Running head: FREQUENT FRAMES IN SPANISH AND ENGLISH

# What's in the input? Frequent frames in child-directed speech offer distributional cues to grammatical categories in Spanish and English

Adriana Weisleder

Stanford University

Department of Psychology

450 Serra Mall

Stanford, CA 94305

Tel: (650) 723-1257

Fax: (650) 725-5699

adrianaw@stanford.edu

*Sandra R. Waxman

Northwestern University

Department of Psychology

2029 Sheridan Road

Evanston, IL 60208-2710

Tel: (847) 467-2293

Fax: (847) 491-7859

s-waxman@northwestern.edu

*corresponding author

Abstract

Recent analyses have revealed that child-directed speech contains distributional regularities that could, in principle, support young children's discovery of distinct grammatical categories (noun, verb, adjective). In particular, a distributional unit known as the *frequent frame* appears to be especially informative (Mintz, 2003). However, analyses have focused almost exclusively on the distributional information available in English. Because languages differ considerably in how the grammatical forms are marked within utterances, the scarcity of cross-linguistic evidence represents an unfortunate gap. We therefore advance the developmental evidence by analyzing the distributional information available in frequent frames across two languages (Spanish and English), across sentence positions (phrase medial and phrase final) and across grammatical forms (noun, verb, adjective). We found that frequent frames in each language did indeed offer systematic cues to grammatical category assignment. Yet we also identify differences in the accuracy of these frames across languages, sentences positions, and grammatical classes.

**What's in the input? Frequent frames in child-directed speech offer distributional cues to grammatical categories in Spanish and English**

To acquire a human language, children must not only learn individual words, but must also discover the distinct *kinds* of words that are represented in their language (or grammatical categories, e.g., nouns, verbs, determiners) and how they map to meaning. Even within their first years, infants make significant headway in this arena. By nine months, they distinguish between two very broad kinds of words: content words (e.g. determiners, prepositions) vs function words (e.g., nouns, verbs, adjectives) (Shi, Werker, & Morgan, 1999). By 13 months, they begin to make finer distinctions among the content words, teasing apart the grammatical form *noun* (e.g., "cat") and mapping this form specifically to objects and object categories (e.g., cats). Over the next several months, they make finer distinctions still, teasing apart the forms *adjective* and *verb*, and mapping each to its associated range of meanings (properties and events, respectively) (Waxman & Lidz, 2006 for a review; Waxman & Booth, 2003).

How do infants accomplish this task? Because languages vary not only in the grammatical forms that they represent, but also in the ways that each form is marked in the 'surface' of the input, infants' accomplishments must rest upon an ability to glean grammatical form class information from the surface structure of the utterances that they hear. But this claim -- that the surface structure of a language provides reliable cues to grammatical category assignment -- has been a controversial one. In the 1950's, structural linguists observed that words from the same grammatical category tend to appear in the same distributional environments (Harris, 1954). For example, words that appear with the morpheme –*ed* in English also tend to appear with the morpheme –*s*. Based on this observation, several researchers proposed that distributional regularities may serve as cues to grammatical form class assignment (Kiss, 1973; Maratsos, 1979, 1988; Maratsos & Chalkley, 1980). However, this proposal was not endorsed universally. At issue was whether there is in fact sufficient structure in the language input to support the acquisition of grammatical categories, and if so, whether infants are in fact sensitive to the kinds of regularities present (Chomsky, 1965; Pinker, 1987).

For decades this challenge appeared insurmountable. In recent years, however, researchers using computational tools to examine the language input have documented that the distributional evidence available in naturalistic, child-directed speech may in fact offer strong cues to grammatical form class assignment (Mintz, Newport, & Bever; 1995; Redington, Chater, & Finch, 1998; Cartwright & Brent, 1997). In a compelling recent demonstration, Mintz (2003) introduced the notion of *frequent frames*, a distributional pattern defined as two words that bracket one intervening word (e.g., *I __ you).* In an analysis of six corpora involving adult-child conversations, Mintz identified the most frequent frames in adults' speech. (See 1 – 6, below, for some representative examples.) He then considered whether within each such frame, the intervening words tended to belong to the same grammatical category. For some frames (e.g., 1-3), the intervening words were predominantly verbs; for others (e.g., 4-6), the intervening words were predominantly nouns. This suggests that *frequent frames* constitute a distributional unit that could, in principle, support the acquisition of distinct grammatical classes. Moreover, recent work reveals that young children are sensitive to distributional patterns like these (Gómez, 2000; Gómez & Maye, 2005; Mintz, 2006). Taken together, these recent demonstrations have breathed new life into the hypothesis that distributional information in the input could support the discovery of distinct grammatical forms.

| | |
|---|---|
| *1. I __ it* | *4. the __ and* |
| *2. you __ to* | *5. a __ on* |

*3. I __ you* 6. *the __ is*

However, evidence to this effect has thus far focused primarily on the distributional information available in English. Because languages differ considerably in how the grammatical forms are marked on the surface of utterances, the scarcity of cross-linguistic evidence represents an important gap to fill. Consider, for example, free word order languages like Turkish, in which the sequences in which words can appear vary freely. In such languages, the relevant distributional units may not be sequences of co-occurrences among words (as in the frequent frames for English); instead they may be sequences of co-occurrences among sub-lexical morphemes (Mintz, 2008). In principle, then, the distributional units that emerge as central in one language may differ from those that emerge in another.

