Language Variation and Change, **5** (1993), 61–76. Printed in the U.S.A.
© 1993 Cambridge University Press 0954-3945/93 $5.00 + .00

Preferred argument structure in spoken French and Spanish

WILLIAM J. ASHBY
University of California, Santa Barbara

PAOLA BENTIVOGLIO
Universidad Central de Venezuela

ABSTRACT

This article uses the quantitative methodology of GOLDVARB to examine the variable distribution of lexical noun phrases representing core arguments of the verb in a corpus of spoken French and a corpus of spoken Spanish. It is shown that this distribution is not random, but instead conforms to a grammatically and pragmatically motivated pattern known as Preferred Argument Structure.

This article uses quantitative methodology to examine the distribution of nouns and noun phrase types in spoken French and Spanish, in light of the hypothesis of Preferred Argument Structure (PAS) developed by Du Bois (1985, 1987). We shall demonstrate that the syntactic distribution of noun phrases (NPs) in these Romance languages is far from random and is strikingly like that of Sacapultec, the ergative Mayan language studied by Du Bois.

Preferred Argument Structure is not conceived as a syntactic structure per se, but rather as a measurable discourse preference for a syntactic structure. Du Bois's theory pertains to the form and discourse role of the "core" arguments of the verb—that is, the subject and the direct object. In order to distinguish the subject of an intransitive verb from the subject of a transitive verb, Du Bois follows the practice initiated by Dixon (1979), whereby the subject of a one-argument verb is denoted by an **S**, the subject of a two-argument verb by an **A**, and the direct object by an **O**.

PAS has both a grammatical and a pragmatic dimension. The grammatical dimension, as explained by Du Bois, can be expressed by two "constraints" relating to the presence or absence of full, lexical NPs in the same clause. The first constraint, called the "One Lexical Argument Constraint," reflects the preponderance of clauses in the Sacapultec data examined by Du Bois in which only one of the core arguments is marked by a full, lexical NP (additional core arguments, however, may be expressed by pronouns or by zero).

An earlier version of this article was given at NWAVE-XX, Georgetown University, 1991, and at the Linguistics Colloquium, University of California, Santa Barbara. We are grateful for the suggestions of discussants at these forums and especially thank Sandra A. Thompson for her comments on the written version.

TABLE 1. *Dimensions and constraints of Preferred Argument Structure*

	Grammar	Pragmatics
Quantity	One-Lexical Argument Constraint	One New Argument Constraint
Role	Non-lexical A Constraint	Given A Constraint

Source: Adapted from Du Bois (1987:829).

A related grammatical constraint is that the single lexical NP tends to occur either in the **S** role or in the **O** role, but not in the **A** role. This second grammatical feature of PAS is termed the "Non-lexical A Constraint."

In its related pragmatic dimension, PAS can likewise be expressed in terms of two constraints. First, clauses tend to contain no more than one piece of new information; Du Bois refers to this tendency as the "One New Argument Constraint." Moreover, the single piece of new information tends to be encoded in the **O** role or in the **S** role, rather than in the **A** role—hence, Du Bois's "Given A Constraint." The hypothesis of Preferred Argument Structure claims, then, that the syntactic distribution of full, lexical NPs is determined by the patterns of information flow in discourse—that is, the introduction of new information and the management of given information. When speakers introduce a new referent, they are more likely to encode it as an **S** or as an **O** than as an **A**. Once the referent has been introduced, it may, of course, continue to be active over an indefinite stretch of discourse. In this case, the referent, now given, may be tracked by a pronoun occurring in any of the syntactic roles or, in some languages, by a null subject or object. The grammatical and pragmatic features of PAS are summarized in Table 1. It is important to note that Du Bois's constraints on the syntactic distribution of lexical NPs do not constitute absolute prohibitions or requirements, but rather strong patterning preferences that can be measured in natural discourse.

In addition to Du Bois's study of spoken narratives in Sacapultec, the patterns of PAS have been shown generally to obtain in spoken Korean (Lee, 1984), Japanese (Downing, 1985; Iwasaki, 1985), Chamorro (Scancarelli, 1985), Hebrew (Smith, 1987), Papago (Payne, 1987), Brazilian Portuguese (Dutra, 1987), and Mandarin (Tao, 1991). Although he gives little quantitative data, Lambrecht (1987, 1988) described the "preferred clause" of spoken French, in terms very similar to those of PAS. On the other hand, O'Dowd (1990), studying a rather unusual setting (oral instructions given during CPR training), suggested that the PAS hypothesis may not describe equally well all genres of spoken language, and no claims have been made for its relevance to written discourse. Nevertheless, its demonstrated applicability to oral narratives and to conversation in the wide variety of languages mentioned here lends credence to Du Bois's suggestion that PAS may be a universal (Du Bois, 1987:837). Kumpf's (1992) finding that second language

learners are sensitive to the constraints of PAS when speaking the target language gives further support to the notion of the universality of PAS.

