Final draft, September 2017; to appear in Ilja A. Seržant and Alena Witzlack-Makarevich (eds.), The Diachronic Typology of Differential Argument Marking. Berlin: Language Science Press 2018.

Reassessing scale effects on differential case marking: Methodological, conceptual and theoretical issues in the quest for a universal

KARSTEN SCHMIDTKE-BODE AND NATALIA LEVSHINA
Leipzig University

ABSTRACT

It is widely believed that when differential case marking depends on the referential properties of the NP in question, it is governed by a well-defined hierarchy or scale of referential categories, and that the resulting systematicity is one of the most robust generalizations in linguistic typology. This view has recently been called into question, with Sinnemäki (2014) and especially Bickel et al. (2015a) claiming that there is now firm typological evidence against such universal scale effects. Since these papers are based on the largest world-wide databases compiled so far, their results are likely to be taken as the current state of the field. In the present paper, we re-examine Bickel et al.'s data from a different perspective and re-evaluate their negative conclusions: First, we complement their analysis in terms of diachronic "family biases" by a more direct inspection of the raw data and an alternative statistical model, both of which afford a clearer understanding of where and how exactly the predicted scale effects are violated. Proceeding from this, we argue for the existence of universal scale effects on case marking, and we embed this argument in a more general discussion on current methodological, conceptual and theoretical issues in postulating these effects.

1. Introduction

An important discovery of typological research is that differential argument marking (DAM) is systematically related to what we may call the "referential properties" of the argument in question. As outlined and exemplified in the introductory article to the present volume, these comprise animacy, definiteness, specificity, nominality, person, kinship and discourse-pragmatic prominence (e.g. topicality). In comparative research since Silverstein (1976), it has been argued that contrasts in referential properties (e.g. animate-inanimate) can be arranged into an implicational hierarchy or scale that predicts asymmetries in argument marking.[1] Two versions of this referential scale are given in (1)[2], and their classic predictions for case marking follow in (2):

(1) a. "extended animacy hierarchy" (Croft 2003: 130)
1,2 Pro > 3 Pro > proper noun > human common noun > non-human animate common noun > inanimate common noun

[1] The term "asymmetric" is adopted from de Hoop & Malchukov (2008) and refers to the kind of differential argument marking in which an overt case exponent alternates with zero marking. We will return to the notion of "markedness" (and a different way of operationalizing asymmetric case marking) in §2.1 below.

[2] Further incarnations of the same idea include, for example, Comrie's (1981) "animacy hierarchy", DeLancey's (1981) "empathy hierarchy", Bickel's (1999) "indexability hierarchy" and Shibatani's (2006) "relevance hierarchy".

b. “individuation scale” (Lazard 1998: 220)
pronoun > human definite > human indefinite/nonhuman definite > nonhuman indefinite > indefinite non-specific

(2) a. If a P argument is unmarked for case for a given referential category in (1a) or (1b), it will also be unmarked for case for all categories to the right.
b. If an A argument is unmarked for case for a given referential category in (1a) or (1b), it will also be unmarked for case for all categories to the left.

The generalizations in (2) have also been referred to as “scale effects” (Bickel et al. 2015a) or “referential effects” (e.g. van Lier 2012) on the distribution of overt case marking. With the compilation of large cross-linguistic databases, it has recently become possible to subject these generalizations to thorough empirical evaluation. And so far, the resulting assessments have been strikingly negative: Thus both Sinnemäki (2014) and Bickel et al. (2015a) identify some clear areal signatures of DAM in case marking, so that the effect might be “first and foremost a pattern prone to diffusion” (Bickel et al. 2015a: 40). When controlling for such areal dependencies, Bickel and his collaborators have argued that there is no evidence for universal effects of the person scale on indexation (Bickel et al. 2015b; Witzlack-Makarevich et al. 2016) and that there is, in fact, direct “evidence *against* universal effects of referential scales on case alignment” (cf. the title of Bickel et al. 2015a).

Importantly in the context of the present volume, Bickel et al.’s assessment is based on the estimation of diachronic “family biases” from synchronic data (Bickel 2011; 2013). In a nutshell, the argument is that when language families produce new generations of offspring, they do not systematically develop into the directions predicted by (1) and (2): Some families are internally diverse with regard to these predictions, and among those that are significantly biased towards certain scale effects on case marking, there is always a substantial number of families that are biased in the opposite direction. In other words, Bickel et al.’s (2015a) finding is that the predictions in (2) are violated too often to qualify as a principle that universally guides the diachronic development of language families.[3] It is *not* our purpose in this paper to take issue with this specific method. However, given that Bickel et al.’s (2015a) conclusion challenges of one of the most prominent and widely cited generalizations in typology since the 1970s, we would like to discuss and expand the empirical assessment of scale effects on case marking.

Specifically, we intend to do three things: Firstly, in the absence of actual diachronic data for most of the world’s language families, the most direct evidence for typological patterns we have inevitably lies in the synchronic data themselves. Therefore, we would first like to be clear about the synchronic picture in its full extent. To this end, we begin (in §2) by complementing Bickel et al.’s analysis by a more direct inspection of the raw data, which lays bare where and how exactly the predicted scale effects are violated.[4] Secondly, given what is at stake, we feel that Bickel et al.’s (2015a) assessment should be cross-validated by other contemporary statistical procedures for typological research,

[3] We provide some more information on the Family Bias Method in the Appendix.

[4] We would like to thank Balthasar Bickel and his collaborators for making their entire data and their algorithms publicly available (cf. also Bickel et al. 2017).

such as those proposed by Cysouw (2010) and Jaeger et al. (2011). We show (in §3) that these mixed-effects regression methods yield robust synchronic evidence for the predicted scale effects on case marking. In view of this result, a more general discussion is in order about methodological, conceptual and theoretical issues in comparative research: To what extent are purely synchronic analyses justified? What does it take for an effect to be called "universal", and what is the role of the referential scale in explaining differential case marking? In discussing these matters, we question some specific assumptions made by Sinnemäki (2014) and Bickel et al. (2015a), but also certain interpretations of the referential scale in formal-generative approaches to differential case marking. In §5, finally, we conclude the paper by summarizing our major points. Our study comes with several supplementary materials (SM1–SM4), which can be downloaded from the authors' websites[5], as well as an Appendix at the very end of the paper.

2. Dissecting the data

2.1 Coding procedure

Bickel, Witzlack-Makarevich and Zakharko (2015a) [henceforth BWZ, for the sake of economy] examine a sample of 435 languages for referential effects on case marking, under which they subsume all kinds of morphology on verbal arguments, regardless of its fusion type (i.e. including adpositional flagging and non-concatenative signals of case) and its host (i.e. including markers that are limited to elements of the NP other than the noun itself, such as case on German determiners). The classic typological predictions with regard to such case exponents were given in (2) above, but we need to refine the notion of *markedness* at this point. The statements in (2) imply a difference between zero and overt case marking, i.e. a contrast in coding material (as in Comrie 1981 or Croft 2003). BWZ, by contrast, frame the predictions in terms of more abstract grammatical relations (as in Silverstein 1976): Low-ranking P arguments (and high-ranking A arguments) are predicted to preferably establish an unmarked grammatical relation, while high-ranking P arguments (and low-ranking A arguments) are predicted to map onto a marked grammatical relation. BWZ take an unmarked grammatical relation to be an alignment set that also includes other syntactic functions beside the one at issue, notably the S role of intransitive clauses: For example, a case formative that applies to (and hence aligns) S and P defines an {S=P} set, while a marker that does not distinguish S, A and P defines a yet more general {S=A=P} alignment set. On this view, case formatives that exclusively target {P} or {A} define very narrow, thus more specific and hence *structurally marked*, sets.

The crucial question, then, is whether P arguments with higher referential prominence (and A arguments with lower prominence) tend to occur in such marked alignment sets. We can illustrate this on the basis of case marking in Chantyal (Sino-Tibetan, Bodic: Nepal), also discussed by BWZ as a representative example of their coding procedure: Speakers of Chantyal consistently mark A arguments by Ergative case and consistently code S by a zero Absolutive. P arguments are split in such a way that pronouns and human NPs always receive overt Dative case, while non-human NPs typically go in the unmarked Absolutive, just like S. However, the marking for non-

[5] Cf. <http://www.kschmidtkebode.de/publications.html> and <http://www.natalialevshina.com/publications.html>.

human NPs actually depends on the degree of empathy felt towards that entity[6], so that the precise point at which the referential scale is cut off is not easy to determine. At any rate, though, it is clear that the higher-ranking P arguments define a narrow, marked alignment set {P}, while the lower-ranking P arguments are mapped onto a more general alignment set {S=P}, and not the other way around. A arguments consistently define a narrow set {A}, i.e. they are not split to begin with.

In Table 1 below, the facts about Chantyal are represented in BWZ's coding format:

Table 1. Coding in Bickel et al. (2015a)

Language	Family	Macro continent	Referential condition	Sub-systems	A	P	Alignment
Chantyal	Sino-Tibetan	Eurasia	N-high	NA	marked	marked	S|A|P
Chantyal	Sino-Tibetan	Eurasia	N-low	NA	marked	unmarked	S=P|A
Chantyal	Sino-Tibetan	Eurasia	Pro	NA	marked	marked	S|A|P

Table 1 displays the three referential conditions that are relevant to case marking in Chantyal, summarizes the alignment pattern in each condition and specifies, for both A and P, whether they establish a marked or an unmarked alignment set in the given referential condition.[7] The contrast between N-high and N-low captures the above-mentioned fact that a more specific referential contrast (such as animate-inanimate) is difficult to establish.

