The discourse basis of ergativity revisited

Geoffrey Haig, Stefan Schnell

Language, Volume 92, Number 3, September 2016, pp. 591-618 (Article)

Published by Linguistic Society of America
DOI: *https://doi.org/10.1353/lan.2016.0049*

THE DISCOURSE BASIS OF ERGATIVITY REVISITED

GEOFFREY HAIG
University of Bamberg

STEFAN SCHNELL
University of Melbourne

Since Du Bois's (1987b) seminal paper, ergative alignment in morphosyntax has been claimed to correlate with a characteristic constellation of argument realization in discourse: both intransitive subjects (S) and transitive objects (P) serve to introduce new referents via full noun phrases (NPs), while transitive subjects (A) are dispreferred for this function and are thus mostly realized as pronouns or zero (e.g. Dixon 1995, Du Bois et al. 2003, Goldberg 2004). This ergative patterning in discourse is generally accounted for in terms of information-management strategies employed by speakers in dealing with the cognitive demands of introducing and monitoring referents in discourse. These claims have recently been questioned by Everett (2009), whose data (English and Portuguese) show no support for the claimed ergative bias in discourse and raise doubts about explanations in terms of information management. The present article subjects the claims of an ergative bias in discourse to more rigorous testing, drawing on the largest database compiled to date (nineteen spoken-language corpora from fifteen typologically diverse languages), and assesses the explanatory frameworks. We find that, with the exception of Du Bois's original Sakapultek data, there is very little evidence for the postulated ergative pattern in natural spoken-language discourse crosslinguistically. Although our findings do confirm low levels of full NPs in the A role (Du Bois's 'Non-lexical A' constraint), we concur with Everett (2009) that the semantic feature [±human] provides an empirically more sound and conceptually more economical account than earlier explanations framed in terms of information management. Finally, we address the plausibility of emergentist claims for a diachronic link between ergative alignment in morphosyntax and information flow in discourse. The raw data used in this article and extensive exemplification of the methodology employed are available as online supplementary materials.*

Keywords: ergativity, preferred argument structure, corpus-based typology, discourse structure, information structure, language variation, emergent grammar

1. INTRODUCTION. In a landmark article published almost thirty years ago, Du Bois (1987b) presented a set of proposals regarding the actual realization of arguments of different types of predicate in connected discourse. Based on an investigation of spoken discourse in Sakapultek, a Mayan language with ergative morphology,[1] Du Bois established that INTRANSITIVE SUBJECTS (S) and TRANSITIVE OBJECTS (P) are freely realized

* The research reported in this article was supported by a Discovery Early Career Researcher Award from the Australian Research Council to Stefan Schnell (grant DE120102017, 2012–2015) and through internal research funding by the University of Bamberg. Schnell also thanks the Vera'a community from North Vanuatu for their support and engagement in the ongoing collaborative documentation of the Vera'a language, which was supported by two grants in the VolkswagenStiftung's DoBeS program (grants II/81 898 and II/84 316, 2006–2012). Most of this research was undertaken while Schnell was a member of the Centre for Research on Language Diversity at La Trobe University; thanks for helpful comments and discussion go to Anthony Jukes, Stephen Morey, and Pavel Ozerov. Earlier versions of this article were presented on various occasions, and we would like to thank the audiences at the following venues: LACITO, Paris Sorbonne (May 2012); Linguistic seminar, School of Languages and Linguistics, University of Melbourne (July 2012); University of Konstanz (May 2013); and the Ludwig-Maximilian University, Munich (2013). We thank the following colleagues for collaboration on the GRAID annotation scheme: Johanna Andrees, Ulrike Mosel, Florian Siegl, Claudia Wegener, Dagmar Jung, Hanna Thiele, and Nils Schiborr. For assistance with data handling and statistical analyses we thank Jenny Herzky and Nils Schiborr. The raw data is deposited at the Multi-CAST website hosted by the Language Archive at the University of Cologne (LAC; https://lac.uni-koeln.de/en/); our thanks to Felix Rau and Jonathan Blumtritt from LAC for technical assistance. The online supplementary materials can be accessed at http://muse.jhu.edu/article/628202/pdf. All remaining errors and shortcomings are our responsibility.

[1] To avoid confusion, we adopt the more recent spelling *Sakapultek* throughout, though some sources cited use *Sacapultec*.

by full lexical noun phrases (NPs). In contrast, TRANSITIVE SUBJECTS (A) tend to be realized by some nonlexical form, either a pronoun or zero. This contrast between S/P and A has its counterpart in the ergative patterns of argument encoding in morphosyntax, and Du Bois postulates a diachronic relationship between the two: ergative argument encoding is historically linked to the ergative distribution of full NPs in discourse. Although originally formulated on the basis of the Sakapultek data, this set of hypotheses has since generated a considerable body of research on genetically and areally diverse languages (see bibliography in Du Bois et al. 2003, Everett 2009, and below).

Du Bois's findings can be interpreted as support for a broadly emergentist approach to grammar, according to which the core structures of morphosyntax are 'shaped by discourse in an ongoing process' (Hopper 1998:156). For example, accusative alignment in morphosyntax, involving the formal unity of the A and the S roles, has a demonstrable counterpart in discourse, expressed by, among others, Chafe's 1994 'light subject constraint': the overwhelming majority of S and A arguments (subjects) in natural English discourse (97% in Chafe's corpus) are pronominal. Most subjects have given referents (see Prince 1981:243, 250 for findings from English discourse data), functioning pragmatically as topics within their proposition. Thus the subject relation in grammar is considered to result from the grammaticalization of the topic function of S and A in discourse (cf. Givón 1979:83–85 for a summary of this view). It is, however, less obvious how ergative alignment in morphosyntax, involving the formal unity of intransitive subjects (S) with transitive objects (P), can be likewise derived from discourse considerations. But Du Bois's research appeared to demonstrate that ergative alignment also has its counterpart in discourse: the formal identity of S and P in ergative systems of morphosyntax is matched by their shared functionality in terms of information management, namely as the preferred hosts for new information. The A role, by contrast, exhibits a distinct development into an 'ergative' category, since it is strongly dispreferred for new information and is thus correspondingly seldom expressed through a full NP. The claim was originally articulated in Du Bois 1987b on the basis of text data from Sakapultek, but in the literature it is regularly cast in more general terms:

> Roughly, the claim is that in spontaneous discourse, the distribution of nominal referential forms (such as full lexical noun phrases or pronouns) across the various syntactic positions (subject, object, oblique) is systematically skewed. Speakers freely realize full lexical noun phrases in intransitive subject position or transitive object position, but strongly avoid placing them in transitive subject position. (Du Bois 2003b:48)

For Du Bois then, the demands of communication in actual discourse constitute the arena for two competing motivations to work themselves out: the tendency that unites S and A as the syntactic anchors for the topic role, and the tendency to link S and P as the preferred hosts for new referents. In some languages (or language families), the topic function that is shared by S and A yields a shared propensity for nonlexical expressions here and crystallizes to yield accusative alignment in grammar; in others, the competition goes the other way, and the shared newness and lexicality of S and P are reflected through the emergence of ergative alignment in morphosyntax. Thus, both major alignment systems attested in the morphosyntax of the world's languages appear to reflect universal traits of natural discourse that ultimately shape 'the most fundamental structures of a language's grammar' (Du Bois 2003b:83). This claim is of considerable theoretical import for research on alignment in morphosyntax, of which the ergative type has proved to be theoretically more challenging for unified approaches to semantic/syntax mapping and to diachronic syntax. Du Bois's work posits a hitherto unnoticed link between statistical patterning in natural discourse and alignment types,

grounded in empirically observable facts of natural spoken discourse. Taken together, the tension between the competing motivations of topic continuity, which favors accusativity, and the introduction of new referents, which favors ergativity, provides a simple and elegant account of the two dominant alignment types in the morphosyntax of the world's languages, and thus represents an undeniably attractive hypothesis.

Just how ergative alignment in morphosyntax might actually grammaticalize out of an ergative bias in discourse remains a contested issue, and solid diachronic evidence in support of such a scenario is hard to come by (cf. Harris & Campbell 1995:251–55, and see Queixalós & Gildea 2010, n. 20, for critical assessment). We address this issue in §6 below. But the main thrust of this article pursues a more fundamental goal: we aim to demonstrate that in natural connected discourse, outside of Du Bois's Sakapultek data, there is very little evidence for a 'discourse basis' of ergativity at all: S and P simply do not cluster to the extent that would justify the assumption of ergative patterning in discourse. A second goal is to examine to what extent the observed tendencies in argument realization in discourse can indeed be related to considerations of information management, or whether other factors provide simpler and more robust explanations.

More recent studies by Haspelmath (2006), Kumagai (2006), Everett (2009), and Kibrik (2011:171–72) have already raised doubts about the proposed unity of S and P; but apart from Kumagai 2006 on English and Everett 2009 on English and Portuguese, these studies bring little additional data to bear on the matter. In what follows, we review the claims through a combination of reassessment of existing studies and the analysis of extensive additional text data from a further five languages, which together constitute the largest database yet compiled on the subject. We demonstrate that within the context of this larger database, the discourse basis of ergativity clearly attested in Du Bois's original Sakapultek data appears to be a very isolated phenomenon, and the original explanation in terms of universal strategies of information management does not stand up to closer scrutiny. Thus although, like Du Bois, we believe it is possible to discern robust crosslinguistic regularities in the way grammatical categories are distributed in discourse, in itself a remarkable discovery, we differ in the details of the patterns and the kinds of explanations that account for them.

The article is organized as follows: first, we outline Du Bois's original findings on ergativity in Sakapultek discourse and his explanations thereof (§2), and then introduce the data set underlying this study and illustrate procedures of data analysis (§3). Section 4 reevaluates the claimed unity of S and P, a pivotal aspect of the proposed ergative nature of discourse, against our data, focusing on the claimed role of S arguments as entry points for new information. In §5, we turn our attention to the tendency to avoid lexical arguments in A function ('Avoid lexical A'). We show that, although this is indeed a robust tendency, simpler accounts given in terms of animacy appear to be empirically superior to those given in terms of information management. Finally, §6 recapitulates the main findings and returns to the broader issue of whether statistical patterning in discourse can be invoked to explain the hard facts of grammar, and which mechanisms may be involved.

2. FRAMING THE QUESTION: PREFERRED ARGUMENT STRUCTURE AND THE DISCOURSE BASIS OF ERGATIVITY. Du Bois's notion of preferred argument structure has been developed in a number of publications over many years; we draw here largely on the more recent formulation in Du Bois 2003a. Du Bois suggests that the realization of arguments in discourse is systematically regulated by the following two constraints, which together conspire to yield a characteristic profile, PREFERRED ARGUMENT STRUCTURE (PAS).

(1) QUANTITY CONSTRAINT: Avoid more than one lexical core argument per clause.
(2) NONLEXICAL A CONSTRAINT: Avoid lexical A, that is, expressing the A function through a lexical NP.

These constraints are claimed to be universally operative in discourse, regardless of the alignment of argument-encoding strategies attested in any given language. Crucially, both are violable, operating at the level of statistically significant tendencies across stretches of connected discourse rather than grammaticality constraints on isolated clauses. Thus, isolated clauses that violate either the quantity constraint or the nonlexical A constraint, or even both, are 'grammatical' in probably all languages. Consider the examples in 3a and 3b.