The primary goal of this article is to demonstrate that spoken French and Spanish also conform to the principles of PAS. A second goal is to buttress the PAS theory with a rigorous quantitative methodology, that of GOLDVARB 2.0. Although all of the previous studies we cited have been based on a count of distributions in natural discourse, none has employed rigorous quantitative methodology, rendering their conclusions open to question.

METHODOLOGY

The data by which we tested the PAS hypothesis were derived from similar corpora of spoken French and Spanish. The French corpus was recorded in and around the French city of Tours (an important center of commerce, industry, and tourism located 225 kilometers southwest of Paris) by a single interviewer (WJA) in 1976. The Spanish corpus was recorded in Caracas, Venezuela, in 1987 by two interviewers (university students specially trained for this task). Both speech samples may be considered "careful" (Labov, 1972) and are typically monologues in the sense that the interviewers kept their participation to a minimum and followed no set questionnaire. The recording was done in the home or workplace of the consultants, and every effort was made to assure as natural a sample as possible. Both corpora include speakers representing various age groups. For this study we selected the youngest group (16 to 30 years). Twelve speakers of the 103 informants from the French corpus and 12 of the 160 from the Spanish corpus were selected. The two corpora are also stratified according to various socioeconomic levels. For this article, we chose in each case the upper and lower extremes, with six speakers at each level and an equal number of men and women. However, we shall not deal with the possible—though, in our view, unlikely—differences deriving from the socioeconomic level or sex of the speakers. Table 2 summarizes the characteristics of the speakers from each corpus.

To examine an equal amount of data from each speaker, we decided to consider the stretch of speech bounded by the beginning of the interview and the 101st main clause produced by the speaker. In addition to the first 100 main clauses, we included any embedded and combined clauses with a finite verb occurring within the stretch of speech considered. Because PAS concerns lexical NPs and pronouns, we did not include finite verbs with sentential (usually infinitival) objects or subjects. Relative clauses were also excluded. In this way, we obtained a total of 1,506 clauses from the French corpus and 1,550 clauses from the Spanish corpus.

Each token was then coded according to several variables including clause type, syntactic role, animacy, and information status. The coded data were then tabulated and analyzed with the help of GOLDVARB 2.0 (Rand & Sankoff,

TABLE 2. *Characteristics of speakers (age group = 16–30)*

	Upper	Lower
French		
Male	3	3
Female	3	3
Total	6	6
Spanish		
Male	3	3
Female	3	3
Total	6	6

1990). This computer software for the Macintosh includes two applications of Sankoff's Variable Rule Program (VARBRUL). One application generates probability weights for each variable, and the other selects the variables whose distributions are statistically significant.

RESULTS

The grammatical dimensions of PAS

We begin with a discussion of the grammatical dimensions of PAS. We shall demonstrate that, in connected French and Spanish discourse, there is a clear and consistent preference for some argument structures over others. We shall continue the practice of referring to the subject of a two-argument (transitive) verb as an **A** and the direct object of such a verb as an **O**. For reasons that will become clear shortly, we shall use **X** when referring to the subject of the copula "to be" (*être* in French and *ser* or *estar* in Spanish), reserving **S** for subjects of other one-argument verbs.[1]

The texts in Appendix 1 and Appendix 2 contain brief excerpts from the French and Spanish corpus, respectively. In each text can be found examples of tokens which illustrate the **A**, **X**, **S**, and **O** roles in both their lexical and non-lexical form. The lexical form, labeled **N**, includes all full NPs in subject and direct object position. For example, on line 7 of Text 1 (in Appendix 1), *les provinciaux* 'people in the provinces' is an NP filling the **A** role (i.e., it is the subject of the two-argument verb *aimer* 'to like'), and it is also a lexical NP; thus, it is coded **A-N**. On line 14, *on* 'we' is filling the **S** role (the subject of the one-argument verb *vivre* 'to live'), and it is a pronoun; thus, it is coded **S-P**. On line 4, *la vie en province* 'life in the provinces' illustrates an **X-N**; on line 8, *les Parisiens* illustrates an **O-N**. To take some examples from Spanish, consider Text 2 (in Appendix 2). On line 5, *mi mamá* 'my mom' is an example of an **A-N**. On line 4, *mi mamá y mi hermana* 'my

TABLE 3. *Distribution of A, X, S, and O in lexical and non-lexical form, in French and Spanish*