Having clarified the basic coding procedure in BWZ, we can now examine the data with regard to the case splits they contain. To this end, the following subsections will take a closer look at the effects of the most important referential dimensions coded in the data. In other words, we here first inspect the effects of individual referential properties that are included in hierarchies like (1), such as animacy or person, before we examine the combined effect of these dimensions in §3. Our major goal for the moment is thus to provide typologists with an idea of how numerous the exceptions to well-known referential subscales are and where these are located, i.e. which languages and stocks show which kinds of counterexamples. Although some of the relevant scales are also tested by BWZ, they do not provide the kind of "raw" information we present here, so the following data can be seen as complementary to the statistical analysis offered by BWZ.

[6] In reference to animals, for example, one can contrast 'I killed the chicken-Ø' with 'I cut the chicken-DAT [so that it bled]', cf. Noonan (2003).

[7] The column "Subsystem" does not apply to Chantyal and is hence coded as "not applicable (NA)". In other languages, it captures situations in which the case-marking system is sensitive to other structural factors, such as the difference between main and dependent clauses, periphrastic and synthetic verb forms, etc. Each of these conditions is then evaluated separately with regard to whether case marking also interacts with referential properties of the NP and which alignment sets result. The overall number of case-marking (sub)systems (N = 462) is thus somewhat higher than the number of languages in BWZ's sample (N = 435). Additionally, it should also be noted that BWZ concentrate on what they call "default verb classes" in their paper, disregarding, for instance, the case marking and alignment of experiencer NPs; in other words, their focus is on canonical transitive and intransitive clauses.

2.2 The global picture

The overall distribution of differential case marking is nicely laid out in BWZ (pp. 24–31), especially from an areal perspective. We will discuss the areal patterns in §4 and hence confine ourselves to the overview of the data given in Table 2:

Table 2. Overview of P- and A-splits in the data[8]

Macro-continent	Family	Split systems		Macro-continent	Family	Split systems	
		P	A			P	A
Africa	Adamawa-Ubangi	1		Eurasia	Austroasiatic	1	
	Benue-Congo	2			Dravidian	7	
	Chadic	2			Indo-European	31	15
	Cushitic	2			Kusunda	1	
	Indo-European	1			Mongolian	4	
	Kwa	1			Nakh-Daghestanian	1	3
	Omotic	2			Semitic	1	
	Semitic	1			Sino-Tibetan	13	8
	South Atlantic	1			Tungusic	1	
					Turkic	7	
					Uralic	3	
Americas	Arawakan	1		Sahul	Austronesian	1	
	Barbacoan	2			Awyu-Dumut	1	
	Haida	1			Kalam	1	
	Macro-Ge	1			Madang	1	
	Máku	1			Mangarayan	1	1
	Nadahup	1			Mirndi	1	
	Pano-Tacanan	1	1		Oksapmin	1	
	Pomoan	1			Pama-Nyungan	26	29
	Siouan	1			Tangkic		1
	Tarascan	1			Timor-Alor-Pantar	3	
	Tsimshianic		1				
	Tucánoan	4					
	Uto-Aztecan	3					
	Zuni	1					

2.3 High-low distinctions: Animacy, definiteness, topicality and the like

Perhaps the best-known kinds of case-marking splits are controlled by animacy (as in Armenian (Indo-European) or Gurung (Sino-Tibetan)), definiteness (as in Amharic (Semitic), Brahui (Dravidian) or Barasano (Tucánoan)), specificity (as in Persian (Indo-European) or Udihe (Tungusic)), kinship (e.g. Gumbaynggir (Pama-Nyungan) or uniqueness (proper versus common nouns (e.g. Gitksan (Tsimshianic))). Iemmolo (2010), among others, additionally points to the importance of topicality in inducing case splits. Overall, such contrasts are relevant to 83 cases (= 60%) of all P-splits and 7

[8] The counts presented here differ very slightly from BWZ's original ones: First, we break up BWZ's "Other" area into Africa and the Americas, in order not to lose this kind of information coded in the data; this holds for all analyses to follow in this paper. Second, BWZ's Table 5 on P-marking fails to list Máku, an isolate of South America. Conversely, our own analysis discards Hindi, for which the original coding was complicated by multiple subsystems with overlapping referential categories that did not allow a straightforward reanalysis.

cases (= 12%) of all A-splits. In BWZ's study, the dimensions of animacy, definiteness, specificity, kinship and uniqueness are recorded as such in the database, while discourse-pragmatic and other language-specific contrasts (cf. Chantyal above) are coded as a more general $\mathrm{N_{high}}$-$\mathrm{N_{low}}$ contrast. For purposes of statistical testing, all of these dimensions can be conflated into a $\mathrm{ProN_{high}} > \mathrm{ProN_{low}}$ scale.[9] In Table 3 below, we have compiled the data that are relevant to this scale and outline to what extent they are in keeping with the predictions for P- and A-marking, respectively. In this and all following tables of the same sort, "fit" indicates that a given system fits the predictions of the scale in question and "vio" indicates that it goes against it. "NA" captures all languages that do not exhibit the relevant split. The figures refer to the number of languages, while the figures in brackets indicate the number of distinct families from which these languages come. Violations are additionally underlined.

Table 3. Systems with 'high-low' splits in case marking

	P-marking				**A-marking**			
	Eurasia	Africa	Americas	Sahul	Eurasia	Africa	Americas	Sahul
fit	55 (9)	3 (3)	11 (7)	13 (4)	2 (2)	0 (0)	0 (0)	4 (2)
vio	0 (0)	1 (1)	0 (0)	0 (0)	0 (0)	0 (0)	1 (1)	0 (0)
NA	15 (4)	9 (7)	8 (6)	23 (6)	24 (3)	0 (0)	1 (1)	27 (2)

For P-marking, the splits virtually always work in the predicted direction, i.e. low-ranking nouns are structurally unmarked while high-ranking ones are marked. The only exception in the entire database is Sheko (Omotic), in which the distribution is reversed. In this language, we find an unspecified high-low contrast in the database; therefore, wherever the more concrete dimensions on animacy, definiteness and specificity are involved, there is no single counterexample to the predicted effects. For A-marking, the high-low distinction is much less relevant than for P-splits, so that the numbers are very small to begin with. Again, however, there is only a single exceptional language in the data: This is Gitksan (Tsimshianic: Americas), where common nouns are unmarked while proper nouns are marked, which is precisely the opposite of the predicted effect (under which specific marking, for example, should preferentially apply to lower-ranking A arguments). The effect from these referential dimensions is thus very robust cross-linguistically.

2.4 Nominality: Splits between pronouns and lexical NPs

A fundamental distinction on the hierarchies in (1), but also all of its further variants in the literature, is that between pronominal and lexical (i.e. full nominal) NPs. On all four macro-continents distinguished in Table 2, there are languages which reserve specific P-marking for pronouns and allocate their nouns to an unmarked alignment set (e.g. Yoruba, Gulf Arabic, Thayorre and many others). The opposite distribution would be expected for A-marking (e.g. Cashinahua or Yukulta). Overall, nominality governs 33 cases (= 24%) of differential P-marking and 17 cases (= 29%) of differential A-marking.

[9] The inclusion of pronouns on the scale is justified by the fact that the split between high and low referential prominence may also (or even exclusively) affect pronouns and not only nouns (e.g. in Central Pomo (Americas), where this applies to the third person pronouns).

Apart from such “clean splits” between the two categories, one may, however, also adopt a broader view of the markedness distributions of pronouns and nouns: If, for example, a language exhibits a split of its pronouns but not its nouns, the question is whether the nouns join the marked or the unmarked alignment set (for P, the prediction would be “unmarked” while it would be “marked” for A). We can thus distinguish four scenarios in the data, and we provide the relevant figures for each of them in turn.

Scenario A: A given case system makes a “clean” Pro-N distinction. As can be seen in Table 4, wherever this happens, there is not a single language going against the predicted direction of the split, neither for P- nor for A-marking:

Table 4. Systems with “clean” Pro-N splits in case marking

	P-marking				**A-marking**			
	Eurasia	Africa	Americas	Sahul	Eurasia	Africa	Americas	Sahul
fit	7 (3)	4 (3)	5 (4)	17 (5)	1 (1)	0 (0)	1 (1)	15 (2)
vio	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
NA	63 (10)	9 (6)	14 (10)	19 (6)	25 (3)	0 (0)	1 (1)	16 (2)

Scenario B: A given case system partitions nouns into marked and unmarked subsets but does not divide up pronouns. There is not a single example of P-marking in which the pronouns join the unmarked set (Table 5):

Table 5. Systems with splits in nouns but not in pronouns

	P-marking				**A-marking**			
	Eurasia	Africa	Americas	Sahul	Eurasia	Africa	Americas	Sahul
fit	46 (9)	4 (3)	8 (5)	10 (4)	1 (0)	0 (0)	0 (0)	2 (2)
vio	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	1 (1)	0 (0)
NA	24 (4)	9 (7)	11 (9)	26 (6)	25 (3)	0 (0)	1 (1)	29 (2)

As can be seen, there is one exceptional system for A-marking: This is Gitksan (Tsimshianic: Americas), in which common nouns are in an unmarked alignment set while proper nouns and pronouns are marked, i.e. we find exactly the opposite distribution from what is predicted for A-marking.

