Analysis and falsifiability in practice

Adam J.R.Tallman
Friedrich Schiller Universität Jena

1 Introduction

Martin Haspelmath (MH) provides a classification of methodological approaches to the study of general linguistics. He contrasts a Boasian/Greenberian approach ,which mediates the relation between general linguistics and particular linguistics with comparative concepts, to a natural kinds programme, which does away with the latter. He critiques a trend in generative research whereby the technical vocabulary of a natural kinds programme is adopted without its ontological assumption, that linguistic kinds are natural kinds. MH also provides a discussion of 'analysis' sections in linguistics papers that make them seem strange or even superfluous for a general audience. In this paper I will focus on this notion of 'analysis' and critique a typical justification found for such sections. It is often claimed that theoretical models developed and fit to data of one language make testable predictions about the next. I argue that this is not true in cases where categories and relations presupposed by the hypothesis do not map deterministically onto the next language. While such a theoretical model can be fit to novel language data, hypotheses couched in the model cannot be meaningfully tested. Stated in another way, 'analysis', as it is typically undertaken in linguistics, renders the hypotheses such the analysis is designed to test unfalsifiable in practice. The remedy is to move in the direction of all other sciences and develop measurements of uncertainty in the mapping of language data to categories of general linguistics. It is unclear how this can be done without engaging in a type of cross-linguistic comparison.

Before addressing 'analysis', I briefly discuss MH's claim that descriptive linguistics should also be regarded as theoretical.

2 Is description "theoretical"?

In contrast to common rhetorical practice (e.g. (Ludlow 2011:82-86)), MH defines "theoretical linguistics" in such a way that encompasses the description and documentation of particular languages: "the study of language(s) that aims primarily at understanding (or explaining) rather than at practical applications". In her commentary on MH's

paper D'Alessandro finds MH's definition too broad suggesting that it implies that "everything is a theory".

If one simply defines "theory" in such a way that it necessarily attributes the patterns one observes in particular languages to properties of human language (or the language faculty), then there is nothing to argue with: descriptive linguistics does not have "theories" in this sense. However, one could also use the label more generally to refer to falsifiable hypotheses, regardless of their explanatory scope. In this sense descriptive linguistics does have "theories". My claim that ergative marking in Chácobo never occurs (or is never overtly realized) after clause-type morphemes (Tallman 2018a) is certainly falsifiable (see Moravcsik (2007)), regardless of whether one believes that this claim helps corroborate Coon's hypothesis about split ergativity (Salanova & Tallman 2020) or not (Tallman 2018b). Much of the impetus for developing richly annotated documentary records is to cross-check descriptive generalizations (Himmelmann 1998, 2012). In other words we are interested in our ability to check whether such descriptive generalizations might be wrong. A good descriptive grammar is not just one that gets the facts right, but one where the statements are explicit enough that they could be falsified by novel data. This is implicitly recognized by everyone regardless of their theoretical leanings.

Even if descriptive linguistics does not have theories because they lack a metalanguage that ultimately attributes the patterns to properties of human language, it could still be regarded as theoretical on the grounds that general linguistics' theories require robust and accurate descriptive generalizations to develop and corroborate claims about human languages. Descriptive linguistics also has a heavy concern with the advantages and pitfalls of certain types of data-gathering techniques, which are relevant for the assessment of the accuracy of more general theories (Woodbury 2007, Tallman 2021a). General linguistics' theories cannot be regarded as corroborated or falsified by novel data, if such data involve sentences which are calques from the contact languages, involve inaccurate grammaticality judgements, or, for example, statements about relative metrical prominence that cannot be corroborated with phonetic evidence. Developing methods for cross-checking descriptive generalizations or detecting potential errors in the primary data itself are crucial aspects of scientific practice, a fact that recent 'new experimentalist' work in the philosophy of science emphasizes (Hacking 1983, Galison 1987, Mayo 1996).

In this sense descriptive linguistics is necessarily theoretical. But what about the "theoretical" status of more general claims? It is common rhetorical practice in our field to refer to any general claims as "theoretical" even if it is unclear if they make any testable predictions at all. Many general linguistics "theories" have been critiqued on the grounds that they amount to covert tautologies. Newmeyer (1998:145-146) refers to many claims about competing motivations in functional linguistics as "vacuous"

since they can gloss over "any conceivable state of affairs".[1] Abels & Neeleman (2012) suggest that Kayne's linear correspondence axiom (LCA) can only be true with an unrestricted theory of movement, but that adopting such a theory renders the LCA "toothless". I have suggested that some current versions of the prosodic hierarchy theory are basically tautologies (Tallman 2021b). If by referring to some claim or practice as "theoretical" we mean to assign it some scientific status, I think we are asking the wrong question when we ask whether descriptive and documentary linguistics passes the bar. Descriptive and documentary linguistics is either atheoretical by definition, or obviously theoretical, because descriptive claims are testable and descriptive methodologies are required to assess the empirical facts on which general claims are based. Rather, the question should be directed at universalistic claims, which have a much more tenuous relationship with falsifiability.

3 Models, diagnostics and (severe) testing

In this section I take issue with an often repeated claim regarding falsifiability in linguistics. Newmeyer makes the case succinctly.

