A Corpus Study of Verb Bias in Spanish

Amelia J. Dietrich and Colleen Balukas
Penn State University

1. Introduction

It has long been established in the psycholinguistic literature that a verb's subcategorization frame – also known as its verb bias – has an effect on sentence production and processing (e.g., Garnsey, Lotocky, Pearlmutter, & Myers, 1997; Hare, McRae & Elman, 2003; MacDonald, 1994; MacDonald, Pearlmutter, & Seidenberg, 1994a, 1994b; Trueswell, Tanenhaus & Kello, 1993; Wilson & Garnsey 2009). In bilinguals, such biases may be simultaneously activated in both languages, manifesting in the form of unique behaviors in bilinguals, especially when cognate verbs with different subcategorization frames are considered (e.g. Chambers & Cooke, 2009; Duyck, van Assche, Drieghe, & Hartsuiker 2007; Libben & Titone, 2009; Marian & Spivey, 2003; Schwartz, Kroll, & Diaz, 2007; van Assche, Duyck, Hartsuiker, & Diependaele 2009; van Hell & De Groot, 2008). For instance, the verb biases of a given translation pair may show the effects of crossover between the two languages in the form of shifting biases that do not resemble those expected of either of the verbs in the monolingual speech of each language.

Dussias, Marful, Gerfen, and Bajo (2010) explored cognate verbs and their subcategorization biases in Spanish-English bilinguals in a recent norming study, for which a large portion of the materials was based on a Garnsey et al. (1997) norming study of 100 active verbs in English. These verbs consisted of a number of cognate and non-cognate verbs that were matched with translated Spanish counterparts by Dussias et al. Reliable verb biases of these English verbs are widely available not only in the norming study in Garnsey et al. (1997) but also in other studies of the same verbs (see the corpus study of American English in Gahl & Roland, 2004). Prior to the Dussias et al. (2010) work, however, no such bias information was available for the Spanish verbs. The present study therefore aims to determine if there is a match between the Dussias et al. (2010) norming data and a corpus-based study of the biases of the Spanish verbs in question.

By identifying the subcategorization biases for these verbs in natural speech production data, we aim to provide greater opportunity for bilingual sentence processing research. We do this by offering verb bias information that can be compared in both English and Spanish, allowing other researchers to determine if their findings in the laboratory correlate with corpus findings.

Thus, our first research question to guide the present study is: What are the verb biases of these 10 highly frequent verbs? We then ask: What contexts might co-occur with or provide a discourse explanation for the subcategorization bias of a given verb? This second analysis, carried out using the multivariate analysis program Goldvarb X (Sankoff, Tagliamonte & Smith, 2005), allows us to examine the underlying linguistic factors that might explain any potential difference between our corpus findings on verb bias and the previous norming study. To this end, we looked to the Transitivity Hypothesis (Hopper & Thompson, 1980) and therefore analyzed the possible effects of animacy of the subject and direct object and presence of an indirect object. We also examined form of the subject and

[*] Our deep thanks go to Rena Torres Cacoullos and Paola E. Dussias for their support and guidance in this study, to the Center for Language Science (CLS) and the Department of Spanish, Italian & Portuguese at Penn State for conference support, and to the audience at the Hispanic Linguistics Symposium 2010. This research was funded in part by a National Science Foundation Graduate Research Fellowship under Grant No. DGE-0750756 to Amelia J. Dietrich.

© 2012 Amelia J. Dietrich and Colleen Balukas. Selected Proceedings of the 14th Hispanic Linguistics Symposium, ed. Kimberly Geeslin and Manuel Díaz-Campos, 258-271. Somerville, MA: Cascadilla Proceedings Project.

direct object in order to test the validity of the standard treatment of proper nouns in Preferred Argument Structure (DuBois, 1987).

2. Background

A *subcategorization frame* is the order and category of constituents which co-occur with a particular verb. For example, the verb *believe* in English can occur in any of the following subcategorization frames:

Direct Object:	NP__ NP	*John believes the story*
Sentential Complement:	NP __SC	*John believes (that) the story was true*
Other Complement:	NP__ PP	*John believes in Mary*

The *subcategorization frequency*, which we will also refer to as *verb bias*, indicates which subcategorization frame is preferable for a given verb. In English, it is accepted that *believe* has a sentential, or clausal, complement (SC) bias (Garnsey et al., 1997). Verb biases can and do differ cross-linguistically, even in cognate verbs, so determining English verb biases says nothing about the biases of verbs in other languages. Recent psycholinguistic research has found that in sentence processing of oral input, reading, and sentence production, such biases can speed up processing if the complement which follows the verb is in accordance with the bias associated with that verb. It also appears that second language learners can acquire verb bias information even when acquiring the L2 while immersed in an L1-dominant community (Dussias & Cramer Scaltz, 2008).

Past studies of bilinguals using picture naming tasks show that lexical access in bilinguals is non-selective in nature, in other words, bilinguals non-selectively activate codes in both of the languages they speak (for more psycholinguistic literature on this topic see Duyck, van Assche, Drieghe & Hartsuiker, 2007; Schwartz & Kroll, 2006; Libben & Titone, 2009)[1]. It has been further observed that cognates tend to be a point at which activation of both languages can speed up sentence processing (Elston-Güttler, Gunter & Kotz, 2005). Conversely, where verb biases are used by a hearer to predict what is coming next in a clause, cognates could actually be a point at which processing is slowed because of conflicting verb bias information in the two languages (Dussias et. al., 2010). Such studies of bilinguals can be very useful to the investigation of sentence parsing, but they cannot be carried out without knowing the verb biases in both languages of study.

