Factors Determining Spanish Differential Object Marking within Its Domain of Variation

Sonia Balasch
University of New Mexico

1. Introduction

In Spanish, the phenomenon of *differential object marking* (DOM), presence or absence of the accusative ***a***, has been addressed by numerous scholars from various perspectives (e.g., Fernández Ramírez 1986; Bolinger 1953; Fish 1967; Comrie 1979 and 1981; Weissenrieder 1985, 1990 and 1991; Dumitrescu 1997; Domínguez *et at* 1998; Torrego Salcedo 1999; von Heusinger and Kaiser 2003). This includes empirical diachronic studies documenting the increasing use of ***a*** (e.g., García and van Putte 1995[1987], Company Company 2002). The common thread in this voluminous literature has been the quest for the rules governing the presence or absence of the direct object (DO) marker ***a***. While there has been some success in identifying such rules operating fairly regularly in particular contexts, especially in formal registers, it is clear there is a broad domain in vernacular oral Spanish in which ***a***+DO and ***Ø***+DO are used interchangeably, as illustrated in (1) and (2).

(1) a. *conocí* ***a*** *la que es mi esposa ahora antes de empezar la visita médica* (MDC3MB)
I met my current wife before I started working as a medical visitor
b. [usted] *no llegó... a conocer...* ***Ø*** *ese... teniente L. que era de la radio* (MDD5FB)
you did not get to meet that lieutenant L. who was part of the radio station staff

(2) a. *entonces ya tengo que estar esperando* ***a****l autobús que, por cierto, tarda un montón* (HME-9-A)
then I already have to be waiting for the bus which, by the way, takes a while to arrive
b. *este año la llevé a ver* ***Ø*** *los monumentos en la iglesia* (MDC3FB)
this year I took her to see the monuments in the church

While a good deal of attention has been paid to explaining the use of the accusative ***a***, its inherent variability in oral speech suggests a multivariate statistical approach to assess the numerous morphosyntactic, semantic and discourse-analytic factors that have been proposed to determine its presence or absence, preceded by a careful delimiting of its contexts of variation. But, as Pensado (1995:39-40) points out, after surveying more than a hundred studies of the accusative ***a***, the study of the variable context of this accusative has been largely ignored in the literature. The present paper aims to contribute to remedying this lack from the variationist perspective (Labov 1969), following the initiative of Tippets and Schwenter (2007).

Based on a sample of 50,000 words from *Corpus de Mérida/Venezuela* (Domínguez and Mora 1995) and 79,000 from *Corpus del habla culta de Madrid/Spain* (Esgueva and Cantarero 1981), I address the following research questions: In which contexts is the variation between ***a***+DO and ***Ø***+DO possible? What factors (DO *definiteness, specificity, topicality*, *grammatical number*) significantly influence the presence or absence of ***a*** within this domain of variability? Is this contextual conditioning identical in Mérida and Madrid?

* Previous versions of this paper were presented at the 38th New Ways of Analyzing Variation conference (University of Ottawa, October 2009) and at the 5th International Workshop on Spanish Sociolinguistics (North Carolina State University, April 2010). I am in debt to the WSS5 reviewers/editors' comments as well as the invaluable feedback of several colleagues. I am solely responsible for any remaining error.

In the process of answering these questions, the present study also deals systematically with several methodological issues. How important is it to exhaustively analyze all the pertinent tokens in a corpus sample? Should *animate* and *inanimate* DOs be considered together for the purposes of multivariate statistical analysis? How important is it to eliminate categorical (non-variable) data in the statistical analysis of the variation?

The empirical results lead to reflection on the discourse theoretical basis of the factor labeled *topicality*. The way in which this has been operationalized, both in the present and previous studies, in terms of *anaphoric* and *cataphoric reference*, could just as well reflect a clarification or disambiguation role.

2. Background

Many of the deterministic "facts" adduced by theories about DOM are at best no more than quantitative tendencies in natural spoken Spanish. For example, DO *animacy* and *definiteness* do not necessarily trigger the use of the accusative ***a***, as has often been claimed (Comrie 1979, Croft 1988, Leonetti 2004, Laca 2006:424, among others). Counterexample (3) is one of many in the corpus. Hopper and Thompson (1980:256) did state that a DO being *animate* and *definite* does not suffice to induce the use of accusative ***a***; it must also be "either human or human like – and furthermore … be referential, as opposed to merely definite". The referential characteristic they invoke pertains to the existence of "a specific and extant referent" that can be associated to the nominal phrase coded as DO. But example (3), and others like it, remain clear counterexamples.

(3) *sí,* ***notaba en falta Ø*** *mi padre y mi madre* (HM-401-A)
yes, I **missed** my father and my mother

Domínguez *et al* (1998) pointed out a tendency for authors to regularize the variable use of accusative ***a*** in terms of a series of complex contexts each accepting only one of the two manifestations of this accusative marker. However, the anecdotal examples or intuitive judgments justifying these rules are not necessarily corroborated in natural language. Therefore, the efforts of García (1993), Dumitrescu (1997), Domínguez *et al* (1998), Laca (2006) and Tippets and Schwenter (2007), among others, are important in that they have attempted to infer real patterns of behavior in the use of the accusative ***a*** through the analysis of Spanish corpora. In previous quantitative work on this subject, researchers have restricted the kind of DO they extracted from the transcriptions, a pitfall the present study will try to avoid. García (1993), in studying the syntactic diffusion of the accusative ***a*** over time, excluded all occurrences of non-human or abstract accusatives (*gratitude*, *love*, and so on), as well as all personal pronoun DOs (i.e., *tú* 'you', *él* 'he'). Tippets and Schwenter (2007), in studying the variable use of accusative ***a*** in Buenos Aires and Madrid Spanish, restricted their analysis of the variation by extracting only transitive clauses occurring with verbs overtly ***a*** marked at least once somewhere in the corpus, excluding those where the verb only occurred with ***Ø***+DO.