> … any claim about universals based on a single language makes a testable falsifiable hypothesis. The more languages that are studied the better the hypothesis is tested. (Newmeyer 2008:59)

I argue that this statement (see also (Chomsky 1972, 1985, Mendívil-Giró 2019, D'Alessandro)) is not necessarily true if we are interested in making reliable inferences about whether the hypothesis is corroborated in the context of a new language. Specific conditions have to hold for a theoretical model retrodicted to fit data of one language to make obviously testable predictions about the next. I will make my point abstractly, before discussing specific examples from my own research. I relate my argument to recent work by Mayo, who argues that falsifiability in principle does not automatically translate to falsifiability in practice if the scientific community does not or cannot develop methods for severely testing a hypothesis against novel data. Mayo and Spanos update Popper's criterion of falsifiability (Popper 1959, 1963) based on literature on the philosophy of statistical inference (Mayo 1996). We have reason to take a hypothesis seriously if it survives severe tests. A severe test is one that a hypothesis would be unlikely to survive if it were false (Mayo & Spanos 2011, Mayo 2018). Logically falsifiable hypotheses are not worth their salt if they are coupled with a methodology that prejudge their correctness at the onset. I think that the process of 'analysis'

[1] Newmeyer does not make this claim as a matter of principle. If the competing factors are quantified and made explicit then they can be controlled and such theories can be made testable in practice, as they have been in more recent years (see papers in MacWhinney et al. (2015))

described by MH tends to overestimate the certainty with which we can make reliable inferences from a single language.

What Newmeyer et al.'s claim ignores is that when we want to test a hypothesis with novel data from a different language, we have to fit the theoretical model to descriptive generalizations of that language.[2] There is no ready made guide for translating the data into the metalanguage of the theory.[3] When I elicit data from a speaker of a lesser known language, I transcribe utterances with translations, ideally on different levels of compositionality (Woodbury 2007).[4] I do not transcribe Ns, vPs, N0s, TPs, left branching constituents, remnant movement, traces, prosodic words, etc. I have to make analytic decisions regarding how to map the technical vocabulary of any theory onto the data and descriptive generalizations that emerge from it.

I will refer to a particular mapping of language data into technical theoretical vocabulary as a "model" of that language data. We do not directly corroborate or falsify our hypotheses with novel data. Rather we test our hypotheses against data mediated by one or more models of the data. The model is a translation of the data into the technical vocabulary of one's theory. I am using the notion of "model" in the sense that it is typically used in the context of statistical inference (McElreath 2020), not necessarily the way the notion of model is understood by linguists. Phrase structure grammar or Optimality Theory are not models in this sense. A model is a particular mapping of OT to specific language data, or a particular account of the phrase structure rules in a specific language. Coon's theory of split ergativity is not a model by itself. When biclausality is mapped to split ergative constructions in Chol (Coon 2013) this is a model of the relevant data in Chol vis-à-vis the theory. When Salanova & Tallman (2020) map biclausality to neutrally case-aligned clauses in Chácobo, in contrast to the descriptive literature (Tallman 2018a), this is a model of the relevant data in Chácobo vis-à-vis Coon's theory of split ergativity.

In an ideal case, a single model (M) emerges from the data (D). There is a unique mapping between the data and the categories and relations presupposed by the hypothesis (H), represented in figure 1.

If there is a unique mapping between the categories and structures presupposed by the hypothesis and the novel data we find, then there is no problem with Newmeyer's

[2]Arguably this formulation is too generous. As a PhD student attending seminars on documentary and descriptive linguistics at UT Austin, a frequent complaint by our professors in reference to "theoretical work" on languages of the Americas was that it involved "fitting the language to the theory", meaning that the theory is mapped onto categories which have no language internal justification at all, or worse, mapped to data that are made up, and that the theoretician either does not realize this or does not care.

[3]In fact, the attempt to develop of such guides or "discovery procedures" are explicitly rejected in generative linguistics (Chomsky 1957, Seuren 2004).

[4]'A level of compositionality' roughly refers to the size (in terms of internal components) of the utterance that is being translated. In documentary linguistics we translate whole sentences and also we probe the translations of their individual parts where we can

Figure 1: A single model mediating a theory in an ideal case

claim. A perhaps weaker situation is where there are multiple possible models, but there is an obviously best one. It is through the unique model or the best model that the hypothesis of interest can be tested with novel data.

Real world examples are not like this. I use a simple example to illustrate the point. In Chácobo, ergative A-argument nominal are marked with a high tone on their final syllable. Absolutive {S,P} are left unmarked as in the example in (1). Ergative marking does not appear if the {A} argument occurs after the clause-type morpheme as in 2 (=kɨ 'declarative past' in examples below). Tallman (2018a) refers to these as C(clause-type)-SUBJ(ect) constructions.

(1) Ergative alignment
honi=́ tʃaʂo tsáya=kɨ
man=ERG deer see=DECL:PAST
'The man saw the deer.'

(2) Neutral alignment
tʃaʂo tsáya=kɨ honi
deer see=DECL:ANT man
'The man has seen the deer.'

Coon's theory predicts that where we get ABS-ABS / neutral alignment as in (2) it is because the A argument is in an intransitive clause (that is, it is covertly an S) and the P argument is in a subordinate clause. On the surface the C-SUBJ construction does not look biclausal because there is a single predicate *tʃaʂo tsaya* 'see the deer' and no overt marking of subordination. However, if we search around we notice certain parallels with non-verbal predicate constructions. Non-verbal predicate constructions also display C-SUBJ order as in (3). They of course have no ergative marking, because ergative marking is only motivated by a local verb, i.e., one in the same clause.

(3) Nonverbal predicate construction
tʃáʂo ʂo honi
deer DECL man
'The man is a deer'

Two models of the C-SUBJ construction emerge. One where the C-SUBJ construction is a monoclausal verbal predicate construction (M(odel)1) and another where the C-SUBJ construction is a biclausal construction with covert subordination of a verbal predicate construction under a nonverbal predicate one (M(odel)2).