Building on the recent Spanish norming study of Dussias et al. (2010), we note that verb bias information is only widely available for English verbs, while other languages remain largely unstudied. The present work expands the information available on Spanish. To our knowledge, the norming study of Spanish verb biases carried out by Dussias and colleagues is in fact the first of its kind that makes available norming information for Spanish verbs which have corresponding verb bias information available in English. During the study, the authors collected norming data via a sentence completion task from 525 native speakers of Peninsular Spanish living in Spain over the course of three years. Using a paper and pencil, participants were given sentences such as "*Juan contestó* ______________," and asked to complete the sentence with something that is both semantically plausible and grammatically correct. The study reveals highly consistent responses across time and participants, and its results are a valid first step in establishing verb bias norms in Spanish. However, the results found in the many English norming studies are not consistent (cf. Gahl & Roland, 2004). The present study seeks to investigate Spanish verb biases by a complementary method so as to more firmly establish the true verb biases of the Spanish verbs in question.

Since English verbs have up to this point been the subject of far more study than Spanish ones, we can benefit from the experiences of those researchers who have gone before us. As Gahl and Roland

[1] These studies cover a broad scope of bilingualism, from simultaneous bilinguals to highly proficient second language learners and demonstrate the same trends in behavior across groups.

(2004) did when creating their study of English verb biases using American English corpus data, we note that norming studies of verb bias information are generally constrained to a very limited genre and context of production. In the case of the Dussias et al. (2010) study, participants were given a paper with a series of sentence fragments headed by a proper name and followed by a verb in the preterit form, i.e.: "*María decidió* _______." Participants were asked to complete the sentence (in writing) with no limitations on their answer except that it be grammatically correct and semantically plausible. A corpus based study will generate more varied contexts, providing the opportunity to examine whether the same verb bias information holds true for a verb embedded in a different linguistic context or in different genre or discourse types (i.e.: oral interview, narrative, etc.).

While norming studies seek to establish which subcategorization frame a verb prefers, and such studies are essentially limited only to that question, the present study seeks also to investigate why verb biases are different both for different verbs in the same language and for translation equivalents of the same verb in different languages. We investigate this through a closer look at transitivity and Preferred Argument Structure.

Based on traditional definitions of transitivity it can be assumed that all of the verbs in this study are transitive because we are looking at the types of complements they take. That is to say that they are all taking complements of some sort, a characteristic of transitive verbs. However, Hopper and Thompson (1980) propose that transitivity is not a categorical feature (transitives take objects, intransitives do not), but rather a more gradient feature of a whole clause which is based on a number of parameters. Among those parameters is Agency, whereby a clause is considered to be more transitive if its agent, or subject, is higher in potency. Humans are especially potent agents, while inanimates are much less potent. Argument animacy in the data for the present study could shed light on whether more animate subjects prefer a certain type of complement.

From Thompson and Hopper (2001) we also learn that more frequent verbs are less likely to have a fixed number of argument structures. The argument structure denotes how many and what kind of objects a verb can take. While low frequency verbs tend to fit cleanly and clearly into the transitive/intransitive categories of traditional syntax, verbs which are used more frequently can appear in a variety of different constructions. This can have implications for verb bias, suggesting that perhaps high frequency verbs are more likely to be EquiBias (that is, to occur equally with a nominal or clausal complement) or have no bias at all because their frequency of usage allows them to appear freely with many different complement structures.

3. Data Extraction: Methodology and Exclusions

A total of 10 high frequency verbs included in the Dussias et al. (2010) study were extracted from the Corpus del Español (Davies, 2002-). These verbs were: *contestó*, *contó*, *decidió*, *encontró*, *escribió*, *oyó*, *pensó*, *preguntó*, *respondió*, and *sintió*. We restricted our search to the 20th century section of the CdE, including all genres, which gave us access to approximately 22.5 million words. The corpus consists of both oral and written genres, the latter coming from a wide variety of news articles, academic writing and works of fiction. Following Dussias et al. (2010), only the 3sg preterit form of each verb was extracted from the corpus. This yielded an initial return of 10,283 tokens from various Spanish dialects, of which only 722 tokens from Peninsular sources were ultimately coded for verb complement biases for the purposes of the present study. The primary reason for the large difference between the initially extracted token count and the final token count is our exclusion according to dialect. Given Dussias et al.'s (2010) exclusive use of Peninsular Spanish speakers in their norming study, as well as the possible sensitivity of verb biases to usage frequencies in different speech communities, we determined that is was necessary to exclude non-Peninsular sources for any meaningful comparison between our study and that of the previous norming study.

In addition to restricting the data to Peninsular Spanish sources, we also excluded any reflexive forms of the verbs in question, based on the understanding that such forms can actually represent entirely different meanings, which often results in two completely different translation equivalents.

Note, for instance, the meaning difference between *encontró* (1) and *se encontró* (2) in the following two examples[2]:

(1) NON-REFLEXIVE
"*...el equipo norteamericano* ***encontró*** *una peculiar población...*"
'the North American team **found** a peculiar population..." (España ABC, news)

(2) REFLEXIVE
"*...un movimiento revolucionario, que pronto* ***se encontró*** *con una fuerte represión...*
'a revolutionary movement, that quickly **met** with a strong repression (Enc: Rusia, academic)

The verb in example (2), for instance, almost categorically selects a prepositional phrase, most often beginning with *con* 'with'. If considered in the data, a reflexive form like *se encontró (con)* may skew the overall verb bias information. Therefore, we excluded any reflexive forms from the analysis. We discuss the effects of reflexives further in the discussion section (§7).