(3) a. Njûchi zi-ná-lu-ma alenje
bees.A 3PL.A-PST-bite-INDIC hunters.P
'the bees bit the hunters' (Chichewa; Bresnan & Mchombo 1987:744)
b. Ngarrka-ngku ka marlu panti-rni
man-ERG.A AUX kangaroo.P spear-NPST
'the man is spearing the kangaroo' (Warlpiri; Hale 1982:221)

Both sentences contain two full NPs ('lexical', in Du Bois's terms) as the transitive subject A and the transitive object P, hence violating the quantity constraint. And in both, A is lexical, hence violating the nonlexical A constraint. But although clauses such as 3a and 3b are undeniably grammatical, all investigations of connected discourse we are aware of reveal that clauses with these characteristics are exceedingly rare in actual usage. Du Bois (2003a:35) reports figures for discourse in five languages where the number of clauses with two lexical arguments does not exceed 10% of the total clauses in any of the languages, a tendency confirmed by Everett (2009). Likewise, with regard to the nonlexical A constraint, clauses with lexical transitive subjects A do not exceed 25% of the total transitive clauses in all data we are aware of to date—with the exception of a single highly unusual text, which we discuss in §5.2. There thus seems little reason to doubt that 1 and 2 are indeed remarkably robust tendencies that together contribute to the way discourse across typologically diverse languages is shaped.

However, the constraints in 1 and 2 are not the whole story. Du Bois (1987b) analyzed a corpus of eighteen Pear story retellings in Sakapultek (see Chafe 1980), coding all realizations of A, S, and P for lexical realizations (full NP) as opposed to nonlexical (pronominal, or zero-anaphor). Table 1 provides the percentages of lexical realizations in A, S, and P, respectively (the percentages represent proportions of lexical expressions in each individual role, rather than the respective proportions of roles across all lexical arguments; see the discussion in §3 below; it is simply coincidence that the percentages here add up to 100%).

	A	S	P
TOTAL ARGUMENTS	180	262	177
% LEXICAL	6.1%	48.1%	45.8%

TABLE 1. Percentages of lexical realizations of core arguments in Sakapultek (Du Bois 1987b:822).

In the Sakapultek texts, only about 6% of the available A arguments are lexical, thus confirming the nonlexical A constraint. The relatively low lexicality of arguments in A function is complemented by an equally high proportion of lexical arguments in both S and P functions. Thus, the tendency for A to be nonlexical is apparently matched by a concomitant tendency for S and P to pattern in a parallel fashion, with both ex-

hibiting a high proportion of lexical arguments. This is crucial for Du Bois's following conclusions:

> From the perspective of discourse distribution of grammatical types, S and P constitute a class which is set off as distinct from A. There is a natural unity in discourse to the absolutive syntactic category {S,P}; it is where full NPs may readily appear. … Thus we can say that, for Sakapultek, DISCOURSE HAS ERGATIVE SURFACE SYNTAX. (Du Bois 1987b:823, emphasis added)

According to Du Bois, this characteristic constellation of A contrasting with a unity of S and P in discourse mirrors the traditional definition of ergativity in morphosyntax. This is the 'discourse basis of ergativity'.

Before we explore this idea further, note that the ergative pattern found in the Sakapultek discourse data cannot actually be explained by the quantity and nonlexical A constraints in 1 and 2. Ergativity involves two components: the marked status of A, and the identity of S and P (Dixon 1995). While the special status of A is covered by the nonlexical A constraint, constraints 1 and 2 do not actually entail the formal identity of S and P: where A is relatively low in lexicality, this leaves various constellations of relative lexicality in S and P function, namely [S = P], [S > P], or [S < P]. In the literature, however, PAS is often equated with the unity of S and P, hence with an ergative/absolutive bias in discourse. Thus, in Du Bois's review of PAS in other languages (Japanese, Hebrew, Quechua, and French, among others), evidence for the quantity constraint and for the nonlexical A constraint is interpreted as demonstrating an 'ergative/absolutive patterning' in discourse (Du Bois 1987b:839). But as noted, the two PAS constraints in 1 and 2, confirmed in the data we have considered, are logically distinct from the discourse ergativity claim. The validity of the PAS constraints does not mean that discourse has an ergative bias.

2.1. ACCOUNTING FOR PREFERRED ARGUMENT STRUCTURE. Du Bois's explanations for PAS build on the foundations of theories of information packaging and accessibility, pioneered by Wallace Chafe and associates in the 1970s and 1980s. Speakers' choices in realizing arguments either as NP (lexical) or as pronoun or zero (nonlexical) are thus mediated by concerns of information status and management: only lexical NPs are used for arguments introducing new referents into discourse, whereas nonlexical form is confined to arguments with given referents. There is thus—it seems—a very obvious connection between argument form and information status (new vs. given).[2]

Central to Du Bois's explanation for the quantity constraint is the process of introducing new referents into discourse, which he considers to be particularly 'cognitively demanding' (2003a:38). It is therefore typically instantiated only once in any given clause, essentially echoing Givón's (1995:358) constraint against introducing more than one chunk of new information per clause.

In order to explain the nonlexical A constraint, the quantity constraint needs to be supplemented by other considerations, because it has nothing to say about which of the two functions A and P would be preferred as the lexical argument in a transitive clause. To account for the attested preference for lexical P as opposed to A, Du Bois appeals to the notion of topicality: connected discourse essentially consists of a sequence of propositions recounting the actions of a central figure or figures—the discourse topics.

[2] It is well known that the match is not perfect; specifically, reference to given discourse entities may be made by lexical expressions, depending on factors like activation status, accessibility, and others (cf. Ariel 1990, Lambrecht 1994, Kibrik 2011). The simplified given/new dichotomy is sufficient for our purposes and is widely adopted in the literature (e.g. Corston-Oliver 2003, England & Martin 2003).

Discourse topics tend to be human, and, as many authors have pointed out, the pragmatic role of topic correlates strongly with the grammatical roles of S and A. Because most mentions of topics are given, rather than new, the association of topicality with the S and A roles favors nonlexical rather than lexical expressions in these roles (Du Bois 1987b:830). Thus, avoidance of lexical A is driven by a more general tendency to avoid lexical topics. The same motivations essentially hold for the S role; however, the demands of introducing new referents into the universe of discourse are assumed to counteract these motivations in the case of S, as is discussed in §2.2. The P role, by contrast, tends to be less commonly human, agentive, and topical, and hence lexical expressions are more likely here. In Du Bois's original Sakapultek data, only 10% of all P arguments were human,[3] compared to a figure of 100% of human arguments in the A role (Du Bois 1987b:841)—we return to this factor in §5.2 below. Thus, when the quantity constraint is coupled with an appeal to topicality, an explanation for the nonlexical A constraint emerges: in a transitive clause, only one core argument will be lexical, and the typical association of A with topicality will conspire to leave A nonlexical and P as the preferred host for lexical arguments. We now turn to the question of why the S role—despite being equally likely to express topic referents—is apparently higher in lexicality than the A role.

2.2. Explaining the high lexicality of S: the role of information pressure. As shown in Table 1, in Sakapultek discourse the S role shares with the P role a common propensity to host lexical NPs. This has since been claimed to hold for discourse crosslinguistically: the S role 'welcomes lexical nouns' (Du Bois 2003a:36). But as mentioned, the high lexicality of S follows neither from the quantity constraint nor from the nonlexical A constraint. Indeed, given the typical association of the S role with topicality, we could reasonably expect the S role to be generally low in lexicality. So what is the explanation for the high lexicality of S in Du Bois's Sakapultek data? One candidate is the fact that Sakapultek has morphological ergativity. Du Bois himself does not consider this a relevant factor, claiming that discourse from nonergative languages such as Japanese, Hebrew, and English reveals patterns similar to those attested for Sakapultek discourse. The ergative morphology of Sakapultek is therefore apparently not a plausible source, and other explanations are required for the high lexicality of S.[4]

Du Bois suggests that the S role is associated with a highly specific function in 'managing information flow' (Du Bois 1987b:830). The S role is the preferred choice for introducing new human protagonists into a narrative:

> Even if a protagonist is to figure in a narrative solely as a thematic agent of actions coded with highly transitive verbs, an immediate introduction in the A role would run into problems with the Given A Constraint. However, narrators know that they do not need to get everything said in the same clause; hence it

[3] In the Multi-CAST data (cf. §3), the percentage of [+hum] P arguments is generally higher, though they do not exceed 35% in any language. The higher figures may result from the richer range of transitive verbs found in the narrative texts as opposed to the overall small number of distinct verb types in the Pear story retellings of the Sakapultek data; this would be a fruitful area for further research.

[4] Note that it is the ergative pattern in argument encoding that is, according to Du Bois, attributable to the observed discourse patterns, not vice versa. So on this view, there would actually be no reason to assume an impact of morphosyntactic alignment on the way arguments are realized in discourse. It is nevertheless undeniable that Sakapultek is an outlier in the extent to which S is realized lexically, and it is notably the only language in our sample with clear morphological ergativity. Thus, although we are inclined to attribute the outlier status of Sakapultek to other factors (see §§4 and 5 below), we cannot rule out the possibility that alignment type may be a contributing factor, as suggested in Durie 2003. More recently, a study based on English Pear story retellings (Kumagai 2006) has reopened the question of whether the ergative alignment of Sakapultek may be implicated; we return to this question in §4.3 below.

> becomes simpler to delay the expression of the transitively coded activities for the space of one clause in order to make an introduction in the S role of an intransitive clause. (Du Bois 1987b:830–31)

In other words, subjects of intransitive clauses serve as ENTRY POINTS FOR NEW REFERENTS, before they are then taken up in the flow of discourse. With the subsequent status 'given', they are regularly expressed nonlexically, as pronominal or zero anaphors, and can thus be freely combined with transitive verbs. Du Bois concludes: 'Evidence from the corpus as a whole suggests that speakers indeed follow a general pattern of intransitive introduction followed by transitive narration' (1987b:831).

This pattern is of course familiar from the introductory sections of traditional narratives, where structures such as 4a–c are more likely than 5.

(4) a. There was once upon a time an old goat (intransitive introduction)
b. who had seven little kids, (transitive narration, pron. A)
c. and loved them … (transitive narration, zero A)
(*Grimm's fairy tales*, 'The wolf and the seven young kids')

(5) Once upon a time an old goat had seven kids
(introduction and narration sandwiched into one clause, apparently cognitively costly)

Thus, although a tendency toward topical reference is common to both A and S, this tendency is counterbalanced for S by its function as an entry point for new referents. The introductory function of S apparently prevails in natural discourse, leaving S, like P, a statistically preferred position for introducing new referents, hence for lexical NPs:

> If a full NP or a new mention appears in an argument position, then it will strongly tend to appear IN EITHER S OR O [= P—*GH&SS*] POSITIONS, but not in the A position. (Du Bois 1987b:834, emphasis added)

Crucially, for Du Bois the cognitive demands of information management are most pressing under conditions of high 'information pressure', characterized by a high density of new referents in a particular stretch of discourse. It is under these conditions that the 'entry point' function of S for hosting new referents will be maximally exploited. In discourse with low information pressure, by contrast, Du Bois (1987b:835) predicts that intransitive clauses, when they are not required for purposes of introducing new referents, will not differ significantly from transitive clauses. Under these conditions, the 'frequency of new and lexical arguments in S can be as low as in A' (Du Bois 1987b: 836), and consequently no ergative discourse pattern emerges at all. However, the few available studies that have investigated the effect of information pressure find little evidence for its impact on the overall lexicality of S (O'Dowd 1990, Kumagai 2006), as discussed in §4 below.