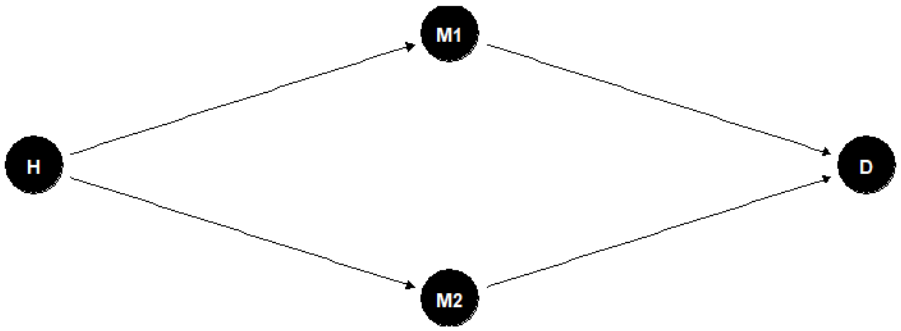

Figure 2: Two candidate models mediating a theory

If we choose M1, the Chácobo data falsify Coon's theory. If we choose M2, Chácobo corroborates Coon's theory. Coon's hypothesis is "better tested" in Newmeyer's words. How do we choose know which model is correct? How do we know whether the Chácobo data corroborate or falsify Coon's structure based theory of case?

Every trained linguist knows the answer to this question. There is by now a long tradition of tallying up 'diagnostics' or 'tests' to support one 'analysis'/'model' against the other. The 'tests' are the epistemic tools we use to *fit* the categories and structures of our hypotheses to the models of the language data we are considering. In generative frameworks they serve to link up data of particular languages to categories and relations presupposed by the hypotheses about the language faculty we are interested in testing. Where the Boasian/Greenbergian approach uses 'comparative concepts', the contemporary natural-kinds programme evokes tests to navigate the space created by the abstractness of the theoretical metalanguage and the facts of individual languages constructions. In the case described above we go fishing for tests that distinguish between monoclausal and biclausal constructions and apply those to the C-SUBJ construction. The hypothesis cannot be directly assessed in relation to the data, rather a hypothesis context must be constructed through bridging principles that link up observants with the theory (Cartwright 2005). Coon's theory of split ergativity does not directly predict anything. Rather, a model is fit that retrodicts the data of interest based on the model we choose.

With all cases of retrodiction, problems emerge when we attempt to make inferences about our hypotheses (Nosek et al. 2018). While the practice of fishing for diagnostics to support one's own hypothesis or to falsify an opponent's is fairly common in linguistics, a more controversial position maintains that the process of fitting mod-

els through diagnostic fishing introduces a confound that limits our ability to make inferences about the validity of our hypothesis based on one or more models (Croft 2001, 2010, Haspelmath 2011, Haspelmath 2018, Haspelmath 2019).

In linguistics we should recognize that the selection of tests proceeds *in tandem* with model fitting, and that this is probably necessarily so. We have not developed a method for sampling diagnostics that reins in the possibility that the diagnostics were chosen *a posteriori* in such a way as to prejudge a desired hypothesis to be correct (or false, depending on our goals). If I am interested in testing M1, I go on a fishing expedition to find evidence for it. I do not sample randomly from a list of ready-made and easily-applicable diagnostics. In other words, the process of fitting a model to linguistic data plausibly has a type of "selection bias" that confounds inferences.

The biclausal / non-verbal predicate model of C-SUBJ construction appears to be supported by a number of unrelated facts; (i) the distribution of the third person plural clitic *=kan*; (ii) the distribution of reportative marking; (iii) the predicate subject ordering mirroring that of the nonverbal predicate construction; (iv) the possibility that some of the clause-type morphemes that appear in this construction can be analyzed as containing dependent clause marking (the same subject marker *=ʔi*). On the other hand other facts point in a different direction; (i) the form of the clause-type morphemes mostly cannot be analyzed as containing any subordinate marking; (ii) there is no evidence that a transitive verb in the C-SUBJ construction is treated as intransitive (like a nonverbal predicate construction) by same subject marking or any transitivity harmonizing morphemes. A more challenging question is how one teases out the possibility that the similarities between nonverbal predicate constructions and the C-SUBJ constructions are due to a diachronic rather than synchronic relationship between these clauses (Gildea 2008).

If all the diagnostics and all the facts that the model seemed to retrodict could be rallied in support of a single model then there would be no problem. In the case reviewed above, it is my view that the data in the aggregate support Coon's hypothesis applied to Chácobo. However, any method that aggregates evidence should also come with a sampling method that can be shown to not bias the outcome. Why should I dismiss the morphological evidence in this case and presume that there there are zero subordinate markers? Note that for other languages the morphological evidence is seen as crucial (e.g. in Basque, Chol and Mebêngôkre). Even if we did have a method that controlled for the confound of selection bias, we currently have no methodology for making inferences when the evidence points in different directions. In the absence of good methods for making inferences with aggregated data we tend to choose one model as the best and bury our uncertainty in assessing the hypothesis in the process. The problem with Newmeyer's claim is not that the categories and relations of the theory are formal and abstract, but that they are noisy and we have no method for distinguishing between the noise and the signal, except through cherry-picking our

data points. Our ability to severely test a hypothesis is therefore dampened by the fact that our methods are confirmationally lax, as they do not make any head-way in teasing out a problematic confound that emerges from the model-fitting process. I now turn to consider another hypothesis that has received much more widespread attention and relate it to this point.