The linguistic factors that we explore in this study are as follows: animacy of the subject and direct object (when present), which provide us insight into the transitivity of a given construction; form of the subject and direct object; and presence of an indirect object. We also code for complement type, which in our multivariate analysis is our dependent variable. Finally, in the discussion of our results in §7, we also consider the possible influence of frequency of similar verb types (e.g. reflexives). Any additional excluded or collapsed token types are discussed in §4 below, organized according to factor group.

4. Hypotheses and Coding of Tokens for Multivariate Analysis

In this section, we discuss the underlying hypotheses for each of the factor groups (independent variables) for which we coded. We also discuss the profile of any excluded tokens within each factor group. As we explain in further detail in §6, the impact of these factor groups on complement type was analyzed using multivariate, or variable-rule, analysis (Sankoff, Tagliamonte & Smith, 2005). This method of analysis allows us to consider all factor groups at once, and at the same time to consider the impact that each individual factor has on complement selection for the ten verbs in question.

4.1. Complement Type

Coding for complement type was initially split into nine categories. For the purposes of direct comparison with the findings of Dussias et al. (2010) only two categories (Direct Object and Sentence Complement) were ultimately considered. This is an approach which has been taken consistently throughout the history of verb subcategorization bias research. Tokens with a Direct Object complement alone (3) or with an additional object or other optional material (4) were collapsed into the Direct Object category. The presence of an indirect object was coded for as a separate factor (see *Indirect Object Complement* in this section for details).

(3) DIRECT OBJECT COMP
"*...en su tiempo libre* ***escribió*** <u>*una novela*</u> *[DO]...*"
'in his/her free time s/he **wrote** <u>a novel</u> [DO]' (Entrevista-ABC, oral)

[2] Any examples provided in this and other sections of the paper originate from the Corpus del Español (Davies, 2002-). Examples are identified according to source and genre.

(4) DIRECT OBJECT COMP, OPTIONAL MATERIAL
“*…**escribió** el conocido editorial [DO] en su revista madrileña [O]…*”
‘s/he **wrote** the well-known editorial [DO] in his/her Madrid-based publication [O]’ (España-ABC, news)

Tokens were categorized as having a Sentence Complement if followed by a complementizer *que* and a finite clause (5). Tokens that were followed immediately by a prepositional phrase (6) were coded separately and in some cases excluded because of the effect the preposition has on verb meaning, as in (7) where *contar* + *con* means ‘to rely on’ as opposed to ‘to tell (as in a story).’ Non-excluded tokens were ultimately collapsed with any other non-DO/SC tokens. Tokens taking an infinitival verb form (8) were also collapsed into this third OTHER group, as were other ambiguous complements, for instance where the token acted as a quotative verb (9).

(5) SENTENCE COMP
“*…Ortega y Gasset **escribió** que [C] lo humano se escapa a la razón físico-matemática…*”
‘Ortega y Gasset **wrote** that [C] that which is human escapes physical-mathematical reason’ (Entrevista-ABC, oral)

(6) PREPOSITIONAL COMP (included)
“*…Federico García Lorca **escribió** sobre [P] sus compañeros de generación…*”
‘Federico García Lorca **wrote** about [P] the peers of his generation” (España-ABC, news)

(7) PREPOSITIONAL COMP (excluded)
“*…siempre **contó** con [P] un grupo de files (sic) acólitos…*”
‘s/he always **relied** on [P] a group of faithful acolytes’ (España-ABC, news)

(8) INFINITIVAL COMP
‘*…Picasso **decidió** participar [I] en aquella aventura…*”
‘Picasso **decided** to participate [I] in that adventure’ (España-ABC, news)

(9) QUOTATIVE
“*…**Preguntó** Alberdi: - ¿ Qué pasó? [Q]...*”
‘Alberdi **asked**, ‘What happened?’ [Q] ’ (*Tiempo de silencio*, fiction)

4.2. Animacy of the subject and direct object

For both the subject and direct object, we coded for whether each was animate or inanimate. We coded all humans and groups of humans as animate, and all non-human objects as inanimate. Given that higher animacy in the subject relative to the object correlates with high transitivity in the clause (Thompson & Hopper, 2001), we hypothesized that higher animacy arguments might motivate DO verb complements. However, the particular lexical types chosen for the present study refer largely to human activities. For this reason it was not surprising to find that our subjects and objects were almost categorically animate and inanimate respectively, as will be presented below. Regardless of complement type or verb, animacy does not seem to be a motivating factor in this particular analysis.

(10) ANIM SUBJ, INANIM OBJ
“*…Lo cierto es que Rimski [A] **escribió** más obras [I] que su « Sherezade »…*”
“…What’s certain is that Rimski [A] **wrote** more works [I] than just his ‘Scheherazade’...” (España-ABC, news)

4.3. Form of the subject and direct object

In the Dussias et al. (2010) study, all sentences began with a full proper noun phrase as a subject, followed immediately by the 3sg-conjugated verb. The *One Lexical Argument Constraint* (DuBois, 1987) is relevant to possible effects of the form of the subject (lexical, pronominal, unexpressed). This constraint of Preferred Argument Structure (PAS) states that even transitive verbs having two core arguments will only express one of those arguments using a full lexical NP. Also, DuBois (2003) argued that the PAS of Spanish actually disfavors the surfacing of a Full NP as the subject of transitive verbs. Rather, Full NPs are more likely to surface as the subjects of intransitive verbs, and as the object arguments of transitive verbs. We therefore hypothesized that the use of Full NPs as the obligatory subjects in the Dussias et al. materials (participants were only able to add information after the subject and verb provided, not before it) may have influenced participants in that study to select more sentential complements than lexical direct object complements. Moreover, this may be the case in natural data as well, in line with DuBois (2003): the heavier the subject constituent, the less likely speakers may be to select lexical DOs as complements. We distinguished five types of subject form: proper name (11), definite NP (12), indefinite NP (13), pronominal subject (14), or null subject (15).