3. Data sources

The *Corpus de Mérida* totals approximately 400 thousand words and the *Corpus de Madrid*, 144 thousand. Both corpora were collected according to the methodological guidelines of the project known as *Estudio de la norma lingüística culta de las principales ciudades de Iberoamérica y de la Península Ibérica* 'Study of the linguistic norm of the main cities in Iberoamerica and the Iberian Peninsula' (cf. Esgueva and Cantarero 1981: VIII).

The *Corpus de Mérida* consists of a total of 80 transcribed sociolinguistic interviews, which aimed at obtaining the maximum participation of the speakers interviewed. As the primary source of spoken data in which to study the variation between ***a***+DO and ***Ø***+DO, twelve of the interviews (approximately 50,000 words) were analyzed.

The 24 transcribed speech samples constituting the *Corpus de Madrid* are more varied: sociolinguistic interviews, free dialogue and secretly recorded spontaneous dialogue (cf. Esgueva and Cantarero 1981:XIII). Out of these, fourteen sociolinguistic interviews (approximately 79,000 words) are analyzed in the present paper.

One of the innovations of the present study rests in the fact that *every one* of the 877 (***a***)+DO tokens in the sample of the Mérida corpus was extracted, in view of a complete accounting of both variable and invariable contexts. For the second data set, every *animate* DO token from the Madrid corpus was extracted.

4. Exclusions protocol

Labov (1969:729) warned that: "one must decide on the number of variants which can be reliably identified, and set aside those environments in which the distinctions are neutralized for phonetic reasons." This prescription is particularly relevant to a corpus of oral speech such as those from Mérida and Madrid. In particular, capturing and transcribing whether or not a speaker uttered the accusative ***a*** before a DO with an initial ***a-*** is highly error-prone. Consequently, and following García (1993), it seems prudent to exclude such tokens, illustrated in (4), from the analysis.

(4) *yo compadezco* ***a aquéllos*** *que no van todo el verano como nosotros* (HM-67-A)
I feel sorry for those who do not go all summer as we do

In contrast to this entirely transcription-related exclusion, it is important to exclude from the analysis of variation, all tokens, and only those, where only one of ***a***+DO or ***Ø***+DO is possible in oral Spanish, due to some linguistic feature of the DO or its environment. Thus, cases such as (5) were excluded, that is, contexts where the DO is the impersonal pronoun *uno* 'one' or is a tonic personal pronoun. In these cases, the use of ***a*** is categorical (required) instead of optional, and has been so since the early stages of Spanish (cf. Martín Zorraquino 1976, according to Ariza (1989:211); García (1993:4); Pensado (1995:19); Company Company (2002:207); von Heusinger 2008).

(5) *toca ir para que lo vean* ***a*** *uno* (MDC3FB)
one must go so that [they] can see you

Clauses like (6), where variation entails change of meaning, where excluded. For example, *¿y* ***a*** *su nieto lo enseña?* 'and do you teach your grandson?' and *¿y* ***Ø*** *su nieto lo enseña?* 'and does your grandson teach you?' are not semantically equivalent and are not interchangeable; obvious when these clauses are translated to English, but equally obvious to Spanish speakers. Likewise, idiomatic expressions like (7) were excluded. There is no room for variation in these clauses: the use of the accusative ***a*** is required or prohibited.

(6) *¿Y* ***a*** *su nieto lo enseña ?* (MDC3MB)
And do you teach your grandson?
(7) *él… jalaba mucho la caña, bebía mucho* (MDD5FB)
he… pulled the cane a lot, he drank quite a lot

Finally, all verb-object lexicalized expressions in which non-tracking (non-referential) DOs are involved in complex predicates were excluded from the analysis (cf. Cano Aguilar (1981:320), Chafe (1994:110), Algeo (1995:204), Thompson 1997 and Traugott 1999) as illustrated in (8). Notice that *tener novia* 'to have a girlfriend' forms a complex predicate in which *novia* 'girlfriend' is part of the predicate instead of being an argument (cf. Thompson 1997:72). The verb *tener* 'to have', in constructions like (8), is one of the verbs frequently used to form complex predicates (cf. Traugott (1999:241-242), Fernández Ramírez (1986:168-169)).

(8) *tener novia es un poco esclavizarse y limitar la actividad de uno* (HM-401-B)
to have a girlfriend is time consuming and it limits the activities one usually does

All the constructions with *inanimate* DOs were grouped, without distinguishing, as Balasch (2009) did, categorically ***Ø*** + *inanimate* DO cases involving *inanimate-specific* DOs (*los helados de vainilla hechos por Juan* 'the vanilla ice creams made by John') or constructions with *tener* ´to have´+ *specific* and *inanimate* DO, or ditransitive constructions. The rationale for bypassing these

considerations is that the single factor of inanimacy by itself virtually categorically prohibits the overt expression of ***a*** in the data. Thus, for example, in the sample from the *Corpus de Mérida,* there are only *five* tokens of ***a*** + *inanimate* DO in almost 699 *inanimate* DOs.