4 Fishing for (dis)confirmation: level-bound lexicalism

Lexicalism is an umbrella term that refers to a host of frameworks and hypotheses that assume that much linguistic phenomena fall out of independently motivated lexical entries, diminishing the descriptive and explanatory scope of syntactic structures or operations (constructional schemas, phrase structure rules, and transformations) that abstract away from lexical items (Müller 2006, Müller & Wechsler 2014). It is also associated with the idea that there is a word-formation component that syntactic rules and operations are blind to or more generally the idea that word-formation rules involve different operations than those involved in building syntactic representations (Bresnan & Mchombo 1995, Williams 2007, Newmeyer 2009). The latter sense of lexicalism could be called 'lexical integrity' (Desjardins 'in prep'), but to make it clear that I am considering lexicalist hypotheses that are contingent on identifying words, I will refer to the latter sense as "level-bound".

Bruening (2018) argues that level-bound lexicalism is false. According to Bruening, level-bound lexicalism implies that word-formation has the following properties, not shared by syntax: (i) parts of words cannot be phrases; parts of words cannot be elided or coordinated; (iii) parts of words cannot be focused; (iv) parts of words have strictly local scope; (v) parts of words do not scope over adjuncts; (vi) if parts of words are ordered head-final, parts of phrases are ordered head-initial and vice-versa; (vii) parts of words are not productive; (viii) part-of-word combinations are not compositional. He argues that either in each case, words do have the property and/or that phrase structure also has it.

Bruening does not consider how one goes about identifying 'words' in the first place. There is no discussion of the possibility that wordhood constituents might motivate different level-bound lexicalist models of the data Bruening is concerned with. In other words, Bruening assumes the situation in 1 for each of the hypotheses about word-formation he considers. His inferences are, therefore, correct insofar as it is actually true that a single model of the language data emerges from the application of wordhood diagnostics. But he does not show this – we are presented with 'words' as given primary data.

It is easy to argue against Bruening's claim that lexicalism has been tested to false-

hood, by mapping the word/phrase distinction to the data he considers differently than he does. As an illustrative case consider compound-internal phrases. Phrases can occur inside x-noun combinations where the noun is the head and x bears an attributive function. In English and some other traditions such combinations are considered to be 'compounds' in the case that x is a noun but are also extended to cases where x is some other string. Compounds are traditionally considered to be 'words', but actually display a host of phrase-like properties as well (Haspelmath & Sims 2010). Assuming that any phrase that appears left-adjacent to a noun and right-adjacent to an adjective phrase is occupying the attributive compound position in English, Bruening argues that examples like that in 4 disconfirm point (i) of level-bound lexicalism.

(4) The old$_{\text{AP}}$ [[dog$_i$-ate-my-homework]$_{\text{CP}}$ excuse]$_{\text{word}}$ won't work because I know you don't have one$_i$.

Davis & Koenig (2020) suggest that such constructions, in fact, involve complement adjective phrases and, thus, the analysis would be as follows, providing no evidence for word-internal phrases and no falsification of lexicalism.

(5) The old$_{\text{AP}}$ [[[dog$_i$-ate-my-homework]$_{\text{CP}}$]$_{\text{AP}}$ excuse]$_{\text{NP}}$ won't work because I know you don't have one$_i$.

Even if one rejects the analysis above, one could also claim that the candidate compound construction is not a compound, but a phrase-level noun-noun combination (Giegerich 2011). Cross-linguistic considerations seem to force us to assume that [NP-X0]XP combinations are needed anyway (see (Lieber & Štekauer 2011) and papers cited therein). Unless a set of convergent criterial properties for wordhood are shown to correlate with the candidate 'compound' in each case (see (Bresnan & Mchombo 1995:n18)), it is not clear why the relevant constructions must be regarded as 'words' with embedded phrases rather than phrases with embedded phrases (Bauer 2017).

Perhaps compounds are problematic and we might wonder whether 'words' are problematic more generally. Haspelmath (2011) has suggested that level-bound lexicalism is problematic because there are no jointly necessary and sufficient criteria for identifying words. Davis & Koenig (2020) seem to shrug off the problem when they state "Haspelmath's point is merely that what is a word is cross-linguistically unclear". Matthews (2002) suggests that wordhood diagnostics might "tend to coincide", even in the face of some unclear examples. However, the problem here is that as the criteria tend not to coincide, level-bound lexicalist hypotheses also tend to become less testable in practice. In other words, the researcher really needs to show that criteria coincide before inferences are made about the hypothesis under consideration (either in favor or against).

Bruening's article does not show that lexicalism is false. He shows that when the baseline categories presupposed by a hypothesis are noisy it is easy to come up with a battery of apparent counterexamples. It is also easy to come up with a battery of confirming examples under such a situation as well.

To illustrate the point I will briefly review the main results of a study on wordhood in Chácobo. Let us say I am interested in testing lexical integrity. 'Words' in Chácobo are encapsulated such that syntactically placed elements cannot occur inside them. First, I have to identify which spans of structure correspond to the word – where syntax ends and where morphology starts. An initial starting point would be to assume a notion of word that makes it as close as possible structurally to that constituent which is assumed to be a word in closely-related languages. The morphosyntactic word in Pano languages is typically described as necessarily containing a single verb root with one or more obligatory inflectional elements (depending on the language), which encode tense, aspect or clause-type (Fleck 2003, Valenzuela 2003, Zariquiey 2018, Neely 2019). Pano languages are also described as '(poly)synthetic' because there are a large number of bound morphemes that can occur between the verb root and the obligatorily inflectional element which are assumed on all accounts to be 'affixes'. The languages all have a single (productive) prefix position which encodes body-part locations with some metaphorical extensions. In all cases the verbal word is defined as a span of structure that extends from the body part prefix to the final obligatory inflectional elements. The model follows the assumption that derivation precedes inflection and thus elements between the root and the final inflection are understood as derivational. In 6, an example is provided from Chácobo where the word is underlined. *ta-* 'foot', *-ɨ* 'intransitive' *tɨkɨ* 'again' are considered derivational affixes and the final obligatory clause-type morpheme is considered an inflectional suffix, conforming to a model of grammar where inflection follows derivation (e.g. Anderson (1985)). Cognate constructions in other Pano languages are not hard to find.