Scenario C: Where systems partition pronouns into marked and unmarked subsets but do not divide up nouns, the data look as follows (Table 6):

Table 6. Systems with splits in pronouns but not in nouns

	P-marking				**A-marking**			
	Eurasia	Africa	Americas	Sahul	Eurasia	Africa	Americas	Sahul
fit	7 (2)	4 (4)	1 (1)	7 (3)	18 (3)	0 (0)	0 (0)	12 (1)
vio	0 (0)	1 (1)	1 (1)	0 (0)	5 (1)	0 (0)	0 (0)	0 (0)
NA	63 (10)	8 (6)	17 (11)	29 (8)	3 (2)	0 (0)	2 (2)	19 (3)

For P-marking, the prediction is that nouns will join those pronouns that are found in an unmarked set, while the opposite is predicted for A-marking. Two languages violate this prediction for P-marking, namely Oromo (Cushitic) and Osage (Siouan). In Oromo, the unmarked set comprises all pronouns in the plural while singular pronouns and all nouns receive P-marking; in Osage, all nouns and third-person pronouns are marked while SAPs are unmarked. For A-marking, we find five aberrant systems, all from Indo-European and specifically Iranian (Roshani and participial clauses in Khufi, Yazgulyâmi, Tarom and Bartangi [10]); in all of them, nouns join an unmarked grammatical relationship.

Scenario D: Where languages partition both nouns and pronouns into marked and unmarked alignment sets, this inevitably results in discontinuities between Pro and N on the referential hierarchy and hence in a violation of the Pro>N subscale. The relevant languages are shown in Table 7:

Table 7. Languages with splits in both pronouns and nouns

	P-marking	**A-marking**
Eurasia	Albanian, German, Vafsi and non-participial clauses in 6 Iranian languages (Tarom, Shahrudi, Dimli, Kirmanjki, Kajali, Eshtehardi)	Qiang
Africa	---	---
Americas	Tsafiki, Tarascan, Máku, Central Pomo	---
Sahul	Kala Lagaw Ya, Gumbaynggir (both Pama-Nyungan)	Kala Lagaw Ya, Yandruwandha (both Pama-Nyungan)

The languages in Table 7 differ in how exactly they implement a Pro-N split, particularly with regard to the distribution of individual referential categories within the pronouns (e.g. singular versus plural pronouns (Albanian), 2PL versus all others (Vafsi), 1+2PL versus the rest (Eshtehardi/Dimli/Kirmanjki main clauses), etc.). Upon closer inspection, however, it turns out that these rather idiosyncratic splits are largely confined to the Iranian languages in Table 7; moreover, there are some principled regularities again: Firstly, in all of the above languages, the nouns are split in such a way that they conform to the predictions of the $N_{high}>N_{low}$ scale, and this applies to both P- and A-marking. (The only exception is German, where the split is according to different noun classes and not referential properties as such.) And secondly, pronouns and nouns may both be split according to the same principle, namely an animacy or definiteness contrast (e.g. Tsafiki, Tarascan, Máku and Central Pomo P-marking and Qiang A-marking); as a result, high-ranking (animate, definite) nouns and pronouns are split off from low-ranking (inanimate, indefinite) nouns and pronouns, thus creating a discontinuity between Pro and N on the referential scale. The observed diversity, therefore, primarily resides in the way that specific person-number categories are organized, and we will turn to these presently.

2.5 Person-conditioned splits

Differential case marking according to person-number constellations is attested for 29 systems (= 21%) for P-marking and 32 systems (= 54%) for A-marking. In the following, we examine person splits separately for singular and non-singular (dual,

[10] These Iranian languages are very closely related; in fact, Roshani, Khufi and Bartangi are sometimes considered dialects of the Shughni language. Similar remarks apply to the Iranian languages which follow in Table 7.

plural) number, in order to capture the empirical picture as precisely as possible. Table 8 shows which person splits are attested in the singular.

Table 8. Person splits in the singular († indicates the number of violating systems)

	P-marking				A-marking			
	Eurasia	Africa	Americas	Sahul	Eurasia	Africa	Americas	Sahul
1-23	3 (1)	1 (1)†	0 (0)	1 (1)†	4 (2)†	0 (0)	0 (0)	2 (1)†
12-3	2 (2)	1 (1)	3 (3)†	6 (2)†	7 (3)	0 (0)	0 (0)	4 (1)†
2-13	4 (1)††††	0 (0)	0 (0)	0 (0)	3 (1)†††	0 (0)	0 (0)	2 (1)††
NA	61 (10)	11 (9)	16 (10)	29 (9)	12 (3)	0 (0)	2 (2)	23 (3)

When languages show a 1-23 split, the predicted direction is 1>23, e.g. a marked alignment set for first-person P. The three Eurasian languages that feature this split for P (all Indo-European) uniformly behave in the predicted direction; in Tera (Chadic: Africa) and Teiwa (Timor-Alor-Pantar: Sahul), by contrast, this scale is violated (23>1). For A, three Eurasian languages (all Sino-Tibetan) fit the predicted direction while an Indo-European system (Tarom participial clauses) goes against it; the two Sahul languages are both Pama-Nyungan and show a violation and a fit, respectively.

At least one taxon from each area exhibits a 12-3 split in P-marking, with one violation of the predicted direction in the Americas (Osage) and in Sahul (Teiwa). For A-marking, the only violation of the scale comes from the Pama-Nyungan language Alyawarra. For the singular, then, the 12>3 scale looks more promising than the 1>23 scale.

What is more difficult to evaluate in terms of scalar predictions is languages that make a 2-13 split, as this split is not predicted by the common versions of the referential hierarchy. BWZ set out to test a hierarchy including 1>2>3 and one including 12>3. If we assume that both of these scales are violated by a 2-13 split, all of the languages in the third row of Table 8 above are problematic and hence constitute counterevidence to the implicational hierarchy in (1a); note that they all come from either Indo-European or Pama-Nyungan.

In the non-singular (conflating plural and dual patterns here), the distribution of person splits is as follows (Table 9):

Table 9. Person splits in the non-singular († indicates the number of violating systems)

	P-marking				A-marking			
	Eurasia	Africa	Americas	Sahul	Eurasia	Africa	Americas	Sahul
1-23	0 (0)	0 (0)	1 (1)	0 (0)	3 (1)	0 (0)	0 (0)	3 (1)
12-3	8 (2)††††††	1 (1)	2 (2)†	5 (2)†	13 (3)	0 (0)	0 (0)	6 (1)
2-13	3 (1)†††	1 (1)†	0 (0)	1 (1)†	2 (1)††	0 (0)	0 (0)	0 (0)
NA	59 (10)	11 (9)	16 (10)	30 (8)	8 (3)	0 (0)	2 (2)	22 (3)

As can be seen, systems with a 1-23 split, despite not being numerous, are consistently organized in the predicted direction, i.e. there is no violation of this scale this time (in contrast to what we saw for the singular above). For 12-3 splits, A-marking is also well-behaved without exceptions, while six Indo-European systems (all from closely related Iranian languages), and again Osage (Americas) and Teiwa (Sahul), violate the 12>3 scale for P-marking. In the latter two languages, then, the violation of the 12>3 scale applies to both singular and non-singular pronouns, whereas in Indo-European, the violations are confined to the non-singular. Finally, we also find some 2-13 splits again; apart from Indo-European (Vafsi, Chali (A- and P-marking), English (P-marking only)), these are now also found in Tsamai (Cushitic: Africa) and Tamambo (Austronesian: Sahul).

The figures provided in this section are not directly comparable to BWZ's, as we examine person effects for the two number categories separately while BWZ intended to home in on one referential dimension at a time (i.e. they tested the robustness of person scales regardless of the number distinction and vice versa). At any rate, however, it is clear that there is quite a bit of diversity with regard to the pronominal splits in question and in view of the small overall numbers and the amount and distribution of exceptions, no straightforward universal appears to emerge from eyeballing the data. The Family Bias estimations involving such person splits (cf. Tables 14 and 15 in the Appendix) yield roughly as many biases in favour of each ranking as against it, and we will have to await our alternative statistical evaluation in §3 to see if the distributions are still robust enough to support the most widespread version of the referential scale, which comprises a 12>3 contrast (as in (1a)).

2.6 Number-conditioned splits

The final split in the data is one of number: According to Bickel's (1999) "indexability hierarchy", "singular and individualized referents are generally easier to point at unambiguously than groups or masses", suggesting that "in many languages, they figure higher on the indexability hierarchy" (Bickel & Nichols 2007: 225). Following this logic, Tables 10a–c below display how the data fit a potential SG>NSG scale. Again, we do this separately for each person category and, in the third person, also separately for nouns and pronouns.