(11) PROPER NAME SUBJ
“…*Guillermo de Torre [PN S]* ***escribió*** *sobre Simplismo (1925) en «Revista de Occidente»…*”
“…Guillermo de Torre [PN S] **wrote** about *Simplismo* (1925) in the ‘Revista de Occidente’…” (España-ABC, news)

(12) DEFINITE NP SUBJ
“*…El mochuelo [DEF NP S]* ***pensó*** *que tal como se habían puesto las cosas, lo mejor era callar…”*
“…The owl [DEF NP S] **thought** that the way things had become, it was better to shut up…” *(El camino,* fiction*)*

(13) INDEFINITE NP SUBJ
“…mucho después de mi regreso a Nueva York (Julio de 1974) un amigo [INDEF NP S] me ***contó*** *la siguiente historia…”*
“...long after my return to New York (June of 1974) a friend **told** me the following story…” (España-ABC, news)

(14) PRONOUN SUBJ
“…Usted [PRO SUBJ] ***escribió*** *un " Poema sinfónico, " para cien metrónomos…”*
“…You [PRO SUBJ] **wrote** a “Symphonic poem” for one hundred metronomes…” (Entrevista-ABC, oral)

(15) NULL SUBJ
“…Y Ø [NULL SUBJ] ***encontró*** *experimentalmente que la frecuencia va con la raíz cuadrada de la tensión…”*
“…And Ø [NULL SUBJ] found experimentally that the frequency is the square root of the tension…”(España Oral: PEDU012C, oral)

It is important to note that we had initially coded proper names as definite NPs, and later separated them out because they accounted for a high proportion of the definite NPs. Although this distinction has not been directly addressed in previous PAS literature to our knowledge, we argue here that proper nouns may behave differently from other definite NPs because they are often discourse given. In our final multivariate analysis, we made three category distinctions according to subject form: definite NP, proper name, and reduced form (which included indefinite NP, pronominal subject, and null form).

Additionally, given that certain object forms like clitic pronouns could not be inserted before the verb in the norming study, participants may have used object forms in the study that do not actually surface at such high rates in natural production. In order to determine if this might have an effect on the type of complement selected, we coded objects for the same forms as those coded for subject, with the exception of the null category and proper name distinction, and with the addition of clitic pronouns. Thus, we coded for proper name, definite NP, indefinite NP, clitic pronoun, or other pronoun type.

It is important to note that the form of the object was only a factor in those tokens in which the verb took an object complement (either direct or indirect). For instance, in the case of sentential complements or quotative complements, coding for the direct object form was not applicable.

4.4. Presence of an indirect object

In coding for the presence of an indirect object, we hypothesized that its inclusion might affect the likelihood of one or the other complement types appearing. In tokens in which the verb took only an indirect object complement, this may reveal a bias toward taking additional complements only after core arguments have been taken and may thus favor direct object complements. We also noted that Dussias et al. (2010) did not provide an opportunity for the indirect object clitic in their norming study's stimuli. We wanted to be sure to examine the effect this construction might have, since in previous studies the stimuli did not include indirect object clitic pronouns (16), and clitic doubling occurs at high rates in all varieties of Spanish.

(16) INDIRECT OBJECT PRONOUN
"*Él mismo* ***me*** *[IO] escribió una carta [DO]…" "*
"He himself wrote **me** [IO] a letter [DO]..." (*Los hombres de a caballo*, fiction)

5. Findings and Analysis: Verb Bias Results

Following Dussias et al. (2010), our nine distinct complement types were collapsed into three general types: Direct Object bias (DO, including both those constructions that had additional complement material like indirect objects and optional material, as well as those that did not); Sentential Complement bias (SC, including both *que* complements and other types of complementizers, i.e.: *sino, si, como*, etc.); and Other bias, which includes all other complement types that we coded for, including infinitivals, prepositional phrases, constructions with indirect objects but no explicit direct objects, and quotatives.

Taking the percentage of occurrences of each of the aforementioned groups for each verb, we determined the verb bias of each verb according to the corpus data. We followed the same method as used in previous studies of verb bias (Trueswell, Tanenhaus, & Kello, 1993; Garnsey et al., 1997; Wilson & Garnsey, 2009; Dussias et al., 2010). In order for a verb to be considered to have a DO-Bias, the percentage of its DO complements had to be at least twice as high as the percentage of its SC complements. The reverse had to be true in order for a verb to be considered SC-Bias, with the percentage of its SC complements being at least twice that of its DO complements. Verbs which met the first requirement but which did not demonstrate at least a 15% difference between the likelihood of occurring with DO or SC were considered to have an EquiBias (EQ_Bias). Verbs which met none of these criteria were listed as having No Bias. The results can be found in Table 1.

Table 1. Verb bias results: percentage of Direct Object (DO) or Sentential Complement (SC) (N = 722)

Verb entry (N)	DO	SC	OTHER	Bias in present study	Agrees with Dussias et. al. (2010)?
contó (31)	68	16	16.2	DO_Bias	YES
encontró (41)	95	4.9	0	DO_Bias	YES
escribió (184)	64	6	30.4	DO_Bias	YES
oyó (11)	100	0	0	DO_Bias	YES
pensó (103)	9.7	46	44.6	SC_Bias	YES
respondió (47)	2.1	6.4	91.5	No Bias	YES
sintió (75)	73	17	9.4	DO_Bias	YES
contestó (17)	5.9	18	76.5	EQ_Bias	No
decidió (101)	4	6.9	89.1	No Bias	No
preguntó (112)	4.5	21	75	SC_Bias	No

We can begin by noting that out of 10 verbs analyzed in our corpus study, seven clearly demonstrate the same subcategorization frequencies as those found by Dussias et al. (2010). This is a successful step forward toward establishing Spanish verb biases for use in future psycholinguistic research. We see no need to discuss this concord any further at this time. It is more important for us to note why *contestó*, *preguntó,* and *decidió* did not demonstrate the same subcategorization bias in our data as was observed in the Dussias et al. norming study.

