness appear to occur at the level of syntax (Haspelmath 2011: 54–58). examples of extended exponence (one meaning with multiple forms) occurring at the level of syntax are negative *ne …pas* in French (Indo-European: Italic, stan1290) and certain negation strategies in Teotitlán del Valle Zapotec (Otomanguean: Zapotecan, teot1238) (Gutiérrez Lorenzo 2018) as in example (2).

(2) Teotitlán del Valle Zapotec

kēd=ba=llu̱b=di	*Jwáyn*	*low-næz*
NEG=COM=sweep=NEG	Juan	RN.face-street

'Juan didn't sweep the road/the street' Gutiérrez Lorenzo (2018)

Examples of cumulative exponence at the level of syntax can be found in function words that encode more than one category and do not appear to be altogether that uncommon. An example is the relative pronoun from Kashmiri (Indo-European: Indo-Iranian, kash1277), which encodes grammatical relations (nominative), number (singular), and gender (masculine), as can be seen from example (3).

(3) Kashmiri

su	*lədkɨ*	*yus*	*dili*	*chu*	*ro:za:n*	*chu*	*m'o:n*
CORR	boy	REL.NOM.SG.M	Delhi	is	live	is	my
bo:y							
brother							

'The boy who lives in Delhi is my brother' (Wali and Koul 1997: 54)

Thus, biunique relations appear in morphology and non-biunique relations appear in syntax. For those who wish to maintain that deviations from biuniqueness are still important for distinguishing between morphology and syntax, there are two ways to respond to the aforementioned counterexamples: (i) deny that there is necessarily a tight relationship between wordhood and morphological status; (ii) posit that the boundaries between morphology and syntax are fuzzy.

With regards to the first response, some linguists conceptualize morphology as autonomous, but do not view morphology as necessarily architecturally encapsulated in words (understood as the smallest unit of syntax). For instance, in Autolexical syntax, the morphological component can project phrase structures that overlap with syntax (or vice versa). In this framework, incorporated nouns in Greenlandic, for example, are simultaneously represented as heads of syntactic noun phrases and as dependents in the morphological structure of the verb (Sadock 1991). If the Autolexicalist perspective is combined with the idea that morphology be signaled by deviations from biuniqueness, we should expect that some syntactic elements display biuniqueness insofar as some of these are also morphological elements.[1]

In Anderson's A-morphous morphology perspective (Anderson 1992, 2005), there is a special component of 'phrasal morphology' that operates over categories in syntax, or post-lexical phonological constituents. Phrasal affixes (or 'special clitics') might be realized at the end or beginning of a syntactic or phonological phrase (the possessive *'s* in English and some Wackernagel clitics), or even, phonological word (e.g. Romance pronominal clitics). If this is a general property of languages, we would except deviations from biuniqueness to be ubiquitous inside words as well as outside of them in 'syntax'.

However, in their current formulations neither Autolexical syntax nor A-morphous morphology help distinguish between morphology and syntax because they provide no diagnostics for distinguishing these phenomena in a consistent

1 Advocates of Autolexical syntax do not actually propose that deviations from biuniqueness can help delimit morphology and syntax as far as we know. The point here is that the architectural properties of the theory seem to require some independent evidence for positing the boundaries between morphological constituents and syntactic constituents on independent grounds. We have not found any proposals in this regard.

fashion. For example, Anderson's arguments for a phrasal morphology depend on identifying 'special clitics' which cannot be placed by the 'normal rules of syntax' (Anderson 1992: 200) or the 'independently motivated syntax of free elements' (Anderson 2005: 31). But the normal rules of syntax are never defined in any general way, nor are clear diagnostics set up that help the linguist decide whether they are dealing with special or normal syntax in specific cases. Tallman (2018b) takes up this problem in the context of clitic phenomena in Pano languages. While previous studies describe or posit certain clitics as 'phrasal affixes' and morphological in some sense, it only takes a minor adjustment (or no adjustment at all) in what one means by 'normal syntax' or 'independently motivated free elements' to incorporate the relevant phenomena into syntactic structure. Similar considerations carry over to Autolexical syntax: all the various modules of grammar are constrained to be 'simple', but it is not clear when complexification in one domain is justified by simplification in another.[2] Currently, there appears to be no way of evading the wordhood issue vis-à-vis the morphology-syntax distinction without lapsing into circularity (Russell 1999).

One might posit that the boundaries between morphology and syntax are fuzzy (Vincent 2011: 434) or that morphological autonomy 'is not absolute' and is a 'matter of degree' (Aronoff 1994: 166; Blevins 2016: 61–62; Cruschina et al. 2013: 2; Maiden 2013: 42–43, 2011: 49; Smith 2013: 248). In such a perspective, formatives and constructions would vary in the extent to which they are morphology-like or syntax-like. Some constructions and formatives are perhaps indeterminate between which systems they belong to, but the overall patterning is such to justify a global distinction between morphology and syntax. For instance, in a discussion of case paradigms in Estonian (Uralic, esto1258), Blevins (2006: 444) suggests that the agglutinative (biunique) suffixes are marginal in some sense when he states that they 'can be seen to be a limiting rather than a normative case'. The normative situation must be for words to be biunique and word-internal morphological elements to display deviations from biuniqueness. Some type of statistical justification ought to be provided to test whether indeterminate cases are really so marginal as to never challenge the idea that morphology is autonomous (Haspelmath 2011).

At first glance it might seem that the problem can be addressed by coding morphemes or constructs as either 'morphological' or 'syntactic' and developing a metric that captures exponence complexity/deviations from biuniqueness. Then one could assess how well exponence complexity correlates with the word-phrase

2 Nor is it clear when complexification of the entire architecture of the model, say by adding a new module like morphophonology, is justified based on its simplification of another module (Woodbury 1996).

distinction. The problem with this approach is that there is no independent way of classifying elements and constructs as morphological or syntactic in the first place. As argued by Haspelmath (2011) and Tallman (2020), language-specific designations such as affix, word, and phrase are not strictly comparable between languages. A solution to this problem is to treat wordhood diagnostics as typological variables and to ask how well exponence complexity is correlated with these variables (Tallman and Epps 2020). After discussing the languages used in this study in Section 3, we develop a set of wordhood criterial variables and a method for calculating exponence complexity, which we will take to be a metric of (non-) biuniqueness in Section 4.

3 The languages and the data of the study

The sample of languages used for this study is based on an expanded data set from that used in Tallman and Epps 2020.[3] However, we limit the current sample to those languages that have at least 50 data points (i.e. 50 morphemes coded for all variables described below), to ensure that they can be used with the statistical methods we focus on.

The languages are drawn from southwestern Amazonia, except for Central Alaskan Yupik, which was chosen because of its status in the field as being understood as canonically polysynthetic. The languages and the number of data points for each language are provided in Table 1.

Table 1: Languages and the number of morphemes coded.

Language	Glottocode	Family	Source	Data points
Movima	movi1243	Isolate	Haude (2006)	153
Wãnsöjöt [Puinave]	puin1248	Isolate	Girón Higuita (2008)	99
Tariana	tari1256	Arawak	Aikhenvald (2003)	119
Ashéninka [Alto] Perené	ashe1272	Arawak	Mihas (2015)	98
Chácobo	chac1251	Pano	Tallman (2018a)	96
Cavineña	cavi1250	Takana	Guillaume (2008)	56
Hup	hupd1244	Naduhup	Epps (2004)	65
Central Alaskan Yupik	cent2127	Eskimo-Aleut	Miyaoka (2012)	81

3 Tallman and Epps (2020) is primarily a qualitative study concerned with illustrating how and why languages of Southwestern Amazonia display a fuzzy boundary between morphology and syntax. This study was narrowly concerned with the 'boundary problem' and only used bi-variate statistical tests to assess the relationship between two variables at a time. The paper was also more narrowly focused on the issue of exponence complexity, rather than the global relationship between all of the variables.

Ideally, our sample would contain a genetically and areally unbiased sample of languages. However, time constraints have prevented us from developing such a sample: it takes weeks to complete a single language. Only closed class morphemes were coded for in our database. The reasons for this are practical: grammars do not (typically) provide a list of all of the open class morphemes, but closed class morphemes usually receive special treatment. Furthermore, the data were gathered from a set of functional domains (valency and argument structure such as case markers, tense/aspect and time, nominal classification and gender, evidentiality and modality) to render them more comparable (cf. Tallman and Epps 2020 for more discussion).[4] This leads to the possibility of sampling bias and is another reason why general theoretical conclusions cannot be drawn from the sample yet. The problem of overcoming sampling bias is discussed again in Section 8. We hope, however, that the present study serves as a reasonable proof of concept that could inform a broader cross-linguistic investigation able to capture cross-linguistic trends. As stated in the introduction, the primary goal of this paper is to formulate the methodological problems and tentative solutions for investigating the morphology-syntax distinction in a variationist perspective. The focus on a small sample of mostly Amazonian languages limits what empirical conclusions can be drawn from our study. Nevertheless, there are a number of reasons why southwestern Amazonian languages are interesting for the questions at hand: (i) the languages do not have inflectional classes and, therefore, arguments concerning the relationship of paradigm complexity to morphological autonomy (Booij 1997b; Stump 2001) do not clearly apply to these languages. If such systems display morphological autonomy, the motivation is different from that typically discussed in the literature; (ii) the languages have been singled out as displaying 'syntax-like morphology', suggesting that they provide an interesting testing ground for basic assumptions about linguistic architecture; (iii) grammars of Amazonian languages often proliferate clitic or other indeterminate categories in their descriptions (Tallman and Epps 2020).

4 It is important to emphasize that 'functional domain' is distinct from language- or grammar-specific grammatical categories or classes used in specific descriptions. Functional domains are extremely rough semantic/pragmatic classifications that are purposely designed to abstract away from wordhood criterial variables. To take an example from English, the past marker d~d~t would be classified under the same domain ('Time') as the temporal frame adverb 'tomorrow' since they both relate to temporal reference. Sampling from language specific 'grammatical categories' would not be a viable method since such grammatical categories often presuppose the morphology-syntax distinction.

4 Morphosyntactic variables and exponence complexity

This section provides a description of the variables used to develop the database. They were chosen based on two criteria: (i) the variable needs to have been described in the literature as definitional or criterial for distinguishing between affixes, clitics, or words (e.g. Dixon and Aikhenvald 2002; Spencer and Luís 2012a; Zwicky and Pullum 1983); (ii) the test, criterion, diagnostic, or definition can be operationalized in a way that allows the variable to be coded in a consistent manner based on the information available in descriptive grammars. The resulting variables are defined and classified in Table 2. Below we provide a brief description of each variable defending its treatment as separate from other variables where appropriate.

Some of the values specified above cannot be coded without relating an element to a base with which it combines. For instance, 'fixedness' refers to a base element against which the permutability of a given element is assessed. The notion of a 'base' is partially a semantic notion based on the notion of 'semantic head' used profitably in other typological work (Anderson 2006; Croft 2001). After a brief description of the variables, the concept of a 'base' is discussed in more detail in Section 4.7.

4.1 Freedom/boundedness

Boundedness has been described as definitional or criterial for an element to be part of morphology. Haspelmath (2011) refers to a "continuum of boundedness",

Table 2: Wordhood and morphological criterial variables, their type, and basic definition.