Interestingly, we need not look as far as Turkish to appreciate the importance of cross-linguistic evidence. Even in languages more closely related to English, certain linguistic features may have consequences on the clarity and force of the distributional evidence for grammatical form classes. For example, Romance languages like French, Spanish and Italian exhibit considerable homophony among key function words including articles and clitic object pronouns. For example, in Spanish, "los" can function either as a determiner (e.g., "*los* niños juegan" ("*the* children play")) or as a clitic (e.g., "Ana *los* quiere así" ("Ana wants *them* like that")). This homophony, and its attendant distributional overlap, is a characteristic of many determiners (*la, los, las, una, unos, unas*) and other function words ("que" could mean *what* or *that*, "como" could mean *how*, *as*, or *eat*). Because such words are so frequent in the input, they often serve as framing elements in frequent frames. This type of homophony could therefore have adverse consequences on the clarity of distributional cues (Pinker, 1987; but see Cartwright & Brent, 1997). Chemla et al. (in press) recently reported that the distributional evidence in frequent frames in French remained robust, even in the face of homophony. However, because this analysis considered the input to only one French-acquiring child, additional cross-linguistic evidence is clearly warranted.

In addition to homophony, another linguistic feature that could have consequences on the clarity and force of a distributional analysis is noun dropping (Torrego, 1987; Snyder, Senghas, & Inman, 2001). Noun dropping refers to a process by which nouns are omitted from the surface of a sentence when their meaning is recoverable from context (e.g., "Quiero *una azul*." ("I want a blue (one).") "*El pequeño* está dormido." ("The little (one) is sleeping.")) Noun dropping, which is more ubiquitous in some languages than others, is relevant to distributional analyses because as a result of this syntactic process, nouns and adjectives often appear within the same frequent frames (e.g., following a determiner and preceding another element) (Waxman & Guasti, in press; Waxman, Senghas & Benveniste, 1997).

In view of these observations, our goal is to advance the cross-linguistic developmental evidence for distributional approaches in three key directions. First, we examine the distributional evidence available to young children acquiring Spanish and compare it to the evidence available to children acquiring English. Second, we consider the relative clarity of frequent frames for identifying the grammatical categories *noun*, *verb* and *adjective*. Third, we consider for the first time a new distributional environment. In addition to the frames described by Mintz (2003), in which two words serve as framing elements (e.g., *"you__ it."*), we also consider phrase-final sequences in which the utterance boundary serves as a framing element (e.g., *"the__."*). Our decision to consider these *'end-frames'* was motivated by evidence that for young learners in particular, ends of utterances have a privileged status (Slobin, 1973). Infants and young children are sensitive to the prosodic cues that signal phrase boundaries (Hirsh-Pasek, Kemler Nelson, Jusczyk, Wright Cassidy, Druss, & Kennedy, 1987), and more successfully identify words presented in utterance-final, as compared to utterance-medial position (Fernald & McRoberts, 1993; Shady & Gerken, 1999). Moreover, in infant- and child-directed speech, key words are often placed in utterance-final position and tend to receive exaggerated pitch peaks and increased durations (Fernald & Mazzie, 1991). Put succinctly, because infants and young children are attentive to phrase boundaries (and especially to those occurring in utterance-final position), if the ends of phrases contain information that is relevant to grammatical form (Gleitman & Wanner, 1982; Morgan & Newport, 1981), then *end-frames* may constitute a potent source of information.

**Method**

*Input Corpora*

We selected six parent-child corpora from the CHILDES database (MacWhinney, 2000); three in Spanish (Irene (Ojea, 2000), Koki (Montes, 1987), and María (López-Ornat, Fernández, Gallo, & Mariscal, 1994)), and three in English (Eve (Brown, 1973), Naomi (Sachs, 1983), and Nina (Suppes, 1974)). We analyzed the utterances of the adult-speakers in all sessions in which the child-speaker was 2;6 years or younger. The English corpora were among those previously examined by Mintz (2003). This insured that our execution of the frequent frames analysis mirrored that reported by Mintz, and provided a point of comparison for Spanish.