5. Data coding for the analysis of variation

All tokens in the sample, even those to be excluded from the final statistical analysis, were coded or categorized according to four linguistic factors: *definiteness*, *specificity*, *topicality* and *grammatical number*. The idea of considering the Spanish DOM as a product of several factors is recurrent among scholars. For example, Leonetti (2008:60) states "DOM in Romance is sensitive to a series of scalar dimensions that make up a multi-dimensional bundle of factors". Kliffer (1995:95) notes the complexity of the use of the accusative ***a*** and wonders whether its study "requeriría probablemente técnicas como las del análisis con reglas variables" 'would rather require techniques such as the ones used in the analysis of variable rules'. Except for the innovative study of Tippets and Schwenter (2007), however, there is no quantitative evaluation of the complexity of this phenomenon from the variationist perspective.

The variationist perspective allows the study of linguistic variation paying close attention to the fact that it is complex, i.e., the result of multiple factors that operate simultaneously in every instance where variation is possible (cf. Labov 2004). It also provides a framework in which to look for the general and recurrent patterns of the variation in those contexts in which it actually occurs. Following the *principle of accountable reporting* (Labov 1966; 1969:737-738, note 20), to accomplish this, every variant form is reported "with the proportion of cases in which the form did occur in the relevant environment, compared to the total number of cases in which it might have occurred" (Labov 1969:738).

5.1. Definiteness

Along with *animacy*, *definiteness* is one of the factors most commonly associated to the use of the accusative ***a***. Indeed, *definiteness* is associated to the use of different markers for DOs in many languages (cf. Comrie 1979 and 1981). *Definiteness* of the DOs is associated with the determiner system (Croft 1988:166) and DOs may be considered more or less definite according to the type of determiner that precedes them. Thus, DOs with definite articles, demonstratives, possessives or numeral adjectives, as well as proper names, are *more definite*. On the other hand, DOs accompanied by indefinite articles, quantifiers (*muchos 'many', poco 'little', algo 'some',* etc.), as well as generic nouns associated with animate entities (*gente 'people', niños 'children', bisnietos 'great-grand children'*, etc.) are "less definite". For brevity, the labels *definite* and *indefinite* are used in this paper.

5.2. Specificity

Among many authors, von Heusinger and Kaiser (2003:42) claim that the use of the accusative ***a***, in Peninsular Spanish, as well as in varieties of American Spanish, is governed by the *specificity* of the DO. This claim is based on the detailed interpretation of particular examples. Whether or not this factor has any independent quantitative effect on the variation is assessed in the present study.

In this paper, DOs are considered *specific* if they refer unique entities, i.e., entities that are not interchangeable with others (cf. Torres-Cacoullos and Aaron 2003:307), as illustrated in (9). Here *ese hombre* 'that man' designates out a specific individual and not anyone else. Otherwise, DOs are considered *nonspecific* when they represent a whole class or set of entities (cf. Ashby and Bentivoglio 1993:69-70), as it is observed in (10). Here no specific people are being referred to, but people in general. It should be mentioned that Leonetti (2003:71 and 2004:80) proposed a category of ambiguous specificity. However, since in the sample only one token was found that could be classified this way, we can simply use the binary distinction *specific/nonspecific*.

(9) *Yo quise mucho* ***a*** *ese hombre* (MDA1FA)
I loved that man very much

(10) con un monumento así, eso ... atraería ***Ø*** *mucha gente* (MDD1MB)
with a monument like that, it … would attract many people

5.3. Topicality as indicated prior or subsequent DO co-reference

Laca (1995:89) typifies many researchers when she indicates that the accusative ***a*** has the function of "indicar la alta «topicalidad» del objeto*"* 'to indicate the high «topicality» of the object'. Weissenrieder (1991:153) reports that in her research, if "[a] DO referent was repeated in a previous or subsequent clause, that NP was considered topical. If it was not repeated, it was considered to be an independent referent."

I coded for the topicality of DOs in terms both of *prior* and of *subsequent co-reference*. These indicators operationalize (cf. Dumont, unpublished) Givón's (1983,1985) notions of *referential distance* and of *topic persistence*, respectively. Every DO that is co-referred to at least once in the ten clauses prior to the DO token is coded for *anaphoric topicality* (AT). DOs that are co-referred to at least once in the following ten clauses of the current studied DO are coded for *cataphoric topicality* (CT).

In this paper, the terminology of «topicality» is used with some reservation. Optional marking of case, number, or numerous other grammatical categories is universally understood to provide possibilities for disambiguation, emphasis, style differentiation and other semantic, discursive and interactional roles. Thus, the opportunity for overt ***a***, in conjunction with *prior or subsequent co-reference*, may serve largely in the interests of clarity as to the syntactic role of the NP in question as well as, or instead of, *topicality*. Thus *prior co-reference* could clear the way for the ***Ø*** variant, whereas *subsequent co-reference* could belatedly disambiguate instances of the ***Ø*** variant. Fortunately, this kind of explanation makes different quantitative predictions about DOM from the *topicality* interpretation, and the results of this paper will able us to test which one is better supported by the data. In the tabulations, this factor group was labeled as *co-reference of DO*.

5.4. Grammatical number

Some authors associate grammatical number (singular or plural) of the DO with its individualization and the use of the accusative ***a*** (cf. Hopper and Thompson 1980; von Heusinger 2008, note 3; Croft 1988). The coding of this factor, together with *definiteness*, *specificity* and *topicality*, allows to test whether in fact it does constrain the use of the accusative ***a***.