Table 10a. Systems with SG>NSG splits in the first person

	P-marking				**A-marking**			
	Eurasia	Africa	Americas	Sahul	Eurasia	Africa	Americas	Sahul
fit	12 (1)	4 (3)	0 (0)	2 (2)	2 (2)	0 (0)	0 (0)	1 (1)
vio	0 (0)	1 (1)	0 (0)	1 (1)	11 (1)	0 (0)	0 (0)	10 (1)
NA	58 (11)	8 (6)	19 (13)	33 (8)	13 (3)	0 (0)	2 (2)	20 (3)

Table 10b. Systems with SG>NSG splits in the second person

	P-marking				**A-marking**			
	Eurasia	Africa	Americas	Sahul	Eurasia	Africa	Americas	Sahul
fit	8 (1)	3 (3)	1 (1)	1 (1)	2 (2)	0 (0)	0 (0)	1 (1)
vio	0 (0)	0 (0)	0 (0)	0 (0)	9 (1)	0 (0)	0 (0)	6 (1)
NA	62 (11)	10 (7)	13 (12)	34 (8)	15 (3)	0 (0)	2 (2)	24 (3)

Table 10c. Systems with SG>NSG splits in the third person

	P-marking				A-marking			
	Eurasia	Africa	Americas	Sahul	Eurasia	Africa	Americas	Sahul
fit.PRO	4 (1)	4 (4)	0 (0)	1 (1)	1 (1)	0 (0)	0 (0)	0 (0)
fit.N	1 (1)	0 (0)	1 (1)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
vio.PRO	1 (1)	0 (0)	0 (0)	1 (1)	4 (1)	0 (0)	0 (0)	2 (1)
vio.N	0 (0)	0 (0)	0 (0)	2 (1)	0 (0)	0 (0)	0 (0)	0 (0)
NA	64 (11)	9 (7)	18 (13)	32 (9)	21 (3)	0 (0)	2 (2)	29 (3)

Again, BWZ seek to assess the number scale as such, without the possible effects of cross-cutting person distinctions. In doing so, they roughly find at least as many violations of the SG>NSG scale as supporting taxa in all areas. The raw but more fine-grained data shown here are complex and suggest a different picture for P- and A-marking. For P-splits, the scale in question mostly (i.e. except for Sahul) receives more support than violations (in raw counts), and it is even exceptionless in the second person. For A-marking, by contrast, there are consistently more violations than fits, yielding BWZ's family-bias results in Table 15 (Appendix). In other words, there is clear evidence against the SG>NSG scale for A-marking while the picture is less straightforward for P-marking. We leave the latter to be explored further by our own statistical model, which will be presented in the next section.

3. Remodelling the data

Now that we have a clearer idea of individual referential dimensions and their behaviour, we can test the robustness of a scale on which they are combined. Perhaps the best-known version of an extended referential hierarchy is the one that recognizes a distinction between speech-act participants and third persons (12>3), a difference between pronouns and full nouns (Pro>N) and a high-low distinction among nouns (which may consist in animacy, definiteness, specificity, topicality and other contrasts). The resulting scale, which is also tested in BWZ, is given in (3) below:

(3) 1,2 Pro > 3 Pro > N_{high} > N_{low}

The relevant predictions for case marking are the previous ones in (2), bearing in mind that "markedness" is defined in terms of alignment sets. For reasons of space and the small number of data points, we will have to confine ourselves to DOM here and exclude differential A-marking from testing. Following BWZ, we will perform two different kinds of statistical evaluation, viz. a conceptually simpler *type model* in §3.1 and a somewhat more complex *rank model* in §3.2.

3.1 Type-based modelling

The basic question in this kind of model is whether the systems that fit the scale in (3) significantly outnumber the systems that violate it, while controlling for genealogical and areal dependencies. The critical issue, therefore, is whether each of the 137 split-P systems in the data is considered a fit to or a violation of (3). In order to be maximally cautious, *any* kind of violation on the following subscales of (3) resulted in the system being coded as "violating":

- *Nominality*: If a language has a "clean" Pro-N split, it fits (3) if the pronouns are marked while the nouns are unmarked; the opposite pattern is a violation. If a language splits only its pronouns, it fits (3) if the nouns join the unmarked sets of pronouns; the opposite pattern is a violation. If a language splits only its nouns, it fits (3) if the pronouns join the marked set of nouns; the opposite pattern is a violation. If a language splits both its nouns and its pronouns, it counts as a violation (cf. our comments in §2.4 above).
- *N_{high}-N_{low}*: All splits according to animacy, definiteness, topicality, kinship and uniqueness are subsumed under the N_{high}>N_{low} distinction (just as in BWZ's test). Since these are usually binary contrasts, they fit the scale in (3) if higher nominals are P-marked while the lower ones are not, while the opposite situation is a violation of (3).
- *Person*:
  - If a language shows a 12-3 split in its pronouns, it fits (3) if speech-act participants (1,2) are marked while 3 is unmarked; the opposite pattern is a violation.
  - If a language shows a 1-23 split, it can be considered a "partial fit" if it takes the direction of 1>23 (i.e. with first person being marked and the others unmarked); in that case, it arguably does obey the proposed 1>3 ranking, while it does not make a distinction between 2 and 3. If the direction of the split is 23>1, it counts as a violation of (3).
  - If a language shows a 2-13 split, it violates the 12>3 part of (3), no matter which direction the split takes (cf. our earlier discussion of this issue).
  - Where a language exhibits different kinds of person splits for singular and non-singular number, each of them was first evaluated separately according to the above criteria, and the values were subsequently combined into a single one. If a system showed a fit in one number category and a partial fit in the other, we coded it as fit; if a system showed a fit and a violation in the other (e.g. Tera and Tsamai), we coded it as partial fit; if a system showed a partial fit in one number category and no split in the other (e.g. Shughni), we also counted it as a partial fit. All other combinations containing some violation were counted as violating systems.

As a result of this coding policy, we obtained the following raw data for the scale in (3) (Table 11):

Table 11. Systems fitting or violating the scale in (3)

	Eurasia	Africa	Americas	Sahul
fit	56 (11)	9 (8)	14 (9)	31 (8)
vio	14 (1)	2 (2)	5 (5)	5 (3)
partial	0 (0)	2 (2)	0 (0)	0 (0)

These figures suggest a rather strong tendency for both systems and families to fit the scale in all macro continents, but in order to control for genealogical relationships and areal dependencies in a rigorous way, a mixed-effects generalized linear model (GLM) is called for. We thus applied a mixed Poisson GLM (also known as mixed loglinear model) to the data at hand. To this end, the data were first cross-tabulated into the format shown in Table 12 (the full dataset is available as supplementary material SM1)[11]:

Table 12. Data coding for the Poisson GLM (segment)

Family	MContinent	Fit	Freq
Adamawa-Ubangi	Africa	fit	1
Adamawa-Ubangi	Africa	vio	0
Benue-Congo	Africa	fit	2
Benue-Congo	Africa	vio	0
Chadic	Africa	fit	1
Chadic	Africa	vio	0

The results of loglinear modelling show that there is no interaction between the fixed effects of *Fit* and *MContinent* ($p = 0.637$): In all areas, there is a strong preference for fitting systems even when genealogical relations are controlled for: $b = 1.43$, $p < 0.0001$ (cf. SM3 for further details). The estimates in a Poisson model represent the multiplicative effect of a variable on the outcome on the log scale, which means that “fit” is about $e^{1.43} \approx 4.2$ times more probable than “violation”.[12]

In short, the type model suggests that there is a strong cross-linguistic tendency for languages to fit the referential scale in (3), independently of macro-continental affiliations. Since the counts were aggregated across language families, the observed cross-linguistic bias towards fitting the scale cannot be attributed to the possible impact of larger families, either.

3.2 Rank-based modelling

In this kind of model, it is tested whether higher-ranking P arguments stand a better chance of being structurally marked than lower-ranking ones. More precisely, we are probing an ordinal relationship by which the odds for marked P arguments should decrease as we proceed down the ranks on the scale (i.e. 1st rank > 2nd rank > 3rd rank, etc.). In order to run an appropriate model, the data were converted into the following long format (Table 13):

[11] For reasons of simplicity, we discarded the two “partial” languages in Table 11 (viz. Tera and Tsamai, both Afro-Asiatic).

[12] An alternative to the above loglinear format is to treat the number of fitting and violating systems as successes and failures in trials within a family, similar to heads or tails when one tosses a coin (where each new language produces either heads or tails). It would then be appropriate to apply logistic binomial regression. We tested whether *MContinent* had a significant influence on the chances of fits as compared to violations within each family. Because of some amount of overdispersion, a quasibinomial GLM was used. This procedure yielded the same result as the one presented above. There is no significant effect of *MContinent* on the chances of fitting or violation. A model with the intercept only has a significant intercept $b = 1.44$, $p < 0.0001$, which means that the odds of fitting are $e^{1.44} \approx 4.2$ times higher than those of violation. This result is almost identical to the one presented above. The two modelling approaches thus converge, which is reassuring.

Table 13. Data coding for the rank-based GLM (segment)

MContinent	Family	System	RefCat	Number	Marking	Rank
Africa	Adamawa-Ubangi	Gbeya	1	SG	marked	12
Africa	Adamawa-Ubangi	Gbeya	1	NSG	marked	12
Africa	Adamawa-Ubangi	Gbeya	2	SG	marked	12
Africa	Adamawa-Ubangi	Gbeya	2	NSG	marked	12
Africa	Adamawa-Ubangi	Gbeya	3	SG	marked	3
Africa	Adamawa-Ubangi	Gbeya	3	NSG	marked	3
Africa	Adamawa-Ubangi	Gbeya	N_{high}	SG	unmarked	N_{high}
Africa	Adamawa-Ubangi	Gbeya	N_{high}	NSG	unmarked	N_{high}
Africa	Adamawa-Ubangi	Gbeya	N_{low}	SG	unmarked	N_{low}
Africa	Adamawa-Ubangi	Gbeya	N_{low}	NSG	unmarked	N_{low}

This format represents each system in the data by 10 rows, allowing us to code each combination of referential category (cf. 4th column, *RefCat*) and number (5th column) separately. This way, we can now also take person differences between singular and non-singular into account. The full data are available as supplementary material SM2.

We fitted a mixed-effects logistic GLM to these data. The response variable was *Marking*, with the values "marked" and "unmarked" (6th column of Table 13). The predictor that represented the position of the arguments on the referential scale was called *Rank* (7th column). We included *Number* and *MContinent* as further fixed effects and tested the interactions between the predictors. The individual tendencies of systems and language families to mark more or fewer referential categories (variables *System* and *Family*) were encoded as random intercepts.[13] Since *System* is nested within *Family*, we are dealing with a multilevel hierarchical model.