2.3. INTERIM CONCLUSIONS. Du Bois 1987b identifies two principles to account for observed regularities in the way lexical forms are distributed across argument positions: the quantity constraint and the nonlexical A constraint ('Avoid lexical A'). Together, they constitute what Du Bois terms 'preferred argument structure', a 'preference in discourse for a certain grammatical configuration of argument realizations' (Du Bois 2003a:53). Based primarily on the data from Sakapultek, Du Bois 1987b further suggests that in discourse, the S and P roles share a common propensity to be hosts for lexical arguments, although this latter claim is not in fact entailed by the two constraints that constitute PAS. From the postulated unity of S and P arises the claimed link to ergative alignment in morphosyntax.

While we see little reason to doubt the special status of A, the corollary claim about the unity of S and P is more contentious. Du Bois's original explanation for the high number of lexical arguments in the S role rests on the claim that the S role is typically exploited as an entry point for new referents, in particular under conditions of high in-

formation pressure. In most subsequent work, however, the finer points of this analysis have fallen by the wayside and are reduced to an 'ergative' pattern in which S and P are united in contrast with A. For example, Du Bois (2003b:48; see also 2003a:36) states that speakers 'freely realize full lexical noun phrases in intransitive subject position or transitive object position, but strongly avoid placing them in transitive subject position', while Dixon (1995:211) states that S and P share an association 'with the introduction of new information'. What is surprising is that the claim of the unity of S and P continues to be maintained, even though several earlier studies have reported very divergent values for S and P (e.g. Kärkkäinen 1996). This is in sharp contrast to the repeated and robust confirmation of the low lexicality of A. There are therefore good reasons to subject the claimed unity of S and P to more rigorous testing; §§3 and 4 take up this challenge.

3. Corpus data and methodology. In this section we provide an overview of the corpus data underlying this study and briefly outline our approach to their quantitative analysis. More detailed exemplification of corpus composition, coding procedures, and quantitative analysis can be found in the online supplementary materials.[5]

Our analysis is based on a cross-language data set comprising approximately 25,000 clauses of spontaneous spoken language, taken from nineteen corpora representing fifteen different languages. For one language, English, we have drawn on five distinct corpora. The data stem from two sources: first, fourteen data sets from previously published research, and second, the Multi-CAST database (Multilingual Corpus of Annotated Spoken Texts; Haig & Schnell 2016), comprising original narrative texts from five languages. Full details of the corpora are available in the online supplementary materials (tables 1 and 2, and the appendices), and details of the coding and annotation procedures are laid out in §3 of the supplementary materials.

Research on PAS has been based on the quantitative analysis of the frequencies of S, A, and P in a given corpus. There are, however, at least two different ways of interpreting the raw data, and it is crucial to distinguish the two, as they yield quite different results. Investigations based on one method are thus not directly comparable with those based on the other, and we briefly outline the differences here.

The first possibility addresses the following question: 'How lexical is each particular argument role (as opposed to the other roles)?'. The question is thus how individual roles differ with regard to their propensity to host lexical, as opposed to nonlexical, arguments. This is the perspective behind the Sakapultek data from Table 1 above: the number of lexical expressions in each syntactic function is considered in relation to the total referential expressions of that function; thus 48.1% of all S arguments and 45.8% of all P arguments were realized as a lexical NP. The second possibility addresses a different question: 'Where do the lexical forms go in a given text?'. Here, the investigator takes the total number of lexical expressions in the entire text and calculates the respective proportions of S, A, and P arguments within that total. In other words, it aims at investigating the respective shares of S, A, and P among the totality of lexical expressions in the text. On this approach, the numbers of nonlexical expressions in the text are irrelevant for the calculations. This perspective on the data is adopted, for example, in Du Bois 2003a:37. Thus the same data can be interpreted in two different ways, each yielding different quantitative results, which are not directly comparable. While both approaches have their respective merits, throughout this study we follow the first approach, pioneered in Du Bois

[5] The online supplementary materials referenced throughout this article can be accessed at http://muse.jhu.edu/article/628202/pdf.

1987b, based on the comparison of lexical versus nonlexical expressions in each syntactic function. Where other cited sources have used the other perspective, we have either recalculated the figures (where possible) or excluded the data from this study. In §3.2 of the online supplementary materials, our approach is justified more extensively, and the quantitative effects of the different approaches are exemplified on a number of corpora.

4. TESTING THE UNITY OF S AND P: HOW ERGATIVE IS NATURAL DISCOURSE? The unity of intransitive subjects S and transitive objects P in opposition to transitive subjects A, a central claim of Du Bois (1987b, 2003a,b), has been evaluated quite critically in recent work by Haspelmath (2006) and Everett (2009). In this section we review the S = P hypothesis against the extended corpus and evaluate various explanations for the regularities we identify. The tendency to avoid lexical expressions in the A role is taken up in §5 below.

4.1. OVERALL FINDINGS. The raw figures from our complete database of nineteen corpora are summed up in Table 2. The significance of S correlating with A, S with P, and A with P is tested pairwise for each language with a Fisher's exact test, and the results are provided in Table 3.

	ROLE								
	A			S			P		
LANGUAGE	*n* LEX	ALL	% LEX	*n* LEX	ALL	% LEX	*n* LEX	ALL	% LEX
Cypriot Greek	38	243	16	88	300	29	258	483	53
English (Kärkkäinen 1996)	15	217	7	27	253	11	102	164	62
English (Kumpf 2003)	22	249	9	107	206	52	145	244	59
English (Kumagai 2006)	85	444	19	218	538	41	373	516	72
English (Everett 2009)	38	392	10	97	921	11	237	397	60
English (Schiborr 2014)	83	422	20	159	688	23	562	1,111	51
French	32	481	7	290	1,025	28	324	481	67
Gorani	16	182	9	83	301	28	100	144	69
Korean	284	2,184	13	541	2,080	26	840	2,153	39
Mapundungun	24	161	15	133	339	39	137	161	85
Northern Kurdish	46	277	17	207	527	39	232	396	59
Portuguese	27	155	17	94	257	37	138	163	85
Roviana	19	151	13	50	231	22	72	151	48
Sakapultek	11	180	6	126	262	48	81	177	46
Spanish	35	571	6	215	979	22	341	571	60
Teop	62	319	19	216	640	34	234	470	50
To'aba'ita	74	358	21	288	712	40	218	376	58
Vera'a	115	795	14	538	2,026	27	580	905	64
Yagua	26	219	12	98	445	22	73	167	44

TABLE 2. Lexicality of A, S, and P in nineteen corpora.

The overall picture is quite varied. In almost half of the corpora (nine), S correlates neither with P nor with A. Evidence for the claimed unity of S and P is clearly evident for only two of the nineteen corpora, where S and P are sufficiently similar for the remaining differences to not be significant beyond what can be expected from random distribution ($p > 0.01$): Du Bois's 1987b Sakapultek and Kumpf's 2003 English data. In four of the corpora, it is S and A that clearly correlate in the sense that differences between the two are not statistically significant in Fisher's exact tests ($p > 0.01$). Figure 1 visualizes the range of data from Table 2 in a box plot.[6] The S values for Sakapultek and English from Kumpf 2003 clearly stand out and are identified as bolded circles.

[6] Explanations for Tukey box plots: bold black line = median value (second quartile); gray box = interquartile range (IQR), extending from first to third quartile (25% of data points above and below median, re-

CORPUS	A/S	S/P	A/P
Cypriot Greek	$p < 0.01$	$p < 0.0001$	$p < 0.00001$
English (Kärkkäinen 1996)	$p < 0.5$	$p < 0.00001$	$p < 0.00001$
English (Kumpf 2003)	$p < 0.00001$	$p < 0.5$	$p < 0.00001$
English (Kumagai 2006)	$p < 0.00001$	$p < 0.00001$	$p < 0.00001$
English (Everett 2009)	$p > 0.5$	$p < 0.00001$	$p < 0.00001$
English (Schiborr 2014)	$p < 0.5$	$p < 0.00001$	$p < 0.00001$
French	$p < 0.00001$	$p < 0.00001$	$p < 0.00001$
Gorani	$p < 0.0001$	$p < 0.00001$	$p < 0.00001$
Korean	$p < 0.00001$	$p < 0.00001$	$p < 0.00001$
Mapundungun	$p < 0.0001$	$p < 0.00001$	$p < 0.00001$
Northern Kurdish	$p < 0.00001$	$p < 0.001$	$p < 0.00001$
Portuguese	$p < 0.01$	$p < 0.00001$	$p < 0.00001$
Roviana	$p < 0.1$	$p < 0.001$	$p < 0.00001$
Sakapultek	$p < 0.00001$	$p > 0.5$	$p < 0.00001$
Spanish	$p < 0.00001$	$p < 0.00001$	$p < 0.00001$
Teop	$p < 0.001$	$p < 0.001$	$p < 0.00001$
To'aba'ita	$p < 0.00001$	$p < 0.01$	$p < 0.00001$
Vera'a	$p < 0.00001$	$p < 0.00001$	$p < 0.00001$
Yagua	$p < 0.01$	$p < 0.001$	$p < 0.00001$

TABLE 3. Fisher's exact test values for pairwise testing of S, A, and P; values from Table 2.

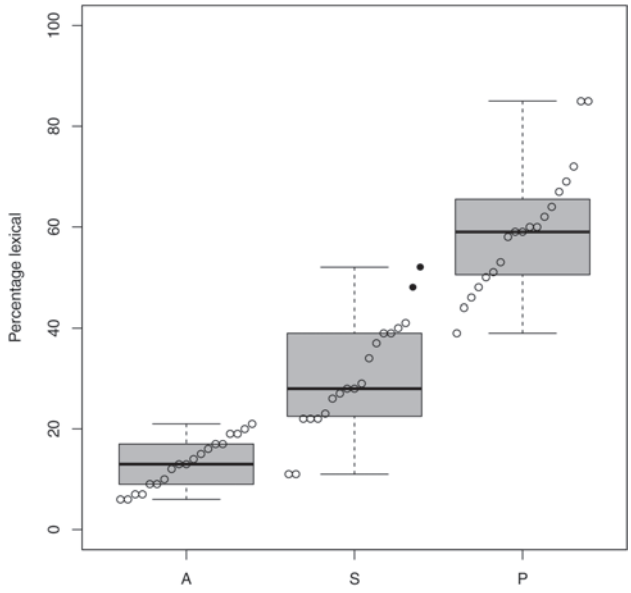

FIGURE 1. Mean percentages of lexical arguments across syntactic functions in nineteen corpora.

These findings echo the conclusions of Everett 2009, based on English and Portuguese. In a similar vein, Haspelmath (2006), in his review of Du Bois et al. 2003, examines published data from eleven languages, focusing on the pragmatic status of A, S, and P (new vs. given).[7] Although his figures are not directly comparable to ours due to

spectively); whiskers = extend to most extreme value still within 1.5*IQR away from first and third quartile respectively; outliers (in Figs. 2 and 3) = data beyond whisker range, here represented by crossed-out horizontal lines.

[7] There are certain problems in interpreting Haspelmath's data; the figures for 'S' in Inuktitut do not tally with those of the source (Allen & Schröder 2003, table 10), while those for Nepali are based on the overall

the focus on pragmatic aspects, his conclusion is strikingly similar: the statistical analysis yields significant deviations between S and P, indicating that S is not grouped with P. Instead, S 'behaves as intermediate between A and O [= P]' (Haspelmath 2006:912). Consideration of the larger database used in this article thus confirms that there is little reason to postulate a universal unity of S and P in discourse crosslinguistically, or at least there is just as much justification for assuming S = A.