6. Results and discussion

Table 1 shows the variable rule analysis, using GOLDVARB X (Sankoff, Tagliamonte and Smith 2005), of all the *animate* DOs in the Mérida corpus (150 tokens), not just the DOs of verbs having at least one instance of ***a***+DO as in Tippets and Schwenter 2007, after excluding error-prone tokens and those determined by categorical factors. It is possible to be most confident in the validity of these results, since the data collection was not biased towards verbs with high rate of ***a***, the distribution of factors is not skewed by the inclusion of inanimate tokens, and the remaining tokens could all take both variants (**a** or ***Ø***) without change of meaning. This is the analysis adopted as correct in this paper, and it was used to assess other ways of handling the data. Note that there was just one token with ambiguous specificity, which was excluded (149 tokens were coded as *specific* or *nonspecific*). Grammatical number never emerged as significant. Therefore, the results for this factor are not shown in *any* tabulation.

Two main purposes guide the discussion that follows. The first is to compare the data analysis with and without categorical contexts. The second is to show the impact, in terms of the accuracy of the results, of considering *animate* and *inanimate* DOs together in the analysis of *the variation*. To ease the comparisons for the reader, the numerous tables are all displayed with the factor groups in the same order, irrespective of their significance or the magnitude of their effects.

Table 1. Variable rule analysis of the factors that favor the use of ***a***+DO, using all *animate* DOs in the Mérida corpus, but excluding all tokens occurring in categorical contexts.

	Probability weight	Marginal % of ***a***	% of factor
Definiteness			
definite	.64	43 (39/90)	60 (90/150)
indefinite	.30	20 (12/60)	40 (60/150)
Range	34		
[Specificity]			
[specific]	[.49]	33 (32/98)	66 (98/150)
[nonspecific]	[.51]	35 (18/51)	34 (51/150)
Co-reference of DO			
neither AT nor CT	.71	50 (24/48)	32 (48/150)
one of AT or CT	.43	28 (21/75)	50 (75/150)
both AT and CT	.31	22 (6/27)	18 (27/150)
Range	40		
Total	.34 (input)	34 (51/150)	100% **N=150**

Results in Table 1 show that *definiteness* and *co-reference of DO* significantly affect the rate of overt ***a***. *Specificity* (and *grammatical number*) did not show significance. In Table 1 *and successive tabulations*: significant factors are highlighted with grey background; no significant factors are shown as square brackets; and the highest range (40, in Table 1) corresponds to the strongest constraint (*co-reference of DO*, in Table 1).

6.1. Definiteness and specificity

Table 1, and the rest of the analyses, confirms the key role of *definiteness* in favoring overt ***a***, while showing no effect of *specificity*, despite the claims of von Heusinger and Kaiser (2003:42), and other authors. Definite referents favor the use of accusative ***a*** while indefinite referents do not.

6.2. Effect of co-reference of DO

This factor has a strong effect, with the absence of both kinds of reference (neither AT nor CT) favoring overt ***a*** (Probability weight = .71), and the presence of both (both AT and CT) disfavoring it (Probability weight = .31). This is the opposite of what it would be expected if *co-reference of DO* were to be considered indications of *topicality*, and overt ***a*** another such indication. The observed pattern is, however, suggestive of a clarification or disambiguation pattern, where overt ***a*** compensates for the lack of any co-reference of the DO in the preceding or prior discourse, while subsequent co-reference compensates for the lack of overt DO marking.

It is possible to carry out a simpler test of the relationship among the two types of DO *co-reference* and presence of ***a***. Table 1a displays a contingency table of overt ***a*** against DO *co-reference*.

Table 1a. Cross-tabulation of DOM and DO co-reference. $\Sigma(O-E)^2/E=8.81$. A chi-squared test with 2 degrees of freedom is significant at the $p<0.02$ level. N=150.

	neither AT nor CT	one of AT or CT	both AT and CT
***a* present**			
O=observed number p_{row}=0.34 (51/150)	24	21	6
E=expected number=N x p_{row} x p_{column}	16.32	25.50	9.18
$(O-E)^2/E$	3.61	0.79	1.10
***a* absent**			
O=observed number p_{row}=0.66 (99/150)	24	54	21
E=expected number=N x p_{row} x p_{column}	31.68	49.50	17.82
$(O-E)^2/E$	1.87	0.41	0.56
p_{column}	0.32 (48/150)	.50 (75/150)	0.18 (27/150)

The significant chi-squared statistic for Table 1a is made up largely of two major components, substantially more ***a*** than expected in contexts where there is neither *anaphoric* nor *cataphoric* DO *reference* (O = 24 instead of E = 16.32) and less ***Ø*** than expected (O = 24 instead of E = 31.68) in the same contexts. Again, this is not what it would be expect under a topicality analysis, but exactly what would be expected under a clarification/disambiguation analysis.

6.3. Exclusion of inanimate DOs

Since in the Mérida sample there are far more *inanimate* DOs (80% = 699/877 tokens) than *animate* ones (20% = 178/877 tokens), it might seem statistically preferable to be able to combine the two kinds in the same analysis. Table 2 shows what happens if this is attempted.

Table 2. Variable rule analysis of the factors that favor the use of ***a***+DO, using both *animate* and *inanimate* DOs in the corpus, but excluding all tokens occurring in categorical contexts.

	Probability weight	Marginal % of ***a***	% of factor
[Definiteness]			
[definite]	[.55]	14 (43/302)	68 (302/443)
[indefinite]	[.40]	9 (12/141)	32 (141/443)
Specificity			
specific	.65	19 (34/179)	40 (179/442)
nonspecific	.40	8 (20/263)	60 (263/442)
Range	25		
[Co-reference of DO]			
[neither AT nor CT]	[.56]	14 (27/200)	45 (200/443)
[one of AT or CT]	[.49]	13 (22/171)	39 (171/443)
[both AT and CT]	[.37]	8 (6/72)	16 (72/443)
Total	.12 (input)	12(55/443)	100% **N=443**

The first observations are that in comparison with Table 1, the effect of *definiteness* and *co-reference* are lost, and *specificity* becomes significant. It is not hard to see how these radical changes came about. There are only five ***a***+DO tokens among the 699 *inanimate* DOs. Moreover, in contrast with the contexts of the *animate* DOs, the *inanimates* are overwhelmingly definite, non-specific, and are referred to neither in prior nor in subsequent contexts. Thus all these factors become associated with lack of overt ***a*** not because they affect ***a*** expression, but because of their association with inanimate DOs.