Variable	Name	Description	Type	Sec.
Boundedness	FREE	Can the morph stand on its own as an (elliptical) utterance?	Binary	4.1
Interruptability	INTERone	Can the morph be separated from the verb/noun/adjective root by a free form?	Binary	4.2
Fixedness	PRfixed	Does the morph display a fixed order with respect to the verb/noun/adjective root?	Binary	4.3
Coding elaboration	CODelab	Does the morph display inflectional elaboration independent of the base with which it combines semantically?	Binary	4.4
Prominence projection	PRM	Does the morph always/sometimes/never project its own stress domain?	Ordinal	4.5
Exponence complexity	EXPcomplex	A metric that aggregates various types of deviations from biuniqueness that a morpheme can display	Continuous	4.6

implying that the concept is graded or that it is a continuous variable. In this study, we narrowly define a variable of freedom/boundedness according to the single criterion of whether the element can stand as a free form (Bloomfield 1933; Haspelmath 2021; Hockett 1958): a morpheme that can stand alone as an elliptical utterance is **free**, one that cannot is **bound**. The clause in example (4) from Chácobo is a free form, but none of the morphemes that make it up are free forms except the negation morpheme in example (5f).

(4) Chácobo
tsaya-ʔaka *=yáma=tɨkɨ (n)=ʔitá=kɨ*
see-PASS =NEG=again=REC:PST=DECL:PST
'He was never seen again.'

(5)
a.	**tsaya* (intended: 'see')	(but *tsaya=kɨ* 's/he saw')
b.	**-ʔaka*	(coded: bound)
c.	**=tɨkɨ(n)*(intended: 'again')	(coded: bound)
d.	**=ʔitá* (intended: 'recently')	(coded: bound)
e.	**=kɨ*	(coded: bound)
f.	*(=)yáma* 'there is nothing/no one '	(coded: free)

4.2 Interruptability/contiguity

Another criterion associated with the morphology-syntax distinction is that of non-interruptability. Words cannot be interrupted by other words, whereas phrases can be. The criterion stated as such is circular since it relies on the very notion of word it is supposed to test (Mugdan 1994). Furthermore, contiguity is a matter of degree, since two elements can be more or less separable from each other depending on how many different classes of elements can intervene simultaneously (Croft 2001). We follow Haspelmath's (2011) suggestion and assume that the interrupting element should be a free form (as defined above in Section 4.1). Interruptability is, therefore, coded as a binary variable. We will refer to elements that can be interrupted from their base by a free form as **interruptable**. Otherwise they are **contiguous**, even if they do not appear directly adjacent to their base – in other words, an element will still be coded as contiguous if it can be interrupted from its base by bound elements. Interruptability/Contiguity does not follow from Freedom/Boundedness. Not all bound elements are contiguous. For instance, four of the bound elements from the Chácobo example (4) are interruptable, as shown in (6a),

and one of them (the passive marker *-ʔaka*), shown by the ungrammaticality of example (6b), is not interruptable. The base in this example is *tsaya* 'see' in bold and the elements *yama* 'negation', *=tiki(n)* 'again', *=ʔita* 'recent past' and *=ki* 'past' are elements whose contiguity are assessed in relation to this base.

(6) Chácobo

a. ***tsaya****-ʔaka* *honi* *siri* *=yama=tiki (n)=ʔitá=ki*
see-PASS man old =NEG=again=REC:PST=DECL:PST
'The old man was not seen again yesterday.'

b. ****tsaya*** *honi* *siri* ***-ʔaka*** *=yama=tiki (n)=ʔitá=ki*
see man old -PASS =NEG=again=REC:PST=DECL:PST
'Intended: The old man was not seen again yesterday.'

Conversely not all free elements are interruptable. Noun classifier systems are sometimes built from elements that are free in our sense because they can stand alone as complete utterances. However, when they combine with a noun base, they cannot be interrupted from that noun by another free form. This is true of the morpheme *-panki* 'long, rigid object' of Ashéninka Perené, as shown in example (7).

(7) Ashéninka Perené
karini-taki *incha-****panki***
smooth-INTNS plant-CLT:long.rigid
'The wood planks are very smooth.' (Mihas 2015: 410)

4.3 Fixed/variable order

Affixes are often described as occurring in a fixed order with respect to their base, whereas words do not necessarily display this property (Dixon and Aikhenvald 2002). We coded morphemes as **fixed** if they always occur either to the left or to the right of the base they combine with. Otherwise they are coded as displaying **variable** ordering. An example of a bound morpheme that displays variable ordering in this sense is the inferential morpheme *=tukwe* from Cavineña (Takana), illustrated in examples (8a) and (8b). The base in example (8a) is *shana* 'leave' (in bold). The base in example (8b) is *ju* 'be' (in bold). The element whose fixedness value is being assess is *tukwe*. These examples show that it can occur on both sides of its base.

(8) Cavineña

a. *tu-wa* *=tukwe* *ekana* *ka-**shana**-ti-na-kware*
there-LOC =CONTEVID 3PL REFL-**leave**-REFL-COME.TEMP-REM:PST
etawiki=kwana
bed=PL
'There (at the tip of a wood), they left their beddings, on their way (to our village fiesta, thinking they would find their beddings back when returning to their community.' (Guillaume 2008: 643)

b. ***ju**-eti-ya* *=tukwe* *=tu-ke* *=e-kwe* *ea-tseweki=ke*
be-COME.PERM-IPFV =CONTEVID =3SG-F =1SG-DAT 1SG-sibling=LIG
'I feel my brother is going to come back.' (Guillaume 2008: 2)

Fixed order is not always associated with boundedness. The morpheme *d'oʔ* 'take, causative' of Hup (Nadahup) always occurs before a base to which it adds a causative function as in example (9b), but it is a free form in our sense and can even serve as the base for inflectional morphemes as in example (9c).

(9) Hup

a. *dĕh* *wɔ́ç-ɔ́y*
water boil-DYN
'The water is boiling' (Epps 2004: 517)

b. *pĕd* *dĕh* *d'oʔ-wɔ́ç-ɔ́y*
Ped water take-boil-DYN
'Ped is boiling water' (Epps 2004: 517)

c. *yág* *ʔãh* *d'óʔ-óy* *ʔãh* *g'et-ni-tæ̃ʔ-ní-h*
hammock 1SG take-DYN 1SG stand-be-CNTRFCT-INFR-DECL
'(…) I took (was given) a hammock; I would have stayed there (but these days it's impossible).' (Epps 2004: 614)

4.4 Complex/simplex or coding elaboration

Words can be elaborated with inflectional morphology independently of the base with which they combine. Most linguists would judge constructions to be syntactic if some type of inflectional marking appeared on more than one element in such a way that suggested that more than one base was being elaborated with inflectional morphology. This is the main reason why periphrastic constructions are considered syntactic (or at least not strictly morphological). If a morpheme was found taking on its own inflection in some construction we coded it as 'yes' for this property and as 'no' otherwise. This can be observed with the distinction between

affixal and analytic causatives in Paresi (Arawak). The analytic causative *moka* can receive coding elaboration by person markers and the affixal causative *-ki* cannot. The distinction is illustrated in example (10).

(10) Paresi

*ha=**moka***	*natyo*	*hoka*	*n=aotya-**ki**-tsa*	*xitso*
2SG=CAUS	1SG-put	CON	1SG=remember-CAUS-TH	2PL
haliti	*ni=rai-ne*			
Paresi	3SG=talk-POSSD			

'You made me teach you all the Paresi language.' (Brandão 2014: 269)

Morphemes that can never combine with another morpheme independent from their base will be referred to as 'simplex'. If we refer to a morpheme as 'complex' we only mean that there are contexts where it appears to combine with another morpheme (e.g. an inflectional marker) independently of its combination with a base.

4.5 Prominence projection

Words are typically described as phonologically or prosodically 'independent', whereas affixes are not. Clitics are often described as phonologically 'deficient' like affixes (Aikhenvald 2002: 42). A form that projects its own phonological word can be considered more word-like than one that does not (Spencer and Luís 2012a: 127). Affixes are canonically understood as not projecting their own phonological word. In this paper, we do not refer directly to the projection of 'phonological words', since we are interested in developing a set of more directly observable variables (see Bickel et al. 2009; Schiering et al. 2012).

Prominence projection refers to whether an element projects a domain of culminative and obligatory prominence ('stress') or not (Hyman 2006, 2009). The stress does not have to occur on the element itself, but the formation of a stress domain has to make crucial reference to it. Furthermore, an element that has its own stress, but does not imply its own special domain of stress from the base it combines with, does not project a prominence domain in our sense (e.g. stressed or stress-attracting affixes).

Prominence projection can take one of three values in our study. Either the element always projects stress, can project stress or never projects stress. The term thus refers to the degree to which an element projects its own independent stress domain. The inverse of prominence projection is **prominence deficiency**, a term that we will use below where necessary. Unfortunately grammars are not always explicit regarding which morphemes can project stress and whether they project

stress in specific examples and one has to follow a complex line of deductions (or inferences) to reach conclusions regarding how a morpheme should be coded (e.g. the morpheme is described as a 'particle', particles are described as 'words', and the linguist refers to words as containing stress).

An example of a morpheme coded for projecting stress is *ita* 'reciprocal' from Urarina. The morpheme is described as a 'particle' (Olawsky 2006: 607), 'particles' are described as 'words' (Olawsky 2006: 84), and 'words' have 'stress' (Olawsky 2006: 121) and (typically) one high tone (Olawsky 2006: 148). We thus infer that *-ita* projects a stress domain since it is described as a particle. We assume, in contrast, that the reciprocal in Cavineña never projects its own stress domain because it is described as an affix in Cavineña, which cannot be phonological words in this language (Guillaume 2008: 41, 268).

Morphemes that vary according to whether they project a stress domain or not are common in our study. An example of a morpheme which can optionally project a stress domain is the future clitic *=utsu* 'go, going, future' from Kokama-Kokamilla. This can be inferred from the fact that 'phonological words' are written separately and each phonological word has a single stress (Vallejos Yopán 2010: 117). A context where the morpheme projects its own stress domain is provided in example (11b) where *=utsu* is separated from its verb base and hosts a bound and phonologically reduced clitic pronoun *y=* 'third person feminine' (Vallejos Yopán 2010: 210).

(11) Kokama

a. *penu* *yawachima-ka-t=**utsu*** *uyarika* *awa=pura*
1PL.F arrive-REI-CAUS=FUT again person=FOC
ukuata-ri-n=pura=nu
pass-PROG-NMLZ=FOC-PL
'We will reach again the people (who are crossing the street)'
(Vallejos Yopán 2010: 603)

b. *yanamata* *kari-ri=tsui* *y=itika-ka* *y=utsu*
bush scrape-PROG=ABL 3SG.F=throw-REI 3SG.F=FUT
'After scarping the bushes, he goes to throw it.'
(Vallejos Yopán 2010: 480)

Stress projection tends to correlate with boundedness such that bound elements do not project their own stress domain (see Section 5). However, there are morphemes that always project stress but which are bound. This is true of the epistemic marker *kará* 'dubitative' in Chácobo. Under no circumstances can it stand as an utterance by itself, however, it always projects its own stress domain, even in positions where other modal elements would not, cf. example (12).

(12) Chácobo
kaa karáˈ =kɨ́
go DUB =DECL:PST
'He went (I think).'