--------------------------------------
Insert Table 1 about here
--------------------------------------

*Distributional Analysis Procedure*

Gathering the frames. Following Mintz (2003), we defined a *frame* as two linguistic elements with one word intervening. We considered two different types of frames. In the case of *mid-*

*frames* (**A__B**)**,** the two framing elements were words (denoted by **A** and **B**), and the intervening word (denoted by **__ )** varies. This is the frame analyzed by Mintz (2003). In the case of *end-frames* (**A__.**), the first framing element was a word, the second an utterance-final boundary, and the intervening word varies. We segmented every adult utterance into 3-element frames. For example, the utterance "Look at the doggie over there" yielded five frames, four mid-frames ("look__ the", "at __ doggie", the __ over", "doggie __ there") and one end-frame ("over __.") We did not include any frames that crossed an utterance boundary.

<u>Selecting the frames</u>. Next, we tabulated the frequency of each frame and selected for analysis the 45 most frequent mid-frames and the 45 most frequent end-frames.[i] These constitute our two groups of *frequent frames*.

<u>Identifying the intervening words</u>. We then listed all of the intervening words (both types (e.g., *dog*, *cat*, *run*) and tokens (e.g., every instance of *dog, cat*, and *run*)) that appeared within each frequent frame; each list constituted a *frame-based category*.

<u>Identifying the grammatical category assignment of intervening words.</u> We assigned each intervening word to a grammatical category (noun, verb, adjective, preposition, adverb, determiner, wh-word, conjunction, or interjection). This step of the analysis was carried out by a native speaker of each language; a native Spanish-speaker categorized all the words in the Spanish corpora and a native English-speaker categorized those in the English corpora. To ensure that the same criteria were applied across the two languages, grammatical category assignments in each language were then checked by a Spanish-English bilingual. For any word that could be assigned to more than one grammatical category (e.g., in English, "party" is both a noun and verb; in Spanish, "vino" is both a noun and a verb), the corpus was consulted to identify the correct assignment.

*Quantitative Evaluation Procedure*

Our quantitative evaluation focuses on the accuracy, or consistency, of the frame-based categories.[ii] To compute Accuracy, we compared every pair of words that occurred within any frame (See Mintz (2003) and Redington et al. (1998) for full details). We identified each pair as either a Hit (the two words were members of the same grammatical category) or a False Alarm (the two words were members of different grammatical categories). We then calculated Accuracy by computing the proportion of Hits [Accuracy = Hits / (Hits + False Alarms)].

*Baseline Categorization: Comparison to Chance*

To obtain a baseline categorization measure, we used a Monte Carlo method. Specifically, we computed accuracy scores for random word categories that were generated for each corpus. To accomplish this task, the intervening words within any of the frame-based categories were randomly distributed to form 'dummy' categories, which matched the frame-based categories in size. This random shuffling of the intervening words was repeated 1000 times, with accuracy computed on each shuffle. Accuracy scores obtained from these 1000 shuffles provided a baseline against which to compare the results from the frame-based categories and compute significance levels. For example, if only 2 out of 1000 shuffles matched the score obtained by the frame-based categorization method, the frame-based category score was said to be significantly above chance with a probability of 0.002.

**Results and discussion**

Table 1 offers an overview of the corpora for each child.

*Comparing frequent frames (actual) to baseline (chance)*

Table 2 presents the Accuracy scores for the frequent frames, along with the corresponding baseline measures. The Accuracy scores for all corpora were significantly higher than baseline (all *p*'s < .001), documenting that there is indeed consistent distributional information within the frames that converges on grammatical categories. This replicates Mintz (2003) and extends the work to a new frame-type (end-frames) and to a new language (Spanish).

--------------------------------------
Insert Table 2 about here
--------------------------------------

*Comparing across languages, frame-types, and grammatical class*

We next asked whether there were systematic differences in accuracy between the two languages, between the two types of distributional environments, and between the grammatical classes. To address this question, we aggregated the frame-based categories for each language and frame-type[iii] and calculated the accuracy score within each. We then categorized each frequent frame by the modal grammatical class of its intervening words (e.g., if the frame contained more nouns than words from any other grammatical class, it was classified as a Noun-frame). We submitted these accuracy scores to a three-way ANOVA: language (English v Spanish) by frame-type (mid-frame v end-frame) by grammatical class (noun-frame v verb-frame v adjective-frame). See Table 3. A main effect of language indicated that accuracy was higher in English (M = .72) than Spanish (M = .60), $F(1,209) = 5.022$, $p = .026$. A main effect for frame-type indicated that accuracy was higher for mid-frames (M = .77) than for end-frames (M = .55), $F(1,209) = 13.374$, $p < .001$. A main effect for grammatical class revealed higher accuracy for verb-frames (M = .76) than for noun-frames (M = .71), and higher accuracy for noun-frames than for adjective-frames (M = .50), $F(2,209) = 5.154$, $p = .007$. All differences among these means were statistically reliable, Tukey's HSD, all *p*'s < .05 (see Fig. 1).