Clearly, there are not enough tokens of ***a*** + *inanimate* DO to do a separate variable rule analysis, and it is a methodological error to combine the animate and inanimate tokens in a single analysis, since the distribution of factors in the inanimate contexts swamps that in the animate contexts. Finally, it is interesting that far more of the *inanimate* DO (46% = 406/699) fall in excluded contexts than do *animate* ones (16% = 28/178).

6.4. Effect of neglecting DOs of verbs with no overt a in the corpus

Collecting only DOs of verbs having at least one overt ***a***, as Tippets and Schwenter 2007 did, is a methodological device for ensuring that all the tokens in the corpus could conceivably take ***a***. However, in the present study the claim is that *all* the DOs, *except those explicitly excluded as categorical*, should be included in the data collection and subsequent analysis, as long as the list of categorical factors is *carefully and exhaustively* constructed. Note that the careful construction of the list of categorical factors conforms to the *principle of accountability*, while the exclusion, a priori, of certain verbs that do not have at least one overt ***a*** has the unfortunate consequence of excluding all rare verbs that do happen to occur, of which there may be many, and exaggerating the effect of any verbs that have a lexical preference for overt ***a***. In Tables 3a and 3b, these problems are illustrated, by re-analyzing the data of Table 1 in such a way that now only DOs of verbs showing at least one overt ***a*** are counted. Table 3a shows the case where no attention is paid to excluding tokens occurring in categorical contexts, while in Table 3b such tokens are excluded from the analysis. The effect in both

cases of censoring verbs with no tokens of ***a*** + DO is the drastic reduction of the number of tokens available to the analysis and the inflation of the overall rate of overt ***a*** (.62 in Table 3a and .55 in Table 3b, while in Table 1 is .34).

A secondary effect of excluding DOs of verbs with no ***a*** tokens is to include a disproportionate number of DOs with no *prior or subsequent co-reference*. Because DO co-reference and overt ***a*** seem to be compensatory disambiguating devices, when contexts with no ***a*** are excluded, then many contexts with DOs that are co-referred to are automatically excluded. This leads to a blurring of the distinction between the overt ***a*** rates in the *co-reference* factor group (note particularly the marginals) and a loss of significance for this factor group.

It follows that abandoning Labor's *principle of accountability*, which calls for the inclusion *of all pertinent tokens in the corpus*, just for the sake of avoiding a detailed analysis of categorical contexts, leads to serious biases and a loss of important information.

Table 3a. Variable rule analysis of the factors that favor the use of ***a***+DO, using only *animate* DOs of verbs exhibiting at least one overt ***a***, without excluding categorical contexts.

	Probability weight	Marginal % of ***a***	% of factor
Definiteness			
definite	.69	79 (49/62)	57 (62/109)
indefinite	.26	38 (18/47)	43 (47/109)
Range	43		
[Specificity]			
[specific]	[.48]	66 (44/67)	62 (67/109)
[nonspecific]	[.53]	54 (22/41)	38 (41/109)
[Co-reference of DO]			
[neither AT nor CT]	[.51]	59 (24/41)	38 (41/109)
[one of AT or CT]	[.49]	59 (30/51)	47 (51/109)
[both AT and CT]	[.50]	77 (13/17)	15 (17/109)
Total	.62 (input)	62 (67/109)	100% **N=109**

Table 3b. Variable rule analysis of the factors that favor the use of ***a***+DO, using only *animate* DOs of verbs exhibiting at least one overt ***a***, but excluding all tokens occurring in categorical contexts.

	Probability weight	Marginal % of ***a***	% of factor
Definiteness			
definite	.71	76 (35/46)	54 (46/85)
indefinite	.26	31 (12/39)	46 (39/85)
Range	45		
[Specificity]			
[specific]	[.50]	58 (30/52)	62 (52/85)
[nonspecific]	[.50]	50 (16/32)	38 (32/85)
[Co-reference of DO]			
[neither AT nor CT]	[.56]	57 (21/37)	43 (37/85)
[one of AT or CT]	[.49]	53 (20/38)	45 (38/85)
[both AT and CT]	[.34]	60 (6/10)	12 (10/85)
Total	.55 (input)	55 (47/85)	100% **N=85**

6.5. Impact of ignoring the exclusion of tokens affected by categorical factors

Tables 4 and 6 show the effect of including tokens that should have been eliminated from the analysis by virtue of being outside the envelope of variation. Table 4 shows little difference from Table 1, due largely to the competing effects of categorical factors requiring overt ***a***, and those prohibiting it. Nevertheless, a weakening of the *co-reference of DO* factor group can be noted: the range of *this factor* effects in Table 1 is 40 while in Table 4 is only 26.

Table 4. Variable rule analysis of the factors that favor the use of ***a***+DO, using all *animate* DOs in the corpus, without attention to categorical factors.