Another way of approaching the issue is to calculate the relative proximity of S and A and of S and P for each individual corpus, something that Fig. 1 does not provide. This yields a single value for each individual corpus, indicating the relative proximity of S and P as opposed to S and A, which we term the DISCOURSE ERGATIVITY INDEX (DEI). To calculate the DEI for a given corpus, we take the difference in percentage points between the values for S and P, and between those for S and A, and then subtract the former from the latter. If the difference between S and P is smaller than that between S and A, we get a negative value, indicating that in this corpus, S clusters more strongly with P in lexicality, thus tending toward discourse ergativity. If, however, the difference between S and A is smaller, then we find a positive value, indicating a tendency toward accusativity. The formula for calculating the DEI is given in 6.

(6) DISCOURSE ERGATIVITY INDEX = (P − S) − (S − A)

The DEI thus provides a direct measure for the proximity of S to A relative to its proximity to P. Plotting the DEI of our nineteen corpora (see Appendix B in the online supplementary materials for absolute values) yields Figure 2.

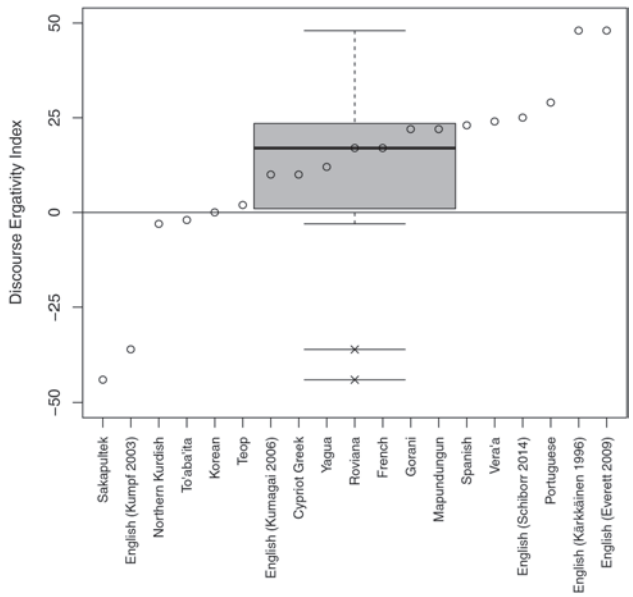

FIGURE 2. Discourse ergativity index for corpora in Table 7.

distribution of new referents, rather than the proportion per role (cf. Genetti & Crain 2003, figure 3), hence are not directly comparable to the other languages in the table. However, the overall trend identified by Haspelmath points in a similar direction to our own, confirming the lack of any general S = P alignment in discourse.

The different DEI values for individual corpora are approximately normally distributed with the mean around 17, showing an overall tendency for S and A (rather than P) to cluster in terms of lexicality.[8] Nonetheless, S = P is a possible configuration, being attested in Sakapultek and the English data of Kumpf 2003. These two corpora diverge radically from all other corpora, however, with extremely low negative DEI values, presumably due to specific, and apparently quite rare, discourse features. In most corpora of natural discourse, the proportion of lexical argument expression shows an accusative rather than ergative pattern. In fact, the English corpora of Everett 2009 and Kärkkäinen 1996 have extreme values for accusativity, but given the focus of this article on the ergativity claim, we forego in-depth discussion of these findings.[9]

The issue of the discourse basis of ergativity can thus be reduced to an investigation of the specific features that characterize a very restricted set of data. In what follows, we take a closer look at the English corpus of Kumpf 2003 and the Sakapultek corpus of Du Bois 1987b. The relevant factors are more straightforwardly identifiable for Kumpf's English corpus, since we have comparative data from other English corpora at our disposal. For Sakapultek, the questions are more challenging; we focus on the issue of information pressure and the entry-point function of the S-role.

4.2. S = P IN ENGLISH. As noted above, Kumpf's 2003 data from English diverge from most other corpora, including the other English corpora. Table 4 summarizes the data from the five English corpora.

	ROLE								
	A			S			P		
SOURCE	*n* LEX	ALL	% LEX	*n* LEX	ALL	% LEX	*n* LEX	ALL	% LEX
televized interviews, all persons (Everett 2009)	38	392	10	97	921	11	237	397	60
classroom interactions, instructors' speech only, all persons (Kumpf 2003)	22	249	8	107	206	52	145	244	60
informal conversation, all persons (Kärkkäinen 1996)	15	217	7	27	253	15	102	164	62
Pear story retellings, third person only (Kumagai 2006)	85	444	21	218	538	41	373	516	72
oral history monologue, third person only (Schiborr 2014)	83	422	20	159	688	23	562	1,111	51

TABLE 4. Lexicality of A, S, and P in five English corpora.

Kumpf's data are the only ones that exhibit anything approaching S = P. In all other English corpora, P exceeds S by at least 25 percentage points. Hence, the ergative profile in Kumpf's data must be related to factors other than structural properties of the En-

[8] An anonymous referee asks whether the corpora from the Multi-CAST data set cluster in any significant way that might reflect particularities of the coding practices. A glance at Fig. 2 shows that this is not so; the five Multi-CAST corpora—Northern Kurdish, Teop, Vera'a, Cypriot Greek, and English (Schiborr 2014)—are distributed fairly evenly among the fifteen corpora in the central data range (−5 to +25), suggesting that the unified coding scheme used for these corpora has not skewed the results in any recognizable way.

[9] Two factors conspire to yield the near-identity of S and A in the two English corpora: the first is that both investigators included first- and second-person pronouns in their counts; the second is that both corpora are based on conversational data. In data of this nature, first- and second-person pronouns are heavily represented, and they are most commonly distributed in the S and A roles, leading to generally low levels of lexicality here as opposed to the P role.

glish language itself. One is the person of the discourse referents: Kumagai (2006) and Schiborr (2014) both explicitly restrict their counts to third-person referents and exclude first- and second-person forms. As would be expected, this leads to an overall increase in the lexicality of A. Additionally, we note that Kumagai (2006) does not include zeros in his counts, which also works toward increasing the overall lexicality of A. The data from Kärkkäinen 1996, Kumpf 2003, and Everett 2009, by contrast, include first- and second-person forms and show correspondingly low levels of lexical A (see below for examples). However, although this factor may account for differing levels of lexical A, it seems to have no consistent impact on S or P.

The most likely explanation for the high percentage of lexical S in Kumpf's data lies in a peculiarity of the discourse type, teachers' explanations in science classes. Kumpf (2003:120–25) points to the exceedingly high level of lexical NPs expressing given information. In the teachers' explanations of scientific content, nouns tend to be repeated verbatim, rather than pronominalized, even though the referent may be given. This discourse strategy serves to ensure that the often complex concepts are grasped by the pupils: 'One way to maximize the salience of an entity is to mention it in full' (Kumpf 2003:123). Kumpf notes that specialist vocabulary items such as *amperage* or *chromosome* are 'often repeated, sometimes in structures that typify their specialized use' (2003:125).

One might conjecture that the frequent repetition of these items as full NPs would lead to an overall high level of lexicality in S, A, and P roles. However, the A role in Kumpf's data remains low in lexicality. We suggest that this can be accounted for by the frequent occurrence of first- and second-person pronouns, which Kumpf includes in her counts, as subjects of HAVE-clauses with presentational sense. Illustrative examples are shown in 7.

(7) a. **we have** the formula (Kumpf 2003:124)
b. **we have** a diagram on page four forty (Kumpf 2003:125)
c. if **we have** four batteries (Kumpf 2003:125)
d. and **I have** this picture (Kumpf 2003:123)
e. or are **you gonna have** a Greek nose (Kumpf 2003:123)
f. so how do **we** calculate the voltage? (Kumpf 2003:124)
g. **we have** the triangle (Kumpf 2003:124)

Clauses such as those illustrated in 7 are considered transitive; hence their pronominal subjects are exemplars of the A role. The frequent use of such clauses thus serves to decrease the overall level of lexicality of A. We surmise that the tendency toward maximal explicitness alluded to above will be most apparent in an increase of levels of lexicality of S and P, while this tendency is neutralized for the A role due to the high frequency of the constructions illustrated in 7.

As we do not have the necessary details regarding Kumpf's 2003 corpus, none of this is ultimately conclusive. However, high information pressure—postulated in Du Bois 1987b—can safely be rejected as an explanation for the near-unity of S and P in these data, since, according to Kumpf (2003), most of the lexical arguments have given, activated referents, for instance *voltage* in 7f. Note that this contrasts sharply with the Pear film retellings of Kumagai 2006, which include a large number of newly introduced referents yet show highly divergent values for S and P. Kumpf's data suggest that in fact more general considerations of genre and the semantics of the referents concerned (in particular human vs. nonhuman) may ultimately be more important than information pressure, a point we take up in more detail below.

In sum, when compared with four other corpora of spoken English, the S = P versus A pattern in Kumpf 2003 emerges as exceptional, suggesting that this pattern cannot be linked to the English language per se, and therefore reflects a conspiracy of factors including both specifics of the texts and aspects of the coding convention and quantitative analysis: first, the overall high lexicality, induced by the instructors' concern with maximal repetition and clarity; second, the inanimate nature of most of the referents (*voltage*, *this diagram*; see §5.2 on the role of animacy); third, the high frequency of presentational transitive *have*-constructions; and fourth, the inclusion of first/second-person pronouns, which evidently push down the overall lexicality of A. Due to the unavailability of the primary data, we are unable to identify the ultimate causes with any degree of certainty, but we have been able to pinpoint several likely candidates. This case further underscores the need for maximum transparency of the data and the coding procedures.

4.3. S = P IN SAKAPULTEK. Du Bois's explanation for the S = P pattern in Sakapultek is in terms of the special function of S as a 'staging area' for discourse-new referents under circumstances of high information pressure, as outlined in §2.2 above. Let us consider some of the specifics of Du Bois's corpus. The Sakapultek data consist of eighteen distinct Pear film retellings, narrated by fifteen male and three female speakers (Du Bois 1987b:812). Du Bois (1987b:826) identifies a total of 177 discourse referents introduced across the total of 458 clauses in his corpus, which means that a new referent is introduced every 2.6 clauses on average. When evaluating the information pressure of his corpus, Du Bois (1987b:834) considers only the seventy human discourse participants, new mentions of which occur every 6.5 clauses. The questions are, first, whether this figure represents 'high information pressure', and second, whether it is indeed the S role that is most responsive (Du Bois 1987b:835), in acting as the entry point for these 177 new mentions or at least the seventy new human mentions.

With regard to the first question, a comparison with Kumagai's 2006 Pear film corpus is instructive: he counts a total of 231 new mentions in the twenty distinct Pear film retellings of his English data, of which an estimated eighty-five have human reference, figures roughly comparable to the seventy in Du Bois's eighteen retellings. But Kumagai's new mentions are distributed over 1,654 clauses, yielding a rate of one new mention per 7.2 clauses, and a rate of one new human mention per nineteen clauses. Although Kumagai (2006:675) himself characterizes this as relatively high information pressure, the speakers in his data were still producing almost three times as many clauses for each newly introduced referent when compared to Du Bois's Sakapultek speakers. In comparison, then, the Sakapultek data show a quite exceptional density of new information: speakers are repeatedly introducing new referents, but scarcely elaborating on them in subsequent clauses.