	N		P			
	n	%	*n*	%	Total	%
French						
A	32	7	449	93	481	24
X	87	19	375	81	462	23
S	203	36	360	64	563	28
O	324	67	157	33	481	24
Total	646	33	1341	67	1987	
		$\chi^2 = 456.05$	$p < .001$	$df = 3$		
Spanish						
A	35	6	536	94	571	27
X	65	21	252	79	317	15
S	150	23	512	77	662	31
O	341	60	230	40	571	27
Total	591	28	1530	72	2121	
		$\chi^2 = 439.75$	$p < .001$	$df = 3$		

mother and my sister' illustrates an **S-N**, and *unas muñecas* 'some dolls' on line 6 is an example of **O-N**. In addition to standard subject and object NPs, our **N** type also includes left- and right-dislocated subject and object NPs,[2] as illustrated on line 4 of Text 1 by *la vie en province*, and NPs following the French existential *il y a* 'there is/there are' or the Spanish equivalent, *hay*. Examples of NPs following these existentials are *un truc* 'one thing' on line 5 of Text 1 and *un balconcito* 'a little balcony' (line 16 of Text 2).[3] The non-lexical form, labeled **P**, includes subject and object pronouns, pronouns in left- or right-dislocations, and object clitics. French (but not Spanish) has subject clitics, as illustrated on line 2 of Text 1 by *on a* 'we have'; and Spanish (but not French) has null subjects, as illustrated on line 1 of Text 2 by *era así* '[I] was like this'. These are also included in our **P** category.

The Non-lexical A Constraint

Table 3 displays the numerical distribution of **A**, **X**, **S**, and **O** tokens in their lexical (**N**) and non-lexical (**P**) forms in the French and the Spanish corpora. Table 3 shows that in both corpora the non-lexical (**P**) form predominates in all three subject types (**A**, **S**, and **X**), but the lexical (**N**) form predominates in the **O** role. It is of particular note that subjects of two-argument verbs (the **A** role) represent roughly one-third of all subjects (481/1506 in the French corpus and 571/1550 in the Spanish corpus). Very few of them, however, are lexical NPs—only 7% (32/481) in the French data and 6% (35/571) in the Spanish data. This contrasts rather strikingly with subjects of the **X** and **S** types. For the former, 19% of the French tokens (87/462) and 21% of the

TABLE 4. *Distribution of lexical arguments among grammatical roles in Sacapultec, Brazilian Portuguese, Spanish, and French*

	Sacapultec[a]		Brazilian Portuguese[b]		Spanish		French	
	n	%	*n*	%	*n*	%	*n*	%
A	11	5.0	–	8.0	35	6.0	32	5.0
S + X	126	57.8	–	39.0	215	36.0	290	45.0
O	81	37.2	–	53.0	341	58.0	324	50.0
Total	218		–		591		646	

[a]Dubois (1987).
[b]Dutra (1987), only percentages given.

Spanish tokens (65/317) are lexical NPs. For the **S** type, 36% of the French tokens (203/563) are full NPs, whereas for Spanish the ratio is 23% (150/662). These distributions are, thus, consistent with Du Bois's Non-lexical A Constraint, summarized by the dictum, "Avoid lexical A's" (Du Bois, 1987:823).

Both the French and the Spanish data shown in Table 3 suggest a continuum in which the ratio of lexical **N** versus non-lexical **P** rises, as one moves down the columns of Table 3 from **A** to **X** to **S** to **O**. We shall return to this finding shortly, when we discuss the semantic and pragmatic dimensions of PAS.

In Table 4, we see that the low incidence of lexical NP tokens fulfilling the **A** role in French and Spanish is consistent with previous findings for Sacapultec (Du Bois, 1987) and Brazilian Portuguese (Dutra, 1987), despite the difference in genre between our data (free-flowing interviews) and those of Du Bois and Dutra (the recounting of a film shown to the consultants).[4]

The One Lexical Argument Constraint

The other grammatical constraint of PAS shown in Table 1 is the One Lexical Argument Constraint. According to this constraint, summarized by Du Bois's dictum, "Avoid more than one lexical argument per clause" (Du Bois, 1987:819), we would expect few tokens in which a lexical subject of the **A** type is found in the same clause with a lexical **O**, that is, few instances in which a transitive verb has both its subject and direct object arguments represented as lexical NPs. This does indeed prove to be the case in both the French and the Spanish data, as shown in Table 5.