When the Other category for the verb *contestó* is examined in greater detail, which in these data shows EQ_Bias but in the norming study was No Bias, we see that approximately 18% of instances of this verb co-occurred with indirect object clitic pronouns. Overall 11.8% of completions found with tokens of this verb are constructions which have an indirect object but no DO. These are largely constructions bearing the clitics *me, te, le,* followed by a sentential complement or quotative material, as in (17) and (18) below:

(17) IO, SENTENTIAL COMPLEMENT

"*…me [IO]* ***contestó*** *con mucha inteligencia que [SC] un edificio con el que se intenta investigar tiene que estar lleno de errores*"

"… He **answered** me with considerable intelligence that a building that attempts to investigate has to be full of errors" (España-ABC, news)

(18) IO, QUOTATIVE MATERIAL

"*…Mira, Fernandito - me [IO]* ***contestó*** *-, lo que yo hice fue recoger un hilo…[Q]*"

" 'Look, Fernandito,' s/he **answered** me, 'what I did was pick up a thread…" (España-ABC, news)

We observe a similar pattern with the verb *preguntó* (which here shows No Bias but in the previous norming study showed a DO Bias), where 4.5% of tokens in the Other category are constructions with an IO and no DO, as seen in examples (17) and (18) above. In both cases IO clitic constructions represent a proportion of occurrences of the verb which is equal to or greater than that of a DO construction with that same verb. The different results for the present study's corpus data and the data collected by Dussias et al. can be accounted for by considering Dussias's norming materials. In that study, participants were given lists of sentence fragments, such as "*María contestó* ________," which they were asked to complete. These materials did not leave room for participants to include an indirect object clitic, so the probability of indirect object constructions for all verbs in the Dussias study is severely limited by the nature of the materials. In the case of the verbs *contestó* and *preguntó* we see that such constructions are fairly common and might have resulted in a different outcome in the norming data had indirect object clitics been permitted by the norming stimulus materials.

In the case of *contestó* and *preguntó* we also see high probability of these verbs occurring in constructions which we refer to as quotatives, as seen in example (9) in the previous section. In quotatives, the complement of the verb is a quotation of speech, such as that found in fiction or non-fiction narratives. For the verbs in question, such complements accounted for 59% and 70% of the data for *contestó* and *preguntó* respectively. Excluding these data, or, as an alternative, treating these tokens as DO complements would result in a significant change in outcome. However, each of these solutions is problematic for different reasons and without being sure how this would interact with the indirect object clitics we choose to leave them as Other for the purposes of this study. It may be argued that in fact, these quotative-selecting verbs are not the main verbs of the clauses at all. As Thompson (2002) argues for complement-taking predicates in English such as *think,* 3sg preterit (our *pensó*), such phrases are often used for epistemic, evidential or evaluative purposes to introduce the main idea which follows it and to show how the speaker is negotiating or aligning their stance with another. *Contestó* and *preguntó*, by indicating what kind of interaction the subject is having with the indirect object, seem to behaving in a similar way – as introducers of ideas rather than main ideas themselves.

We also observe that the results for *decidió* do not match the results found in Dussias et al. (2010). In the Dussias et al. data the probabilities meet the threshold to identify *decidió* as a SC Bias verb, but still a full 53% of the responses in those data were infinitive forms assigned to the Other category (P.Dussias, personal communication, December 22, 2010). In our data we find 89.1% of completions are identified as Other, and in fact, all 89.1% are infinitival constructions, as well. Still, the difference between the percentage of infinitival completions in Dussias et al.'s work as compared to our own is large. This overwhelming tendency for *decidió* to appear with a verb in its infinitive form may bring to light a need for more than just a DO/SC bias distinction. Speakers' high sensitivity to complement distributions surely do not ignore all complements that fall outside of the DO/SC binary relationship. Therefore, it would be instructive to take other additional complement types into consideration when determining verb biases, as well.

6. Findings and Analysis: Factors contributing (or not) to complement selection

Below in Table 2, we lay out each of the factor groups considered in our multivariate analysis, run using the Goldvarb X program (Sankoff, Tagliamonte & Smith, 2005). In variable-rule analysis, multiple factor groups can be simultaneously taken into consideration, while at the same time allowing for a determination of the relative contributions of individual factors. The difference between weights for variables within a given factor group (see, e.g., Presence of IO in Table 2 below) reflects its statistical significance within the multivariate analysis.

We compared only two complement types in our analysis: DO and SC, excluding all cases of Other. Due to the heterogeneity of the Other category, which collapses diverse complement types into a single group, its inclusion in the multivariate analysis would make little sense. This reduced the analyzable number of tokens to 289.