--------------------------------------
Insert Figure 1 about here
--------------------------------------

These findings provide support for the hypothesis that the clarity of the distributional information available in frequent frames varies across languages, and within languages, it varies across different distributional environments and grammatical form classes. In particular, the analyses reported here, coupled with a glance at Table 3, suggest that in both languages, frequent frames contain robust cues for identifying the grammatical forms *noun* and *verb*, but weaker cues for the form *adjective*. This outcome for adjectives, although consistent with our proposal, should be interpreted with some caution. Our analysis indentified numerous noun- and verb-frames in each language, and the accuracy of these frames tended to be high. By contrast, we identified only a handful of adjective-frames (three and six, for Spanish and English, respectively). Because these also included a number of nouns and verbs, their accuracy was relatively low. Interestingly, and as predicted, although this pattern was evident in both languages, a glance at table 3 suggests it was especially pronounced in Spanish

--------------------------------------
Insert Table 3 about here
--------------------------------------

*General Discussion*

The current evidence provides support for the claim that the distributional information in frequent frames contains cues that could be useful to young children as they establish the main grammatical categories of their native language. This work extends previous evidence in three ways. First, it broadens the empirical base, examining the input available to children acquiring Spanish as their mother tongue, and comparing it to the input available to children acquiring English. We demonstrate that even in the face of linguistic features that may render the distributional evidence in Spanish less clear on the surface (including homophony and noun drop), frequent frames nonetheless offer robust cues to grammatical form class.

Second, this work casts a wider distributional net, considering not only frames that occur in utterance-medial positions (mid-frames), but also frames that occur in utterance final position (end-frames). We demonstrate for the first time that in both English and Spanish, end-frames carry distributional cues to grammatical form class. Perhaps not surprisingly, the distributional information for end-frames (which, by definition, are less constrained by their surrounding elements than are mid-frames) is less accurate than that for mid-frames. Yet end-frame information, however noisy, may be especially useful to young children (Slobin, 1973), as they are better able to recognize words that occur at the ends of utterances (Fernald & McRoberts, 1993; Shady & Gerken, 1999).

Third, to the best of our knowledge, this is the first investigation to consider the relative accuracy of the distributional evidence available in frequent frames for the discovery of each of these grammatical categories. We found that frequent frames contained robust cues for the grammatical forms noun and verb, but weaker cues for adjectives, and that this pattern, although evident in both languages, appeared to be more pronounced in Spanish than in English.

On the whole, there were more commonalities than differences between the two languages, suggesting that this linguistic unit (the frame) stands as a powerful source of information for children acquiring either English or Spanish. At the same time, differences between the languages did emerge, which are likely due to linguistic features of the input. For example, in Spanish (as in other Romance languages) there is considerable homophony among function words (e.g., the word "la" can function either as a determiner (e.g., "*la niña* juega" ("*the* girl plays")) or as a clitic (e.g., "ella *la* puso aquí" ("she put *it(f)* here")). These function words, which are highly frequent, often emerge as framing elements, and because they are homophonous, they affect the clarity of the distributional cues. Consider, for example, the mid-frame *la* ____ *a*, which occurred frequently in Spanish. When *la* was used as a determiner, this frame picked out primarily nouns (e.g., *Lleva la muñeca a tu cuarto* (Take the doll to your room)), but when *la* was used as a clitic, the very same frame picked out primarily verbs (e.g., *No la vuelvas a tocar*. (Don't *it-clitic* go to touch again)). Examples like this, considerably more common in Spanish than English, compromised frame accuracy.

Noun-drop also appears to have compromised accuracy in Spanish, especially in cases in which determiners emerged as framing elements. Consider, for example, the end-frame *un* ___. in which the intervening words included nouns (e.g., *Dame un dulce*. (Give me a candy.)) and adjectives (e.g., *Eres un presumido.* (You are a vain (one).)) The distributional overlap between nouns and adjectives as a consequence of noun-drop was also evident, though less frequent, in mid-frames (e.g., *un poco de* (a little of)).

These observations suggest that certain features, including homophony and noun-drop, may indeed compromise the clarity of the distributional evidence available in frequent frames in Spanish. However, this is not to say that as a whole, the distributional evidence available to Spanish-acquiring children is weaker, less informative or impoverished, relative to English. After all, children acquiring Spanish and English exhibit comparable developmental milestones. For example, between 21 and 24 months, infants learning either English or Spanish begin to distinguish adjectives from nouns, mapping the former primarily to object properties (e.g., color, texture) and the latter to object categories (e.g., dog, animal) (Waxman, Braun & Weisleder, in prep). Instead, we suggest that children acquiring different languages will rely on different kinds of distributional information to establish the grammatical categories of their language.