4.6 Exponence complexity

As stated above, for many advocates of morphological autonomy, exponence complexity is one of the defining features of morphology. There are a number of different types of exponence complexity. Below we review those types we were able to code and describe how we aggregated these into a single metric of exponence complexity.

4.6.1 Number of allomorphs

For each morpheme, we coded for the number of segmental allomorphs. Minimally, each morpheme has one (allo-)morph. Morphemes with a high degree of allomorphy are not very common in our database, but they do occur. For instance, Olawsky (2006: 609–622) describes two causatives in Urarina: a 'causative 1' *-a* that encodes 'direct personal involvement' and a more productive 'causative 2' *-eratia~ratia~rate~tçate* that encodes indirect causation. His discussion shows that the direct causative has one realization, whereas the indirect causative has five realizations. Thus, the direct causative receives a score of 1, whereas the indirect causative receives a score of 5. In Chácobo, we find the reverse situation: the direct causative *-ʔak~ʔa~ak~a* displays more allomorphy (4 allomorphs) than the indirect causative *=wa* (1 allomorph) (Tallman 2018a: 656).

4.6.2 Suppletive allomorphy

We coded for whether or not the morpheme displays suppletive allomorphy. It is often difficult to tell the difference between allomorphy produced by phonological rules and allomorphy that is genuinely suppletive. The judgement was, therefore, somewhat subjective and required making decisions concerning whether the morphophonological rules reported in the grammar could account for the observed allomorphy or not – this is not a trivial task, because grammarians do not always report the precise span of structure that a morphophonological rule operates over. As a default rule we followed statements of the authors concerning whether allomorphy was suppletive or not.

A clear example of suppletive allomorphy is the variation between *-sha~-mere* in the Cavineña (Takana) causative. Guillaume's description of these forms implies that the variation is suppletive, because the difference in form is attributable to the transitivity of the verb base, rather than a phonological rule (Guillaume 2008: 285–299). However, the variation in form of the Urarina causative direct causative *-eratia~ratia~rate~tçate* refers to phonological environments, making a suppletive designation unlikely (Olawsky 2006: 609–622).

Note that suppletive allomorphy was coded as a separate variable from the number of allomorphs. For a morpheme to receive a 'yes' coding for suppletive allomorphy only requires that one of the allomorphs be recognized as suppletive. Therefore, the variable is unfortunately coarse grained and future research is needed to develop a more complex and rigorous diagnostic for gauging the degree of opacity in variation in form.

4.6.3 Multiple exponence

Another type of deviation from biuniqueness is multiple exponence: where a meaning maps onto multiple forms in the same structure. Instances of multiple exponence were extremely rare in the languages of our study (there are only nine examples in total). An example of extended exponence is the causative *a-…-kiẽ-…-ki* in Paresi (Arawak), as in *a-koeza-ki-tsa* 'make someone laugh' from *koeza* 'laugh' (Brandão 2014: 260–267). Morphemes were coded as 'yes' if they displayed such multiple exponence and 'no' otherwise.

4.6.4 Fossilization/empty roots

Another type of deviation from biuniqueness is where a given form corresponds to no consistent meaning or no meaning at all. As described in the introduction, such formatives, and the 'morphomic' patterns that they involve, constitute an important aspect of research into morphological autonomy. However, for the languages of our study it was difficult to find examples of such cases. We speculate that the reason for this is that grammarians are inclined to describe morphomic patterns as being instances of homophony, rather than attributing morphomic patterns to special autonomous morphological structure. For instance, Epps (2004) discusses the issue of homophony versus polysemy throughout her grammar of Hup, but the possibility that formatives might be 'morphomic' in Aronoff's (1994) sense is not addressed. It is possible that the issue is not (very) relevant to Amazonian languages, but the gap could also reflect the fact that analyses of 'morphomic structure' have not yet been incorporated into the analytic arsenal of Amazonian linguists.

It was easier to identify cases in which a morpheme is described as combining with an empty root or where the meaning of the root is opaque or non-compositional in the relevant context. Such fossilized base-affix combinations provide evidence for morphological autonomy just as well because they are examples of deviations from biuniqueness (candidate empty morphs). Thus, we coded morphemes that were described as combining with empty or fossilized (i.e. synchronically meaningless or semantically opaque) bases at least in some cases as 'yes', and those that were not as 'no'.

Examples of morphemes that combine with empty or fossilized bases are the nominal classifiers *aa* and *ki* in Ashéninka Perené (Arawak). While the normal pattern for nominal classifiers is to be highly productive and compositional, Mihas (2015: 416) notes that these two forms occur in 'frozen, conventionalized units'. An example of one of these forms is *-ki* 'small round' in the context of *ki* 'bathe' where the meaning is 'administer plant juice into eyes', cf. example (13). The morpheme *-ki* is therefore coded as 'y(es)'. If no evidence is found in the relevant sources that a morpheme combines with an empty base or that it occurs in non-compositional contexts, then the morpheme is coded as 'no' on this parameter.

(13) Ashéninka Perené
y-a-ak-i-ro *o-ishi-paye=nta*
3.M.A-take-PFV-REAL-3.NM.OBJ 3.NM.POSS-leaf-PL=DEM
*i-kaa-**ki**-t-ak-i-ro* *ina*
3.M.A-bathe-CLT:small.round-EP-PFV-REAL-3.NM.OBJ mother
'He took leaves and administered them [their juice] into my mother's eyes.' (Mihas 2015: 416)

4.6.5 Exponence complexity as a metric

In order to reduce the number of dimensions in our data set we combine the values associated with deviations from biuniqueness into a single metric of exponence complexity. All categorical values of exponence complexity are converted to numerical ones, such that 'yes' receives a score of 1 and 'no' one of 0. Then we calculate exponence complexity with the following simple formula

$$\mathbf{ec} = a + s + m + f$$

where *ec* refers to exponence complexity, *a* to the number of allomorphs, *s* to the presence of suppletive allomorphy, *m* to the presence of multiple exponence, and *f* to whether the formative combines with an empty or semantically opaque base. This results in a minimum value of 1, since the minimum number of allomorphs is non-zero, or in other words we do not posit zero morphs. We note that this

calculation of exponence complexity is provisional and theoretically crude. It is partly based on our desire to have a single overarching dependent variable of exponence complexity, but there is a possibility that results obtained from the data transformation provided above are partly a spurious result of the transformation itself (see Cysouw 2002 on this point with respect to information loss in typological indices).[5]

The idea that exponence complexity can be treated as a dependent variable in measures and models of the morphology-syntax distinction finds support from many current descriptions of morphological complexity, which view the concept as crucial, cf. Section 4.6. We view the wordhood criterial variables not considered in this formula to be independent variables, when a distinction between dependent and independent variables is necessary for the application of a specific statistical method. Such a case is provided in Section 6.

4.7 The population sampled: morphemes and base-morpheme pairs

In Haspelmath's (2011) discussion of wordhood diagnostics he points out the possibility of developing an empirical test for the notion of 'fuzzy word' based on quantifying grammatical units on a boundedness scale and testing whether they demonstrate a 'clustering distribution' (a point we return to in Section 7). In a footnote in the same section, he notes that such an empirical test would be difficult because the population to be sampled is an 'open-ended set' by which he means that it is unclear how to define or constrain which grammatical units should be sampled (since these could be affixes, clitics, words, and phrases).

The methodology we present here does not directly address the problem of 'open-ended unit size', but to a certain extent evades it. The wordhood criterial tests ask whether a base-morpheme pair is more like a combination of a word and an affix or a word and another word. As we pointed out in Section 3 in some cases it makes little sense to inquire about the wordhood properties of individual forms abstracted from their base. Non-interruptability and fixedness require a base to be interpreted at all.

The methodological move we take in this paper depends on defining a consistent notion of 'base'. The notion of base in this study is close to the notion of semantic head discussed in Croft (2001: 259) and Anderson (2006: 22). Roughly, the

5 An anonymous reviewer points out, correctly we think, that future researchers should consider calculating the relative weight of each factor in *ec* by using confirmatory factor analysis. We do not have enough data to apply such an analysis at this point.

notion of semantic head combines profile determinant with specificity. When profile determinant and specificity cannot be evoked, the base can be inferred based on its distribution (is the element in a structural position typically shared by bases?). In the expressions *the broken vase* and *the vase broke*, the forms *vase* and *broke* are the profile determinants respectively because they are what is symbolized by the expression as a whole (or more simply if in a combination *x*–*y*, the expression can unequivocally be regarded as a type of *x*, then *x* is the profile determinant). When two profile equivalents are in combination the one which is more semantically specific is understood to be the semantic head. We diverge from this notion of semantic head when one of the formatives appears to be semantically empty or when each formative is a profile determinant and neither is obviously more specific than the other. In such cases we defer to distributional facts, asking whether the candidate empty base in question has the same distributional properties as base elements in the language.[6]

The other criteria such as allomorphy and boundedness etc. do seem to single out properties of morphemes rather than base-morpheme pairs, but they can also be interpreted in light of the latter. If a morpheme is free, it is independent of its base and, therefore, we assume that base-morpheme combinations with that morpheme are more like phrasal combinations according to this criterion. The fact that a morpheme displays high allomorphy can similarly be translated into higher allomorphy for the set of base-morpheme pairs that contain this morpheme.

Nevertheless Haspelmath's (2011) point is important to keep in mind. This study is not concerned with the issue of wordhood at the level of constituency *per se* because this requires the application of wordhood diagnostics over larger

6 A semantically empty/opaque base can be identified because it occurs in a structural position which typically contains the profile determinant and is combined with an element that can be identified as meaningful based on other combinations. To illustrate this point, consider the transitive marker *-a* in Chácobo. When we compare forms such as *niṣ- i* 'tie oneself, be tied' with *niṣ-a* 'tie someone/something' and *ta-niṣ* 'tie foot' it is plausible to posit that *niṣ* 'tie' is a root, and *-a* marks transitivity. However, we also find *-a* with candidate 'fossilized roots' such as in *yon-a* 'drive, make object work', where the root *yon* 'work' cannot be straightforwardly classified as a root independent of *-a*. We posit that *yon* is the base here based on the patterning of *-a* and the base is coded as fossilized. In this case, *yon* would be coded as our fossilized base. It is difficult to claim that it is a profile equivalent on its own, however it displays near distributional equivalence with roots that do have this property. Cases where each formative is a candidate for being a base on grounds of specificity is resolved in a similar fashion by deferring to distributional facts (see Croft 2001 on the notion of 'primary information bearing unit'). In our estimation such cases were marginal in our data. However, future research could show that a different notion of base is needed. It may be desirable to determine how to quantify the degree to which some formative is a base or not according to a set of applicable criteria or to develop a method which excludes ambiguous cases so as to avoid researcher bias.

structures. Rather it is concerned with the extent to which base-morpheme combinations pattern into base-affix versus base-word combinations. Finding that some base-morpheme combination is more base-affix like rather than base-word-like does not tell us anything with respect to whether some base-affix combination is a word constituent itself. Rather it suggests that there is something (relatively more) morphological about the combination of these elements. Consider the causative *-car~-caar~-caara~-caarar* of Central Alaskan Yupik, illustrated in example (14).