The analyses reveal a significant main effect of *Rank* as well as two significant interactions between the predictors: one between *Rank* and *Number*, and the other between *Rank* and *MContinent*. In the presence of multiple interactions, it is best to explore the results visually. Figure 1 displays the average probabilities of "marked" P arguments in the singular and the non-singular on the vertical axis. The horizontal axis represents the four ranks on the scale, from left to right. The different colours and lines correspond to the four macro continents, which are explained in the legend.

In the singular, we observe very little if any difference between the first two positions on the scale (12 and 3). Figure 1 thus confirms our earlier observation that the difference between speech-act and third-person (singular) pronouns is not very relevant for P-marking overall, but also that there are hardly any violations of the predicted effect where it occurs. In Africa and Sahul, the most obvious decrease in the chances of P being marked is found between the pronouns and the nouns. In contrast, the Americas and Eurasia have a large difference in the probability of marking between all high-prominence arguments (pronouns and high-prominence nouns) and low-prominence nouns.

[13] We also tested models in which we additionally allowed for the rank effect to vary between the families in the sample, i.e. by adding random by-family slopes. Where such models were feasible given the present sample size per family, they did not make a significant contribution to the model (and were hence discarded in the stepwise modelling process), nor did they affect the stability of the rank effect.

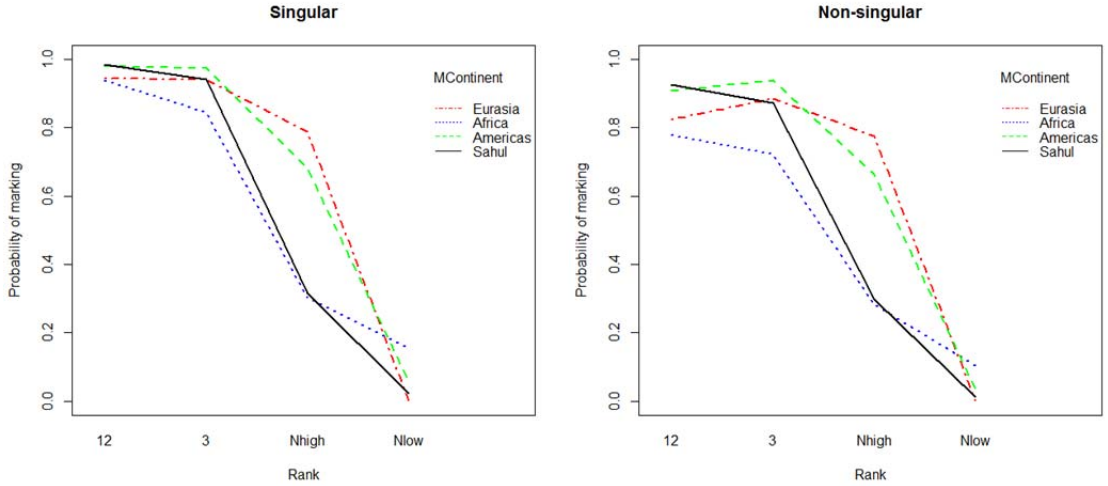


Figure 1. Influence of ranks on the probability of marked Ps in the singular (left) and the non-singular (left), by macro continent.

In the non-singular, Figure 1 nicely reflects what we saw in Table 9 above: In the Americas (specifically, Osage) and particularly in Eurasia, there is a certain number of languages that violate the 12>3 part of the scale, leading to a slight positive rather than the expected negative slope of the relevant curves in Figure 1. We saw above that these exceptions are virtually all located in Iranian languages, and their effect is not strong enough to yield significant counterevidence (post-hoc tests of P-marking: Eurasia: $b = 0.22$, $p = 0.284$, Americas: $b = 0.13$, $p = 0.767$). By contrast, all other ranking effects in Figure 1 are negative and significant (cf. SM3 for further technical details of the model).

In sum, what we can take from this model is the following:

- We do not find any significant violations of the referential hierarchy in (3).
- Singular and non-singular number behave slightly differently with regard to the effects of the 12>3 subscale (the effect is largely irrelevant in the singular and mixed though not significantly contradictory in the plural). However, as there is also a significant main effect of *Rank* in the data, the hierarchy in (3) is robust enough across the number categories as well.
- The macro continents behave differently with regard to the average cut-off point that is most relevant on the hierarchy.

There are thus evidently areal patterns and restrictions in DOM, but the predicted effect of the referential hierarchy in (3) is uniform enough in our model to assume that it is universally valid, after all. Therefore, while BWZ argue “against universal effects of referential scales on case marking” (cf. the title of their paper), we would argue for the universality of precisely the *effect*, no matter which particular dimensions of referential prominence are the most relevant ones in individual languages or macro areas. This and further issues of interpretation deserve more elaborate discussion, provided in the next section.

4. Interpreting the data

In assessing the alleged universality of scale effects on case marking, a number of fundamental questions arise that will influence one's conclusion on the matter. In the following subsections, we are going to discuss a selection of these, notably assumptions about methodological choices, geographical distributions, counterexamples, and the ontological status of scales, i.e. what they represent and what they are supposed to do.

4.1 Methodological approaches to typological data

At first glance, the most striking difference between BWZ's approach to modelling the data and ours is that the former is framed in terms of what Greenberg (1978) calls "the dynamicization of typology": As BWZ (p. 24) put it, "any evaluation" of alleged universal pressures "needs to target trends in diachrony rather than current distributions". The Family Bias Method attempts to model such diachronic trends by investigating whether genealogical taxa tend to develop in keeping with the alleged universal (here: in the direction predicted by a given referential scale) or not. Other dynamic approaches are based on estimating and comparing transition probabilities from the genealogical structure (i.e. family trees) of individual taxa (cf., e.g., Dunn et al. 2011, Cysouw 2011, Bickel et al. 2015c). All of these dynamic methods are, of course, promising developments in linguistic typology. But it should be borne in mind that they are not based on diachronic data, but on particular inferences drawn from synchronic distributions and/or genealogical relations. And such inferences, in turn, usually involve delicate decisions on uncertain issues, such as the branch lengths in family trees, the threshold for defining diachronic biases or the way in which one extrapolates from large to small families.

Again, we do not wish to call these methods into question, but in the absence of *world-wide* data on actual diachronic developments, we believe that densely sampled synchronic data are still a viable, legitimate and powerful source of evidence in linguistic typology. Instead of throwing out the synchronic baby with its bathwater, then, we have here followed equally recent methodological proposals by Cysouw (2010) and Jaeger et al. (2011) to model synchronic distributions by means of mixed-effects regression procedures. These are standard ways of modelling variation in other disciplines, and while they cannot, by definition, target any diachronic trends, they are powerful means of staking out the room for universal pressures once family- and area-internal variation is controlled for. In fact, just like the Family Bias Method, they examine the number of "fits" and "violations" taxon by taxon (cf. Table 12 again). The difference is that our models end up taking all taxa in the data on board (including those that the Family Bias Method would have excluded as "internally diverse") and that they always operate with the actual values of all isolates rather than estimating them based on extrapolation procedures. What we can obtain from this is a classic Greenbergian statement that "with overwhelmingly greater than chance frequency" (e.g. Greenberg 1966: 79), systems of differential case marking tend to obey the referential hierarchy in (3) rather than going against it.

Ultimately, then, it is fair to say that, at the current stage of research, synchronic and diachronic methods of modelling typological data have complementary advantages and drawbacks. And as long as that is the case, we see no reason to trust carefully sampled and analyzed synchronic data any less than diachronic inferences drawn from them.

4.2 Geographical universality

A common assumption since at least Bossong (1985) has been that differential argument marking, and its systematic correlation with referential categories, is "extremely widespread" (Aissen 2003: 439) and independent of macro-areal affiliations.

In all fairness, these claims refer to differential P-marking only, and BWZ's data suggest that they would, indeed, be plainly wrong for A-marking. Although we do not know the principles according to which BWZ selected their sample languages, it seems safe to say that differential A-marking is generally dispreferred and its occurrence is skewed heavily towards Eurasia and Sahul, and here again towards Indo-European, Pama-Nyungan and perhaps Sino-Tibetan. For differential P-marking, on the other hand, the picture is less clear. The two largest distributional studies, namely BWZ and Sinnemäki (2014), appear to yield somewhat different results, which we set out and discuss for interested readers in the supplementary materials (SM4); from the facts presented there, it seems to us that when languages develop case marking for direct objects, the differential marking type is indeed more likely, across the world's linguistic macro areas, than unsplit marking.

But the overall distribution of DOM is actually less vital than another point raised in Sinnemäki (2014): He argues that the individual referential dimensions underlying DOM exhibit conspicuous areal contours. While animacy is distributed fairly evenly across the globe, definiteness/specificity shows a strong skewing towards Africa and the Old World more generally. In our model, too, we found some significant areal differences in the preferred cut-off points on the hierarchy in (3). However, we opine that such areal skewings do not invalidate the basic insight of the referential scales in (1) and in (3). As far as we can see, all versions of referential scales proposed in the typological literature are intended to be cross-linguistic generalizations over referentially-conditioned splits in individual languages, no matter which of the referential categories on a given scale are actually relevant in those languages. In other words, the hierarchy aims to capture a language with a particular person split in the pronouns just as much as a language with an animacy split among full NPs. Therefore, the requirement for the universality of scale effects is not that each individual subscale or referential dimension needs to be attested throughout the world, but that wherever referentially-conditioned splits *do* occur, they will strongly tend to obey the referential hierarchy rather than going in the opposite direction. Crucially, this latter issue is not addressed in Sinnemäki's (2014) paper: He asks which referential (or other structural) dimensions are responsible for differential object marking in the sample languages and how these dimensions are distributed geographically. He does not, however, look at the directionality of the effect, i.e. whether a language that has an animacy split actually works in the predicted direction. To the extent that these effects are uniform (cf. §4.3 below), we do not see any reason to question the validity of referential scale effects on purely geographical grounds.