Table 5 shows the distribution of nominal and pronominal arguments of two-argument verbs. Four types are possible: **A-N/O-N** (both arguments are expressed as full lexical NPs), **A-N/O-P** (the subject is a lexical NP and the object is a pronoun), **A-P/O-N** (the subject is a pronoun and the object is a lexical NP), and **A-P/O-P** (both arguments are pronouns). In the French data, we see that only 5% of tokens (23/481) with two-argument verbs have

TABLE 5. *Distribution of nominal and pronominal arguments of two-argument verbs in French and Spanish*

Type	Count	%
French		
A-N/O-N	23/481	5
A-N/O-P	9/481	2
A-P/O-N	301/481	63
A-P/O-P	148/481	31
Spanish		
A-N/O-N	13/571	2
A-N/O-P	22/571	4
A-P/O-N	328/571	57
A-P/O-P	208/571	36

TABLE 6. *Distribution of A and O according to their form, in French and Spanish*

	O-N		**O-P**			
	n	%	*n*	%	Total	%
French						
A-N	23	72	9	28	32	7
A-P	301	67	148	33	449	93
Total	324	67	157	33	481	
Spanish						
A-N	13	37	22	63	35	6
A-P	328	61	208	39	536	94
Total	341	60	230	40	571	

both a lexical subject and a lexical object. In Spanish, the figure is even lower—only 2% (13/571).

Whereas the low overall incidence of tokens with both **A-N** and **O-N** suggests that both languages are consistent with the predictions of the One Lexical Argument Constraint, French may be less sensitive to the constraint than Spanish, as shown in Table 6, which charts the distribution of tokens containing **A** and **O** in both the lexical (**N**) and non-lexical (**P**) forms. Note that for French, when the **A** role is filled by a full NP (**A-N**), 72% of the tokens (23/32) also have a lexical NP direct object (**O-N**). For Spanish, this ratio is only 37% (13/35). An examination of the 23 French tokens containing both lexical **A** and lexical **O** reveals, however, that many of the NPs coded as **O-N** are not true objects, but adjuncts of what have been called "support verbs" (Danlos, 1992; Giry-Schneider, 1987), as illustrated by *les Parisiens*

TABLE 7. *Components of high and low transitivity*

	High	Low
A. Participants	2 or more participants, **A** and **O**	1 participant
B. Kinesis	action	nonaction
C. Aspect	telic	atelic
D. Punctuality	punctual	nonpunctual
E. Volitionality	volitional	nonvolitional
F. Affirmation	affirmative	negative
G. Mode	realis	irrealis
H. Agency	**A** high in potency	**A** low in potency
I. Affectedness of **O**	**O** totally affected	**O** not affected
J. Individuation of **O**	**O** highly individuated	**O** nonindividuated

Source: Hopper and Thompson (1980:252).

ont tendance . . . 'Parisians tend [to] . . .' (lit. 'Parisians have tendency [to] . . .') in Text 1, lines 9–10. In such support verb constructions, the verb itself does little more than mark tense and aspect and introduce the predicate noun, which carries most of the semantic content. Of the 23 French tokens with both lexical **A** and lexical **O**, six contain similar locutions involving the support verbs *avoir* 'to have' or *faire* 'to do/make' + NP, as illustrated further in (1) and (2).

(1) . . . *la plupart des filles qui font médecine* ***ont envie d'être*** *pédiatres* . . . (30)[5]
'. . . most girls who study medicine **want** to be pediatricians . . .' (lit. 'have desire')

(2) . . . *si les élèves* ***feraient grève*** . . . (13, *sic*)
'. . . if the students **went on strike** . . .' (lit. 'would make strike')

It is questionable whether the NPs following these low-content verbs are **O**s at all, although they are coded as such for the purpose of this study.

Sixteen of the remaining 17 tokens with both lexical **A** and lexical **O** have relatively "low transitivity" according to the criteria set forth by Hopper and Thompson (1980:252), reproduced in Table 7. Hopper and Thompson described transitivity as a "global property" (251) of a clause that is determined by a complex of semantic, syntactic, and pragmatic properties, which "allow clauses to be characterized as MORE or LESS Transitive . . ." (253), depending on how many of the components of transitivity shown in Table 7 are present in the clause. Applying this transitivity scale to (3) and (4), for example, we see that these examples have seven of the nine features of low transitivity proposed by Hopper and Thompson. They convey nonaction and are atelic (i.e., the action is "not viewed from an endpoint"); they are nonpunctual, nonvolitional, and irreal; their **A**s are low in potency (there is no agent that is transferring an action to a patient); and their **O**s are not affected

by the action and are nonindividuated. Examples (3) and (4) exhibit only two features of high transitivity: there are two participants and they are affirmative.

(3) . . . *l'eau va* ***absorber le goût du poisson.*** (56)
'. . . the water is going to **absorb the flavor of the fish**'
(4) *Toutes les familles ont* ***un médecin généraliste.*** (30)
'Every family **has a general practitioner**'

Only one of the 23 tokens, (5), is relatively high in transitivity, exhibiting at least seven of the 10 features of high transitivity proposed by Hopper and Thompson: there are two participants, the verb denotes an action that is telic, punctual, volitional, affirmative, and real, and the **A** is relatively high in potency (Hopper & Thompson, 1980:252).