(14) Central Alaskan Yupik
*Qavar-**caar**-yuk-aanga* *May'a-mun*
sleep-CAUS-think-IND.3SG.1SG mayaq-ALL.SG
"He thinks Mayaq is trying to make me sleep." (Miyaoka 2012: 1055)

The causative is coded as bound (it cannot stand as a full utterance), contiguous (it cannot be interrupted by a free form), fixed (it is a suffix), and simplex (it cannot occur with its own inflection or elaboration of any type). It also displays relatively high allomorphy (containing 4 allomorphs). Thus, it is very affix-like according to the wordhood criterial variables (or more precisely the base-*-car~-caar~caara~-caarar* combination is base-affix like). This does not imply that such a combination should be understood as a 'word'. Rather, example (14) suggests that it is part of word-internal structure. The base-morpheme combination where the morpheme is this causative is more like a morphological combination than a syntactic one.

The methodology thus investigates whether base-morpheme combinations display more morphology-like versus syntax-like properties according to the variables discussed above. To the extent that morphology is a system independent from syntax, we expect that indeterminacy in this regard should be statistically marginal. A different methodology needs to be employed to assess criterial convergence over larger structures (Tallman 2020, 2021; Tallman et al. 2019).

5 Measuring morphological autonomy: correlation matrices

The idea that morphology is autonomous from syntax but with some indeterminacy or fuzziness introduces the question as to whether languages vary in the degree to which their morphological systems are autonomous (or, how large their space of indeterminacy or fuzziness is). This section presents a few proposals of how to visualize and measure the degree of morphological autonomy as a typological

index. It also provides an overview of some the associations between the wordhood and morphological criterial variables discussed in the previous section.

Given that morphological autonomy is associated with exponence complexity in the literature, one might think that statistical summaries of exponence complexity could be taken as approximate measurements of the degree of morphological autonomy. However, as we argued in Section 4.6, what is important is the way that exponence complexity correlates with other wordhood criterial properties. The correlation matrix in 1 provides a global overview of the correlation between the wordhood criterial variables and exponence complexity. Variables were recoded as numerical such that the syntax/word-like result was set as 0 and the morphology/affix-like result was coded as 1. With fixedness, for example, elements considered to be affixes or part of morphology are expected to have a fixed position, while words or syntactic elements are expected to be able to variably order. Thus, we coded fixed as 1 and variable as 0. Exponence Complexity already is a numeric variable and thus does not need to be recoded. Table 3 provides an overview of the recoding.

Figure 2 presents correlation matrices for each language separately. The code and further details can be found in the Supplementary materials (cf. 9). It is immediately apparent that languages vary in terms of whether their variables correlate at all. When they do exhibit significant correlations, these are not necessarily in the expected direction. In Chácobo and Central Alaskan Yupik all variables except exponence complexity are positively correlated with each other. In Central Alaskan Yupik, we find perfect correlations between stress deficiency and boundedness (i.e. a morpheme that is bound never projects stress), stress deficiency and coding elaboration (i.e. all morphemes that can be modified independently from their base project stress), boundedness and coding elaboration (i.e. all morphemes that can be modified independently from their base are free) and fixedness and contiguity (i.e. all contiguous morphemes display a fixed order with respect to their base). The other languages show weaker correlations across variables. For example, Hup displays no perfect correlations. In fact, it is unclear whether there is an overall tendency for the wordhood criterial variables to

Table 3: Recoding of variables as numeric.

Variable	Coded as 0 (synt.)	Coded as 1 (morph.)
Interruptability	Interruptable	Not interruptable
Coding elaboration	Present	Absent
Fixedness	Variable	Fixed
Boundedness	Free	Bound
Prosodic prominence	Present	Absent = 2, both = 1 (clitics)

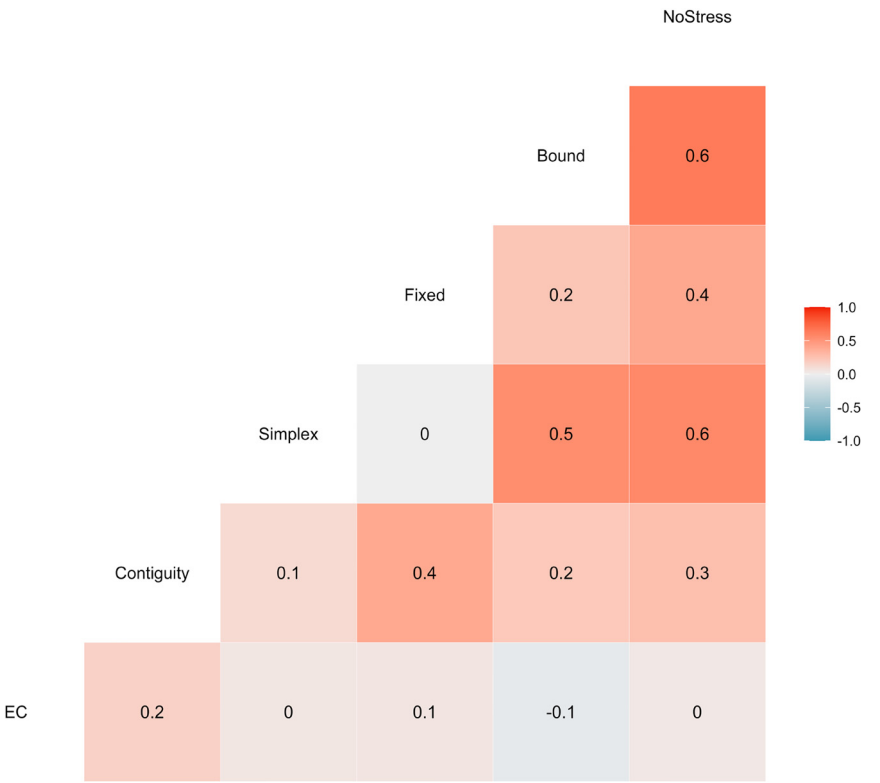

Figure 1: Correlation matrix of the wordhood criterial variables using Kendall's tau over all eight languages.

be correlated with one another evidenced by the values close to zero. Other languages like Cavineña and Tariana occupy a middle ground, with only a few variables showing strong correlations. In Cavineña, most morphemes that have coding elaboration also have stress. In Tariana, most bound morphemes do not have stress and do not exhibit coding elaboration.

It is not necessarily the case that the variables are positively correlated with one another. As suggested in Tallman and Epps (2020), the particular structure of Movima's classifier system results in negative correlations between boundedness and exponence complexity.[7] In Tariana there is a negative correlation between

7 The reason for this is that classifying morphemes are not distinguished from their free form nouns in our study (it is not clear how to distinguish them in a principled way) and these forms sometimes display variations in form according to position.

fixedness and coding elaboration. Contrary to expectation, in both Tariana and Movima, there is a negative correlation between the absence of coding elaboration (simplex in the graphs above) and occurring in a fixed order in relation to the base. This can also be related to the semi-open class classifier systems of these languages. Classifiers occur in a fixed order with relation to a head-noun, but can also occur with their own morphology when they function as roots.

The question then arises as to whether some of the variables should be discarded in one or all of the languages. Considering, for example, interruptability in Tariana, it seems implausible that this variable could be used to distinguish between morphological and syntactic elements, given how low the correlations are. The method described in the next section (Section 6) addresses the issue concerning the composition of morphological autonomy more explicitly.

Mean correlation across all variables are listed in Table 4. It is quite low overall, with CAY exhibiting a mean correlation of 0.57. Hup, on the other hand, shows a much lower mean correlation at 0.13 and that of Movima is even overall slightly negative. To the extent that the global correlations between wordhood criterial variables can be used as a metric for determining morphological autonomy, Central Alaskan Yupik has a much stronger distinction than Hup.

6 The composition of morphological autonomy: random forests

In this section, we use Random Forests (RF) as a way of assessing the composition problem, i.e. the question of whether or not languages vary in terms of which variables are important for distinguishing between morphology and syntax. A Random Forest is a classification algorithm that aggregates over a multitude of decision trees (Louppe 2014). The number of variables out of all dependent variables that are tried at each split in each decision tree for best classifying the data has to be defined beforehand. This is usually determined by running multiple RF models with different numbers of variables to find the one producing the best results. RF is a powerful tool for classification tasks and variable selection (i.e. which variables should be considered in a study) and has also been used for

Table 4: Mean correlation coefficients for each language.

Movima	−0.03	Puinave	0.24	Tariana	0.15
Asheninka Perene	0.32	Chácobo	0.55	Cavineña	0.17
Hup	0.13	Central Alaskan Yupik	0.57		

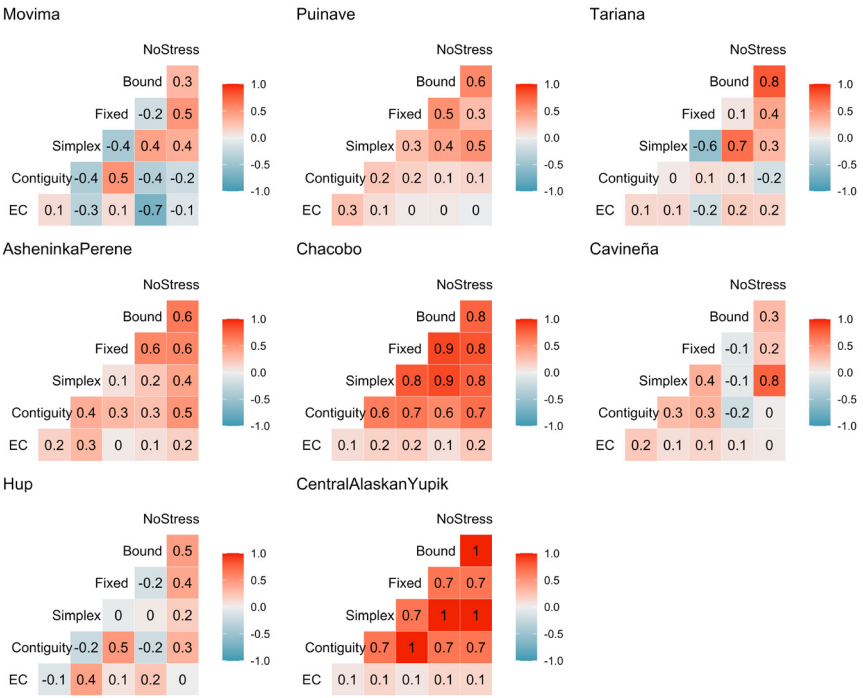

Figure 2: Correlation matrices per language.

addressing problems in linguistics (Levshina 2016; Tomaschek et al. 2018). The variables of our study are strongly correlated with one another in many cases, which makes interpretation of the results under regression very challenging (Belsely et al. 1980). Random Forests do not suffer from this issue, since they do not make any assumptions about the input. RFs still require specification of a dependent variable and it is not clear how to determine this *a priori*. The variation in the strength and direction in correlations between the variables show that this problem is not trivial. For instance, we pointed out that interruptability/contiguity is likely not an important variable in Movima. If this variable was chosen as the dependent variable in the classification model (i.e. asking to what extent can the other variables predict contiguity/interruptability), we are likely to get a very different picture of the relative importance of the variables compared to say, if we considered bounded/free as the dependent variable. The selection of the dependent variable should be guided by theoretical considerations, but there is no general agreement in the field regarding whether there is a crucial determinant of the morphology-syntax distinction (Haspelmath 2011).