	Probability weight	Marginal % of ***a***	% of factor
Definiteness			
definite	.65	51 (54/105)	59 (105/178)
indefinite	.30	23 (17/73)	41 (73/178)
Range	35		
[Specificity]			
[specific]	[.49]	41 (46/112)	63 (112/178)
[nonspecific]	[.52]	37 (24/65)	37 (65/178)
Co-reference of DO			
neither AT nor CT	.68	53 (29/55)	31 (55/178)
one of AT or CT	**.42**	**32 (28/88)**	49 (88/178)
both AT and CT	**.42**	**40 (14/35)**	20 (35/178)
Range	26		
Total	.40 (input)	40 (71/178)	100% **N=178**

It would be desirable to check the effect of including tokens for each of the categorical factors separately, so that they could not cancel each other out. However, it is known that the inclusion of contexts with few tokens will not have any significant impact on the analysis of the variation. Non-variable contexts with many tokens, however, may well distort it. Table 5 shows that ***a*** + personal pronoun and general pronoun *uno* were the most frequent excluded contexts (15 tokens) in the sample from Mérida.

Table 5. Distribution of exclusions in the Mérida sample

Type of exclusion	
a + personal pr/general pr *uno*	**15 (54%)**
phonological neutralization	6 (21%)
verb-object lexicalized exp.	5 (18%)
semantic change	2 (7%)
Total	28 tokens

To evaluate the specific impact of the pronominal DOs, the variation in a dataset integrating the 150 tokens in Table 1 and the 15 categorical contexts of ***a*** + pronominal DO entered in Table 5 was analyzed. The results of the analysis of these 165 tokens are in Table 6.

Table 6. Variable rule analysis of the factors that favor the use of ***a***+DO, using all 150 *animate* DOs in the Mérida sample, including pronominal DOs.

	Probability weight	Marginal % of ***a***	% of factor
Definiteness			
definite	.63	51 (52/103)	62 (103/165)
indefinite	.29	23 (14/62)	38 (62/165)
Range	34		
[Specificity]			
[specific]	[.49]	40 (44/110)	67 (110/164)
[nonspecific]	[.51]	39 (21/54)	33 (54/165)
Co-reference of DO			
neither AT nor CT	.67	52 (26/50)	30 (50/165)
one of AT or CT	.42	33 (26/80)	49 (80/165)
both AT and CT	.43	40 (14/35)	21 (35/165)
Range	24		
Total	.40 (input)	40 (66/165)	100% **N=165**

Because of a disproportionate number of pronominal DOs with both *anaphoric* and *cataphoric reference*, the disfavoring effect of this factor (probability weight of .31 in Table 1) is now attenuated (probability weight of .43 in Table 6), blurring the role of this factor group.

6.6. Stability of DOM across corpora

When the data from the Madrid corpus are analyzed in the same way as Table 1 the results (Table 7) are similar, except for a sizable increase in the overall rate of ***a*** (34% (51/150) in Mérida *vs.* 63% (85/136) in Madrid) and a concomitant increase in the input probability (.32 in Mérida *vs.* .63 in Madrid). This might be ascribed to diverging rates of DOM in the two communities, but it may well only reflect the sociological composition of the two samples or the somewhat different interview styles in the two corpora.

The *co-reference of DO* factor group does not attain statistical significance in the Madrid corpus, although the direction of the effects (*no co-reference* favoring overt ***a***) is the same as in Mérida. Combining the two corpora with *Corpus* as a factor group shows the stability of the results (Table 8) across the two sets of data, when compared with Table 1.

Table 7. Variable rule analysis of the factors that favor the use of ***a***+DO, using all *animate* DOs in the Madrid corpus, excluding all tokens occurring in categorical contexts.

	Probability weight	Marginal % of ***a***	% of factor
Definiteness			
definite	.62	74 (63/85)	63 (85/136)
indefinite	.30	43 (22/51)	37 (51/136)
Range	32		
[Specificity]			
[specific]	[.43]	66 (40/61)	45 (61/136)
[nonspecific]	[.56]	60 (45/75)	55 (75/136)
Co-reference of DO			
neither AT nor CT	[.53]	67 (29/43)	32 (43/136)
one of AT or CT	[.49]	61 (39/64)	47 (64/136)
both AT and CT	[.48]	59 (17/29)	21 (29/136)
Total	.63 (input)	63 (85/136)	100% **N=136**

Table 8. Variable rule analysis of the factors that favor the use of ***a***+DO, using all *animate* DOs in the Madrid and Mérida corpora, excluding all tokens affected by categorical factors.

	Probability weight	Marginal % of ***a***	% of factor
Definiteness			
definite	.61	58 (102/175)	61 (175/286)
indefinite	.33	31 (34/111)	39 (111/286)
Range	28		
[Specificity]			
[specific]	[.46]	45 (72/159)	56 (159/285)
[nonspecific]	[.55]	50 (63/126)	44 (126/285)
Co-reference of DO			
neither AT nor CT	.64	59 (54/92)	32 (92/286)
one of AT or CT	.45	43 (59/138)	48 (138/286)
both AT and CT	.39	41 (23/56)	20 (56/286)
Range	25		
Corpus			
Madrid	.66	63 (85/136)	48 (136/286)
Mérida	.35	34 (51/150)	52 (150/286)
Range	31		
Total	.48 (input)	48(136/286)	100% **N=286**

7. Conclusion

This paper attempts to contribute to the understanding of differential object marking in Spanish by returning to the *principle of accountability* for variationist analysis. Based on samples of interviews from the *Corpus de Mérida/Venezuela* and the *Corpus del habla culta de Madrid/Spain*, the first step was to delineate in exactly which contexts the variation between ***a***+DO and ***Ø***+DO is possible.