(6) *ina hoʂo ta niʂ ɨ tɨkɨ́ kɨ*
dog white foot tie INTR AGAIN DECL:PAST
'The white dog got its feet tied up again.'

However, in contrast to other Pano languages, the verbal "word" in Chácobo can be interrupted by a full subject noun phrase *ina hoʂo* 'white dog'.

(7) *ta niʂ ɨ ina hoʂo tɨkɨ kɨ*
foot tie INTR dog white AGAIN DECL:PAST
'The white dog got its feet tied up again.'

Lexical integrity can be saved by identifying a smaller constituent as the verbal word. In order to motivate such an analysis we would have to find independent evidence via wordhood diagnostics. Relying on the fact that the constituent fails noninterruptability alone would be circular. Such a constituent can be found if we consider

reduplication facts in Chácobo. In Chácobo there happens to be a domain of structure that corresponds precisely with the domain that cannot be interrupted by any free forms. The basic facts are illustrated in 8 and 9.

(8) *tá niꞩ *(í) *(tá) niꞩ ɨ kɨ ínaka*
foot tie ITR foot tie ITR DECL:ANT dog
'The dog had been getting his leg(s) tied up'

(9) *ta niꞩ *(mís) ta niꞩ mís kɨ*
foot tie (ANTIPASS) foot tie ANTIPASS DECL:PAST
'S/he always ties people up.'

Elements outside of this domain either do not copy under reduplication, as in =kɨ 'declarative (past)' or copy optionally as in 10 below.

(10) *niꞩ-a (yáma) niꞩ-a yáma kɨ*
tie-TR (NEG) tie-TR NEG DECL:PAST
'S/he was (not) tying and tying.'

Thus, if we exclude the clause-type morpheme and much of the other material whose cognate forms would typically be considered affixes, the word can be identified according to a four-slot template. The span of structure identified by what is obligatorily targeted under reduplication is four positions long, from the prefix to the antipassive marker (prefix-root-transitivity marker-(antipassive)). These 4 positions are occupied by morphemes that are all bound, cannot be variably ordered, cannot be interrupted by any candidate syntactic elements, display deviations from biuniqueness, and occupy a domain of minimal bimoraicity. If we model Chácobo such that the 'word' corresponds to this span of structure, lexical integrity is confirmed.

This line of argument makes level-bound lexicalism a fairly weak hypothesis. In principle, we can simply redefine words' boundaries until we find one congenial to the hypothesis. If we ever run into too much trouble, we can simply declare that the language in question is isolating and the problem disappears. On the other hand, we have found convergent pieces of evidence for the word in this case. Lexical integrity appears to be at least weakly corroborated.

A more serious problem emerges when we consider what alternative models of the Chácobo word could be supported by attempting to rally more diagnostics (Tallman 2018a, 2020). Tallman (2020) develops a sentence-level template designed to display and code constituency test results. One challenging aspect of applying wordhood tests is how open-ended and vague they are. An obvious case is the test of noninterruptability (which both serves as a prediction of lexical integrity, but is also claimed to be a diagnostic of wordhood). It claims that 'words' cannot be interrupted, but does not specify by what (other words?). The test of fixedness gives us different results depending on whether we assume that morphemes must always occur in a rigid order or that

they can vary locally and produce differences of scope (?). Even the language-internal reduplication diagnostic mentioned above is ambiguous. Why interpret the domain that *must* be copied as the word domain, rather than the domain of structure that can optionally be copied? One strategy that could be used to avoid biasing our results in one direction would be to report all interpretations of our diagnostics. Tallman (2020) presents a total of 16 morphosyntactic diagnostics using this method. When they are displayed over the sentence-level template as in Figure 3, the support for our four-slot word model looks much more tenuous.