(5) *Le directeur* ***a envoyé des lettres*** . . . (10)
'The principal **sent letters** . . .'

The pragmatic dimensions of PAS

Thus far, we have been concerned primarily with the grammatical dimensions of PAS. We shall now turn to the pragmatic dimension (cf. Table 1). Here, we shall be concerned only with lexical NPs, inasmuch as pronouns are not sensitive to the pragmatic variables that reflect the pragmatic dimensions of PAS. There were a total of 646 lexical NPs that represented one of the core arguments of the clause (the **A**, **X**, **S**, and **O** roles) in the French corpus and 591 in the Spanish corpus. Each of these NPs was coded for the following semantic and pragmatic variables.

Activation state (with two possible factors: new vs. non-new). We considered as new only those full NPs characterized by Prince (1981) as "brand new." That is, those that are mentioned for the first time in the discourse and are neither evoked by a frame (Du Bois & Thompson, 1991), nor suggested by a schema (Chafe, 1987). An example of a new NP is seen on line 6 of Text 2: *unas muñecas* 'some dolls'. The dolls have not been mentioned in the previous discourse, and there is nothing in the text or in the situation that may hint at this referent. On lines 10 and 11 are two examples of non-new tokens: *esas muñecas* and *las muñecas*.

Generalizability. Following Du Bois and Thompson (1991), we coded the NPs according to whether they are generalizing or particularizing. According to Du Bois and Thompson:

> Generalizing NPs are used to refer to a class whose members are considered to be interchangeable or any instance of a substance interchangeable with any other instance . . . or to characterize certain types of verbal meanings (in cases where the NP is Predicating) . . . Particularizing NPs are used to refer to entities or a set of entities which are not considered to be interchangeable.

Consider Text 1, line 9, where *les Parisiens* is generalizing because it clearly refers to "a class whose members are . . . interchangeable." On the other hand, *Paris* (line 5) is particularizing because it has a unique referent.

Animacy. The distinction between animate and inanimate is obvious. In all but a few cases, animate NPs in our corpora refer to humans. An example of an animate NP is *mi mamá* 'my mom' on line 9 of Text 2. *Un balconcito* 'a little balcony' on line 16 is an example of an inanimate.

The Given A Constraint

Our hypothesis, which refines the seminal insight of Du Bois (1987), was that speakers do not encode new participants randomly across the four core roles of **A**, **X**, **S**, and **O**, but rather that they avoid encoding new participants in the **A** and **X** roles, preferring **O** and **S** as sites for the introduction of new participants. To test this hypothesis, we established newness of the referent (new vs. non-new) as the dependent variable for the variable rule analysis performed by GOLDVARB. The results are presented in Table 8.

Both the distributions and the probability weights established by GOLDVARB (Table 8) show that new participants tend to be generalizing rather than particularizing, inanimate rather than animate, and that the syntactic role of these new referents tends to be either **O** or **S** rather than **X** or **A**. Note that the probability weights are distributed in the same direction in both languages, with relatively small differences between the results for French and those for Spanish.[6] These results confirm our hypothesis. We can now affirm that the encoding of new information across the core syntactic roles may be represented as a continuum, represented in Figure 1.

The continuum extends from **A** (the syntactic role where new information is least often encoded) to **O** (the role where new information is most often encoded). The overall conclusion seems to be that new information is hardly ever encoded in **A**,[7] as Du Bois's Given A Constraint correctly predicted, but neither is new information usually encoded in the **X** role. Rather, it is the **O** role, followed by the **S** role, that favors new information.

The splitting of single-argument verbs, usually denoted by **S**, into two subtypes (our **X** and **S**) appears justified by the distributions that obtain in our sample. From a viewpoint of discourse, this decision also seems a sensible one because of what Dutra (1987) has already noted and named "the hybrid S category." In effect, **S** is a hybrid if it is analyzed as a unity, but **S** is, in fact, not a unified category like **A**. It is obvious that a copulative clause and a truly intransitive one accomplish very different discourse functions. Predicate nominals typically are used to talk about an entity that has already been introduced into the discourse, but speakers rarely introduce and predicate about an entity at the same time. The operation of predicating thus seems to entail the speaker's assessment of the relative ease with which the hearer can process the information about the predicated subject. It is, we believe, for this reason that **X** subjects are similar to **A** subjects—in fact, to such an extent that we would like to paraphrase Du Bois (1987) by saying, "Avoid new **X**s."