With these problems in mind, we suggest two ways in which the importance and relevance of variables can be assessed cross-linguistically. One option is to use author classifications (simplified to be either 'affix', 'clitic' or 'word') as the dependent variable. This reflects an intuition that many researchers have, namely that grammar authors are mostly consistent in applying wordhood criteria, even if these criteria vary by author. The other option is to select exponence complexity as the dependent variable, which we take to be 'theoretically grounded' in the sense that it is considered to be particularly important for advocates of morphological autonomy (cf. Section 2). These choices in dependent variables reflect two different perspectives on how cross-linguistic variation in the morphology-syntax distinction should be understood, one grounded on the partly intuitive classifications of experts on the specific languages, and the other grounded in certain linguistic theories concerning morphological autonomy. We applied both options. The exponence complexity option is presented in this paper (see Supplementary materials for the author classification option.)

All the following RF models are implemented in R with the randomForest package (Liaw and Wiener 2002). We aggregate over 10,000 trees for each of them with two variables tried at each split. All plots were produced with ggplot2 (Wickham 2016). More details and the code can be found in the Supplementary materials. Each model outputs error rates for each level of the dependent variable in a confusion matrix and an overall error rate for the model (called out-of-bag error or estimate, OOB for short). The former provides the classification accuracy with respect to the dependent variable and the latter how accurate the classification is overall. The model also provides a plot displaying Mean Decrease in Accuracy (MDA), which can be related to the relative importance of variables. Broadly speaking, certain variables are better and more consistent at classifying the dependent variable correctly than others across all the decision trees and this is reflected by MDA. Note that the actual value does not matter, since it is not comparable across models. What is important is the ranking of the independent variables and whether or not there are clear breaks between them.

To this we add a calculation of the baseline, the accuracy, and the difference between these two. The idea is to account for the skewness of the data, since the more skewed the data are, the easier it is to get a correct classification by chance. For example, imagine two languages X and Y with 100 data points each. Language X has 34 affixes, 33 clitics, and 33 words, while language Y has 15 affixes, 5 clitics, and 85 words. If we classify all elements as words in language Y, we get 85% right. But no matter what category we choose for language X, we do not achieve over 34% correct classification. The relative skewness by language needs to be taken into account when assessing whether the RF model performed well or not. The calculation of each of these measures is very simple. The baseline is exactly what we just discussed. More simply put, it represents the maximal row-sum of the confusion matrix divided by the number of data points. The accuracy represents the sum of

correct predictions (i.e. the diagonal in the confusion matrix) divided by the number of data points, i.e. it reflects the proportion of correct predictions. The difference is calculated by subtracting the baseline measure from the accuracy, i.e. it reflects how much better the RF model performed over chance.[8]

While Random Forests over author classifications provide an intuitive idea of cross-linguistic variation concerning the composition of the morphology-syntax distinction, they are problematic. It is possible that the author's conceptions of what an affix, clitic, and word are vary in arbitrary ways and according to theoretical biases that are not directly relevant to the morphology-syntax distinction.[9] The results of applying RFs with author classification as a dependent variable are thus provided only in the Supplementary materials. They suggest that authors vary substantially in which variables they consider to be important in classifying elements as affixes, clitics and words.

A theoretically more grounded way of assessing the composition problem using RF might be to use exponence complexity as a dependent variable. For these models we treat exponence complexity as a factor by considering a value of 1 as 'low' and all values larger than 1 as 'high'. We thus ran RF models for each language with exponence complexity as the dependent variable and all others (prominence projection, coding elaboration, fixedness, and interruptability) as predictors. More formally, the models are specified as follows:

$$\text{Exponence Complexity} \sim \text{Fixedness} + \text{Boundedness} + \text{Interruptability} + \text{Prominence} + \text{Coding Elaboration}$$

The OOB error, baseline, accuracy, and difference values are presented in Table 5. In the majority of languages, the model either performs at chance level or even slightly below. This is due to the model classifying nearly all elements as having low exponence complexity, reflecting the skewed distribution in the data. Only in Movima, which has a higher number of elements with high exponence complexity is the model able to classify well and outperform chance. The MDA plots for this language is shown in Figure 3. We see that free occurrence is by far the most important variable for classification in Movima, followed by coding elaboration. The other three variables play virtually no role in the classification. The other MDA

8 We thank Marc Tang for suggesting these additional measures.

9 Furthermore, even if author classifications can be used, we are classifying over affixes, clitics, and words. But descriptive grammars vary in terms of whether they consider clitics morphological or syntactic elements. For instance, compare Guillaume (2008: 54) with Zariquiey (2018: 159): Guillaume considers clitics to be 'grammatical words' whereas Zariquiey considers clitics to be 'phrasal suffixes'. That there are striking differences between clitics in these languages overall is not clear and in neither case is there an explicit defense of the syntactic (Cavineña) or morphological (Kakataibo) treatment of the category (see Tallman 2018b for further discussion).

Table 5: Confusion matrices, baseline, and accuracy measures for RF models with exponence complexity as the dependent variable.

Movima						**Puinave**					
OOB estimate of error rate: 6.43%						OOB estimate of error rate: 29.29%					
	High	Low	class.error	bl.	57.14		High	Low	class.error	bl.	72.73
High	52	8	0.134	acc.	93.57	high	0	27	1.0	acc.	70.71
Low	1	79	0.013	diff.	36.43	low	2	70	0.028	diff.	−2.02
Tariana						Ashéninka Perené					
OOB estimate of error rate: 10.08%						OOB estimate of error rate: 11.22%					
	High	Low	class.error	bl.	90.76		High	Low	class.error	bl.	88.78
High	0	11	1	acc.	89.92	High	0	11	1	acc.	88.78
Low	1	107	0.009	diff.	−0.08	Low	0	87	0	diff.	0
Chácobo						Cavineña					
OOB estimate of error rate: 35.42%						OOB estimate of error rate: 19.64%					
	High	Low	class.error	bl.	67.71		High	Low	class.error	bl.	80.36
High	0	31	1.0	acc.	64.58	High	0	11	1	acc.	80.36
Low	3	62	0.046	diff.	−3.13	Low	0	45	0	diff.	0
Hup						Central Alaskan Yupik					
OOB estimate of error rate: 13.85%						OOB estimate of error rate: 16.05%					
	High	Low	class.error	bl.	90.77		High	Low	class.error	bl.	83.95
High	0	6	1	acc.	86.15	High	0	13	1	acc.	83.95
Low	3	56	0.051	diff.	−0.046	Low	0	68	0	diff.	0

plots can be found in the Supplementary materials for reference, but one should abstain from interpreting them altogether due to the poor model performance.

Dixon and Aikhenvald (2002) proposed that some notion of fixedness was one of the most important criteria for identifying words. Our results, however, do not confirm this idea insofar as exponence complexity is taken as the most important property for characterizing morphology: In Movima, fixedness has the lowest variable importance and in Chácobo the second lowest one, i.e. it does not contribute much to classifying elements into high versus low exponence complexity. The results suggest that the criterion of fixed/variable order is not important in distinguishing morphology and syntax at least in some languages. In Movima, boundedness is the main contributor to the classification, while in Chácobo it is prominence. Such discrepancies suggest that languages vary in terms of which criteria are important for distinguishing morphology and syntax if exponence complexity is taken to be an important criterion for division.

A general problem with this approach is that in the languages of our sample, exponence complexity displays very weak correlations overall with any of the other wordhood criterial variables, as shown by Figure 1. It seems that RFs can be used to describe variation in the use of wordhood criterial variables in relation to some base classification. They cannot help us determine the baseline classification (what base-element pairs should be categorized as morphological versus syntactic constructions) and they cannot help us determine whether a morphology-syntax distinction is motivated to begin with.

To the extent that the morphology-syntax distinction is valid and we agree on some way of grounding the distinction empirically, the variables also suggest that languages may vary concerning whether and the degree to which wordhood

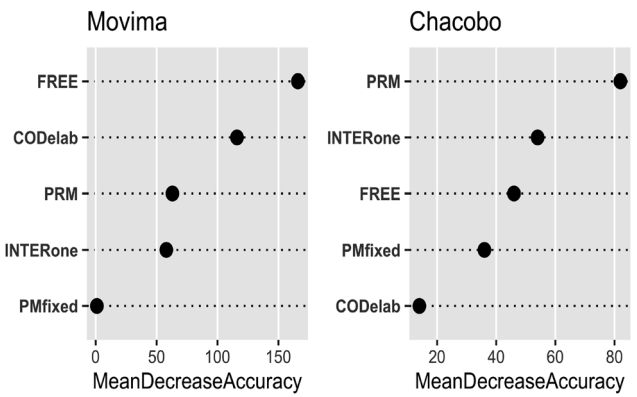

Figure 3: Variable importance of RF models for Movima and Chácobo with exponence complexity as the dependent variable.

criterial variables reflect important structural generalizations in the structuring of the morphology-syntax distinction.

7 Testing morphological autonomy: cluster validation

In his discussion of wordhood diagnostics, Haspelmath (2011) addresses the possibility that a fuzzy notion of word might be valid. Conceptualizing constructs and formatives as positioned on a unidimensional continuum of boundedness, he contrasts two hypothetical situations. In one situation grammatical units are 'randomly distributed' on the continuum and in another situation elements display a 'clustering distribution'. Variations on Haspelmath's illustrative figures are produced with the strip plots in Figure 4.[10]

Haspelmath (2011: 63) argues that "If the dimension along which the units differ (the boundedness scale) can be quantified, the clustering can be demonstrated by

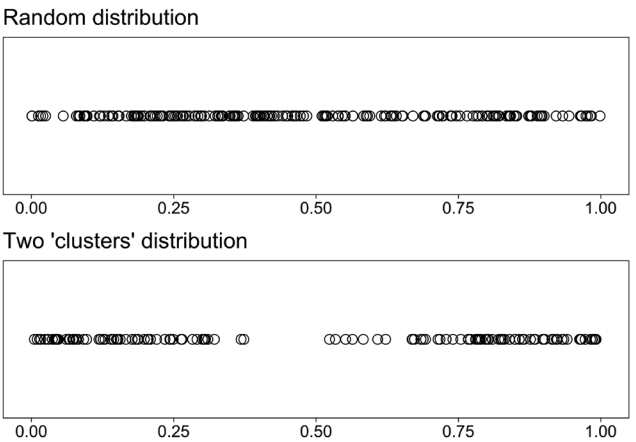

Figure 4: The affix-word continuum: two hypothetical situations.

10 The random distribution was produced with the runif function in R. The clustered distribution was produced by the rnorm() function in R for distributions with means 0.09 and 0.91 each with standard deviations of 0.15. Note that in Haspelmath's original text, the 'clustered distribution' contained three clusters, representing hypothetical clitics. However, based on our review of the literature on morphological autonomy, a cluster for clitics *should* represent challenges to the morphology-syntax distinction, rather than confirmatory evidence, because their validation as an independent cluster would show that a dual partition is not motivated. Clitics should constitute statistically marginal deviations from a global distinction between morphology and syntax.

statistical techniques." In a footnote he describes problems in quantifying the "boundedness scale" including the problem that the sample population would be an "open-ended set" (addressed in Section 4.7), that it is unclear how to weigh the diagnostics, and finally, "finding clusters in a multidimensional space is even more difficult". With respect to the weighting problem, at our current stage of knowledge, the issue really reduces to whether we should assign *a priori* weights to any of the variables, but there appears to be no clear principle for deciding on this at this point. If anything, the data reviewed above suggest that any *a priori* weighting based on theoretical considerations (such as the idea that exponence complexity should be weighted higher) may reflect a bias towards the languages on which those theoretical considerations are primarily based (Sneath and Sokal 1973).[11] At present there is simply no clear empirical basis to give one category a heavier weighting than any other.