We see that overall, DO complements are more frequent than SC, as reflected in the input value of .72 (and overall rate of 71%). We find one factor group to be a statistically significant predictor of a Direct Object complement: the presence of an indirect object favored a Direct Object complement with a factor weight of .75. Note, though, that occurrences of the verbs with an IO and any complement type accounted for only 14.9% of the data (43/289 tokens).[3]

All other factor groups were found to be statistically non-significant. However, for the sake of discussion, we have additionally included the factor group Form of Subject in Table 2 below. Though not statistically significant, the direction of effect for the form of the subject is in the direction that we expected based on DuBois' (1987) Preferred Argument Structure. We find that proper names fall

[3] One caveat of this finding is that one verb, *sintió*, never appears with an IO in our data. Thus, the presence or absence of an IO cannot be predictive for this particular verb.

somewhere between full, definite NPs and reduced forms (subject pronoun or unexpressed). With a DO complement rate of 71%, proper names demonstrate behavior more similar to that of reduced forms (76%) than they do to that of the definite NPs, with a rate of 56%, disfavoring the selection of a DO complement. Note, too, that proper names make up a substantial proportion of the data, approximately one-third.

Table 2. Factors contributing to the selection of DO complements over SC complements in Peninsular Spanish (non-significant factor groups within brackets)

Factor Group	Factor Weight	% DO Complement	Total *N*	% Data
Presence of an IO				
Present	0.75	88.4%	43	14.9%
Not Present	0.45	68.3%	246	85.1%
Range:	30			
Form of Subject				
Definite NP	[.34]	56.4%	39	13.5%
Proper Name	[.49]	70.9%	103	35.6%
Reduced	[.55]	75.5%	147	50.9%
Range:	---			

N = 289/722	DO-complement: 71.3%	[non-significant factors]
Log likelihood: -173.291	Input: 0.722	

In addition to the analysis of the pooled data above, we also conducted separate analyses of these factors for each of our ten verbs alone, hypothesizing that perhaps certain factor groups that were not found significant across verbs might have been significant within verbs. However, almost every factor group was not selected as significant. Although for several of the verbs (e.g. *contó*), the relatively low number of tokens gathered may have been an issue, overall we do observe other tendencies in the data that corroborate findings of previous research. For instance, our subjects are nearly all animate and our objects nearly all inanimate. This marked difference in animacy between these two argument types is expected (Thompson & Hopper 1980), as humans tend to be the most common "actors" that speakers refer to. However, the near-categorical animacy of subjects held true across complement type, and thus had no effect on complement selection.

7. Discussion

We find that our corpus-based data align both with the norming results found by Dussias et al. (2010) and with the general findings on Preferred Argument Structure (DuBois 2003; Hopper & Thompson 2001), at least with respect to the complement biases of our verbs and with the form of their arguments. One of the caveats of our study, mentioned previously in §2, is that the effect of factors like the presence of a corresponding reflexive verb (e.g. *encontrar/encontrarse*) was not considered.

As with most corpus studies, the lack of information available about individual speakers may be a confounding factor. Granted, our goal is not to look at individual differences but rather the general tendency of speakers of a given population – in this case, monolingual Peninsular Spanish speakers. We cannot know for sure if any of the speakers that contributed to this corpus are in any way multilingual or in contact with other dialects. However, since our findings do align with the monolingual measures of Dussias et al.'s study, for which there *is* language history information, we can be relatively certain that our study is sound in this respect.

Further, Vázquez Rozas and García-Miguel (2006) observe that in a mixed-genre Spanish corpus, the most frequent classes of verbs across all grammatical persons are Mental (e.g., *sentir, pensar,* *decidir*), Relation (*tener, dar)* and Material Processes (*poner, crear*) verbs (macroclasses defined by the ADESSE Project, Albertuz, 2004). When looking specifically at the third person singular forms of the verbs, it was observed that the Mental Processes class (37.78%), which includes verbs such as *sentir, pensar* and *decidir*, is approximately twice as frequent as the next two most common classes, Relational (22.28%) and Material (19.95%). Since the corpus used in the present study, the Corpus del Español (Davies, 2002-), is composed of a similar variety of both written and oral data, we can expect a similar distribution of these classes in our overall data. It is important to note, however, as stated by Vázquez Rozas and García Miguel (2006), that verbs of mental processes are used much more frequently in the first person singular than in any other grammatical person. Of the verbs examined in this study, four are of the Mental macroclass: *decidió, oyó, pensó, sintió* (Proyecto ADESSE). Table 3 shows the token frequency per million words of the Corpus del Español of each verb in the 1sg and 3sg present and preterit forms.

Table 3. Token frequency of verbs of the Mental semantic macroclass, according to tense and person

	DECIDIR	OIR	PENSAR	SENTIR
Present				
1sg	1.40	9.29	**70.37**	**61.78**
3sg	**12.49**	**78.70**	53.94	48.94
Total	13.89	87.99	124.31	110.72
Preterit				
1sg	14.2	12.27	39.65	36.32
3sg	**50.83**	**17.92**	**53.94**	**56.26**
Total	65.03	30.19	93.59	92.58

*More frequent grammatical person indicated in **bold**

Since our data are limited to the third person singular form of the selected verbs, we were concerned that specific results for those types of verbs may not be able to be generalized to the verb as a whole because the frequency of the form and its behavioral characteristics could be skewed by third-person-only information. What we find instead is that within the corpus used for the present study, the 3sg form of the verb is more common than the 1sg form of the verb within the preterit tense for all four verbs. The frequency findings from the corpus search support the choice by Dussias et al. (2010) and the present study to use 3sg preterit forms to examine verb biases because the 3sg is more common within preterit tense of these verbs. While *oír* is two to three times as likely to occur in the Present as in the Preterit and *pensar* and *sentir* are just about equally likely to occur in the Preterit and the Present, *decidir* occurs four times more often in the Preterit than the Present. The survey illustrated in Table 3 demonstrates that the preterit tense shows more consistency in favoring the 3sg grammatical person over the 1sg across the different verb types, making it a preferable form to examine in the present study of verb biases.