TABLE 8. *Syntactic roles, animacy, and generality as a function of newness of referent*

Factor	Count	% New	Weight
French			
Syntactic role			
O	143/324	44	.629
S	64/203	32	.517
X	11/87	13	.241
A	0/32	0	—
Animacy			
Inanimate	179/467	38	.537
Animate	40/179	22	.406
Generalizability			
Generalizing	167/438	38	(.542)
Particularizing	55/208	26	(.412)[a]
	n = 646 12 speakers	input .314	
Spanish			
Syntactic role			
O	142/341	42	.558
S	44/150	29	.506
X	12/65	18	.372
A	2/35	6	.197
Animacy			
Inanimate	175/430	41	.562
Animate	25/161	16	.339
Generalizability			
Generalizing	139/331	42	.565
Particularizing	61/259	24	.418
	n = 591 12 speakers	input .311	

[a]Not selected as significant by GOLDVARB for factor groups in parentheses.

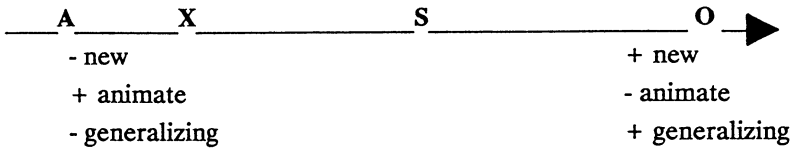

FIGURE 1. Continuum of syntactic roles according to newness, animacy, and generality.

Just as **A** and **X** are similar in their strong aversion to new referents (**A** being the more marked of the pair), **S** and **O** also have some common features, as already noted by others, in primis Du Bois. We have demonstrated that they are much more favored as sites for encoding new information than are **X** and **A**. If one accepts the arguments presented thus far, one is tempted to ask, "On what basis does a speaker choose **S** over **O** to encode a new par-

TABLE 9. *Animacy, generality, and newness of referent as a function of S versus O role*

Factor	Count	% S	Weight
French			
Animacy			
Animate	85/130	65	.746
Inanimate	118/397	30	.413
Newness			
Non-new	139/320	43	.542
New	64/207	31	.437
Generalizability			
Particularizing	65/147	44	(.559)
Generalizing	138/380	36	(.477)[a]
	$n = 527$ 12 speakers	input .382	
Spanish			
Animacy			
Animate	58/103	56	.733
Inanimate	93/388	24	.434
Newness			
Non-new	106/305	35	(.551)
New	44/186	24	(.417)[a]
Generalizability			
Particularizing	87/194	45	.644
Generalizing	63/297	21	.405
	$n = 491$ 12 speakers	input .288	

[a]Not selected as significant by GOLDVARB for factor groups in parentheses.

ticipant?" In other words, what factors are at work in the selection of an **S** or **O**?

To answer this question, we took into consideration all lexical NPs occurring in the **S** and **O** roles (eliminating tokens with NPs representing the **A** and **X** roles). The results are shown in Table 9. From the distributions and probability weights given in Table 9, one factor is clearly decisive for both French and Spanish in favoring the encoding of information as **S** rather than **O**: the animacy of the referent. This result was expected, as Du Bois and others have found the same correlation.

In addition to the factor of animacy, which met GOLDVARB's test of significance in both the French and Spanish data, the factors of generalizability and newness also appear to be determinants of the choice of **S** versus **O**. For both languages, the probability weights favoring **S** over **O** are higher with particularizing NPs than with generalizing NPs, although generalizability was selected by GOLDVARB as significant only for the Spanish data. It was expected that **O** would have a higher likelihood than **S** of encoding new (vs. non-new) NPs (cf. Figure 1), and indeed, we see in Table 9 that the probability weights for **S** are higher for both languages when the referent of the NP is non-new, even though this distribution was selected as significant only for the French data.

We see, then, an identical pattern in the distributions and probability weights for both languages. This pattern suggests that the speaker tends to encode a referent in the **S** role, rather than in the **O** role, when it is animate and particularizing, and when its information status is given. The role of animacy was selected as significant by GOLDVARB for both French and Spanish, but the significance of generalizability for French and newness for Spanish was not demonstrated. It should be noted, however, that if a given variable fails the test of significance, this does not necessarily mean that the variable has no effect on the variation; the data may simply be insufficient. To this extent, our conclusions about the role of newness and generalizability remain tentative.

CONCLUSION

In this article, we have demonstrated that Du Bois's Preferred Argument Structure pertains to both French and Spanish discourse. From a syntactic point of view, in both languages there are few clauses containing transitive verbs in which both the subject and the direct object appear as lexical NPs. When a full NP does occur, it is much more likely to appear as the subject of an intransitive verb or as the object of a transitive verb than as the subject of a transitive verb. We have found, moreover, that it is useful to distinguish true intransitives from copulatives; as with transitive clauses, full NPs rarely occur in subject position with copulas. From a pragmatic point of view, as the PAS theory predicted, clauses in both French and Spanish rarely contain more than one piece of new information. New information has the highest probability of occurrence in nouns found in object position and the next highest probability in nouns serving as the subject of intransitives. Based on the distributions found in our corpora, we have postulated that speakers may choose to encode new referents in either the **S** or in the **O** role, based on the animacy of the referent, with **S** strongly favoring animacy. Other semantic or pragmatic factors that may play a part in this choice are generalizability and activation status of the referent. A definitive statement on the effect of the latter must await further research. We hope to have demonstrated here the fruitfulness of an approach to discourse that relies on quantitative analysis.