The issue of multidimensionality is not specific to the morphology-syntax distinction, but is true of problems in typology (or even linguistics) in general (Baayen 2013; Bickel 2010b; Round and Corbett 2020). Thus, in principle, there is no *a priori* reason why the multivariate structure of the data present a problem insofar as we are willing to import methods used successfully in other domains of linguistics.

The first real problem arises when we consider the fact that there is no obvious way in which one of the variables can be considered the dependent variable. Haspelmath's (2011) formulation, however, suggests that such a dependent variable is not needed, because we could cluster using "statistical techniques". We simply need to show that the data cluster into *x* number of clusters beyond chance or randomness. Haspelmath (2011: 64) also states that "We should be open to the possibility that other kinds of clusters, e.g. 'affixoid', 'clitic group', 'tight phrase', or 'stems' will turn out to be more significant than word clustering. This would have the consequence that the primary division within morphosyntax would not be between morphology and syntax, but along other lines."

There is, however, a problem with this formulation apart from those mentioned by Haspelmath (2011: 64): Aldenderfer and Blashfield (1984: 33) point out that there "is no standard or even useful definition of the term 'cluster', and many have argued that it is either too late or irrelevant to create one". There are a variety of different clustering algorithms and models that depend on different

11 Consider the view of Sneath and Sokal (1973: 109) on the question of weighting variables for biological taxonomy: "We are therefore discussing *a priori weighting*, before a classification is commenced, and what we feel is objectionable is to presuppose knowledge that is not yet available, either about the classification of the organisms or about the presumed significance of their characters". We agree with Sneath and Sokal on this point in the context of studying cross-linguistic variation in the morphology-syntax distinction.

properties that the researcher may wish to assign to a cluster (Hartigan 1975; Jain 2010). For instance, is similarity between elements of the same cluster more important than their dissimilarity with elements outside of it? What is the correct shape for a cluster in some *n*-dimensional space of variables (Kaufman and Rousseeuw 2005: 44)? Clustering is an exploratory method for hypothesis development not a method of inferential statistics. Clusters arrived at through clustering algorithms need to be validated, to ensure that they are not arbitrary partitions of the data. However, cluster validation techniques are currently extremely domain specific and it is not clear whether they are generally applicable across disciplines. It is beyond the scope of this paper to provide any sort of comprehensive review of clustering methods and validation techniques applied to the morphology-syntax distinction. Rather we illustrate one clustering method and one validation technique that we think engages with Haspelmath's (2011) formulation of the problem as we understand it. We use hierarchical clusters and assess height difference between the first and second partition relative to the same difference in a simulated 'random' language and a set of simulated languages, a method we illustrate below.[12] We suggest that this method can help us determine whether a partition into two clusters for a given language accounts for the data better than chance. To the extent that a dual partition of the data can be interpreted as motivating a morphology-syntax distinction, this serves as a test for a morphology-syntax distinction that does not rely on a dependent variable. However, the results of clustering models beg another question concerning the structure of the languages cross-linguistically. This concerns what the optimal number of clusters is for a given language, a problem which Haspelmath (2011) alludes to when he mentions "other kinds of clusters". This question is much more difficult to answer and cluster models do not offer a clear interpretation as far as we can tell.

Following Jain and Dubes (1988), we assume that cluster validation means that the researcher should overcome "the clustering tendency problem". This "refers to the problem of deciding whether data exhibit a predisposition to cluster into

12 A reviewer suggests that *k*-means clustering would be more appropriate for the research question addressed in this section. In a previous phase of research on this topic, the authors of this study considered and implemented this, but ultimately rejected using *k*-means clustering on conceptual grounds. *K*-means clustering coupled with the elbow method for cluster validation consistently suggested that the languages of our study had between 4 and 9 optimal clusters. At face value one might think that such results provide evidence against the morphology-syntax distinction, however, no such conclusion is warranted. As *k*-means clusters are not hierarchically taxonomized, maintaining such a conclusion would be akin to claiming that a morphology-syntax distinction is only motivated just in case there are no distributional subclasses under the supra-classes 'affix' (morphological element) and 'word' (syntactic element). But, no linguist of any theoretical stripe has ever advocated such a position to our knowledge.

natural groups without identifying the groups themselves. Clustering algorithms will create clusters whether the data are naturally clustered or purely random" (Jain and Dubes 1988: 201). Following a suggestion in Jain and Dubes (1988), we address this issue by developing a data set in which results for the variables or our study are simulated and we then compare the results of the cluster models on individual languages to the simulated data set. Since a single simulated language might not necessarily be representative of randomness, we construct a data set of 1000 simulated languages for comparison.[13] Details of how the simulated data were constructed can be found in the Supplementary materials.

For each language, we constructed a distance matrix using the Gower distance metric appropriate for mixed data types (Gower 1971) – recall that our variables are binary, ordinal, or continuous. These distance matrices were then used as the input for hierarchical cluster models, using Ward's minimum variance method with the cluster package in R (Maechler et al. 2021). The dendrogram for the simulated data set is provided in Figure 5 and for the languages of our sample in Figure 6.

For some of the languages, visual comparison with the simulated language (SL) lends some support to a morphology-syntax distinction. For others, the distinction is less obvious. Consider the difference between SL in Figure 5 with that of Chácobo and Movima in Figure 6. Compared to the dendrogram of the SL, both languages exhibit a striking height difference between the first two partitions. The height difference reflects the distance between clusters, and thus, shows that in Chácobo and Movima a partition into two groups creates clusters that are much better separated than in the SL. Central Alaskan Yupik, Ashéninka Perené, and Tariana display roughly the same difference compared with the SL, although to a lesser degree. That Wãnsöjöt/Puinave, Cavineña, and Hup cut base-morphemes

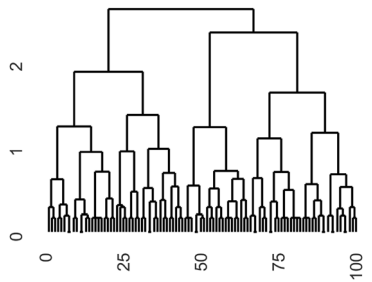

Figure 5: Hierarchical clustering on a simulated language.

13 We thank an anonymous reviewer for suggesting this to us.

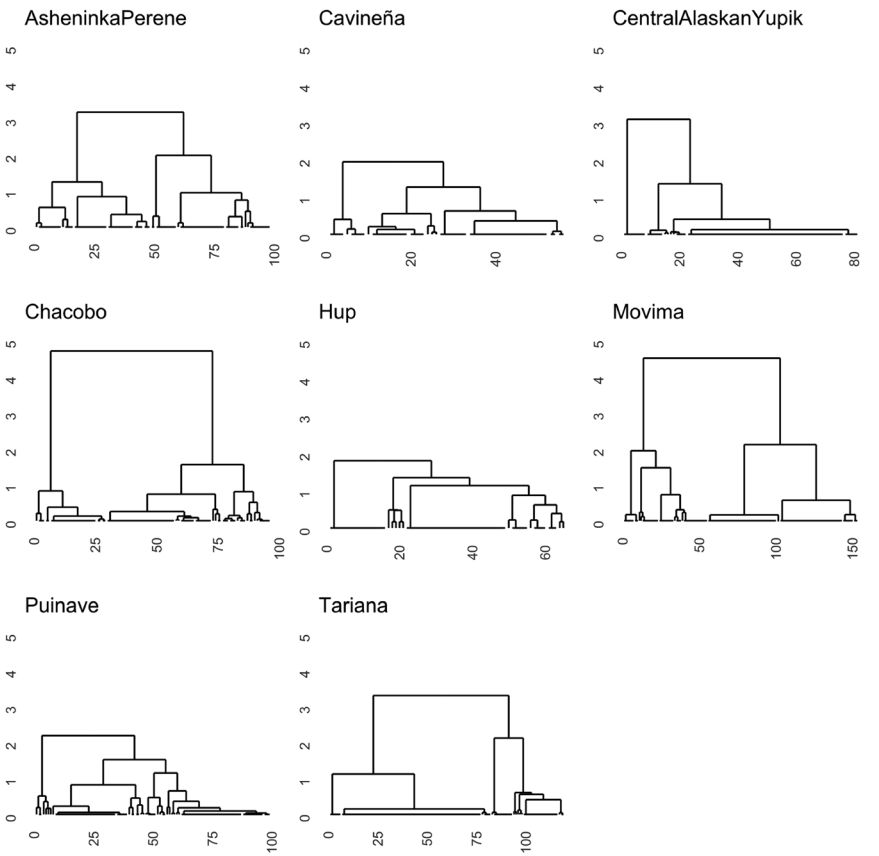

Figure 6: Hierarchical clustering on the eight languages of the sample.

into morphological and syntactic types better than chance is less obvious. The first versus second partition difference are closer to that of the SL. In fact, in Hup the height difference between the first partition and the second is smaller than that of the SL.

If visual inspection of the dendrograms serves as a basis for assessing the motivation of partition into morphological and syntactic constructions, the hierarchical cluster models suggest that languages vary according to how much better than chance they cluster elements into two partitions (morphology vs. syntax). However, this should **not** be interpreted as meaning that the wordhood criterial variables pattern no better than chance as a whole. Once we move beyond considering the first partition the situation becomes more complex, a point we now turn to with reference to the metric of cophenetic distance.

Table 6 provides an overview of the height differences between the first partition and the second across the languages of the study and also the correlation between cophenetic distance and Gower distance between the elements coded in each of the languages. We also include the values for the simulated language displayed in 5 which we use for illustration. Cophenetic distance is a measure of how (dis)similar two elements need to be in order to be grouped into the same cluster (Sokal and Rohlf 1962). It can be calculated by the height of the dendrogram where the branches of the clusters that contain the elements meet. The cophenetic correlation is used to assess how well the clusters imposed on the data fit the distances between those elements. The correlation varies between 0 and 1, with higher numbers reflecting a better fit of the model to the data. Studies on data sets simulated by Monte Carlo methods have shown that the cophenetic correlation coefficient can vary between 0.55 and 0.65 with (pseudo-)random data depending on the number of variables and data points (Rohlf and Fisher 1968: 408). In our case the cophenetic correlation coefficient of the SL varies from 0.45 to 0.60 over 1,000 simulations. We suggest that the height difference between the first and the second partition could serve as a basis for assessing the evidence for a morphology-syntax distinction. If the distinction was not motivated and elements were not clustering into two groups better than chance, as Haspelmath suggests that they might be, we would expect the height differences from the first and second partitions to be close that of the simulated language or the mean of the 1,000 simulated languages. Thus, we can assess how much evidence there is for a morphology-syntax distinction in a given language based on their position in relation to the simulated languages of our study. Figure 7 plots the first versus

Table 6: Cophenetic-distance correlations and the height differences between the first and second partition in hierarchical cluster models of 8 natural languages, the simulated language, and the mean of 1000 simulated languages.