Preliminary investigations of new mentions in the narrative texts of the Multi-CAST corpora indicate that the number of new mentions per clause is significantly lower than in the Pear story texts, suggesting that Pear story retellings are characterized by high information pressure when compared to, for example, narrative texts. We investigated the number of new referents in two narrative texts from the Gorani corpus (West Iranian, Iranian, Indo-European; Mahmoudveysi et al. 2012), which together made up a total of 483 clause units, a figure comparable to the Sakapultek corpus. There were a total of just seventeen introductions of new human referents in the two narratives, yielding an overall rate of around twenty-eight clauses for each new human referent. In narratives of this nature, the bulk of the text consists of strings of clauses recounting the activities of a small number of central participants—the text topics—interspersed with a relatively small number of new introductions. A corpus consisting of eighteen Pear film retellings,

by contrast, basically involves a sequence of multiple new introductions, with comparatively little narrative elaboration on the participants after they have been introduced. It should be clear that, on balance, Pear film retellings will display quite a different discourse profile from other kinds of discourse. And in the case of Sakapultek, it seems that the generally high level of information pressure triggered by the genre was pushed to the extreme, because speakers evidently recounted the story in an almost 'telegraphic' fashion, using roughly four times fewer clauses overall than the English speakers of Kumagai 2006. We assume then that, in part, the unusual findings from Sakapultek do indeed result from high information pressure operative in these texts. However, the second question to be resolved is whether—under the conditions of high information pressure evidently prevailing in the Sakapultek data—it really is the S role that is disproportionately deployed in the function of introducing the new human referents.

4.4. QUANTITATIVE IMPACT OF THE 'ENTRY-POINT' FUNCTION OF S. In Du Bois 1987b, the entry-point function is advanced as the main motivation for the high lexicality of S. Speakers apparently 'often select an intransitive verb, not necessarily for its conceptual content or semantic one-placeness, but for its compatibility with constraints on information flow' (Du Bois 1987b:831). This leads to a characteristic pattern in discourse, referred to by Du Bois as 'intransitive introduction followed by transitive narration'. Intransitive verbs, then, are functionally specialized for the function of introducing new referents. Thus, in texts where high numbers of new referents are introduced (high information pressure), we expect to find a high proportion of new, hence lexical, expressions in the S role.

The actual data available on this issue, however, while supporting a difference between S and A in terms of proportion of new referents, do not confirm a general 'entry-point' function for the S-role. Of the 177 new mentions in the Sakapultek data, fifty-eight occur in the S-function (Du Bois 1987b:826, table 6). In other words, approximately two thirds of new mentions occur outside of the S-role (most prominently in oblique roles), and within the S-role itself, less than a quarter of all arguments are new mentions. Under the assumption that fifty-eight new mentions are all expressed by lexical NPs in the S role, these account for less than half of the 126 lexical S arguments (Du Bois 1987b:822). Lichtenberk (1996) conducted a detailed study of referent tracking in narrative texts in To'aba'ita (Oceanic, Solomon Islands), which included an investigation of the syntactic distribution of entry points. Among the 712 instances of the S role in his data, just fourteen involved a new mention—less than 2% of all the S exemplars (Lichtenberk 1996: 399–401). Preliminary observations from the narrative texts in the Multi-CAST corpus suggest similarly low levels.

This suggests that the exceptionally high level of lexical S in the Sakapultek data is not entirely explainable in terms of the 'entry-point function for new referents'. Everett (2009:17–18) notes that 'no unequivocal quantitative data supporting this claim have been offered', and our tentative investigation likewise fails to yield robust support for the claim that high information pressure is primarily reflected in an increase in the lexicality of S. After all, as Thompson and Hopper (2001) point out, in English conversational discourse, the bulk of the story line is carried by intransitive verbs, and this presumably remains an important function of S even under conditions of high information pressure. As has frequently been observed in the literature, the introductory function may often be carried by functions other than S, for instance objects, left-dislocation, or nonverbal predicates (see Danes 1974, Lambrecht 1994:176ff., Prince 1998). It is puzzling that the Sakapultek Pear film retellings should differ so significantly from those of the only other Pear film retellings in our data set, the English data of Kumagai 2006. Above we noted

that Kumagai's individual retellings were, on average, four times longer than the Sakapultek Pear stories, suggesting that this may have been a source of difference. In order to test this, we analyzed the Pear film retellings of German, published in Himmelmann 1997, which yielded the results shown in Table 5.[10]

	A	S	P
TOTAL ARGUMENTS	137	212	148
% LEXICAL	8.0%	39.2%	73.0%

TABLE 5. Lexical realizations in German Pear film retellings (Himmelmann 1997).

With a DEI of 3, the German Pear film retellings fall within the expected range of weak accusativity exhibited by most of our other corpora (see Fig. 2 above). The values for S and P are close to those of Kumagai's English Pear film retellings. The main difference from Kumagai's results lies in the lower figure for the lexicality of A in the German data, which we surmise arises from Kumagai's decision to ignore all zeros in his counts. Significantly, the average length of each German retelling is around sixty-five clause units, more than twice as long as the Sakapultek figure, though not quite as long as the average of Kumagai's English data. On the whole, the comparison with the German data confirms the exceptionality of the Sakapultek data in terms of the length of the individual retellings, and we suggest that this is the most promising explanation for the remarkably high level of lexicality of the S role. While the genre of Pear film retellings itself is likely to lead to higher levels of lexicality generally, this effect would be amplified if the texts themselves were shorter. An additional factor is discussed by Stoll and Bickel (2009) in their investigation of differences in referential density and lexicality between Pear film retellings in Russian and Belhare (Sino-Tibetan, Nepal), namely culture-specific habits of narration, which impact the way referents are introduced and taken up in discourse (Stoll & Bickel 2009:553). It is therefore possible that the extreme values in the Sakapultek data reflect features of a distinct indigenous narrative culture. It is also possible that the ergative alignment of Sakapultek morphology, which distinguishes Sakapultek from the other languages of our data set, is relevant; we take up the issue of alignment type in §6 below. Finally, we should not rule out specifics of the coding practice as a further contributing factor impacting the results, or a combination of all of the above. At this point we can only speculate; a reanalysis of the original data or a replication of the original experimental procedure with other Sakapultek speakers appears to be the only principled means for resolving the issue.

4.5. THE UNITY OF S AND P IN DISCOURSE: CONCLUSIONS. This section began with a survey of the available evidence in favor of the claimed unity of S and P in discourse. Our data set revealed little evidence for this claim: if anything, we find that S correlates with A, though weakly. But for most of the languages in the data set, S cannot be meaningfully correlated with either A or P. If natural discourse displays an affinity with any alignment type, then split-intransitive would appear to be the more appropriate match (see below), not ergativity. Nevertheless, two corpora do exhibit the claimed unity of S and P. In the case of Kumpf's 2003 English data, we note that the S = P pattern is also an outlier when compared to the four other English corpora in our data set, suggestive of some highly specific factors involved in precisely this corpus. We identified particular features of the genre (instructors' explanations of scientific concepts), involving a

[10] The six retellings were produced by five female speakers and one male speaker, with an overall clause number of 382 (all taken from the appendix of Himmelmann 1997). The corpus was annotated using GRAID (Haig & Schnell 2014); that is, all S, A, and P arguments were coded for their lexical status: full NP, pronoun, or zero.

high number of inanimate NPs, which, despite being informationally given rather than new, were repeated verbatim rather than pronominalized. We suspect that coding decisions have also had an effect, but the raw data were not available to us to check. What can be said with some certainty, however, is that the high levels of lexical S do not result from high information pressure in the data, as expressly noted by Kumpf 2003.

The Sakapultek data pose a very different puzzle. We investigated the issue of information pressure, initially comparing the rates of new mentions in the Sakapultek data with narrative data, noting that the former exhibited a far higher rate of new mentions. We noted, however, that high information pressure does not necessarily lead to an increase in the lexicality of S. If these considerations are on the right track, then we can hypothesize that the postulated unity of S and P in discourse is primarily an artifact of a highly marked discourse type involving an extremely high density of introductions of new referents. Seifart 2011 shows that the beginnings of narratives are characterized by an overall high proportion of lexical NPs when compared to later stretches of the same narrative. A collection of very condensed Pear film retellings like the Sakapultek ones comes close to a collection of such 'beginnings'. These texts evidently lack the epic dimensions required for topic continuity to develop and to leave its trace in low levels of lexicality. Why the Sakapultek speakers adopted this telegraphic style of narration remains an open question.

Perhaps the most striking aspect of the S category is its stubborn resistance to consistently aligning with either A or P. Du Bois (1987b) had already noted the flexibility of S, suggesting that the differing values of S were primarily triggered by variation in information pressure. However, the results of this section suggest that this cannot be the whole story. Our own view is that the high range of values for the S category primarily reflects the lack of semantic restrictions on the S role. Intransitive predicates express a far broader range of event types than transitives, including states, inchoatives, dynamic events, and so on, so the S role is open to a broad range of referent types (to some extent this also applies to P, as pointed out in Fauconnier & Verstraete 2014). The fundamental difference between transitive and intransitive clauses is aptly summed up by Say (n.d.): 'transitive verbs are all alike, every non-transitive verb is nontransitive in its own way'. If semantics impacts lexicality, then we would expect S to display a greater range of values than A, which is what we generally find. The considerable variation in the S values could then simply reflect different subject matter, or genre, rather than information pressure. In general, the impact of semantics has not been given sufficient prominence; in the final sections we investigate these factors more closely, focusing on the A role. Since, however, our data were not analyzed for information status (new vs. given, etc.) but only for lexicality, the finer interactions between information status and animacy cannot be reliably extracted from our data and await further research.

5. EXPLANATIONS FOR THE NONLEXICAL A CONSTRAINT. Unlike the unity of S and P, the tendency to avoid lexical arguments in the A function ('Avoid lexical A') is a remarkably robust effect in all studies we are aware of (cf. Table 2 above), with the exception of one text to be discussed below. We concur with Everett (2009) that 'Avoid lexical A' is a good candidate for a quantitative discourse universal. In the literature, two possible causes have been invoked as explanations for this effect (cf. Goldberg 2004).

(8) Avoid lexical A is the result of discourse pressure and information management (Du Bois 1987b, 2003a).

(9) Avoid lexical A is a side effect of animacy and topicality (Everett 2009).

In the remainder of this section we consider the relative adequacy of each explanation in accounting for the available data.

5.1. Discourse pressure and information management. An explanation in terms of discourse pressure is linked to the more general tendency to avoid two lexical arguments in one clause. This is the suggestion that is regularly alluded to in the PAS literature. The idea would be as follows.

(10) Hypothesized link between nonlexical A and the quantity constraint: As P is more likely to be lexical generally, then if a transitive clause already contains a lexical P argument, the A is unlikely to be lexical.

This idea yields a testable prediction: in those transitive clauses where P is lexical, A arguments should be significantly less frequently realized lexically than in those transitive clauses where P is not lexical, because the quantity constraint would conspire to prevent the realization of a second lexical argument. This prediction was tested by Haspelmath (2006) and Everett (2009), who find no significant effect such that an A is more likely to be lexical when P is not lexical. Rather, the tendency to avoid lexical A holds regardless of whether P is lexical. We replicated this investigation on a much larger data set, with more than 1,800 transitive clauses from five languages (the Multi-CAST corpus); the results are provided in Table 6.

VERA'A	LEXICAL	NONLEXICAL	TOTAL	
A in clauses with nonlexical P	39	231	270	Fisher's
A in clauses with lexical P	63	416	479	$p > 0.5$
all A arguments	102	647	749	
ENGLISH	LEXICAL	NONLEXICAL	TOTAL	
A in clauses with nonlexical P	32	139	171	Fisher's
A in clauses with lexical P	37	150	187	$p > 0.5$
all A arguments	69	289	358	
TEOP	LEXICAL	NONLEXICAL	TOTAL	
A in clauses with nonlexical P	26	135	161	Fisher's
A in clauses with lexical P	29	107	136	$p > 0.5$
all A arguments	55	242	297	
NORTHERN KURDISH	LEXICAL	NONLEXICAL	TOTAL	
A in clauses with nonlexical P	15	96	111	Fisher's
A in clauses with lexical P	24	122	146	$p > 0.5$
all A arguments	39	218	257	
CYPRIOT GREEK	LEXICAL	NONLEXICAL	TOTAL	
A in clauses with nonlexical P	15	77	92	Fisher's
A in clauses with lexical P	18	117	135	$p > 0.5$
all A arguments	33	194	227	

Table 6. Lexicality of A in clauses with lexical P in the Multi-CAST corpus.