NOTES

1. On splitting the S category, see Dutra (1987) and Bentivoglio (1988, 1989).
2. We do not deny that dislocated subjects and objects are pragmatically different from regular subjects and objects, but we do not have enough tokens to consider these differences here. See Ashby (1988) on dislocations in French.
3. The NP following these existentials is traditionally considered the object of 'to have' because in Latin it was marked for the accusative case. The French existential derives from Vulgar Latin *ibi habet* (with the impersonal pronoun *il* added beginning in Old French) and the Spanish from *habet ibi*. Indeed, this NP exhibits some features of the direct object in the grammar of modern French and Spanish (in pronominalization or interrogation, for example). Nevertheless, we

consider it to be an S, not an O, because it is the sole argument of the verb. In spoken French, *il* is optional, and no other pronominal or nominal argument is possible. Likewise, in Spanish, no other subject-like argument can be used with existential *hay* (e.g., *Hay muchas muñecas* 'There are a lot of dolls', but not **Ello hay muchas muñecas*).

4. In Table 5 the S and X roles are combined to permit comparison.
5. Numbers in parentheses give speaker identification.
6. The fact that generalizability for French was not selected as significant by GOLDVARB does not necessarily mean that this factor group is irrelevant to the variation in that language. It may simply be a reflection of insufficient data.
7. Indeed, in the French sample there were *no* new As.

REFERENCES

Ashby, William J. (1988). The syntax, pragmatics, and sociolinguistics of left- and right-dislocations in French. *Lingua* 75:29–46.

Bentivoglio, Paola. (1988). La posición del sujeto en el español de Caracas: Un análisis de los factores lingüísticos y extralingüísticos. In Robert M. Hammond & Melvyn C. Resnick (eds.), *Studies in Caribbean Spanish dialectology*. Washington, DC: Georgetown University Press. 13–23.

_____ (1989). La posición del sujeto en las cláusulas copulativas en el español de Caracas. *Actas del VII Congreso Internacional, Asociación de Lingüistica y Filología de la America Latina* [ALFAL], tomo II. Santo Domingo, RD: ALFAL, Filial Dominicana. 173–196.

Chafe, Wallace. (1987). Cognitive constraints on information flow. In Russell Tomlin (ed.), *Coherence and grounding in discourse*. Amsterdam: John Benjamins. 21–51.

Danlos, Laurence. (1992). Support verb constructions: Linguistic properties, representation, translation. *Journal of French Language Studies* 2:1–32.

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

Downing, Pamela. (1985). *Classifier constructions and referentiality marking in Japanese*. Paper presented at Conference on Japanese Language and Linguistics, University of California, Los Angeles.

Du Bois, John W. (1985). Competing motivations. In John Haiman (ed.), *Iconicity and syntax*. Amsterdam: John Benjamins. 343–365.

_____ (1987). The discourse basis of ergativity. *Language* 63:805–855.

Du Bois, John W., & Thompson, Sandra A. (1991). *Dimensions of a theory of information flow*. Unpublished manuscript, University of California, Santa Barbara.

Dutra, Rosalia. (1987). The hybrid S category in Brazilian Portuguese: Some implications for word order. *Studies in Language* 11:163–180.

Giry-Schneider, Jacqueline. (1987). *Les prédicats nominaux en français: Les phrases simples à verbe support*. Geneva and Paris: Droz.

Hopper, Paul, & Thompson, Sandra A. (1980). Transitivity in grammar and discourse. *Language* 56:251–299.

Iwasaki, Shoichi. (1985). The "given A constraint" and the Japanese particle *ga*. In S. Delancy (ed.), *Proceedings of the first Annual Pacific Linguistics Conference*. Eugene: Department of Linguistics, University of Oregon. 152–167.

Kumpf, Lorraine. (1992). Preferred argument structure in second language discourse: A preliminary study. *Studies in Language* 16: 369–403.

Labov, William. (1972). *Sociolinguistic patterns*. Philadelphia: University of Pennsylvania Press.

Lambrecht, Knud. (1987). On the status of SVO sentences in French discourse. In Russell Tomlin (ed.), *Coherence and grounding in discourse*. Amsterdam: John Benjamins. 217–261.

_____ (1988). Presentational cleft constructions in spoken French. In John Haiman & Sandra A. Thompson (eds.), *Clause combining in grammar and discourse*. Amsterdam: John Benjamins. 135–179.