More specifically, the most informative distributional cues in one language may differ from the most informative cues in another. For fixed word order languages like English, distributional evidence based on co-occurrences of words may be quite informative. In languages with rich inflectional morphology (e.g., Spanish, French, Italian), we suspect that additional cues to grammatical form class may be gleaned from distributions of morphemic, phonetic, or prosodic properties. For example, our results suggest that differentiating between the roles of different homophones (particularly when they are function words) might be important for grammatical class assignment. Therefore one potential avenue for future research may be to explore whether there are differences in the phonetic or prosodic characteristics of homophonic pairs.

It will also be informative to focus on developmental matters. There is considerable evidence documenting that features of parental input vary not only across languages but also across development. In particular, the complexity of infant-directed speech changes over the first few years. How might these developmental changes affect the clarity of the distributional evidence? Perhaps the earliest input, like the input we have analyzed here, offers clearer distributional evidence to support the discovery of grammatical form classes. However, it is also possible that the early input, which is characterized by short utterances and exaggerated intonational contours (Fernald & Simon, 1984), serves to regulate infant emotion (Fernald, 1992) and facilitate the identification of individual words in the continuous speech stream, but contains relatively little information to support the discovery of distinct grammatical forms.

In sum, we have shown that there are distributional regularities in the linguistic input to children acquiring either English or Spanish that could, in principle, support the acquisition of distinct grammatical categories. Of course, because infants are sensitive to distributional regularities in domains other than language (Fiser & Aslin, 2002), the evidence presented here has broader implications for development. The challenge facing infants, and infancy researchers, is to discover how to make use of these regularities across development and across languages.

## References

Brown, R. (1973). A first language: the early stages. Cambridge, MA: Harvard University Press.

Cartwright, T. A., & Brent, M. R. (1997). Syntactic categorization in early language acquisition: Formalizing the role of distributional analysis. *Cognition, 63*(2), 121-170.

Chemla, E., Mintz, T.H., Bernal, S., & Christophe, A. (in press). Categorizing words using frequent frames: What cross-linguistic analyses reveal about distributional acquisition strategies. *Developmental Science.*

Chomsky, N. (1965). *Aspects of the Theory of Syntax*. Boston, MA: MIT Press.

Fernald, A. & Simon, T. (1984). Expanded intonation contours in mothers' speech to newborns. *Developmental Psychology, 20,* 104-113.

Fernald, A. & Mazzie, C. (1991). Prosody and focus in speech to infants and adults. *Developmental Psychology, 27(2)*, 209-221.

Fernald, A. (1992). Human maternal vocalizations to infants as biologically relevant signals: An evolutionary perspective. In J.H. Barkow, L. Cosmides & J. Tooby (Eds.) The adapted mind: Evolutionary psychology and the generation of culture. Oxford, England: Oxford University Press.

Fernald, A. & McRoberts, G. (1993). Effects of prosody and word position on infants' lexical comprehension. Paper presented at the Boston University Conference on Language Development, Boston.

Fiser, J. & Aslin, R.N. (2002). Statistical learning of new visual feature combinations by infants. *Proceedings of the National Academy of Sciences, 99(24)*, 15822-15826.

Gleitman, L.R. & Wanner, E. (1982). Language acquisition: The state of the art. In L.R. Gleitman & E. Wanner (Eds.), *Language acquisition: The state of the art.* Cambridge, England: Cambridge University Press.

Gómez, R. L. (2002). Variability and detection of invariant structure. *Psychological Science, 13(5)*, 431-436.

Gómez, R. L., & Maye, J. (2005). The Developmental Trajectory of Nonadjacent Dependency Learning. *Infancy, 7(2)*, 183-206.

Harris, Z. S. (1951). *Structural Linguistics*. Chicago, IL: University of Chicago Press.

Hirsh-Pasek, K., Kemler Nelson, D.G., Jusczyk, P.W., Cassidy, K.W., Druss, B., & Kennedy, L. (1987). Clauses are perceptual units for young infants. *Cognition, 26(3)*, 269-286.

Kiss, G. R. (1973). Grammatical word classes: A learning process and its simulation. *Psychology of Learning and Motivation, 7*, l-41.

López-Ornat, S., Fernández, A., Gallo, P., & Mariscal, S. (1994). *La Adquisición de la Lengua Española* [The Acquisition of the Spanish Language], Siglo XXI, Madrid.

MacWhinney, B. (2000). The CHILDES Project: Tools for anlyzing talk. 3rd Edition. Vol. 2: The Database. Mahwah, NJ: Lawrence Erlbaum Associates.

Maratsos, M. (1979). How to get from words to sentences. In D. Aaronson & R. Rieber (Eds.), *Perspectives in Psycholinguistics.* Hillsdale, NJ: LEA.

Maratsos, M. P., & Chalkley, M. A. (1980). The Internal Language of Children's Syntax: The Ontogenesis and Representation of Syntactic Categories. In K. E. Nelson (Ed.), *Children's Language, Volume 2* (pp. 127-214): Gardner, New York.