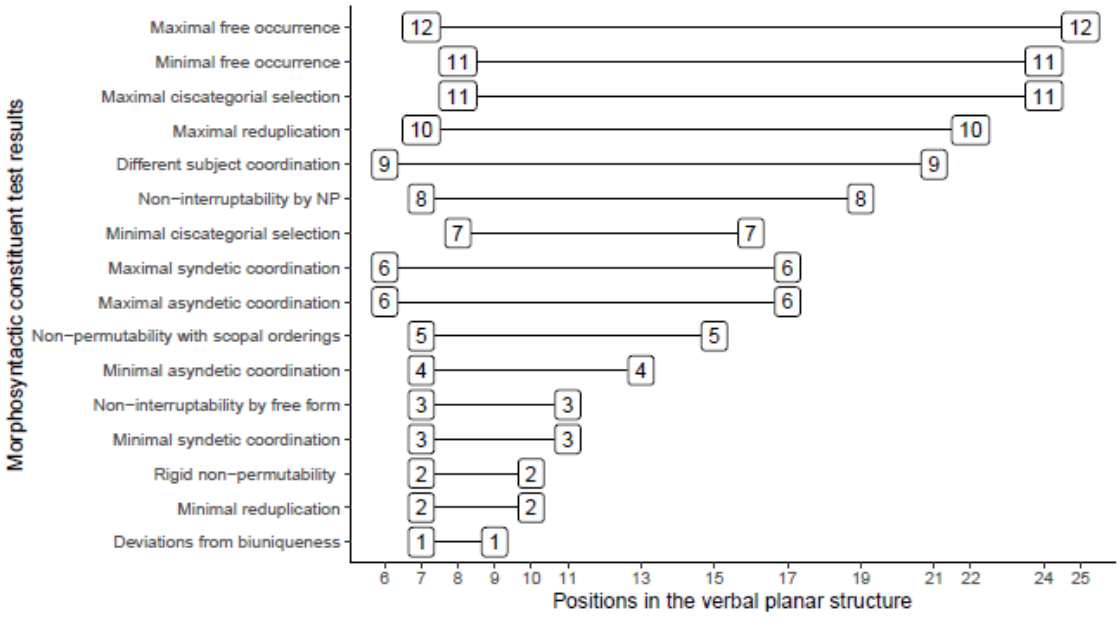

Figure 3: Wordhood tests over a template for the verbal predicate construction in Chácobo. The template begins at number 6 because, elements prior to this position are not identified by any constituency test that also overlaps with the verb

There are a *lot* of issues that arise at this point. Perhaps some of the diagnostics should be dismissed as unreliable. Perhaps some of the diagnostics pick out constituents at a higher level of structure. It is not entirely clear how to engage with such questions language-internally in a way that does not prejudge the question we are interested in answering. The only way to engage with such questions systematically, without studying nonconventional aspects of language, would be to engage in a typological study of the diagnostics themselves – to see which diagnostics tend to converge cross-linguistically, and which diagnostic results can be dismissed as noise.

What of our original question about lexical integrity? My point here is not that lexical integrity is false. Bruening did not present evidence that level-bound lexicalism

is false, because he never presented a version of the theory that could ever be tested. The relationship between the formal category "word" and the diagnostics or principles that could be rallied to its defense is too noisy for the anecdotal literary style of argumentation typically employed by linguists to draw any reliable inferences. While linguistic argumentation implicitly relies on aggregation of data points (argument 1, argument 2, argument 3, etc.) we have no method of drawing inferences if the data point in more than one direction, beyond impressionistic majority rules judgements, *ab auctoritate* arguments about the superiority of some diagnostics over others, or simply by unconsciously suppressing problematic data throughout the fitting process (selection bias). Level-bound lexicalist hypotheses are falsifiable in principle. This is perhaps why Davis & Koenig (2020) shrug off Haspelmath's criticism of the notion of 'word'. However, there are too many degrees of freedom in the interpretation of the baseline theoretical categories for us to easily test the hypothesis in practice. We might corroborate or falsify the theory 'by accident' by virtue of our analysis. A more rigorous approach will involve reconceptualizing the baseline wordhood diagnostics themselves as comparative concepts. Such a project is necessarily typological as there is no guarantee that a diagnostic developed for one language will easily translate to the next (Tallman et al. in prep).

5 Final comments on 'Analysis'

Haspelmath summarizes much of current linguistic practice in a way that is either jarringly or pleasantly frank.

> ... It is quite common for research articles to consist of two parts: One part lays out the phenomena in a way that is generally comprehensible to any linguist, and another part (typically called "analysis") describes the phenomena a second time, using the highly technical metalanguage of current mainstream generative grammar ... (Haspelmath 2021:12)

As MH describes it, the 'analysis' appears as some type of notational ornamentation glossed over the descriptive facts meant to symbolically ratify the researcher's belief in Universal Grammar. Indeed, if, as I have argued for the prosodic hierarchy hypothesis (Tallman 2021b), the hypothesis of interest is a tautology, it is hard to imagine that an 'analysis' couched in its metalanguage could amount to anything else but a redescription of the data in different terms. D'Alessandro takes issue with this caricature of linguistic methodology on the grounds that the analysis section is where one shows that the theory makes predictions. I do not disagree with this in principle, but would emphasize that at least a few conditions should hold for us to make that determination. First, it has to be shown that the theory in question is not a self-sealing

tautology. Can it be shown that the theory is testable in principle given its relationship to the data it purports to predict or can every conceivable counterexample be analyzed away through remapping the theoretical categories or evoking some extra non-informative auxiliary hypothesis? The second condition is that the researcher be able to show that the data retrodicted to fit and test the theory have been gathered based on a methodology that does not prejudge the correctness or falseness of the theory. In cases where the theory presupposes categories and relations that display a multidimensional but nondeterministic relationship with diagnostics and predictions, I fail to see how this can be done without a broad typological study that clarifies how theoretical models that the hypothesis is couched in will map to novel data in cases where the baseline categories of the theory are 'abstract' (and thus nondeterministic).

I think incorporating some of the methodological insights of distributional typology into generative research would rein in some of the problems. I perhaps disagree with MH in that I am less sure that selection bias (diagnostic fishing) is an insurmountable obstacle to research based on hypotheses developed under the auspices of generative grammar. For instance, while ostensibly adopting a generative perspective, Guardiano & Longobardi (2017) presents a methodology where parameters are systematically coded, which if driven to its logical conclusion could rein in the problem of selection bias in hypothesis testing. Furthermore, typological work could potentially provide support for particular generative hypotheses. In such cases it is not a matter of assuming natural-kinds, but rather adopting a methodology that is agnostic about them. In my view, Bickel et al. (2009) did test the prosodic hierarchy hypothesis (or a version of it) with a rigorous method that did not presuppose its correctness or falseness.