It was then possible excluded to exhaustively tabulate all the DOs in these corpora, not only by the traditional factor groups of *animacy*, *definiteness, specificity, topicality*, *grammatical number* but also by whether or not variation could be meaningfully detected for each of the factors in these groups. This represents the main methodological improvement claimed over previous studies. In the course of the analysis, it was noticed that the *inanimate* DOs and *animate* ones should not be conflated within the same statistical analysis because their great difference in overt marking rates (near categorical lack of ***a***-marking for *inanimate* DOs) and in how the tokens are distributed. In fact, the data contain little information on how marking is patterned for inanimate DOs; this would require a corpus many times larger than the one collected in the present paper.

Among the other factor groups only *definiteness* and *co-reference of DO* significantly influence the presence or absence of ***a*** within this domain of variability. There is no quantitative evidence that *specificity* or *grammatical number* have any effect.

In the process, it was demonstrated how important it is to exhaustively analyze all the pertinent tokens in a corpus sample and to eliminate error-prone and categorical (non-variable) data in the statistical analysis of the variation. Otherwise significant distortion of the results is likely and the significance of certain factors may be obscured by the noise introduced. In the words of Labov (1969:729), ignoring the identification of the contexts in which the variation is effectively possible and those where only one of the variants is always (categorically) employed may "obscure a number of important constraints on variability."

The empirical results lead to reflection on the discourse theoretical basis of the factor labeled *topicality*. The way in which this has been operationalized, both in the present and previous studies, in terms of *anaphoric* and *cataphoric reference*, seems to reflect a reference clarification or disambiguation role more than the establishment or maintenance of topicality.

Finally, it was found that contextual conditioning is identical in Mérida and Madrid; though overall overt ***a*** rate is much higher in the latter.

References

Algeo, John. 1995. Having a look at the expanded predicate. In Bas Aarts and Charles F. Meyer (eds.). *The verb in contemporary English. Theory and description*. 203-217. Cambridge: Cambridge University Press.

Ariza, Manuel. 1989. La preposición *a* de objeto. Teorías y panorama. *Lexis* 13.2. 203-222.

Ashby, William J. and Paola Bentivoglio. 1993. Preferred argument structure in spoken French and Spanish. *Language Variation and Change* 5. 61-76.

Balasch, Sonia. 2009. The importance of the variable context in analyzing differential object marking (DOM) in contemporary Spanish. NWAV 38. University of Ottawa.

Bolinger, Dwight L. 1953. Verbs of being. *Hispania* 36.3. 343-345.

Cano Aguilar, Rafael. 1981. *Estructuras sintácticas transitivas en el español actual*. Madrid: Editorial Gredos.

Chafe, Wallace. 1994. *Discourse, consciousness, and time*. Chicago: The University of Chicago Press.

Company Company, Concepción. 2002. Grammaticalization and category weakness. In Ilse Wishcer and Gabriele Diewald (eds.). *New reflections on grammaticalization*. 201-15. Amsterdam: John Benjamins.

Comrie, Bernard. 1979. Definite and animate direct objects: a natural class. *Linguistica Silesiana* 3.13-21.

Comrie, Bernard. 1981. *Language universals and linguistic typology*. Chapter 6. Chicago: University of Chicago Press.

Croft, William. 1988. Agreement vs. case marking and direct objects. In Michael Barlow and Charles A. Ferguson (eds.). *Agreement in natural language. Approaches, theories, descriptions*. 159-179. Stanford, CA.: Center for the study of Language and Information (CSLI).

Domínguez, Carmen Luisa and Elsa Mora. 1995. *Corpus sociolingüístico de la ciudad de Mérida*. Mérida: Universidad de los Andes. Departamento de Lingüística.

Domínguez, Carmen Luisa; Blanca Guzmán, Luis Moros, Maryelis Pabón, Lis Morelia Torres and Roger Vilaín. 1998. Observaciones sobre el uso de la preposición ***a*** en el objeto directo: un estudio sobre el español de Mérida. *Letras* 59:89-120. Caracas: Universidad Pedagógica Experimental Libertador (UPEL).

Dumitrescu, Domnita. 1997. El parámetro discursivo en la expresión del objeto directo lexical: español madrileño vs. español porteño. *Signo y Seña* 7. 305-354.

Dumont, Jennifer. Unpublished. Full NPs in conversation and narratives: The effects of genre on information flow and interaction. PhD Dissertation. University of New Mexico

Esgueva, M. and M. Cantarero (editors). 1981. *El habla de la ciudad de Madrid: Materiales para su estudio*. Madrid: Consejo Superior de Investigaciones Científicas.

Fernández Ramírez, Salvador. 1986. *Gramática española*. Madrid: Arco Libros.

Fish, Gordon T. 1967. 'A' with Spanish direct object. *Hispania* 50.1:80-35.

García, Erica C. 1993. Syntactic diffusion and the irreversibility of linguistic change: personal *a* in Old Spanish. In Schmidt-Radefeltd, Jürgen and Harder, Andreas (eds.). *Sprachwandel und Sprachgeschichte: Festschrift für Helmut Lüdtke zum* 65. 33-50. Tübingen: Narr.

García Erica C. and Florimón van Putte. 1995 [1987]. La mejor palabra es la que no se habla. In Carmen Pensado (ed.). *El complemento directo preposicional*. 113-131. Madrid: Visor Libros.

Givón, Talmy. 1983. Topic continuity in spoken English. In Talmy Givón (ed.). *Topic continuity in discourse: A quantitative cross-language study*. 347-363. Amsterdam: John Benjamins.