Mintz, T. H., Newport, E. L., & Bever, T. G. (2002). The distributional structure of grammatical categories in speech to young children. *Cognitive Science, 26*(4), 393-424.

Mintz, T. H. (2003). Frequent frames as a cue for grammatical categories in child directed speech. *Cognition*, *90*(1), 91-117.

Mintz, T. H. (2006). Finding the verbs: distributional cues to categories available to young learners. In K. Hirsh-Pasek & R. M. Golinkoff (Eds.), *Action Meets Word: How Children Learn Verbs* (pp. 31-63). New York: Oxford University Press.

Mintz (2008). *Distributional Analysis as a Method for Categorizing Words in Infancy.* Paper presented at the International Conference for Infant Studies, Vancouver, Canada.

Montes, R. (1987). *Secuencias de clarificación en conversaciones con niños.* [Clarification sequences in conversations with children] *Morphe 3-4*: Universidad Autónoma de Puebla.

Morgan, J.L. & Newport, E.L. (1981). The role of constituent structure in the induction of an artificial language. *Journal of Verbal Learning and Verbal Behavior, 20,* 67-85.

Ojea, A. (2000). Llinàs-Ojea corpus: Irene. In MacWhinney, B. (Ed.), The CHILDES database.

Pinker, S. (1987). The bootstrapping problem in language acquisition. In B. MacWhinney (Ed.), *Mechanisms of language acquisition*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Redington, M., Chater, N., & Finch, S. (1998). Distributional information: A powerful cue for acquiring syntactic categories. *Cognitive Science, 22*(4), 425-469.

Sachs, J. (1983). Talking about the there and then: the emergence of displaced reference in parent-child discourse. In K. E. Nelson (Ed.), (4). *Children's language*, Hillsdale, NJ: Lawrence Erlbaum Associates.

Shady, M. & Gerken, L. (1999). Grammatical and caregiver cues in early sentence comprehension. *Journal of Child Language, 2,* 163-175.

Shi, R., Werker, J. F., & Morgan, J. L. (1999). Newborn infants' sensitivity to perceptual cues to lexical and grammatical words. *Cognition, 72,* B11-B21.

Slobin, D. I. (1973). Cognitive prerequisites for the acquisition of grammar. In C. A. Ferguson & D. I. Slobin (Eds.), *Studies of child language development.* New York: Holt, Rinehart & Winston.

Snyder, W., Senghas, A., and Inman, K. (2001). Agreement morphology and the acquisition of Noun-drop in Spanish. *Language Acquisition,9:2,*157-173.

Suppes, P. (1974). The semantics of children's language. *American Psychologist, 29*, 103-114.

Torrego, E. (1987). Empty categories in nominals. Unpublished manuscript. University of Massachusets, Boston.

Waxman, S.R., Senghas, A., & Benveniste, S. (1997). A cross-linguistic examination of the noun-category bias: Its existence and specificity in French- and Spanish-speaking preschool-aged children. *Cognitive Psychology, 32(3),* 183-218.

Waxman, S.R., & Booth, A.E. (2003). The origins and evolution of links between word learning and conceptual organization: New evidence from 11-month-olds. *Developmental Science, 6(2)*, 130-137.

Waxman, S.R. & Lidz, J. (2006). Early word learning. In D. Kuhn & R. Siegler (Eds.), *Handbook of Child Psychology*, 6th Edition, Volume 2 (pp. 299-335). Hoboken NJ: Wiley.

Waxman, S. R. & Guasti, M. T. (in press). Linking nouns and adjectives to meaning: New evidence from Italian-speaking children. *Language Learning and Development.*

Waxman, S.R., Braun, I.E., & Weisleder, A. (in prep) The breadth of adjective learning in English- and Spanish-acquiring infants.

Author Note

This research was supported by NIH R01 HD30410 to Waxman. Portions of this research were presented at the meetings of the Society for Research in Child Development (2007). We are indebted to E. Leddon for extensive discussions, and to M. Kaufman and T. Piccin for methodological expertise.

Address correspondence and reprint requests to Sandra Waxman, Department of Psychology, Northwestern University, 2029 Sheridan Rd., Evanston, IL 60208-2710. E-mail: s-waxman@northwestern.edu.

Footnotes

[i] All of the frames in the analyzed set surpassed a frequency threshold of 5% in proportion to the total number of frames in each corpus. Moreover, when we restricted the analysis to the set of 25 most frequent frames we obtained comparable results, suggesting that the results are robust even under more restricted conditions.

[ii] Here we report Accuracy for token frequencies. In fact, several different patterns converged on the same conclusion as those we report here. First, independent analysis on word *types* and *tokens* revealed the same pattern of results. Second, in addition to Accuracy, we also analyzed the Completeness of the frame-based categories. This measure considers the proportion of word pairs from the same grammatical category that were grouped together in the same frame (see Mintz, 2003 for details). Although we computed Completeness as well as Accuracy for every analysis, the results of these analyses were complementary. We therefore report exclusively on Accuracy because this measure better reflects our goal to determine whether frequent frames are equally accurate across different languages and distributional environments. Finally, note that an analysis that strives for high Accuracy (even at the expense of Completeness) would result in categories with high internal consistency. Once this is achieved, some of the categories could be merged (based, for example, on their degree of overlap) in order to achieve higher Completeness (Mintz, 2003).