While MH suggests that comparative concepts are an epistemic tool linked inextricably to the Boasian/Greenbergian programme, I am less sure that the distinction between natural-kinds programme and the Boasian/Greenbergian programme is so stark in principle. I think the relationship between ontological positions and methodology that MH fleshes out, might be normative rather than necessary. For one, there is a certain affinity between diagnostics and comparative concepts. In actual practice, comparative concepts are partially based on diagnostics. For instance, Bickel (2010)'s study of clause-linkage reconceptualizes all of the diagnostics used to distinguish between coordination, subordination and cosubordination as typological variables and seeks to assess the correlations between the typological variables directly arriving at the conclusion that the abstract categories are partially supported as prototypes. (Good 2016) repurposes HPSG nested feature structures for a multivariate typological study of templates. This suggests that minimally some of the formalisms developed in the natural-kinds programme will have an epistemic value in coding and modelling typological variation (even if we reject the idea that they allow us to directly commune with the universal).

Secondly, in certain limiting research contexts, the notion of a 'diagnostic' might collapse into that of a comparative concept. As natural-kinds linguists finds themselves in a research context where the categories and relations of their theory can be identified by necessary, sufficient, and cross-linguistically applicable diagnostics, the distinction between tests and comparative concepts tends to zero. For instance, a comparative concept of 'tense' might be nearly identical with what is supposed to head a T(ense)P(hrase). This affinity between diagnostics and comparative concepts makes it possible (and I would argue fruitful) for researchers to shift approaches and present the same data from different perspectives. Fitting a model on a language can sharpen one's research questions in the field (Rice 2006). Natural-kind theories and their impulsion to lead the researcher to fish for diagnostics can have an important epistemic value in description and documentation. You learn things from fishing. I often organize elicitation sessions around whether I can meaningfully test some hypothesis implied in the 'analysis' section because it made me realize that there was a gap in my understanding of the language I am studying. There is no harm in this as long as the fieldworker is not so spell-bound by the formalism that they lose track of the fact that fitting a model does not necessarily translate into a tested hypothesis.

References

Abels, Klaus & Ad Neeleman. 2012. Linear Asymetries and the LCA. *Syntax* 15. 25–74.

Anderson, Stephen. 1985. Inflectional morphology. In T. Shopen (ed.), *Language typology and syntactic description: Grammatical categories and the lexicon*, 150–201. CUP.

Bauer, Laurie. 2017. *Compounds and Compounding*. Cambridge: CUP.

Bickel, Balthasar. 2010. Capturing particulars and universals in clause-linkage. In I. Brill (ed.), *Clause Linking and Clause Hierarchy*, 51–104. Amsterdam: John Benjamins.

Bickel, Balthasar, Kristine A. Hildebrandt & René Schiering. 2009. The distribution of phonological word domains: A probabilistic typology. In J. Grijzenhout & B. Kabak (eds.), *Phonological Domains: Universals and Deviations*, 47–75. De Gruyter Mouton.

Bresnan, Joan & Sam A. Mchombo. 1995. The lexical integrity principle: Evidence from Bantu. *Natural Language and Linguistic Theory* 13. 181–254.

Bruening, Benjamin. 2018. The lexicalist hypothesis: Both wrong and superfluous. *Language* 94.

Cartwright, Nancy. 2005. *The Dappled World: A Study of the Boundaries of Science*. Cambridge: CUP.

Chomsky, Noam. 1957. *Syntactic structures*. The Hague: Mouton.

Chomsky, Noam. 1972. *Language and mind*. New York: Harcourt Brace Jovanovich.

Chomsky, Noam. 1985. *Knowledge of Language: Its Nature, Origins and Use*. New York: Praeger.

Coon, Jessica. 2013. *Aspects of Split Ergativity*. Oxford: OUP.

Croft, William. 2001. *Radical Construction grammar*. Oxford: OUP.

Croft, William. 2010. Ten unwarranted assumptions in syntactic argumentation. In Boye & Engberg (eds.), *Language Usage and Language Structure*, 313–350.

D'Alessandro, Roberta. ???? Not everything is a theory.

Davis, Anthony & Jean-Pierre Koenig. 2020. The nature and role of the lexicon in HPSG. In R.D. Borsley et al. (ed.), *Head-driven Phrase Structure Grammar*, 51–104. Berlin: Language Sciences Press.

Desjardins, Jared. 'in prep'. A Cross-theoretical and Cross-linguistic survey of lexical integrity and the nature of the morphology-syntax interface.

Fleck, David. 2003. *A Grammar of Matses*: Rice University PhD dissertation.

Galison, Peter. 1987. *How Experiments End.* Chicago and London: The University of Chicago Press.

Giegerich, Heinz. 2011. Compounding and Lexicalism. In R. Lieber & P. Štekauer (eds.), *Oxford Handbook of Compounding*, OUP.

Gildea, Spike. 2008. Explaining similarities between main clauses and nominalized phrases. *Amerindia* 32. 57–75.

Good, Jeff. 2016. *The Linguistic Typology of Templates.* Cambridge: CUP.

Guardiano, Cristina & Giuseppe Longobardi. 2017. Parametric Theory and Parametric Comparison. In I. Roberts (ed.), *The Oxford Handbook of Universal Grammar*, 377–400. OUP.

Hacking, Ian. 1983. *Representing and intervening: introductory topics in the philosophy of science.* Cambridge: CUP.

Haspelmath, Martin. 2011. The indeterminacy of word segmentation and the nature of the morphology and syntax. *Folia Linguistica* 45.