In future studies, it may be illustrative to consider the reflexive forms excluded in the present study in comparison with their non-reflexive counterparts, given that the former sometimes have a higher token frequency. In such cases, the correlation between verb biases of the two forms (and the potential shifts in verb biases due to analogy in one or the other direction) may allow for an even more fine-tuned analysis of the factors that influence a given verb's bias. That is, while such reflexive and non-reflexive verbs generally have different semantic functions, the similarities in form between pairs may serve to maintain certain semantic links between them and lead to a mental representation of the two that is not as distinct as we assumed here.

The form of the subject displayed trends that may have interesting implications for PAS. As predicted, proper names behaved more like reduced subject forms, usually discourse-given information, than like definite NPs (the type of NPs with which they are usually classified). This is not surprising because proper names are more likely to be discourse-given than other lexical NPs, even definite ones. The data in the present study would therefore suggest a need to take a closer look at the behavior of proper names in argument structure/verb bias selection.

With respect to the form of the object, we acted on the assumption that all objects were equal. However, based on the literature on tracking NPs and verb transitivity (see Hopper & Thompson 1980:711; DuBois 2003:208), we know that this is in fact not true. Not all apparent arguments of a verb phrase are actually referential. In coding for object form the way that we did, we did not distinguish full NP objects that were actually acting as non-tracking NPs (and that are therefore semantically bleached) from those that carry greater semantic weight.

(19) TRACKING NP
"…Chopin ***escribió*** *sus dos conciertos [T] para piano cuando era igual de joven que yo ahora…"*
"…Chopin **wrote** his two concerts for piano when he was as young as I am now…"
(Entrevista ABC, oral)

(20) NON-TRACKING NP
"…Pero contempló el paisaje que le rodeaba y ***sintió*** <u>*miedo*</u> *[NT]…"*
"…But he contemplated the landscape that surrounded him and he **felt** afraid…" (*La melodía prohibida,* fiction)

Tracking NPs, like that in example (19) above, are used to refer (in this case) to an object that is in and of itself semantically and discourse meaningful (DuBois 2003:208). In other words, tracking NPs are manipulable discourse participants. Non-tracking NPs, like that in example (20), are not meant to be referred back to or "tracked" within the discourse. Rather, they act as orienting predications or sometimes have a classifying function. In this case, a predicating NP forms a predicate with the verb, for example the verb + NP construction *sentir miedo* has the same meaning as the verb *asustarse*, 'to be scared.' In future research it would be interesting to investigate whether the form of a subject is modulated by the presence of a DO that forms a prefabricated segments or "prefabs" (e.g. Erman & Warren, 2000) with the verb, resulting in what is in actuality an intransitive clause with only a single argument (as demonstrated by the *asustarse* example above).

8. Conclusion

To summarize, the data found in the present corpus study replicate the subcategorization frequency results of seven Spanish complement-taking verbs found in the Dussias et al. (2010) norming study. For those three verbs resulting in different results in this study, we point out that limitations in the norming stimuli and the restriction to a binary SC- or DO-Bias only can account for the different results. Given that the corpus used represents a wider and more varied sample of written and spoken language than that employed for the laboratory study, we believe that these results represent a solid and important step toward establishing accepted verb-bias information for highly frequent Spanish verbs.

This study has also investigated whether form of the subject, animacy of the subject, form of the direct object, animacy of the direct object and presence of an indirect object co-occur or may cause a verb to choose a particular complement as its preferred subcategorization frame. We found one factor to be significant: presence of an indirect object. Direct object complements were favored when an indirect object was present. The present study can also have implications for Preferred Argument Structure, given the fact that proper names do not appear to behave in exactly the same way as full NPs.

In sum, the findings of this study are important in two ways. First, we were able to contribute to the information available on Spanish verb biases through this corpus study and to also validate the findings of the study carried out by Dussias et al. (2010). Second, we were able to determine across verbs what factors favor one complement type over another. In future studies, we plan to look at individual verbs in greater detail (with more tokens, as well) in order to determine if these factors interact in different ways depending on the verb, as well as if different biases result from different factors in other tenses, moods or grammatical persons.

References

Albertuz, Francisco J. (2004). *Syntaxis, semántica y clases de verbos: Clasificación verbal en el proyecto ADESSE, VI Congreso de Lingüística General*. Santiago de Compostela, 3-7 de mayo de 2004.

Chambers, Craig G. & Cooke, Hilary. (2009). Lexical competition during second-language listening: sentence context, but not proficiency, constrains interference from the native lexicon. *Journal of experimental psychology: Learning, memory, and cognition*, **35**, 1029-40.

Davies, Mark. (2002-) *Corpus del Español* (100 million words, 1200s-1900s). Available online at http://www.corpusdelespanol.org.

DuBois, John W. (1987). The discourse basis of ergativity. *Language*, **63**, 805-855.

DuBois, John W. (2003). Discourse and grammar. In Michael Tomasello (ed.), *The new psychology of language: Cognitive and functional approaches to language structure*, vol. 2, 47-88. Mahwah, NJ: Lawrence Erlbaum Associates.

Dussias, Paola E., & Cramer Scaltz, Tracy R. (2008). Spanish-English L2 speakers' use of subcategorization bias information in the resolution of temporary ambiguity during second language reading. *Acta Psychologica,* **128***,* 501-513.

Dussias, Paola E., Marful, Alejandra, Gerfen, Chip, & Bajo Molina, María Teresa. (2010). Usage frequencies of complement-taking verbs in Spanish and English: Data from Spanish monolinguals and Spanish-English bilinguals. *Behavior and Research Methods*, **42 (4)**, 1004-1011.