Language	1st/2nd part. height diff.	Cophenetic Correlation
Central Alaskan Yupik	1.7257	0.9865
Tariana	1.1836	0.9353
Chácobo	3.1473	0.9219
Cavineña	0.6766	0.8740
Movima	2.3942	0.8290
Wānsöjöt/Puinave	0.6677	0.7728
Hup	0.4552	0.7726
Ashéninka Perené	1.2021	0.7104
Simulated Language (Figure 5)	0.2179	0.5134
Mean of 1,000 simulated languages	0.5754	0.5195

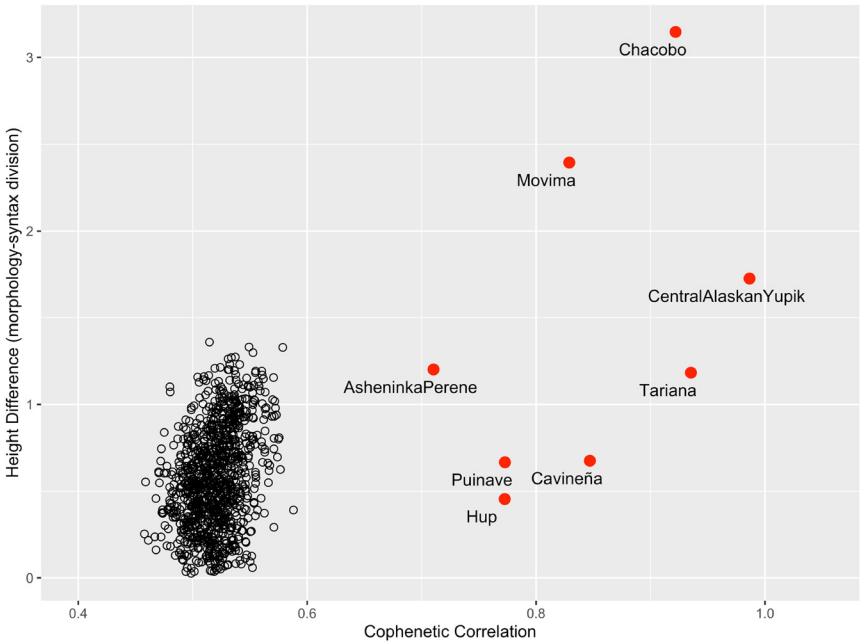

Figure 7: Cophenetic correlation and first/second partition height differences of 1,000 simulated data sets (black circles) and the languages of our sample (red dots).

second partition height difference and cophenetic correlation of the 1,000 simulated languages and the languages of our study on the *x* and *y* axes respectively. We can see from this plot that all real languages have a higher cophenetic variation than all the simulated languages. This is to be expected because real languages have more groupings of morphosyntactic categories regardless of whether they display a morphology-syntax distinction. The situation is different when we consider languages in terms of the first versus second partition height difference (the *y*-axis). Five out of the eight languages of our study overlap with the distribution of SLs (Asheninka, Tariana, Cavineña, Puinave and Hup) on the *x*-axis. It is difficult to know what the strict cut off point should be for assessing whether a language has a morphology-syntax distinction, but distributional overlap with a set of random languages suggests a relatively weak division between morphology and syntax at the very least. Visual inspection of Figure 7 suggests that in Puinave, Cavineña and Hup, in particular there is not very strong evidence for a morphology-syntax distinction. These languages have first versus second partition height differences that are relatively close to the mean value of the SLs in the aggregate. Tariana and Asheninka are beyond the third quartile of the distribution

of SLs (0.7933), and thus it is more likely that these languages display a morphology-syntax distinction.

Importantly, the languages vary in terms of how well distinguished they are from the SLs, with, Chácobo, Central Alaskan Yupik and Movima appearing with a more motivated distinction between morphology and syntax than the other languages. Chácobo, Movima, and Central Alaskan Yupik stand out in that they exhibit height differences outside the range of the simulated languages.

What these results suggest is that languages may vary in terms of whether they display a morphology-syntax distinction. Furthermore, we can measure the degree to which such a distinction is valid. We have, thus, moved towards providing a method that can engage with Haspelmath's critique of the notion of fuzzy wordhood.

8 Conclusions and future research

This paper has provided an overview of three problems that a variationist perspective on the morphology-syntax distinction could address. The problems were illustrated and explored using different statistical methods. Correlation matrices can be used to provide an overall profile of associations between wordhood criterial variables. To the extent that strong positive correlations can be understood as reinforcing a more discrete distinction between morphology and syntax, correlations matrices can provide a good overview of linguistic variation in the extent to which morphology and syntax is being distinguished between two variables at a time. The differences in the associations between the variables suggest that some variables are more important than others and that languages might vary in this regard. Random Forest models provide a method for assessing this question, but they depend on assuming a dependent variable or base classification. Clustering methods do not require a dependent variable, but they require choice of a clustering algorithm, a clustering method and a validation technique. It was suggested here that the simulation of a (pseudo-)random data set that included variables of the same type and the same range of that of the languages of the study could be used to ground a clustering validation technique. We showed that some languages do indeed provide strong evidence for a binary partition of their data into morphological and syntactic constructions, but some do not. The fact that these hierarchical models still fit the latter languages better than chance invites the possibility that in these languages different types of global divisions are more important.

Throughout the study, we pointed out a number of problems that need to be overcome for variationist studies on the morphology-syntax distinction to proceed.

The most obvious one is that a larger sample of languages with a larger sample of morphemes needs to be gathered in order to make empirically responsible generalizations. For the assessment of the composition problem some theoretically motivated baseline classification has to be given to the data for the relative importance of variables across languages to be appropriately tested. It is not clear that exponence complexity is the right answer in this case, because correlations between exponence complexity and other word variables were so low in the languages of our study. The architecture problem requires an assessment of the data that overcomes the 'clustering tendency problem'. Visual inspection of the data are not enough. Rather the results of clustering models applied to different languages should be compared with some null or random distribution.

We think that an emphasis on the issue of global architecture might help bridge the gap between two broad ideas present in the literature on the morphology-syntax divide, one which emphasizes continuity and the other which emphasizes discontinuity.

On the continuity side, the grammaticalization literature emphasizes gradience between affixal and word-like elements (Hopper and Traugott 2003: 7; Traugott and Trousdale 2010: 25). However, it is not entirely clear why a language should display even a statistically motivated distinction between morphology and syntax in such a perspective. Should we not expect there to be just as many languages that are no better than random than not? In grammaticalization theory the emphasis on the history of individual elements or specific constructions typically occurs without an eye towards how such elements are embedded in a larger ecology of morphosyntactic groupings (e.g. Kuteva et al. 2019 and papers cited therein). We have established that elements and constructions can be placed on a cline from syntactic independence to morphologically integrated, but now, with the methods explored in this paper, we can move towards exploring the distribution of elements and constructions along this cline in a given language. On the discontinuity side, we find some discussions of morphological complexity emphasizing a lack of continuity between morphology and other domains and liminal elements are taken to be marginalia (see Section 1). Despite the fact that many of these researchers have been able to show that there are domains of grammar that display patterns distinct from combinatorial syntax and phonology, the idea that such patterns support an autonomous morphological representation remains a matter of controversy (Luís and Bermúdez-Otero 2016). Apart from adding more languages and data points to the sample, and developing a more rigorous sampling technique, future research will involve further developing the wordhood criterial properties, perhaps elaborating a less arbitrary and more theoretically motivated metric of exponence complexity, and exploring different types of clustering methods (both exploratory and validational). The upshot of

such a project would be the development of a data base and a set of methodologies that could help in the development of more testable theories of linguistic architecture. Rather than simply emphasizing that the boundaries between morphological and syntactic structure (or whatever other domains we wish to propose) are fuzzy to some unknown degree, we could move towards developing theories that posit quantifiable constraints on the variation and (dis)continuity.

References

Aikhenvald, Alexandra Y. 2002. Typological parameters for the study of clitics, with special reference to tariana. In R. M. W. Dixon & Alexandra Y. Aikhenvald (eds.), *Word: A cross-linguistic typology*, chap. 2, 42–78. Cambridge: Cambridge University Press.

Aikhenvald, Alexandra Y. 2003. *A grammar of Tariana*. Cambridge: Cambridge University Press.

Aldenderfer, Mark S. & Roger K. Blashfield. 1984. *Cluster analysis* (Quantitative Applications in the Social Sciences 44). London: SAGE Publications.

Anderson, Gregory D. S. 2006. *Auxiliary verb constructions*. Oxford University Press.

Anderson, Stephen R. 1992. *A-morphous morphology*. Cambridge: Cambridge University Press.

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

Anderson, Stephen R. 2015. Dimensions of morphological complexity. In Matthew Baerman, Dunstan Brown & Greville Corbett (eds.), *Understanding and measuring morphological complexity*, 11–28. Oxford: Oxford University Press.

Aronoff, Mark. 1994. *Morphology by itself: Stems and inflectional classes*. Cambridge: MIT Press.

Baayen, Harold. 2013. Multivariate statistics. In Robert J. Podesva & Devyani Sharma (eds.), *Research methods in linguistics*, chap. 16, 337–372. Cambridge: Cambridge University Press.

Baker, Mark. 1988. *Incorporation: A theory of grammatical function changing*. Chicago: University of Chicago Press.

Belsely, David A., Edwin Kuh & Roy E. Welsch. 1980. *Regression diagnostics. Identifying influential data and sources of collinearity*. New York: Wiley.

Bickel, Balthasar. 2010a. 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. 2010b. Distributional typology: Statistical inquiries into the dynamics of linguistic diversity. In Bernd Heine & Heiko Narrog (eds.), *The oxford handbook of linguistic analysis*, chap. 37, 901–922. Oxford: Oxford University Press.

Bickel, Balthasar, A. Hildebrandt Kristine & René Schiering. 2009. The distribution of phonological word domains: A probabilistic typology. In Janet Grijzenhout & Kabak Baris (eds.), *Phonological domains: Universals and deviations*. Berlin, New York: De Gruyter Mouton.

Blevins, James P. 2006. Word-based morphology. *Journal of Linguistics* 42(3). 531–573.

Blevins, James P. 2016. *Word and paradigm morphology*. Oxford: Oxford University Press.

Bloomfield, Leonard. 1933. *Language*. New York: Holt, Rinehart and Winston.

Booij, Geert. 1997a. Allomorphy and the autonomy of morphology. *Folia Linguistica* 31. 25–56.

Booij, Geert. 1997b. Autonomous morphology and paradigmatic relations. In Geert Booij & Jaap van Marle (eds.), *Yearbook of morphology 1996*, 35–53. Dordrecht: Springer.