Haspelmath, Martin. 2018. The moving parts and the fixed parts of our theories: Why functional-adaptive explanations are more testable. `https://dlc.hypotheses.org/1167`.

Haspelmath, Martin. 2019. Ergativity and depth of analysis. *Rhema* 4. 108–130.

Haspelmath, Martin. 2021. General linguistics must be based on universals (or nonconventional aspects of language. *Theoretical Linguistics* .

Haspelmath, Martin & Andrea Sims. 2010. *Understanding Morphology.* Routledge.

Himmelmann, Nikolaus P. 1998. Documentary and descriptive linguistics. *Linguistics* 6. 161–195.

Himmelmann, Nikolaus P. 2012. Linguistic Data Types and the Interface between Language Documentation and Description. *Language Documentation and Conservation* 6. 187–207.

Lieber, Rochelle & Pavol Štekauer. 2011. Introduction: Status and Definition of Compounding. In R. Lieber & P. Štekauer (eds.), *The Oxford Handbook of Compounding*, OUP.

Ludlow, Peter. 2011. *The philosophy of generative linguistics.* Oxford: OUP.

MacWhinney, Brian, Andrej Malchukov & Edith Moravcsik (eds.). 2015. *Competing motivations in grammar and usage.* Oxford: OUP.

Matthews, Peter. 2002. What can we conclude? In R.M.W. Dixon & Alexandra Aikhenvald (eds.), *Word: A Cross-Linguistic Typology*, 266–281. OUP.

Mayo, Deborah G. 1996. *Error and the Growth of Experimental Knowledge.* Chiacgo: The Uni-

versity of Chicago Press.

Mayo, Deborah G. 2018. *Statistical Inference as Severe Testing: How to get beyond the statistics wars*. Cambridge: CUP.

Mayo, Deborah G. & Aris Spanos. 2011. Error Statistics. In P. S. Bandyopadhyay & M.R. Forster (eds.), *Handbook of the Philosophy of Science. Volume 7: Philosophy of Science*, 153–198. Elsevier.

McElreath, Richard. 2020. *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. Oxon: CRC Press.

Mendívil-Giró, José-Luis. 2019. How much data does linguistic theory need? on the tolerance principle of linguistic theorizing. *Frontiers in Communication, Language Sciences* 3(62). 1–6.

Moravcsik, Edith A. 2007. What is universal about typology? *Linguistic Typology* 11(1). 27–41.

Müller, Stefan. 2006. Phrasal and lexical constructions. *Language* 82. 850–883.

Müller, Stefan & Stephen Wechsler. 2014. Lexical approaches to argument structure. *Theoretical Linguistics* 40. 1–76.

Neely, Kelsey. 2019. *The Linguistic Expression of Affective Stance in Yaminawa (Pano, Peru)*: Berkeley PhD dissertation.

Newmeyer, Frederick J. 1998. *Language Form and Language Function Language*. Cambridge: MIT Press.

Newmeyer, Frederick J. 2008. Universals in syntax. *The Linguistic Review* 25. 35–82.

Newmeyer, Frederick J. 2009. Current challenges to the lexicalist hypothesis: An overview and critique. In W.D. Lewis et al. (ed.), *Time and again: Theoretical perspectives on formal linguistics in honor of D. Terence Langendoen*, 91–117. Amsterdam: John Benjamins.

Nosek, Brian A., Charles R. Ebersole, Alexander C. DeHaven & Daviod T. Mellor. 2018. The preregistration revolution. *PNAS* 115(11). 2600–2606.

Popper, Karl. 1959. *The Logic of Scientific Discovery*. Abingdon-on-Thames: Routledge.

Popper, Karl. 1963. *Conjectures and Refutations*. New York: Routledge.

Rice, Karen. 2006. Let the language tell its story? The role of linguistic theory in writing grammars. In F. Ameka et al. (ed.), *Catching Language: The Standing Challenge of Grammar Writing*, 235–268. OUP.

Salanova, Andrés & Adam J.R. Tallman. 2020. Nominalizations, case domains, and restructuring in two Amazonian languages. In A. Alexiadou & H. Borer (eds.), *Nominalization: 50 years on from Chomsky's Remarks*, 363–389. OUP.

Seuren, Pieter A. 2004. The importance of being modular. *Journal of Linguistics* 40. 593–635.

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

Tallman, Adam J.R. 2018b. VS constituent order and split ergativity in Chácobo (Pano).

Tallman, Adam J.R. 2020. Constituency and Coincidence. *Studies in Language* doi:https://doi.org/10.1075/sl.19025.tal.

Tallman, Adam J.R. 2021a. Documentación y descripción lingüística. In G. Camacho-Rios & G. Gallinate (eds.), *Introducción a la Lingüística en el contexto Boliviano*, .

Tallman, Adam J.R. 2021b. Review of Caroline, Féry, Intonation and prosodic structure. *Journal*

of Linguistics 1–6.

Tallman, Adam J.R., Sandra Auderset & Hiroto Uchihara (eds.). in prep. *Constituency and Convergence in the Americas*. Berlin: Language Sciences Press.

Valenzuela, Pilar M. 2003. *Transitivity in Shipibo-Konibo in Grammar*: University of Oregon PhD dissertation.

Williams, Edwin. 2007. Dumping lexicalism. In G. Ramchand & C. Reiss (eds.), *The Oxford handbook of linguistic interfaces*, 353–381.

Woodbury, Anthony C. 2007. On thick translation in linguistic documentation. *Language Documentation and Description* 4. 120–135.

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