4.3 Structural universality

Bossong (1985: VIII) voices a common opinion among comparative linguists when he claims that the patterns of differential object marking are "structurally uniform […] around the earth" (our translation), in the sense that whenever DOM is driven by referential properties, it follows the direction given in (2) above rather than going against it. BWZ extend this assumption to differential A-marking as well and ask whether “there exists one or more *universal* scale(s) on which all [split] systems fit” (p. 22), and we already know that their conclusion is negative.

There are two issues involved here. The first and more important one pertains to the number of weight of counterexamples. The figures above suggest that splits in terms of animacy, definiteness and other high/low-contrasts are almost without exception, for both A- and P-marking (Tables 3 and 5 above). The same holds when languages make “clean” splits between nouns and pronouns (Table 4). From this perspective, the lower end of the traditional referential hierarchy, as well as its global ranking of pronouns and nouns, can be considered structurally uniform, indeed (cf. also Levshina 2017+ for further statistical corroboration). What is more problematic is the internal ranking of person and number distinctions, i.e. particularly the upper part of the referential hierarchy. Here, Tables 7–10 suggest considerable language-specific variation and thus idiosyncratic historical developments (cf. also Filimonova 2005 on this point). Therefore, when BWZ test for scales involving particular pronominal splits (e.g. 1>2>3>N or 12>3>N), and with the cross-cutting number distinctions being disregarded, it is not surprising that they find a number of exceptions; in fact, they even find roughly as many family biases in favour of and against these scales (cf. Appendix). By contrast, our alternative regression analysis of the $12>3>N_{high}>N_{low}$ scale still showed a robust enough effect for this particular person split (in both the type model and the rank model), even when number is taken into account as a separate variable. Taken together, our analyses suggest that referential effects on case marking are sufficiently homogeneous to be considered universal, at least by typologists who (unlike Bickel et al.) accept purely synchronic evidence as a valid basis for establishing universals.

A second point about structural homogeneity relates to BWZ’s finding (p. 34) that no single scale they tested fits A and P *simultaneously*. As with Sinnemäki’s argument about the areal restrictedness of animacy or definiteness, one may object here that it actually does not matter whether the high-low distinction is less important for A-marking than for P-marking. In fact, it has recently been emphasized that A and P are not simply “each other’s mirror-image” (Fauconnier & Verstraete 2014) in a number of ways, and hence also differ in regard to the referential properties that are relevant when they are case-split. It may thus very well be the case that the referential hierarchies in (1) are poorer predictors for A- than for P-marking because they miss some of the crucial dimensions (e.g. particular kinds of focus) and overstate others (e.g. animacy and definiteness). However, to the extent that they *are* applicable, it is again the predicted *effects* that are at stake here. And as we saw above, the effect is strikingly homogeneous as far as high-low distinctions and the clean Pro-N splits are concerned. Where A and P may respond very differently is the referential dimension of number, as was shown in Table 10, so that we see opposing rather than uniform effects of the alleged SG > PL scale. This is certainly worth further investigation, but given that most versions of the referential hierarchy are not even concerned with number contrasts, we do not see this as a serious challenge to referential scale effects in general.

4.4 The status and purpose of referential scales

In this final section, we would like to comment on two remarks by BWZ on the usefulness of scales in typological research. The first one relates to the fact that by far most languages work in terms of a specific binary opposition[14], which is why BWZ explicitly reject the terms “scale” or “hierarchy” to capture such simple splits. In our view, this issue is largely terminological in nature: In so far as binary oppositions (like Pro > N) are implicational statements as we see them in other typological domains (e.g. like SG > PL or VOICED PLOSIVES > VOICELESS PLOSIVES), we are not averse to calling them “(implicational) scales” or “(implicational) hierarchies”. The more important issue is the second one, relating to the level of abstraction at which comparative scales are formulated. Recall that BWZ find positive evidence for their Pro/N_{high} > N_{low} scale, but they question the usefulness of such a scale precisely because it seems too heterogeneous to reflect a single underlying principle (p. 36 of their paper). The same kind of criticism may actually be levelled against the extended hierarchies in (1a) and (3), which also conflate a number of logically distinct dimensions (e.g. a person contrast within the pronouns, a split in nominality and various other properties). The question is, therefore, to what extent the postulation of more abstract (i.e. extended, multidimensional or more general) hierarchies is justified.

In general, the motivation behind postulating referential scales is to capture constraints on cross-linguistic variation. Mapping diverse language-specific oppositions onto more abstract comparative scales firstly serves the purpose of increasing the scope of the constraint; as compared to individual scales, it is thus arguably a more elegant way of formulating cross-linguistic generalizations. It does, however, also suggest that there is a unified explanation for the phenomenon in question. Gildea & Zúñiga (2016), for example, note that the referential hierarchy has often been taken to reflect a coherent cognitive phenomenon, a “representational constraint” in the sense of Haspelmath (subm.) or Elman et al. (1996). For example, Kiparsky (2008: 39–40) characterizes his version of the referential hierarchy as an “inviolable […] part of the design of language”, i.e. of “U[niversal] G[rammar]”. In so far as such representational principles directly constrain the possible shapes of case-marking systems, the postulated hierarchy is said to *explain* the cross-linguistic patterns we observe.[15]

In functional-typological work, referential hierarchies are not inviolable “top-down” principles of cognition; the correlations they capture (i.e. between an argument’s referential prominence and its likelihood of receiving special case marking) are typically given more probabilistic explanations in terms of language usage and change.[16] Now, if one believes that these correlations fall out entirely from local processes of grammaticalization and can be fully explained by reference to the respective source construction (e.g. Cristofaro 2013), there is really no gain in postulating an extended or more abstract hierarchy beyond individual referential dimensions. By contrast, for typologists who argue that these individual dimensions can

[14] Exceptions to this are languages that make a certain kind of split in the pronouns (e.g. 12>3) and a different one in the nouns (e.g. N_{high}>N_{low}, cf. Table 7 above), or languages that use multiple cases or different case allomorphs differentially, depending on referential properties.

[15] A formal account of a very different kind is presented in Aissen (2003), but the conclusion ultimately also reads like an UG-based representational constraint: “[T]he principles underlying DOM” may be “part of core grammar”, implemented by a “universally fixed […] ranking of constraints” (Aissen 2003: 439–40).

[16] There are, of course, also attempts in the typological literature to link implicational universals and semantic maps to “conceptual spaces”, i.e. coherent “regions” of the human mind (cf. Croft 2003). But this sort of cognitive interpretation does not seem to be prominent for the referential hierarchy. For a general critique of this approach, see Cristofaro (2010).

receive a unified explanation, such an abstraction is more useful. Perhaps the best-known line of argumentation in this direction is that of communicative efficiency (e.g. Dixon 1979; Comrie 1981; Newmeyer 2005; Haspelmath 2008; Hawkins 2014): Speakers tend to mark those A and P arguments whose syntactic function is relatively unexpected (or surprising) given their referential properties, while expected role-reference constellations are left unmarked (cf. also Haspelmath 2018 for a systematization of this proposal). Crucially, this account is said to work for all kinds of referential splits in the same way, whether they are based on animacy, definiteness or other kinds of prominence in particular languages. While still in need of further corroboration, there is mounting evidence from frequency data (e.g. Dahl 2000; Fry 2003; Jäger 2007; Lee 2006), psycholinguistic experimentation (e.g. Kurumada & Jaeger 2015; Fedzechkina et al. 2012) and computer simulations (e.g. Lestrade, this volume) in favour of this approach, at least for DOM (cf. also Levshina 2018).[17]

In sum, then, the postulation of more abstract or multidimensional referential hierarchies is not just an elegant way of formulating cross-linguistic generalizations about case splits. It is also useful if one believes that a unified explanation can be given to those splits. With regard to the latter, we currently see little, if any, evidence for an innate, inviolable referential hierarchy in Kiparsky's sense, but accumulating evidence in favour of functional explanations that operate with probabilistic constraints on usage and diachronic change.[18]

5. Conclusion

In this paper, we have attempted to re-present and reanalyze Bickel et al.'s (2015a) typological data on differential case marking. Their database, along with Sinnemäki's (2014), constitutes the largest current repository for gauging case-marking patterns in the world's languages, and we would thus like to acknowledge again the tremendous amount of cross-linguistic groundwork that these colleagues have carried out. Moreover, Bickel's (2011; 2013) Family Bias Method is a valuable addition to the toolkit of quantitative typology, as it starts out from considering how possibly universal pressures on language should play out in the diachronic development of families. It is thus *conceptually* different from the kinds of regression models that we have used in the present paper, although it operates with exactly the same kind of synchronic typological data. The most important technical difference is that its final results are based on

[17] As we saw earlier, differential A-marking is generally rarer, geographically and genealogically more restricted, and no parallel evidence from psycholinguistic experimentation is currently available. Moreover, there is compelling evidence that differential A-marking involves additional motivations that do not apply to P-marking in the same way (de Hoop & Malchukov 2008, Fauconnier & Verstraete 2014). For these reasons, it is presently rather difficult to estimate just how much of differential A-marking is amenable an account in terms of communicative efficiency.