Again, quantity effects fail to materialize in these data: there is no evidence that the lexicality of A is dependent on the lexicality of P in the same clause. In other words, the tendency to avoid lexical A holds regardless of the presence or absence of an additional lexical argument in the same clause.[11] This is strong evidence against the effects of discourse pressure in motivating the nonlexical A constraint; the preference for nonlexical expression of A thus appears to be independent of the quantity constraint on information management.

[11] An additional test of the (ir)relevance of the quantity constraint for 'Avoid lexical A' would be to investigate the overall levels of lexicality, including noncore arguments. For reasons of space we restrict ourselves here to S, A, and P.

5.2. ANIMACY AND TOPICALITY. Another obvious candidate for the source of the low lexicality of A is animacy considerations, more specifically humanness [±hum], as suggested by Everett (2009). The tendency for transitive subjects to be [+hum] is well known (cf. Dahl 2000) and is reflected systematically in the grammar of inverse systems and certain kinds of split-ergative alignment (Silverstein 1976). However, it has generally not been appreciated just how pervasive this effect is crosslinguistically in discourse. Table 7 gives the percentage of [+hum] A arguments among the A arguments from those languages in our sample for which the relevant figures could be extracted.

	[+hum] A	TOTAL A	% [+hum] A
Vera'a	744	795	94
English (Schiborr 2014)	355	422	84
Teop	303	319	95
N. Kurdish	263	277	95
Cyp. Greek	220	243	91
English (Everett 2009)	360	392	92
Portuguese	135	155	87
Roviana	117	121	97
Korean	2,047	2,184	94

TABLE 7. Percentages of [+hum] arguments in A function.

It is evident that the tendency for A to be [+hum] is at least as strong as the tendency for A to be nonlexical, in fact more so. Given the obvious close match between the semantic feature of [+hum] and the syntactic A role, the question arises of whether the nonlexical A constraint may in fact be merely epiphenomenal of a much broader tendency to cast [+hum] arguments with nonlexical forms, rather than being a consequence of the A role itself.

Teasing out the effects of syntactic role from those of the semantics of [±hum] is not straightforward, given that the two correlate so strongly. We applied two methods for isolating the effects of animacy from role. First, we investigated the relative rates of lexicality in those exemplars of the A role that are NOT human. If role was the crucial factor, we would expect similarly low rates of lexicality here; if, by contrast, [±hum] was decisive, we could expect these A arguments to display significantly different rates of lexicality compared to the overall ones. This hypothesis can only be tested for statistical significance on relatively large corpora, due to the extremely low proportion of nonhuman A arguments in our data. Table 8 provides the figures from the largest corpus in the Multi-CAST data set, that of Vera'a.

	LEXICAL	NONLEXICAL	TOTAL
human	89	655	744
nonhuman	26	25	51
TOTAL	115	680	795

TABLE 8. Lexicality of human vs. nonhuman A in Vera'a.

The Vera'a data strongly suggest that it is not the A role itself that is driving the nonlexical A constraint; rather, it is the high proportion of human arguments within that role. A nonhuman A is more likely to be lexical (approximately 50%) than a human A (approximately 12%). Thus it appears that the overall strong tendency for A to be nonlexical is largely neutralized in those (few) examples where the A is nonhuman.

Further evidence for the role of human vs. nonhuman within the A role comes from a German text analyzed in Andrees 2012. The text is the commentary to an animal docu-

mentary film describing the wildlife of Finland. Due to its content, the text exhibits an exceptionally high number of nonhuman protagonists in the A role, such as wolves, bears, and birds of various kinds. This is obviously a very unusual kind of text; in the spontaneous narratives of the Multi-CAST corpus, animals do figure quite prominently, but they tend to be anthropomorphized, meaning they are capable of planned action, speech, and empathy, and were thus counted as [+hum]. In this text, however, the animals remain quite obviously animals. The text therefore provides an interesting test case for investigating the relative impacts of the feature [±hum] and of the syntactic role A.

There are numerous examples of transitive clauses in the text with nonhuman lexical As. The following are typical.

(11) a. **Über 180.000 Seen** formen dieses Land am Polarkreis,
over 180 000 lakes form this country at.the Arctic.Circle
'**More than 180,000 lakes** form this country at the Arctic Circle'
b. in dem **Bären und Wölfe** eine Zuflucht finden.
in which bears and wolves a sanctuary find
'in which **bears and wolves** find sanctuary.'
(German; Andrees 2012:Appendix 1, WS3)

(12) **Eine Braunbärenmutter** führt ihre Jungen durch den Sumpf.
a brown.bear.mother leads her cubs through the swamp
'**A brown bear mother** leads her cubs through the swamp.'
(German; Andrees 2012:Appendix 1, WS102)

The animal documentary text contains a total of 370 arguments in A, S, and P roles; Table 9 gives the levels of lexicality for the three roles in this text.

A			S			P		
n LEX	ALL	% LEX	*n* LEX	ALL	% LEX	*n* LEX	ALL	% LEX
60	89	67	125	183	68	80	98	82

TABLE 9. Lexicality of A, S, and P in German animal documentary (Andrees 2012:21).

Note first the generally high levels of lexicality in all roles (well over 50%). The most striking finding, however, is that high levels of lexicality also obtain for A (67%). Thus, there is no evidence for the nonlexical A constraint here at all. The figure 67% is of an altogether different scale from the figures of < 25% in our data (cf. Fig. 1) and in all other studies we are aware of. Nor does it appear to reflect a general feature of German, as opposed to other languages: Table 5 above shows that German Pear film retellings exhibit the typical < 25% lexical A found in all other texts. It could of course be argued that the text is not a spontaneous narrative but a scripted commentary, thus not really comparable to other texts we have considered. However, the nonlexical A constraint is robustly present in all text genres so far investigated, both written and spoken: conversational, narratives, Pear film retellings, and child language. The complete absence of this effect in the animal documentary text is therefore unlikely to be an artifact of text genre or of information pressure. Rather, it reflects the highly unusual subject matter: a connected text in which virtually all protagonists are [−hum]. We therefore conclude that the nonlexical A constraint itself is overwhelmingly, if not entirely, a consequence of the basic fact that in most kinds of discourse, transitive subjects are [+hum]. When this condition is not met, as in the rare case of the animal documentary film, no effects of the nonlexical A constraint can be observed.

The second method we apply in assessing the interaction of semantics and role is by comparing S with A, but differentiated for [±hum]. S and A obviously share some discourse functions, for example, as the favored role for topics. However, they differ con-

siderably in the relative proportion of [+hum] exemplars: as seen in Table 7, for A the figures approach 100% in the texts we have examined, while in the S role, human participants make up from 50–80% (see Appendix B of the online supplementary materials for the raw figures). There are thus sufficient exemplars of [−hum] S for a meaningful comparison based on humanness.

Recall that Du Bois's assumption has been that the fundamentally different lexicality profiles of S and A result directly from the nature of the two respective functions: 'Subject position welcomes lexical nouns [reference omitted], as long as the predicate is one-place—that is, if the subject is S rather than A' (Du Bois 2003a:36). If that is indeed the case, then we should still find significant differences between S and A when both are [+hum]. But if humanness is a significant factor, we should find human S and A patterning in a similar manner. We tested this hypothesis on the five languages from the Multi-CAST corpus as well as the Portuguese and English corpora in Everett 2009; the results are provided in Figure 3.

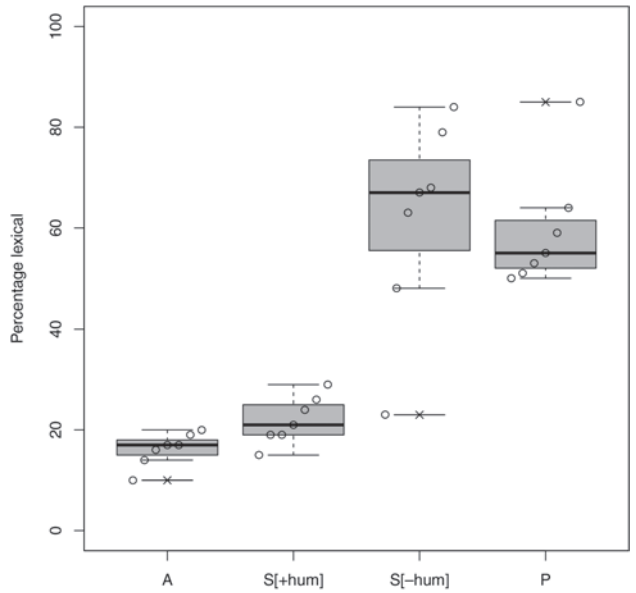

FIGURE 3. Percentage ranges for A, S[+hum], S[−hum], and P in seven corpora.

The difference between S and A is scarcely significant, once the feature of humanness is taken into account: an argument in S function that is [+hum] is just as unlikely to be lexical as an A argument is. Again, this is strong evidence against any special role of syntactic function: lexical expression is avoided with almost equal force for [+hum] S arguments as it is for A.

Figure 3 shows a very clear split of the S category, with nonhumans tending toward P, while humans tend toward A. This is obviously reminiscent of split-intransitive (or active, or semantic) alignment (Donohue 2008). Data from Acehnese (Malayo-Polynesian; Aceh, Indonesia), a language with split-intransitive alignment in morphosyntax, is relevant here. In Acehnese, the S category is split into S-Undergoers, that is, those S that are 'affected, non-volitional and non-controlling', and S-Actors, those that express a 'volitional causer or controller of states and events' (Durie 2003:177). In Acehnese morphosyntax, an S-Actor is formally treated like an A, while an S-Undergoer is treated

like a P. While the actor/undergoer distinction is not fully coextensive with our [+hum] vs. [−hum] distinction, consideration of Durie's examples suggests that the match is very close. Durie (2003) investigates levels of lexicality for the A, S-Actor, S-Undergoer, and P roles in the text 'Mouse-deer' in Acehnese, yielding the percentages of lexicality given in 13; see also figure 7 in Durie 2003.

(13) A = 18% S-Actor = 18% S-Undergoer = 57% P = 56%

It is evident that the S-Actors pattern like A, while the S-Undergoers pattern much like P. Durie (2003:190) attributes the observed patterns in discourse to the split-intransitive alignment in morphosyntax, concluding that PAS in discourse is 'fine-tuned' to the alignment type of the individual language. However, the figures for Acehnese given in 13 fall squarely in the range of lexicality illustrated in Fig. 3 above, where the S category is split into [+hum] and [−hum]. Acehnese is thus not actually very different from the other languages in our sample, despite its split-intransitive alignment. What this suggests is that this pattern is not necessarily linked to any particular alignment type, but appears to be a robust crosslinguistic effect in discourse, clearly evident in our corpus of languages with mostly nominative/accusative alignment. Although more data from languages with different alignment types are necessary to clarify the possible impact of alignment type, we can state with some confidence that the brute semantics of humanness is sufficiently powerful to impact levels of lexicality of S, regardless of alignment type.