Brandão, Ana Paula Barros. 2014. *A reference grammar of Paresi-Haliti (Arawak)*. University of Texas at Austin dissertation.
Croft, William. 2001. *Radical construction grammar*. Oxford: Oxford University Press.
Cruschina, Silvio, Martin Maiden & John Charles Smith. 2013. Introduction. In Silvio Cruschina, Maiden Martin & John Charles Smith (eds.), *The boundaries of pure morphology: Diachronic and synchronic perspectives*, 1–8. Oxford: Oxford University Press.
Cysouw, Michael. 2002. Interpreting typological clusters. *Linguistic Typology* 6(1). 69–93. https://doi.org/10.1515/lity.2002.003.
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*, chap. 1, 1–41. Cambridge: Cambridge.
Epps, Patience. 2004. *A grammar of Hup*. Berlin: Mouton de Gruyter.
Girón Higuita, Jesús Mario. 2008. *Una gramática del Wãnsöjöt (Puinave)*. Utrecht: LOT.
Gower, John C. 1971. A general coefficient of similarity and some of its properties. *Biometrics* 27(4). 857–871.
Guillaume, Antoine. 2008. *A grammar of Cavineña*. Berlin: Mouton de Gruyter.
Gutiérrez Lorenzo, Ambrocio. 2018. *Negative constructions in Teotitlán del Valle Zapotec. Qualifying paper*. Austin, Texas: University of Texas at Austin.
Hartigan, John A. 1975. *Clustering algorithms*. New York: John Wiley & Sons.
Haspelmath, Martin. 2011. The indeterminacy of word segmentation and the nature of morphology and syntax. *Folia Linguistica* 1(45). 31–80.
Haspelmath, Martin. 2021. Towards standardization of morphosyntactic terminology for general linguistics. In Luca Alfieri, Giorgio Francesco Arcodia & Paolo Ramat (eds.), *Linguistic categories, language description and linguistic typology*. 35–57. Amsterdam: John Benjamins.
Haude, Katharina. 2006. *A grammar of Movima*. Nijmegen: Radboud Universiteit.
Hockett, Charles. 1958. *A course in modern linguistics*. New York: Macmillan.
Hopper, Paul J. & Elizabeth Closs Traugott. 2003. *Grammaticalization*, 2nd edn. Cambridge: Cambridge University Press.
Hyman, Larry M. 2006. Word-prosodic typology. *Phonology* 23. 225–257.
Hyman, Larry M. 2009. How (not) to do phonological typology: The case of pitch-accent. *Language Sciences* 31. 213–238.
Jain, Anil K. 2010. Data clustering: 50 years since k-means. *Pattern Recognition Letters* 31. 651–666.
Jain, Anil K. & Richard Dubes. 1988. *Algorithms for clustering data*. Englewood Cliffs: Prentice Hall.
Jakobson, Roman. 1939. Signe zéro. *Mélenades de linguistique offerts á Charles Bally*, 143–152. Genève: George.
Kaufman, Leonard & Peter J. Rousseeuw. 2005. *Finding groups in data: An introduction to cluster analysis*. Hoboken: Wiley & Sons.
Kuteva, Tania, Bernd Heine, Bo Hong, Haiping Long, Keiko Narrog & Seongha Rhee. 2019. *World lexicon of grammaticalization*, 2nd edn. Cambridge: Cambridge University Press.
Levshina, Natalia. 2016. Why we need a token-based typology: A case study of analytic and lexical causatives in fifteen European languages. *Folia Linguistica* 50(2). 507–542.
Liaw, Andy & Matthew Wiener. 2002. Breiman and Cutler's random forests for classification and regression. *R News* 2(3). 18–22.

Louppe, Giles. 2014. *Understanding random forests: From theory to practice*. Liège: University de Liège (Faculty of Applied Sciences Department of Electrical Engineering & Computer Science) dissertation.

Luís, Ana R. & Ricardo Bermúdez-Otero. 2016. Introduction. In Ana R. Luís & Ricardo Bermúdez-Otero (eds.), *The morphome debate*, 1–11. Oxford: Oxford University Press.

Maechler, Martin, Peter J. Rousseeuw, Anja Struyf, Mia Hubert & Hornik Kurt. 2021. *cluster: Cluster analysis basics and extensions*. https://CRAN.R-project.org/package=cluster. R package version 2.1.2—For new features, see the 'Changelog' file (in the package source).

Maiden, Martin. 2011. Morphomes and 'stress-conditioned allomorphy' in Romansch. In Martin Maiden, Charles John Smith, Maria Goldbach & Marc-Olivier Hinzelin (eds.), *Morphological autonomy: Perspectives from Romance inflectional morphology*. Oxford: Oxford University Press.

Maiden, Martin. 2013. 'Semi-autonomous' morphology? A problem in the history of the Italian (and Romanian) verb. In Silvio Cruschina, Martin Maiden & John Charles Smith (eds.), *The boundaries of pure morphology: Diachronic and synchronic perspectives*, 26–44. Oxford: Oxford University Press.

Matthews, Peters H. 1991. *Morphology*, 2nd edn. Cambridge: Cambridge.

Mel'čuk, Igor. 1993. *Cours de morphologie générale (théorique et descriptive), introduction et première partie: Le mot.* Montréal: Presses de l'Université de Montréal.

Mel'čuk, Igor. 2006. *Aspects of the theory of morphology*. Berlin: Mouton de Gruyter.

Mihas, Elena. 2015. *A grammar of Alto Perené (Arawak)*. Berlin: Mouton de Gruyter.

Miyaoka, Osahito. 2012. *A grammar of Central Alaskan Yupik (CAY)*. Berlin: Mouton de Gruyter.

Mugdan, Joachim. 1994. Morphological units. In Ron Asher (ed.), *The encyclopedia of language and linguistics*, 2543–2553. Oxford: Pergamon Press.

Olawsky, Knut J. 2006. *A grammar of Urarina*. Berlin: Mouton de Gruyter.

Peterson, John. 2006. *Kharia: A South Munda language*. Universität Osnabrück dissertation.

Pike, Kenneth L. 1972. A problem in morphology-syntax division. In Ruth M. Brend (ed.), *Kenneth L. Pike selected writings*, 74–84. Mouton. First published in: *Acta Linguistica* 5:3 (1949), 125–138. Reprinted by permission.

Rohlf, F. James & David R. Fisher. 1968. Tests for hierarchical structure in random data sets. *Systematic Zoology* 17(4). 407–412. https://doi.org/10.1093/sysbio/17.4.407.

Rosen, Sara Thomas. 1989. Two types of noun incorporation: A lexical analysis. *Language* 65(2). 294–317.

Round, Erich R. & Greville G. Corbett. 2020. Comparability and measurement in typological science: The bright future for linguistics. *Linguistic Typology* 24(3). 489–525.

Russell, Kevin. 1999. What's with all these long words anyways? In Leora Bar-el, Rose-Marie Dechaine & Charlotte Reinholtz (eds.), *MIT Occasional Papers in Linguistics*, 119–130. Cambridge, Massachusetts: MITWPL.

Sadock, Jerrold M. 1980. Noun incorporation in Greenlandic: A case of syntactic word formation. *Language* 56(2). 300–319.

Sadock, Jerrold M. 1991. *Autolexical syntax: A theory of parallel grammatical representations*. Chicago: Chicago University Press.

Sapir, Edward. 1921. *Language: An introduction to the study of speech*. New York: Harcourt, Brace & World.

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

Smith, John Charles. 2013. The morphome as a gradient phenomenon: Evidence from Romance. In Silvio Cruschina, Maiden Martin & John Charles Smith (eds.), *The boundaries of pure morphology: Diachronic and synchronic perspectives*, 247–261. Oxford: Oxford University Press.

Sneath, Peter H.A. & Robert R. Sokal. 1973. *Numerical taxonomy: The principles and practice of numerical classification*. San Francisco: W.H. Freeman and Company.

Sokal, Robert R. & F. James Rohlf. 1962. The comparison of dendrograms by objective methods. *Taxon* 11(2). 33–40.

Spencer, Andres & Ana Luís. 2012a. The canonical clitic. In Dunstan Brown, Marina Chumakina & Greville G. Corbett (eds.), *Canonical morphology and syntax*, 123–150. Oxford: Oxford University Press.

Spencer, Andrew. 1995. Incorporation in Chukchi. *Language* 71(3). 439–489.

Spencer, Andrew & Ana R. Luís. 2012b. *Clitics: An introduction*. Cambridge: Cambridge University Press.

Stewart, Thomas W. 2016. *Contemporary morphological theories: A user's guide*. Edinburgh: Edinburgh University Press.

Stump, Gregory T. 2001. *Inflectional morphology: A theory of paradigm structure*. Cambridge: Cambridge University Press.

Stump, Gregory T. 2016. *Inflectional paradigms: Content and form at the syntax-morphology interface*. Cambridge: Cambridge University Press.

Tallman, Adam J.R. 2018a. *A grammar of Chácobo, a southern Pano language of the northern Bolivian Amazon*. University of Texas at Austin dissertation.

Tallman, Adam J. R. 2018b. There are no special clitics in Chácobo (Pano). In Megan Keough, Natalie Weber, Andrei Anghelescu, Sihwei Chen, Erin Guntly, Khia Johnson, Daniel Reisinger & Oksana Tkachman (eds.), *Proceedings of the workshop on structure and constituency in the languages of the americas 21*, 194–209. University of British Columbia Working Papers in Linguistics 46.

Tallman, Adam J. R. 2020. Beyond grammatical and phonological words. *Language and Linguistics Compass* 14(2). e12364.

Tallman, Adam J. R. 2021. Constituency and coincidence in Chácobo (Pano). *Studies in Language* 45(2). 321–383.

Tallman, Adam J. R., Eric W. Campbell, Hiroto Uchihara, Ambrocio Gutiérrez, Dennis Wylie, Eric Adell, Natalia Bermudez, Gladys Camacho-Rios, Javier Carol, Patience Epps, Michael Everdell, Cristian R. Juárez, Willem de Reuse, Kelsey Neely, Andrés Pablo Salanova, Anthony C. Woodbury, Magdalena Lemus & Denis Bertet. 2019. A new typology of constituency and convergence. *Paper presented at the 13th Conference of the Association of Linguistic Typology*.

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

Tomaschek, Fabian, Peter Hendrix & R. Harald Baayen. 2018. Strategies for addressing collinearity in multivariate linguistic data. *Journal of Phonetics* 71. 249–267.

Traugott, Elizabeth Closs & Graeme Trousdale. 2010. Gradience, gradualness and grammaticalization: How do they intersect. In Elizabeth Closs Traugott & Graeme Trousdale (eds.), *Gradience, gradualness and grammaticalization*, 19–45. Amsterdam: John Benjamins.

Vallejos Yopán, Rosa. 2010. *A grammar of Kokama-Kokamilla*. University of Oregon dissertation.

Vincent, Nigel. 2011. Non-finite forms, periphrases, and autonomous morphology in Latin and Romance. In Martin Maiden, John Charles Smith, Maria Goldbach & Marc-Olivier Hinzelin (eds.), *Morphological autonomy: Perspectives from Romance inflectional morphology*, 417–435. Oxford: Oxford University Press.

Wali, Kashi & Omkar N. Koul. 1997. *Kashmiri: A cognitive-descriptive grammar*. London: Routledge.

Wickham, Hadley. 2016. *ggplot2: Elegant graphics for data analysis*. New York: Springer.

Wolfart, Christopher C. 1973. Plains Cree: A grammatical study. *Transactions of the American Philosophical Society, New Series* 63(5). 1–90.

Woodbury, Anthony C. 1996. On restricting the role of morphology in Autolexical syntax. In Eric Schiller, Barbara Need & Elisa Steinberg (eds.), *Autolexical syntax: Ideas and methods*, 319–366. Berlin: Mouton de Gruyter.

Zariquiey, Roberto. 2018. *A grammar of Kakataibo*. Berlin: Mouton de Gruyter.

Zwicky, Arnold. 1985. How to describe inflection. *Berkeley Linguistics Society* 11. 372–386.

Zwicky, Arnold M. & Geoffrey K. Pullum. 1983. Cliticization vs. inflection: English n't. *Language* 59(3). 502–513. https://doi.org/10.2307/413900.

Supplementary Material: The supplementary material is available at (https://zenodo.org/record/6008054#.YgJYTPiCHIU).