[18] A reviewer of the paper remarked that this formulation, and the efficiency explanation in general, is basically diachronic in nature, which s/he sees as a contradiction to the kind of synchronic typology we have practised here. But these are actually two independent issues. Efficiency explanations are first and foremost about the choices, however subconscious, that individual speakers make for or against overt case marking in online production (and hence "synchronically", in a sense); these necessarily have to propagate in time and space to conventionalize into a grammatical pattern, which adds a diachronic component to the explanation. But since we cannot sample these processes in the same way that we can sample their results across the world's languages, we believe that the synchronic states that we have investigated here are still a viable data source for typologists. This is hence a purely methodological point and does not contradict the fact that usage-based explanations involve diachrony.

statistically significant biases in large families and their extrapolation to small taxa and isolates; it thus neglects large families without biases and introduces some noise into the data from small taxa (cf. Appendix again). The major goal of the present paper was to complement these Family Bias estimations with a look at the actual "raw" data on various referential dimensions and to present an alternative statistical model of the data that relies on widely used regression procedures on the full data set.

In doing so, we found less counterevidence than BWZ's results and their rhetoric suggest. The global structure of the classic hierarchies (pronouns > nouns) and all high-low prominence distinctions (animacy, definiteness, topicality, kinship) are almost without exception, and while there is more variation within the pronominal domain, a closer look at the data reveals that the number of counterexamples is not significant enough to override the strong support that the referential hierarchy in (3) receives from our statistical models.

Therefore, our conclusion is the opposite of BWZ's, namely that there *is* evidence for universal scale effects on case marking. We can subscribe to this view for the following reasons:

- Unlike BWZ, we accept purely synchronic evidence for postulating universal preferences (provided it is as statistically robust as in the present case).
- Unlike Sinnemäki (2014), we do not require that the individual referential properties need to be involved in DAM in all macro areas to the same degree; what matters is that the direction of the effect is uniform, regardless of which specific referential dimensions it comes from.
- Unlike BWZ, we obtain a positive statistical signal even when several referential dimensions are combined into a larger scale.
- Unlike BWZ, we have no reservations to apply the label "scale" even to binary oppositions (which is how most languages work to begin with). That is, even if we did not wish to operate with extended scales such as (1) or (3), we would argue for the existence of "scale effects".

As laid out in §4, we believe that working with multi-term or abstract scales can be useful if one has an explanatory account that unites the various referential dimensions under a single principle. While we reject the view that such a referential hierarchy constitutes an innate representational constraint, we are sympathetic to a functional view that relates different referential contrasts to a common principle of efficient information processing.

Acknowledgements

The research for this paper was carried out in the context of the project *Form-frequency correspondences in grammar* at Leipzig University. The support by the European Research Council (ERC Advanced Grant 670985) is gratefully acknowledged. We would also like to thank Alena Witzlack-Makarevich, Ilja A. Seržant, Martin Haspelmath, Kaius Sinnemäki and an anonymous reviewer for very helpful comments and suggestions. The usual disclaimers apply.

References

Aissen, Judith (2003). Differential object marking: Iconicity vs. economy. *Natural Language and Linguistic Theory* 21: 435–483.

Bickel, Balthasar (1999). Indexability effects in Himalayan languages. Paper presented at the *Workshop on Himalayan Languages*, Santa Barbara, June 1999.

Bickel, Balthasar (2011). Statistical modeling of language universals. *Linguistic Typology* 15: 401–414.

Bickel, Balthasar (2013). Distributional biases in language families. In Balthasar Bickel, Lenore A. Grenoble, David A. Peterson & Alan Timberlake (eds.), *Language typology and historical contingency*, 415–444. Amsterdam: Benjamins.

Bickel, Balthasar & Johanna Nichols (2002). Autotypologizing databases and their use in fieldwork. In Peter Austin, Helen Dry & Peter Wittenburg (eds.), *Proceedings of the International LREC Workshop on Resources and Tools in Field Linguistics*, Las Palmas, 26–27 May 2002. <http://www.uzh.ch/spw/autotyp/ download/canary.pdf>.

Bickel, Balthasar & Johanna Nichols (2007). Inflectional morphology. In Timothy Shopen (ed.), *Language typology and syntactic description, Vol. III: Grammatical categories and the lexicon*, 2nd ed., 169–240. Cambridge: Cambridge University Press.

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

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

Bickel, Balthasar, Alena Witzlack-Makarevich, Kamal K. Choudhary, Matthias Schlesewsky & Ina Bornkessel-Schlesewsky (2015c). The neurophysiology of language processing shapes the evolution of grammar: Evidence from case marking. *PLoS ONE* 10(8): e0132819. doi:10.1371/journal.pone.0132819.

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

Bossong, Georg (1985). *Differentielle Objektmarkierung in den neuiranischen Sprachen.* Tübingen: Narr.

Comrie Bernard (1981). *Language universals and linguistic typology.* Chicago: University of Chicago Press.

Cristofaro, Sonia (2010). Semantic maps and mental representation. *Linguistic Discovery* 8: 35–52.

Cristofaro, Sonia (2013). The referential hierarchy: Reviewing the evidence in diachronic perspective. In Dik Bakker & Martin Haspelmath (eds.), *Languages across boundaries: Studies in memory of Anna Siewierska*, 69–93. Berlin and New York: Mouton de Gruyter.

Croft, William (2003). *Typology and universals.* 2nd ed. Cambridge: Cambridge University Press.

Cysouw, Michael (2010). Dealing with diversity: Towards an explanation of NP-internal word order frequencies. *Linguistic Typology* 14: 253–286.

Cysouw, Michael (2011). Understanding transition probabilities. *Linguistic Typology* 15: 415–431.

Dahl, Östen (2000). Egophoricity in discourse and syntax. *Functions of Language* 7.1: 37–77.

DeLancey, Scott (1981). An interpretation of split ergativity. *Language* 57: 626–657.

Dixon, R. M. W. (1979). Ergativity. *Language* 55: 59–138.

Dunn, Michael, Simon Greenhill, Stephen Levinson & Russell Gray (2011). Evolved structure of language shows lineage-specific trends in word-order universals. *Nature* 473: 79–82.

Elman, Jeffrey L., Elizabeth A. Bates, Mark H. Johnson, Annette Karmiloff-Smith, Domenico Parisi & Kim Plunkett (1996). Rethinking innateness. Cambridge, MA: MIT Press.

Fauconnier, Stefanie & Jean-Christophe Verstraete (2014). A and O as each other's mirror image? Problems with markedness reversal. *Linguistic Typology* 18: 3–49.

Fedzechkina, Maryia, T. Florian Jaeger & Elissa L. Newport (2012). Language learners restructure their input to facilitate efficient communication. *Proceedings of the National Academy of Sciences of the United States of America*, 1–6. <http://dx.doi.org/10.1073/pnas.1215776109>.

Filimonova, Elena (2005). The noun phrase hierarchy and relational marking: problems and counterevidence. *Linguistic Typology* 9: 77–113.

Fry, John. (2003). *Ellipsis and* wa-*marking in Japanese conversation*. London: Routledge.

Gildea, Spike and Fernando Zúñiga (2016). Referential hierarchies: A new look at some historical and typological patterns. *Linguistics* 54.3: 483–529.

Greenberg, Joseph H. (1966) [1963]. Some universals of grammar with particular reference to the order of meaningful elements. In Joseph H. Greenberg (ed.), *Universals of language*, 2nd edn., 73–113. Cambridge, MA & London: The MIT Press.

Greenberg, Joseph H. (1978). Diachrony, synchrony and language universals. In Joseph H. Greenberg, Charles A. Ferguson & Edith Moravcsik (eds.), *Universals of human language,* vol. 1: *Method and theory*, 61–92. Stanford: Stanford University Press.

Haspelmath, Martin (2008). Creating economical morphosyntactic patterns in language change. In Jeff Good (ed.), *Language universals and language change,* 185–214. Oxford: Oxford University Press.

Haspelmath, Martin (2018). Role-reference association and the explanation of argument coding splits. Ms., Leipzig University. Available at https://zenodo.org.

Haspelmath, Martin (subm.). Can cross-linguistic regularities be explained by change constraints? In Karsten Schmidtke-Bode, Natalia Levshina, Susanne Maria Michaelis & Ilja A. Seržant (eds.), *Explanation in linguistic typology: Diachronic sources, functional motivations and the nature of the evidence.* [Under review]

Hawkins, John A. (2014). *Cross-linguistic variation and efficiency.* Oxford: Oxford University Press.

de Hoop, Helen & Andrej Malchukov (2008). Case-marking strategies. *Linguistic Inquiry* 39: 565–587.

Iemmolo, Giorgio (2010). Topicality and differential object marking: Evidence from Romance and beyond. *Studies in Language* 34: 239–272.

Jaeger, T. Florian, Peter Graff, William Croft & Daniel Pontillo (2011). Mixed effect models for genetic and areal dependencies in linguistic typology. *Linguistic Typology* 15: 281-320.

Jäger, Gerhard (2007). Evolutionary game theory and typology: A case study. *Language* 83.1: 74–109.

Kiparsky, Paul (2008). Universals constrain change; change results in typological generalizations. In Jeff Good (ed.), *Linguistic universals and language change*, 24–53. Oxford: Oxford University Press.

Kurumada, Chigusa & T. Florian Jaeger (2015). Communicative efficiency in language production: Optional case-marking in Japanese. *Journal of Memory and Language* 83: 152–178.

Lazard, Gilbert (1998). *Actancy.* Berlin, New York: Mouton de Gruyter.

Lee, Hanjung. (2006). Parallel optimization in case systems: Evidence from case ellipsis in Korean. *Journal of East Asian Linguistics* 15: 69–96.