5.3. Conclusions: what drives the nonlexicality of A? The results of the preceding sections lend further support to those of Everett (2009): first, we find no evidence for the predicted effect of the quantity constraint in explaining the low lexicality of A. We then turned our attention to the role of the humanness feature: given that over 90% of A arguments in natural texts are [+hum], it can reasonably be asked whether this fact alone accounts for the low lexicality of A. We brought two kinds of evidence to bear on this question: first, we showed that for those (few) A arguments that are not [+hum], the effects of the nonlexical A constraint fail to materialize. Second, we compared [+hum] S with A and found no significant difference. In other words, the postulated differences between S and A appear to be an artifact of the higher rates of [+hum] arguments in the A role, and therefore need not be related to any particular discourse functions apparently associated with these roles. These results suggest that there is no 'avoidance' of or 'constraint' against the expression of A as lexical in the sense of an online strategy of information management in discourse.

In conclusion, it appears that the low lexicality of A follows quite naturally from more general tendencies related to subjecthood (Chafe's 1994 'light subject constraint') and the semantics of [±hum]. Thus the apparently marked behavior of the A role, another cornerstone of the ergativity claims, does not arise through the demands of information management. Instead, it is an epiphenomenal by-product of two well-documented and robust tendencies: the pervasive tendency for transitive subjects to be [+hum], and the pervasive tendency for all subjects (S or A) to be topical, hence given information. We propose that rather than assuming a constraint against avoiding lexical arguments in the A role, we can formulate a more general tendency that also accounts for the impact of the [±hum] feature.

(14) A and S, if they refer to human referents, are seldom lexical.

Or, for those languages with an S/A subject relation ('pivot', in Dixon's 1995 terminology), it can be formulated as in 15.

(15) Human subjects are rarely lexical.

6. CONCLUSIONS AND OUTLOOK. The claim that, crosslinguistically, connected discourse exhibits a characteristically ergative profile, such that intransitive subjects and objects pattern alike, has been maintained in a number of publications over the last twenty-five years. The observed patterns have been interpreted as evidence for the grammaticalization of strategies of information management and avoidance of the cognitive costs involved in the introduction of new referents. Following up on Everett's (2009) critical assessment, we bring a typologically more diverse sample to bear on the issue, comprising approximately 25,000 clauses of spoken discourse. Our investigations support Everett's conclusions: there is very little evidence in favor of the claimed ergative profile in natural discourse.

With regard to the existence of a discourse basis of ergativity, our results are thus almost entirely negative: only two corpora known to us clearly display the claimed unity of S and P: Kumpf's 2003 English data, and Du Bois's 1987b Sakapultek data. While the existence of these two data sets lends credence to the possibility of the S = P unity in discourse, we were able to show that this pattern is rare and dependent on highly specific factors. Where the pattern has been claimed for broader data sets, we show that this is largely an artifact of a particular perspective on quantifying the data (cf. §3, and supplementary materials), which skews the results in favor of an increased lexicality of S. Once a broader selection of data is taken into consideration and analyzed uniformly in the manner we have suggested, much of the evidence in favor of a discourse basis for ergativity disappears.

We further investigated the original explanation for the high levels of lexical NPs in the S role, which was couched in terms of the entry function of the S role. We show that in longer texts, the entry function of S has only minimal quantitative impact in the overall sum of S exemplars, though the paucity of suitably annotated data renders this a largely unexplored avenue for future research.

We then turned our attention to explanations for the nonlexical A constraint, perhaps the most robust effect within PAS. We confirmed Everett's (2009) view that there is no significant effect of discourse pressure (the 'quantity constraint') in 'Avoid lexical A'. We then investigated the role of the feature [±hum]. Two different kinds of evidence yielded convergent results: the low rate of lexical expressions in the A role can more simply be accounted for by the high proportion (almost 100%) of [+hum] exemplars in the A role. There is no necessity to ascribe to the A role itself any particular discourse function, and thus no constraint on its formal realization as such. Rather, the low degree of lexicality of A can be interpreted as the cumulative effects of two basic factors: the general avoidance of lexical expressions for [+hum] referents, and—connected to this—the general propensity for subjects to be realized pronominally or as zero (Chafe's 'light subject constraint'). This hypothesis correctly predicts that an S that is [+hum] is also likely to also be nonlexical (cf. Fig. 3).

Finally, recall that the low proportion of lexical expressions in the A role has, with the notable exception of the animal documentary film commentary discussed in §5.2, consistently been confirmed in crosslinguistic studies of discourse. Thus one might expect that if any aspect of PAS should be reflected in the morphosyntax of natural languages, it would be the low lexicality of A. We could thus expect to find a language in which a sentence like 16a, with a nonlexical A, is grammatical, while 16b, with a lexical A, is not.

(16) a. **He** closed the door.
b. **The farmer** closed the door.

Since at least as early as Givón 1978 it has been well known that there are languages that do not permit indefinite referential NPs in the subject role, most prominently Man-

darin. But crucially, this pragmatic constraint affects both S and A. We are not aware of any language that has grammaticalized a constraint on the form, or information status, of NPs in the A role. However, constraints on the semantics of A are well attested, for example, in Acehnese (Durie 2003) or Jakaltek (Craig 1977). In Jakaltek, unlike English, it is ungrammatical to have a simple transitive clause with an inanimate A. On the assumption that the soft constraints of discourse will be reflected in the hard constraints of grammars (Bresnan et al. 2001), this again suggests that the universal factor at stake is not that of information flow/management, but the feature of [±hum] (see also Dahl 2008). But we certainly acknowledge the existence of inverse systems, where the relative animacy of A and P, generally in terms of person distinctions (first/second versus third) but in some cases also topicality, is reflected through special verbal morphology when the least-expected constellation occurs (e.g. a third-person P acting on a first-person A; see Bickel 2011:409–10). Such systems are certainly suggestive of the impact of discourse factors in shaping morphosyntax, and we expect this to be an exceedingly fruitful avenue for future research in discourse-based typology.

Where does all this leave the original claims about the role of discourse structure in shaping alignment systems in morphosyntax toward ergativity? These aspects have remained tantalizingly vague and have never been seriously tested against appropriate diachronic evidence. Du Bois 1987a provides a brief rationale for assuming a connection between discourse and morphosyntax, based on the ergative morphology of Sakapultek. In Sakapultek, one set of person-indexing affixes is used for both S and P, while a distinct set is used for A. Within the paradigm, only the indexing of S/P (Du Bois's 'absolutive' category) shows zero exponence. He suggests that this 'absolutive zero' in the paradigm is functionally motivated by the fact that S and P are precisely those arguments in discourse that are (apparently) most frequently expressed through lexical NPs, rather than by pronouns or zero. Thus 'absolutive zero' is on the one hand motivated by considerations of communicative economy, because it avoids a redundant additional indexing of third-person S/P for NPs already locally present in the clause. On the other hand, the high rate of lexical expressions means a corresponding paucity of free pronouns that might serve as the source for the relevant processes of grammaticalization into an agreement marker for S/P.

In paradigms of person indexing, zero forms for third persons are crosslinguistically common, in both ergative and accusative languages, though the strength of this tendency has yet to be established. It is conceivable that these tendencies relate to the possibility of lexical expression for third-person NPs. But whether levels of lexicality are sufficiently high in discourse to shape processes of grammaticalization is completely speculative, and, as we have shown, in most languages it is actually not the case that S is high in lexicality; in narrative texts, high lexicality (at least 50%) is generally restricted to P. Nor does the appeal to discourse considerations answer the fundamental question of why in the majority of the world's languages the shared topicality of S and A, pulling toward accusativity, should win out over the apparently competing motivation to link S and P. Finally, even if the discourse explanation has some appeal for head-marking languages such as Sakapultek, where ergative alignment is primarily reflected in indexing morphology on the predicate, it is difficult to see how this account can be applied to languages where the exponents of ergative alignment are ergative case markers on the arguments themselves (e.g. Dyirbal). What motivation is there in terms of grammaticalization of information flow/management that could lead to the A receiving special additional morphology (an ergative case marker)?

In some language families, historical data is available at sufficient time depths to reliably reconstruct alignment changes. In Iranian, for example, a shift from accusative to

ergative alignment (restricted to past tenses) can be traced across 2,500 years of historical attestation (Haig 2008). It is noteworthy that the mechanisms involved fall within the realm of commonplace morphological and phonological changes to the case and agreement systems of the languages concerned. It was a particular, and highly contingent, combination of such changes that conspired to yield ergative alignment: loss of finite verb forms in the past tenses, leaving participles as the sole carriers of past-tense propositions; various syncretisms in the case system; and the existence of noncanonical subject constructions. Together these yielded ergative structures in the past tenses. Whether this account is correct in all details is an open question, but the fact remains that the attested changes can be accounted for in terms of relatively simple and crosslinguistically widely attested morphological changes, with no obvious necessity to resort to considerations of information management in discourse. Indeed, in several Iranian languages, further shifts in the morphology have led to the reinstatement of accusative alignments, leaving even closely related languages with distinct alignments. These and similar diachronic developments speak of a more contingent approach to ergativity, according to which ergativity arises as an epiphenomenal and construction-specific constellation, through the combination of essentially independent morphological and phonological processes (cf. Haig 2010, Bickel 2011).

These remarks should not be interpreted as a general critique of emergentist or, more generally, functionalist approaches to grammar. Rather, we advocate a more discerning view, according to which certain aspects of grammatical structure can be fruitfully analyzed in terms of emergent discourse patterning, while others are less amenable to this kind of reasoning. Explanations for grammatical structure in terms of communicative functions are undoubtedly relevant for numerous phenomena, but as Newmeyer 2005 points out, functional pressures do not shape grammar directly, but only via the mediation of minimal incremental shifts that, over many generations of speakers may tip grammars toward certain constellations. Currently, a fruitful synthesis of empirical methodologies from variationist sociolinguistics (Meyerhoff 2000, 2002), corpus linguistics, and typology (Ariel 2000, 2008, Bickel 2003, Bybee 2006, 2007, Bybee & Thompson 2007 [1997], Kibrik 2011, Bickel et al. 2015) is yielding new insights into the interaction of statistical patterning in discourse, for example, with regard to the emergence of grammatical agreement. However, quantitative crosslinguistic investigations of discourse, such as the present study, are still in their infancy.

With regard to the topic of this study, namely the nature of argument structure and alignment, we are led to the conclusion that historically stable and culturally and biologically salient semantic features such as humanness provide a simpler and empirically more adequate ultimate source for both the observed discourse patterns and the crosslinguistic variation in morphosyntax. If grammar does indeed emerge through the entrenchment of discourse regularities over countless generations, we would expect only those that are sufficiently stable across different discourse types to leave any imprint, while constellations that occur only marginally, and in highly specific discourse types, are unlikely to gel in morphosyntax. Such a constellation is the postulated unity of S and P in discourse, which surfaces only under extreme conditions, as we have shown. Thus while we cannot ultimately answer the question of 'where ergativity comes from', we hope to have shown that putative regularities of information management in discourse are an implausible source.

REFERENCES

ALLEN, SHANLEY, and HEIKE SCHRÖDER. 2003. Preferred argument structure in early Inuktitut spontaneous speech data. In Du Bois et al., 301–38.