Lee, Hyo Sang. (1984). *The distribution of preferred argument structure*. Unpublished manuscript, University of California, Los Angeles.

O'Dowd, Elizabeth. (1990). Discourse pressure, genre and grammatical alignment – After Du Bois. *Studies in Language* 14:365–403.

Payne, Doris. (1987). Information structuring in Papago narrative discourse. *Language* 63:783–804.

Prince, Ellen. (1981). Toward a taxonomy of given-new information. In Peter Cole (ed.), *Radical pragmatics*. New York: Academic. 223–256.
Rand, David, & Sankoff, David. (1990). GOLDVARB 2.0. Program and documentation obtained from authors.
Scancarelli, Janine. (1985). Referential strategies in Chamorro narratives. *Studies in Language* 9:335–362.
Smith, Wendy. (1987). *Preferred argument structure in Hebrew discourse.* Unpublished manuscript, University of California, Los Angeles.
Tao, Hongyin. (1991). *Functional units and organizing principles of Mandarin oral discourse.* Unpublished manuscript, University of California, Santa Barbara.

APPENDIX 1

Text 1: Excerpt from French corpus, with examples of syntactic roles in lexical NP (**N**) *and pronoun* (**P**) *form*

Interviewer:	1		Elle est jolie.
			'It's pretty [about Speaker 30's house].'
Speaker 30:	2	**A-P**	Mais, c'est . . . *On* a
	3	**O-N**	*de jolis meubles* mais . . . Mais les,
	4	**X-N**	*la vie en province c'*est tout-à-fait différent de celle
	5	**S-N**	à Paris. Et puis les . . . Il y a *un truc*, je sais pas si
	6	**X-P**	*c'*est vrai pour moi,
	7	**A-N**	mais les, *les provinciaux* en général aiment pas
	8	**O-N**	beaucoup *les Parisiens.*
	9	**A-N**	Oui, quoi . . . C'est-à-dire que les, *les Parisiens* ont
	10	**O-N**	*tendance* à croire
	11	**S-N**	que les, que les, *les Français* en général sont là
	12		pour les accueillir en vacances pendant les
	13		weekends, faire tout ce qu'ils veulent et
	14	**S-P**	puis . . . euh . . . En réalité, *nous on* vit très bien
	15		sans eux.
			'But, it's . . . We have pretty furniture but . . . But, life in the provinces is completely different than in Paris. And then the . . . There's one thing, I don't know if it's true for me, but people in the provinces in general don't like Parisians very much.' 'Yes [filler] . . . That is, Parisians tend to believe that, that French people in general are there to welcome them on weekend vacations, to do everything they want and, uh . . . The truth is, we live very well without them.'

A = subject of a two-argument verb (e.g., *aimer* on line 7)
X = subject of a copulative verb (e.g., *être* on line 4)
S = subject of a one-argument verb (e.g., *vivre* on line 14)
O = direct object of a two-argument verb (e.g., *aimer* on line 7)

APPENDIX 2

Text 2: Excerpt from Spanish corpus, with examples of syntactic roles in lexical NP (**N**) *and pronoun* (**P**) *form*

Speaker 37:	1	**X-P**	Siempre *0* era así traviesa.
	2		Que una vez también que . . .
	3		que yo estaba en mi casa no?,
	4	**S-N**	estaba (sic) *mi mamá y mi hermana.*
	5	**A-N**	Entonces, *mi mamá* tenía
	6	**O-N**	*unas muñecas* en la casa, en la . . . en la cama,
	7		unas muñecas.
	8		Entonces: "¡Ay, qué muñeca tan bonita!"
	9		Y mi mamá: "No, cuidado con esas muñecas,
	10		que esas muñecas me las regaló mi mamá".
	11	**A-P**	Entonces, *yo* . . . t . . , las muñecas, *0* las
	12		tiraba pa'arriba,
	13	**S-N**	*y las muñecas* caían
	14	**X-P**	y *eso* era tiradera de muñecas,
	15		mi hermana y yo, *0* tirábamos las muñecas.
	16	**S-N**	Entonces, o . . . hay *un balconcito* en mi casa,
	17		que . . . hace voladero, así y
	18	**A-P**	*yo* las tiré las muñecas por ahí.

'I've always been naughty like this. Once . . . I was at home, right? And there were my mom and my sister. Then my mom had some dolls in the house, on the . . . on the bed, some dolls. Then: "Oh, what a pretty doll!" And my Mom: "No, careful with those dolls, because those dolls my mother gave them to me!" And then I . . . the dolls, I used to throw them up, and the dolls would fall and that was a throw of dolls, my sister and I, we used to throw the dolls. Then, oh . . . there's a little balcony in my house, that is suspended like this and I threw the dolls from there.'