[iii] There was a high degree of consistency in the *frequent frames* that were found in the three corpora from each language. This speaks to the validity of aggregating across corpora. We note here there were eight English frames that contained only three to five word-types. These frames were excluded from further analysis so that they would not artificially inflate Accuracy. Their inclusion would only increase the size of our effect.

**Table 1: Descriptive statistics for each corpus**

| | *Corpus* | *Number of utterances* | *Word tokens categorized* | *% corpus categorized*[a] | *Word types categorized* | *% corpus represented*[b] |
|---|---|---|---|---|---|---|
| ***English*** | Eve | 13,743 | 3,315 | 5.91% | 393 | 64.89% |
| | Nina | 14,478 | 6,296 | 8.58% | 451 | 64.61% |
| | Naomi | 7,148 | 1,814 | 6.24% | 359 | 53.09% |
| | **Mean** | **11,790** | **3,808** | **7%** | **401** | **61%** |
| ***Spanish*** | Koki | 4,585 | 815 | 0.51% | 220 | 44.36% |
| | Irene | 22,087 | 3,315 | 6.09% | 393 | 75.68% |
| | María | 10,916 | 2,079 | 4.35% | 433 | 66.39% |
| | **Mean** | **12,529** | **2,070** | **4%** | **349** | **62%** |

**[a] # of tokens categorized in frequent frames / total # of tokens in the corpus**
**[b] percentage of tokens in the corpus whose *types* were categorized by frequent frames**

**Table 2: Accuracy scores (token frequency) for each frame-type in**

| | | ***Mid-frames* (A___B)** | ***End-frames* (A___.)** |
|---|---|---|---|
| **English** | Eve | .93 | .83 |
| | (Baseline) | (.48) | (.40) |
| | Nina | .98 | .92 |
| | (Baseline) | (.46) | (.50) |
| | Naomi | .96 | .85 |
| | (Baseline) | (.47) | (.43) |
| | **MEAN** | **.96** | **.87** |
| | **(Baseline)** | **(.47)** | **(.44)** |
| **Spanish** | Koki | .82 | .75 |
| | (Baseline) | (.24) | (.38) |
| | Irene | .68 | .52 |
| | (Baseline) | (.33) | (.34) |
| | Mar’a | .75 | .66 |
| | (Baseline) | (.34) | (.33) |
| | **MEAN** | **.75** | **.64** |
| | **(Baseline)** | **(.30)** | **(.35)** |

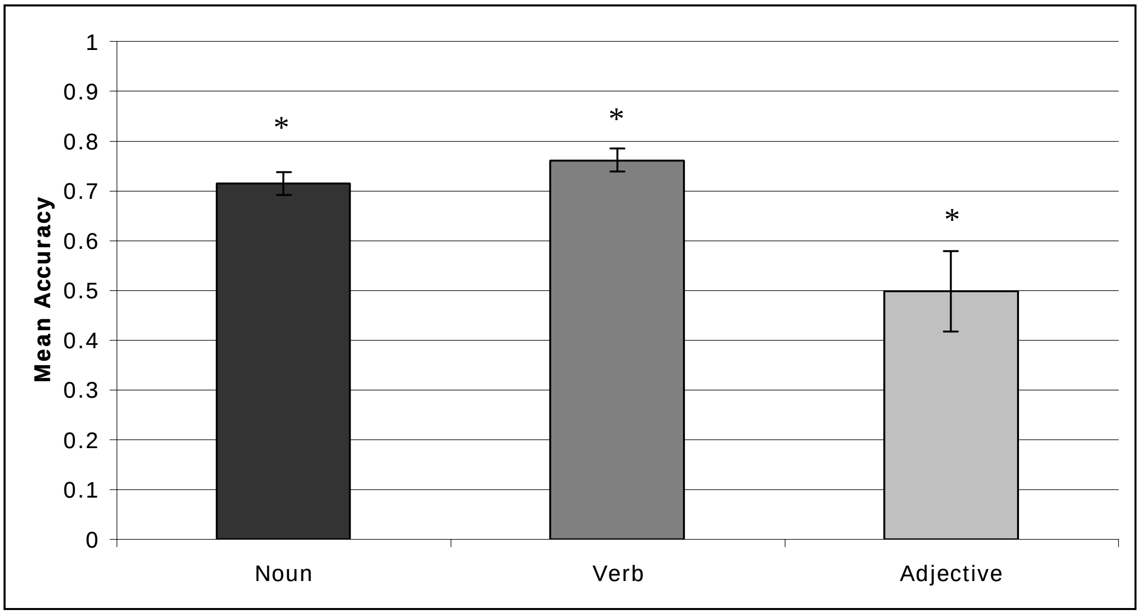
**Table 3: Accuracy scores for each Language, Frame-type, and Grammatical class**