Andrees, Johanna. 2012. Motivationen für bevorzugte Argumentstruktur im Deutschen. Kiel: Kiel University bachelor's thesis.
Ariel, Mira. 1990. *Accessing noun-phrase antecedents*. London: Routledge.
Ariel, Mira. 2000. The development of person agreement markers: From pronouns to higher accessibility markers. *Usage-based models of language*, ed. by Michael Barlow and Suzanne Kemmer, 197–260. Stanford, CA: CSLI Publications.
Ariel, Mira. 2008. *Pragmatics and grammar*. Cambridge: Cambridge University Press.
Bickel, Balthasar. 2003. Referential density in discourse and syntactic typology. *Language* 79.708–36. DOI: 10.1353/lan.2003.0205.
Bickel, Balthasar. 2011. Grammatical relations typology. *The Oxford handbook of linguistic typology*, ed. by Jae Jung Song, 399–444. Oxford: Oxford University Press.
Bickel, Balthasar; Alena Witzlack-Makarevich; Taras Zacharko; and Giorgio Iemmolo. 2015. Exploring diachronic universals of agreement: Alignment patterns and zero marking across person categories. *Agreement from a diachronic perspective*, ed. by Jürg Fleischer, Elizabeth Rieken, and Paul Widmer, 29–51. Berlin: De Gruyter Mouton.
Bresnan, Joan; Shipra Dingare; and Chris D. Manning. 2001. Soft constraints mirror hard constraints: Voice and person in English and Lummi. *Proceedings of the LFG 01 Conference*, 13–32. Online: http://web.stanford.edu/group/cslipublications/cslipublications/LFG/6/pdfs/lfg01bresnanetal.pdf.
Bresnan, Joan, and Sam A. Mchombo. 1987. Topic, pronoun, and agreement in Chicheŵa. *Language* 63.741–82. DOI: 10.2307/415717.
Bybee, Joan L. 2006. From usage to grammar: The mind's response to repetition. *Language* 82.711–33. DOI: 10.1353/lan.2006.0186.
Bybee, Joan L. (ed.) 2007. *Frequency of use and the organisation of language*. Oxford: Oxford University Press.
Bybee, Joan L., and Sandra Thompson. 2007 [1997]. Three frequency effects in syntax. In Bybee 2007, 269–78.
Chafe, Wallace L. (ed.) 1980. *The Pear stories: Cognitive, cultural, and linguistic aspects of narrative production*. Norwood, NJ: Ablex.
Chafe, Wallace L. 1994. *Discourse, consciousness, and time: The flow and displacement of conscious experience in speaking and writing*. Chicago: University of Chicago Press.
Corston-Oliver, Simon. 2003. Core arguments and the inversion of the nominal hierarchy in Roviana. In Du Bois et al., 273–300.
Craig, Colette. 1977. *The structure of Jacaltec*. Austin: University of Texas Press.
Dahl, Östen. 2000. Egophoricity in discourse and syntax. *Functions of Language* 7.1.37–77. DOI: 10.1075/fol.7.1.03dah.
Dahl, Östen. 2008. Animacy and egophoricity: Grammar, ontogeny and phylogeny. *Lingua* 118.141–50. DOI: 10.1016/j.lingua.2007.02.008.
Danes, Frantisek. 1974. Functional sentence perspective and the organization of the text. *Papers on functional sentence perspective*, ed. by Frantisek Danes, 106–28. The Hague: Mouton.
Dixon, R. M. W. 1995. *Ergativity*. Cambridge: Cambridge University Press.
Donohue, Mark. 2008. Semantic alignment systems. *The typology of semantic alignment*, ed. by Mark Donohue and Søren Wichmann, 24–75. Oxford: Oxford University Press.
Du Bois, John W. 1987a. Absolutive zero: Paradigm adaptivity in Sacapultec Maya. *Lingua* 71.203–22. DOI: 10.1016/0024-3841(87)90072-6.
Du Bois, John W. 1987b. The discourse basis of ergativity. *Language* 63.805–55. DOI: 10.2307/415719.
Du Bois, John W. 2003a. Argument structure: Grammar in use. In Du Bois et al., 11–60.
Du Bois, John W. 2003b. Discourse and grammar. *The new psychology of language: Cognitive and functional approaches to language structure*, vol. 2, ed. by Michael Tomasello, 47–88. Mahwah, NJ: Lawrence Erlbaum.
Du Bois, John W.; Lorraine E. Kumpf; and William J. Ashby (eds.) 2003. *Preferred argument structure: Grammar as architecture for function*. Amsterdam: John Benjamins.
Durie, Mark. 2003. New light on information pressure: Information conduits, 'escape valves', and role alignment stretching. In Du Bois et al., 159–96.
England, Nora C., and Laura Martin. 2003. Issues in the comparative argument structure analysis of Mayan narratives. In Du Bois et al., 131–57.

Everett, Caleb. 2009. A reconsideration of the motivations for preferred argument structure. *Studies in Language* 33.1–24. DOI: 10.1075/sl.33.1.02eve.
Fauconnier, Stefanie, and Jean-Christophe Verstraete. 2014. A and O as each other's mirror image? *Linguistic Typology* 18.3–49. DOI: 10.1515/lingty-2014-0002.
Genetti, Carol, and Laura D. Crain. 2003. Beyond preferred argument structure: Sentences, pronouns and given referents in Nepali. In Du Bois et al., 197–203.
Givón, Talmy. 1978. Referentiality and definiteness. *Universals of human language: Syntax*, ed. by Joseph Greenberg, Charles Ferguson, and Edith Moravcsik, 291–330. Stanford, CA: Stanford University Press.
Givón, Talmy. 1979. From discourse to syntax: Grammar as a processing strategy. *Syntax and semantics, vol. 12: Discourse and syntax*, ed. by Talmy Givón, 81–112. New York: Academic Press.
Givón, Talmy. 1995. *Functionalism and grammar*. Amsterdam: John Benjamins.
Goldberg, Adele. 2004. Discourse and argument structure. *Handbook of pragmatics*, ed. by Laurence R. Horn and Gregory Ward, 427–41. Malden, MA: Blackwell.
Haig, Geoffrey. 2008. *Alignment change in Iranian languages: A construction grammar approach*. Berlin: Mouton de Gruyter.
Haig, Geoffrey. 2010. Alignment. *A companion to historical linguistics*, ed. by Vit Bubenik and Silvia Luraghi, 250–69. London: Continuum.
Haig, Geoffrey, and Stefan Schnell. 2014. Annotations using GRAID (Grammatical Relations and Animacy in Discourse): Introduction and guidelines for annotators. Version 7.0. Online: https://lac.uni-koeln.de/en/multi-cast-annotations-background-and-resources/.
Haig, Geoffrey, and Stefan Schnell (eds.) 2016. Multi-CAST (Multilingual Corpus of Annotated Spoken Texts). Online: https://lac.uni-koeln.de/multicast/.
Hale, Kenneth L. 1982. Some essential features of Warlpiri verbal clauses. *Papers in Warlpiri grammar: In memory of Lothar Jagst*, ed. by Stephen S. Swartz, 217–315. Darwin: SIL International.
Harris, Alice C., and Lyle Campbell. 1995. *Historical syntax in cross-linguistic perspective*. Cambridge: Cambridge University Press.
Haspelmath, Martin. 2006. Review of Du Bois et al. *Language* 82.908–12. DOI: 10.1353/lan.2006.0203.
Himmelmann, Nikolaus P. 1997. *Deiktikon, Artikel, Nominalphrase: Zur Emergenz syntaktischer Struktur*. Tübingen: Niemeyer.
Hopper, Paul. 1998. Emergent grammar. *The new psychology of language: Cognitive and functional approaches to language structure*, vol. 1, ed. by Michael Tomasello, 155–75. Mahwah, NJ: Lawrence Erlbaum.
Kärkkäinen, Elise. 1996. Preferred argument structure and subject role in American English conversational discourse. *Journal of Pragmatics* 25.675–701. DOI: 10.1016/0378-2166(95)00010-0.
Kibrik, Andrej A. 2011. *Reference in discourse*. Oxford: Oxford University Press.
Kumagai, Yoshiharu. 2006. Information management in intransitive subjects: Some implications for the preferred argument structure theory. *Journal of Pragmatics* 38.670–94. DOI: 10.1016/j.pragma.2006.02.003.
Kumpf, Lorraine E. 2003. Genre and preferred argument structure: Sources of argument structure in classroom discourse. In Du Bois et al., 109–30.
Lambrecht, Knud. 1994. *Information structure and sentence form: Topic, focus, and the mental representation of discourse referents*. Cambridge: Cambridge University Press.
Lichtenberk, František. 1996. Patterns of anaphora in To'aba'ita narrative discourse. *Studies in anaphora*, ed. by Barbara Fox, 379–411. Amsterdam: John Benjamins. DOI: 10.1075/tsl.33.12lic.
Mahmoudveysi, Parwin; Denise Bailey; Ludwig Paul; and Geoffrey Haig. 2012. *The Gorani language of Gawraju (Gawrajuyi), a village of West Iran: Texts, grammar and lexicon*. Wiesbaden: Reichert.
Meyerhoff, Miriam. 2000. The emergence of creole subject-verb agreement and the licensing of null subjects. *Language Variation and Change* 12.203–30.
Meyerhoff, Miriam. 2002. Formal and cultural constraints on optional objects in Bislama. *Language Variation and Change* 14.323–46. DOI: 10.1017/S0954394502143031.

Newmeyer, Frederick. 2005. *Possible and probable languages: A generative perspective on linguistic typology*. Oxford: Oxford University Press.
O'Dowd, Elizabeth. 1990. Discourse pressure, genre and grammatical alignment—after Du Bois. *Studies in Language* 14.365–403. DOI: 10.1075/sl.14.2.05odo.
Prince, Ellen F. 1981. Toward a taxonomy of given-new information. *Radical pragmatics*, ed. by Peter Cole, 223–55. New York: Academic Press.
Prince, Ellen F. 1998. On the limits of syntax, with reference to left-dislocation and topicalization. *Syntax and semantics, vol. 29: The limits of syntax*, ed. by Peter W. Culicover and Louise McNally, 281–302. San Diego: Academic Press.
Queixalós, Francesc, and Spike Gildea. 2010. Manifestations of ergativity in Amazonia. *Ergativity in Amazonia*, ed. by Spike Gildea and Francesc Queixalós, 1–26. Amsterdam: John Benjamins.
Say, Sergey. n.d. Bivalent verb classes in the languages of Europe: A quantitative typological study. St. Petersburg: Russian Academy of Sciences, ms. Online: https://www.academia.edu/4574509/Bivalent_Verb_Classes_in_the_Languages_of_Europe_A_Quantitative_Typological_Study, accessed March 8, 2015.
Schiborr, Nils Norman. 2014. English. In Haig & Schnell 2016. Online: https://lac.uni-koeln.de/en/multicast-english/.
Seifart, Frank. 2011. Cross-linguistic variation in the noun-to-verb ratio: The role of verb morphology and narrative strategies. Poster presented at the University of Hong Kong, July 21–24, 2011.
Silverstein, Michael. 1976. Hierarchy of features and ergativity. *Grammatical categories in Australian languages*, ed. by R. M. W. Dixon, 112–71. Canberra: Australian Institute of Aboriginal Studies.
Stoll, Sabine, and Balthasar Bickel. 2009. How deep are differences in referential density? *Crosslinguistic approaches to the psychology of language: Research in the tradition of Dan Isaac Slobin*, ed. by Elena Lieven Guo, Nancy Budwig, Susan Ervin-Tripp, Keiko Nakamura, and Seyda Özçaliskan, 543–55. London: Psychology Press.
Thompson, Sandra A., and Paul Hopper. 2001. Transitivity, clause structure, and argument structure: Evidence from conversation. *Frequency and the emergence of linguistic structure* (Typological studies in language 45), ed. by Joan L. Bybee and Paul Hopper, 27–60. Amsterdam: John Benjamins.

[geoffrey.haig@uni-bamberg.de]
[stefan.schnell@unimelb.edu.au]

[Received 22 June 2013;
revision invited 21 May 2014;
revision received 19 February 2015;
revision invited 12 July 2015;
revision received 7 August 2015;
accepted 9 September 2015]