Levshina, Natalia (2018). Differential argument marking and referential scales: Bayesian multilevel regression modeling of typological data. Ms., Leipzig University; Available at http://www.natalialevshina.com/publications.html.

van Lier, Eva (2012). Referential effects on the expression of three-participant events across languages – An introduction in memory of Anna Siewierska. *Linguistic Discovery* 10(3): 1–16.

Newmeyer, Frederick J. (2005). *Possible and probable languages.* Oxford: Oxford University Press.

Noonan, Michael (2003). The Chantyal language. In Graham Thurgood & Randy J. LaPolla (eds.), *The Sino-Tibetan languages*, 315–335. London, New York: Routledge.

R Development Core Team (2016). *R*: *A language and environment for statistical computing*. Vienna: *R* Foundation for Statistical Computing. <http://www.r-project.org>.

Schmidtke-Bode, Karsten (subm.). Attractor states and diachronic change in Hawkins's 'Processing Typology'. In Karsten Schmidtke-Bode, Natalia Levshina, Susanne Maria Michaelis & Ilja A. Seržant (eds.), *Explanation in linguistic typology: Diachronic sources, functional motivations and the nature of the evidence.* [Under review]

Shibatani, Masayoshi (2006). On the conceptual framework for voice phenomena. *Linguistics* 44: 217–269.

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

Sinnemäki, Kaius (2014). A typological perspective on Differential Object Marking. *Linguistics* 52.2: 281–313.

Witzlack-Makarevich, Alena, Taras Zakharko, Lennart Bierkandt, Fernando Zúñiga & Balthasar Bickel (2016). Decomposing hierarchical alignment: Co-arguments as conditions on alignment and the limits of referential hierarchies as explanations in verb agreement. *Linguistics* 54.3: 531–561.

Appendix: Bickel et al.'s (2015a) Family Bias estimations

In this appendix, we provide some of Bickel et al.'s (2015a) results for comparison with our own analysis. Readers familiar with the Family Bias Method may thus jump ahead to Tables 14 and 15 below; for uninitiated readers, we first provide some comments on how to interpret the figures. For a more detailed introduction to the Family Bias Method as such, such readers are referred to Bickel (2013) or to the supplementary materials in Schmidtke-Bode (subm.).

The key question that Bickel et al. (2015a) [henceforth BWZ, as in the main text] seek to address is whether a given referential scale shapes the diachronic evolution of language families. BWZ take the synchronic internal composition of each family as indicative of such directed diachronic processes: If a family is significantly biased (on synchronic grounds) towards fitting a scale rather than in the opposite direction, this may be indicative of the family having developed in the predicted direction, either by continually retaining the fit on each evolutionary trial (i.e. with each new daughter language) or by "correcting" a non-fitting case system at the next cladogenetic juncture (i.e. with a new daughter language). A universal signal for scale effects would then amount to most families in a representative sample being significantly biased in the predicted way, again independently of geographical affiliations.

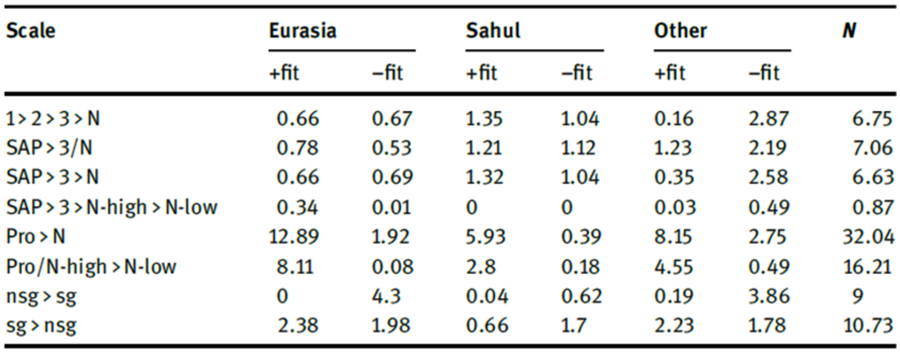
It is obvious that such biases can only be estimated for sufficiently large families (here: N ≥ 5 members). Bickel's method thus extrapolates these estimations to smaller families and isolates. As a consequence, the synchronic data for a language isolate are not simply taken at face value, but as surviving traces of an erstwhile family that itself may or may not have had a principled bias in differential argument marking. In other words, one reckons with the possibility that a given isolate can be the survivor of a family with the opposite bias, or no bias at all. Depending on how strong and uniform the biases are in large families, the method may thus deliberately introduce some "noise" to the data from small families and isolates, rather than always taking their actual values as we find them in the synchronic data. Because of such "interventions" with the data, the extrapolation process is repeated hundreds or even thousands of times and the average results of all estimations are then taken as the final basis for exploring universal trends.

It is against this background that BWZ's Family Bias estimations need to be interpreted. Therefore, the following things need to be kept in mind when looking at the figures below:

- The figures always pertain to *taxa* (i.e. genealogical units) rather than languages.
- The figures exclude taxa that have been estimated to be *diverse* (rather than biased), as internally diverse taxa are argued not to yield conclusive evidence for the family to be shaped by a given referential scale.
- The figures contain non-integer numbers, as the extrapolation to small families and isolates is repeated many times and averaged over; the results thus display the *means* of several hundreds of runs of bias estimations.

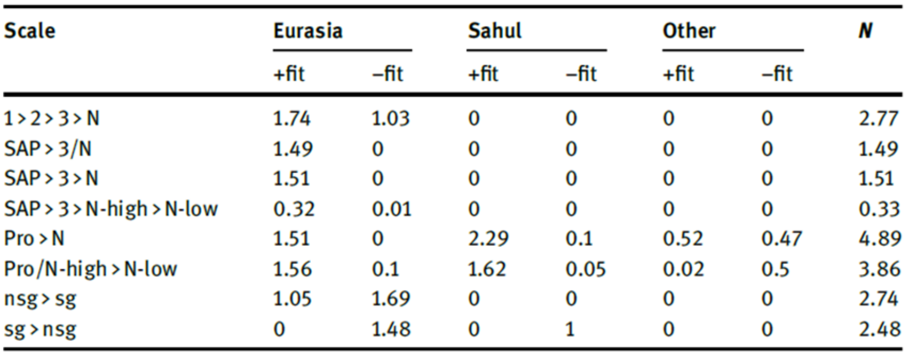
In Tables 14 and 15, we present the results of BWZ's *type model* (cf. our §3.1 for comparison).

Table 14. Results of Bickel et al.'s (2015a: 34) *type-model* analysis of P-splits



Scale	Eurasia		Sahul		Other		*N*
	+fit	−fit	+fit	−fit	+fit	−fit	
1 > 2 > 3 > N	0.66	0.67	1.35	1.04	0.16	2.87	6.75
SAP > 3/N	0.78	0.53	1.21	1.12	1.23	2.19	7.06
SAP > 3 > N	0.66	0.69	1.32	1.04	0.35	2.58	6.63
SAP > 3 > N-high > N-low	0.34	0.01	0	0	0.03	0.49	0.87
Pro > N	12.89	1.92	5.93	0.39	8.15	2.75	32.04
Pro/N-high > N-low	8.11	0.08	2.8	0.18	4.55	0.49	16.21
nsg > sg	0	4.3	0.04	0.62	0.19	3.86	9
sg > nsg	2.38	1.98	0.66	1.7	2.23	1.78	10.73

Table 15. Results of Bickel et al.'s (2015a: 34) *type-model* analysis of A-splits



Scale	Eurasia		Sahul		Other		*N*
	+fit	−fit	+fit	−fit	+fit	−fit	
1 > 2 > 3 > N	1.74	1.03	0	0	0	0	2.77
SAP > 3/N	1.49	0	0	0	0	0	1.49
SAP > 3 > N	1.51	0	0	0	0	0	1.51
SAP > 3 > N-high > N-low	0.32	0.01	0	0	0	0	0.33
Pro > N	1.51	0	2.29	0.1	0.52	0.47	4.89
Pro/N-high > N-low	1.56	0.1	1.62	0.05	0.02	0.5	3.86
nsg > sg	1.05	1.69	0	0	0	0	2.74
sg > nsg	0	1.48	0	1	0	0	2.48

The first column of Tables 14 and 15 lists the scales that were tested as possible candidates for universal referential hierarchies. As can be seen, each of these scales requires that the manifold language-specific referential categories (like the 3SG.PRO.NHUM category from above) are subsumed under a more general category (like "3" in the first scale or "3/N" in the second). The figures in the remaining columns indicate how many taxa (large and small) were estimated to be significantly biased in the direction predicted by each scale ("fit") or against it ("-fit"). As far as we can tell from the raw data, there is a total of 80 taxa in BWZ's database that show some kind of P-split, so the figures in the last column of Table 2 should be compared against this overall number. For example, out of the 80 taxa, only about 7 show a significant bias towards being driven by the SAP > 3/N scale, i.e. where speech-act participants (= SAP or 1,2) behave differently with regard to case marking from third-person referents (3/N). Conversely, this means that the vast majority of taxa were estimated not to show a significant bias along this scale. Crucially, for the 7 taxa that *are* estimated to be biased, there is no clear signal in favour of the proposed scale, as in each of the three macro areas compared here, the number of scale-conforming taxa is counterbalanced by a roughly equal (or even higher) number of scale-violating taxa. According to BWZ, then, this provides clear evidence against a universal effect of an alleged SAP > 3/N scale, and similar conclusions carry over to most other scales they test: The overall number of biased taxa is extremely small in each case, and the counterevidence is in the same range as the fitting cases (except for Pro > N and for Pro/N_{high} > N_{low}).