Duyck, Wouter, van Assche, Eva, Drieghe, Denis, & Hartsuiker, Robert J. (2007). Visual word recognition by bilinguals in a sentence context: Evidence for non-selective access. *Journal of Experimental Psychology: Learning, Memory and Cognition*, **33**, 663-679.

Elston-Güttler, Kerrie E., Gunter, Thomas C., & Kotz, Sonja A. (2005). Zooming into L2: Global language context and adjustment affect processing of interlingual homographs in sentences. *Cognitive Brain Research*, **25**, 57-70.

Erman, Britt & Warren, Beatrice. (2000). The idiom principle and the open choice principle. *Text*, **20**, 29-62.

Gahl, Susanne, Jurasky, Daniel, & Roland, Douglas. (2004). Verb subcategorization frequencies: American English corpus data, methodological studies and cross-corpus comparisons. *Behavior Research Methods, Instruments & Computers*, **36**, 432-443.

Garnsey, Susan M., Lotocky, M.A., Pearlmutter, Neal J., & Myers, E. (1997). Argument structure frequency biases for 100 sentence-complement-taking verbs. Unpublished manuscript, University of Illinois at Urbana-Champaign.

Hare, M., McRae, Ken, & Elman, J. L. (2003). Sense and structure: Meaning as a determinant of verb subcategorization preferences. *Journal of Memory and Language*, **48**, 281-303.

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

Libben, Maya & Titone, Debra. (2009). Bilingual Lexical Access in Context: Evidence From Eye Movements During Reading. *Journal of Experimental Psychology: Learning, Memory & Cognition*, **35**, 381-390.

MacDonald, Maryellen C. (1994). Probabilistic constraints and syntactic ambiguity resolution. *Language and Cognitive Processes*, **9**, 157-201.

MacDonald, Maryellen C., Pearlmutter, Neal J., & Seidenberg, Mark S. (1994a). The lexical nature of syntactic ambiguity resolution. *Psychological Review*, **101**, 676-703.

MacDonald, Maryellen C., Pearlmutter, Neal J., & Seidenberg, Mark S. (1994b). Syntactic ambiguity resolution as lexical ambiguity resolution. In Keith Rayner (ed.), *Perspectives on sentence processing,* 123-153. Hillsdale, NJ: Erlbaum.

Marian, Viorica & Spivey, Michael. (2003). Competing activation in bilingual language processing: Within- and between-language competition. Bilingualism: Language and Cognition, **6**, 97-115.

Proyecto ADESSE: Base de Datos de Verbos, Alternancias de Diátesis y Esquemas Sintáctico-Semánticas del Español. (1.5 million words). Available online at http://adesse.uvigo.es.

Sankoff, David, Tagliamonte, Sali A., & Smith, Eric. (2005). Goldvarb X: A multivariate analysis application [Computer program]. Retrieved 31 March 2010 from http://individual.utoronto.ca/tagliamonte/Goldvarb/GV_index.htm.

Schwartz, Ana I. & Kroll, Judith F. (2006). Language comprehension in bilingual speakers. In Matthew Traxler & Morton Ann Gernsbacher (eds.), *Handbook of Psycholinguistics*, 2nd Edition, 967-999. Amsterdam: Elsevier.

Schwartz, Ana I., Kroll, Judith F., & Diaz, Michele. (2007). Reading words in Spanish and English: Mapping orthography to phonology in two languages. *Language and Cognitive Processes,* **22**, 106-129.

Thompson, Sandra A. (2002). 'Object Complements' and conversation: Towards a realistic account. *Studies in Language, 26(1)*: 125-163.

Thompson, Sandra A. & Paul J. Hopper. (2001). Transitivity, clause structure, and argument structure: Evidence form conversation. In Joan Bybee & Paul J. Hopper (eds.) *Frequency and the emergence of linguistic structure*, 27-59. Amsterdam/Philadelphia: John Benjamins.

Trueswell, John C., Tanenhaus, Michael K., & Kello, Christopher. (1993). Verb-specific constraints in sentence-processing: Separating effects of lexical preference from garden-paths. *Journal of Experimental Psychology: Learning, Memory & Cognition*, **19**, 528-553.

van Assche, Eva, Duyck, Wouter, Hartsuiker, Robert J., & Diependaele, Kevin. (2009). Does bilingualism change native-language reading? Cognate effects in a sentence context. *Psychological science,* **20**, 923-7.

van Hell, Janet G. & De Groot, A. M. (2008). Sentence context modulates visual word recognition and translation in bilinguals. *Acta psychologica*, **128**, 431-451.

Vazquez Rozas, Victoria & García-Miguel, José M. (2006). Transitividad, subjetividad y frecuencia de uso en español. *VII Congrés de Lingüística General*. Barcelona, 18 al 21 de abril de 2006.

Wilson, Michael P. & Garnsey, Susan M. (2009). Making simple sentences hard: Verb bias effects in simple direct object sentences. *Journal of Memory and Language*, **60**, 368-392.

Selected Proceedings of the 14th Hispanic Linguistics Symposium

edited by Kimberly Geeslin and Manuel Díaz-Campos

Cascadilla Proceedings Project Somerville, MA 2012

Copyright information

ISBN 978-1-57473-450-8 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:

Dietrich, Amelia J. and Colleen Balukas. 2012. A Corpus Study of Verb Bias in Spanish. In *Selected Proceedings of the 14th Hispanic Linguistics Symposium*, ed. Kimberly Geeslin and Manuel Díaz-Campos, 258-271. Somerville, MA: Cascadilla Proceedings Project. www.lingref.com, document #2670.