Givón, Talmy. 1995. *Functionalism and grammar*. Amsterdam: John Benjamins.

von Heusinger, Klaus. 2008. Verbal semantics and the diachronic development of differential object marking in Spanish. *Probus* 20. 1-31.

von Heusinger, Klaus and Kaiser, Georg. 2003. The interaction of animacy, definiteness and specificity in Spanish. In Klaus von Heusinger and Georg Kaiser (eds.). *Proceedings of the Workshop Semantic and Syntactic Aspects of Specificity in Romance Languages*. 41-65. Konstanz: Universität Konstanz.

Hopper, Paul J. and Sandra A. Thompson. 1980. Transitivity in grammar and discourse. *Language* 56.2. 251-99.

Kliffer, Michael D. 1995. El «***a***» personal, la kinesis y la individuación. In Carmen Pensado (ed.). *El complemento directo preposicional*. 93-111. Madrid: Visor Libros.

Labov, William. 1966. *The social stratification of English in New York City*. Arlington, VA: Center for Applied Linguistics.

Labov, William. 1969. Contraction, deletion, and inherent variability of the English copula. *Language* 45. 4. 715-62.

Labov, William. 2004. Quantitative reasoning in linguistics. In Ultich Ammon, Norbert Dittmar, Klaus J. Mattheier, and Peter Trudgill (eds.). *Sociolinguistics/Soziolinguistik: an international handbook of the science of language and society.* Volume I. 6-22. Berlin: Mouton de Gruyter.

Laca, Brenda. 1995. Sobre el uso del acusativo preposicional en español. In Carmen Pensado (ed.). *El complemento directo preposicional*. 61-91. Madrid: Visor Libros.

Laca, Brenda. 2006. El objeto directo. La marcación preposicional. In Concepción Company (ed.). *Sintaxis histórica del español. Primera parte: La frase verbal*. Vol. 1. 423-475. México: Fondo de Cultura Económica/Universidad Autónoma de México.

Leonetti, Manuel. 2003. Specificity and object marking: the case of Spanish ***a***. In Klaus von Heusinger and Georg A. Kaiser (eds.). *Proceedings of the Workshop Semantic and Syntactic Aspects of Specificity in Romance Languages*. Arbeitspapier 113. 67-101. Konstanz: Universität Konstanz.

Leonetti, Manuel. 2004. Specificity and differential object marking in Spanish. *Catalan Journal of Linguistics.* Available in: http://ddd.uab.es/pub/linguistics/16956885v3p75.pdf

Leonetti, Manuel. 2008. Specificity in clitic doubling and in differential object marking. *Probus* 20. 1. 33-66.

Martín Zorraquino, María A. 1976. A + objeto directo en el *Cantar de Mio Cid*. *Melanges Gossen*. II. 555-566. Berna.

Pensado, Carmen. 1995. El complemento directo preposicional: Estado de la cuestión y bibliografía comentada. In Carmen Pensado (ed.). *El complemento directo preposicional. 11-59. Madrid: Visor Libros.*

Sankoff, David, Sali A. Tagliamonte and Eric Smith. 2005. GOLDVARB X: A multivariate analysis application for Macintoch and Windows. Department of Linguistics. University of Toronto and Department of Mathematics. University of Ottawa <individual.utoronto.ca/Tagliamonte/Goldvarb/GV_index.htm>

Thompson, Sandra A. 1997. Dircourse motivations for the core-oblique distinction as a language universal. In A. Kamiro (ed.). *Directions in functional linguistics.* 59-82. Amsterdam: John Benjamin.

Tippets, Ian and Scott Schwenter. 2007. Relative animacy and differential object marking in Spanish. NWAV 36. Philadelphia.

Torrego Salcedo, Esther. 1999. El complemento directo preposicional. In Ignacio Bosque and Violeta Demonte (eds.). *Gramática descriptiva de la lengua española.* Vol. 2. 1780-1805. Madrid: Espasa-Calpe.

Torres-Cacoullos, Rena and Jessi Elana Aaron. 2003. Bare English-origin nouns in Spanish: rates, constraints, and discourse functions. *Language Variation and Change*. 15. 289-328.

Traugott, Elizabeth Closs. 1999. A historical overview of complex predicate types. In Laurel J. Brinton and Minoji Akimoto (eds.). *Collocational and idiomatic aspects of composite predicates in the history of English*. 239-274. Amsterdam: John Benjamins.

Weissenrieder, Maureen. 1985. Exceptional uses of the accusative A. *Hispania* 68.2. 393-398.

Weissenrieder, Maureen. 1990. Variable uses of the direct-object marker A. *Hispania* 73.1. 223-231.

Weissenrieder, Maureen. 1991. A functional approach to the accusative A. *Hispania* 74.1. 146-156.

Selected Proceedings of the 5th Workshop on Spanish Sociolinguistics

edited by Jim Michnowicz and Robin Dodsworth

Cascadilla Proceedings Project Somerville, MA 2011

ISBN 978-1-57473-443-0 library binding

Ordering information

Orders for the library binding edition are handled by Cascadilla Press.
To place an order, go to www.lingref.com or contact:

Cascadilla Press, P.O. Box 440355, Somerville, MA 02144, USA
phone: 1-617-776-2370, fax: 1-617-776-2271, sales@cascadilla.com

Web access and citation information

This entire proceedings can also be viewed on the web at www.lingref.com. Each paper has a unique document # which can be added to citations to facilitate access. The document # should not replace the full citation.

This paper can be cited as:

Balasch, Sonia. 2011. Factors Determining Spanish Differential Object Marking within Its Domain of Variation. In *Selected Proceedings of the 5th Workshop on Spanish Sociolinguistics*, ed. Jim Michnowicz and Robin Dodsworth, 113-124. Somerville, MA: Cascadilla Proceedings Project. www.lingref.com, document #2511.