| | | Mid-frames | End-frames | Gram. class Mean | Language Mean |
|---|---|---|---|---|---|
| ***English*** | Noun | .940 | .615 | .778 | **.722*** |
| | Verb | .961 | .612 | .787 | |
| | Adjective | .723 | .488 | .601 | |
| ***Spanish*** | Noun | .643 | .659 | .651 | **.592*** |
| | Verb | .824 | .646 | .735 | |
| | Adjective | .497 | .285 | .391 | |
| **Frame-type Mean** | | **.765**** | **.551**** | | |

*** Significantly different from each other, $p < .05$.**
**** Significantly different from each other, $p < .001$.**

## Figure Captions

Figure 1: Mean Accuracy scores for Noun-, Verb-, and Adjective-frames




* All means significantly different from each other, $p < .05$.

# Appendix

Representative examples of Noun-, Verb- and Adjective-frames for each language (English and Spanish) and frame-type (mid- and end-frames)

## English mid-frames

*the __ is* (139 tokens, 78 types)
baby(2), bag(1), barn(1), basket(1), bear(2), blanket(1), book(1), bookcase(1), box(2), boy(3), brush(1), bug(3), bus(1), car(2), chicken(1), clip(1), cord(1), couch(1), cow(2), cream(1), crib(1), crumb(1), dinosaur(1), dog(4), doggie(3), doll(1), donkey(1), door(2), duck(4), egg(1), elbow(1), elephant(3), family(1), fish(2), floor(1), flower(1), fox(2), girl(2), gob(1), grass(1), head(1), horse(8), house(2), ice(1), juice(1), kangaroo(1), kite(1), kitty(3), lady(2), lamb(2), lipstick(1), lock(1), man(5), mommy(2), monkey(1), moon(6), mouse(3), neck(1), nest(1), nurse(1), ocean(1), paper(1), pig(2), powder(1), rabbit(4), radio(1), rest(1), roof(1), rooster(1), sleeper(1), smoke(2), squirrel(2), sun(6), this(1), tray(1), truck(4), window(1), zoo(1)

*you __ the* (446 tokens, 103 types)
arranging(1), at(1), ate(1), bite(1), boo(1), break(1), bring(7), broke(2), catch(3), chase(1), cleaning(1), close(6), closed(3), closing(3), cook(1), count(1), cover(1), crack(5), cracked(1), cracking(2), cut(1), cutting(4), do(1), draw(2), drawing(1), drink(1), drop(1), dropped(6), eat(1), eating(3), fed(4), feed(8), find(11), finish(1), fit(4), fix(3), for(3), found(2), get(14), give(6), giving(1), got(6), hang(1), have(3), hear(5), held(1), hide(1), hitting(1), hold(6), hugging(1), hurt(1), keep(1), knock(1), know(1), leave(1), let(1), like(27), loving(1), make(14), making(6), mean(1), moving(1), on(4), open(4), pat(2), patting(2), paying(1), pick(1), popped(1), pull(1), pulling(3), push(3), pushing(3), put(73), putting(21), read(9), reading(1), remember(4), roll(4), rolling(1), run(1), say(1), see(14), shut(2), smell(1), spilled(1), take(17), taking(4), tearing(2), think(4), throw(5), throwing(1), to(2), took(4), tore(1), turn(6), unscrew(1), unsticking(1), unwind(1), unwrapping(1), use(1), want(30), washing(1), with(4)

*a __ one* (50 tokens, 17 types)
big(5), black(2), blue(4), brown(2), chocolate(1), fine(1), grape(2), green(7), little(7), new(4), nice(3), nother(1), purple(5), red(2), touch(1), white(1), yellow(2)

## Spanish mid-frames

*la __ de* (247 tokens, 121 types)
actuacion(1), amiga(3), barba(1), basura(1), boca(1), bolsa(4), botella(1), busque(1), cabeza(2), caca(2), caja(1), cajita(1), cama(3), cámara(1), camisa(1), camiseta(1), canción(20), cancioncilla(1), cantabas(1), cantidad(1), cara(15), carita(2), casa(9), casita(12), chaqueta(1), cinta(1), cola(1), colita(1), comida(3), cosita(1), cremallera(1), cuna(2), cunita(2), electricidad(1), encima(1), era(3), espalda(3), esponja(2), esquina(1), falda(1), flor(2), foto(2), gorra(1), granja(1), habitacion(4), hermana(2), historia(7), hoja(2), hojita(1), hora(2), jefa(1), llave(1), madrastra(3), madre(1), maestra(2), mama(12), manina(1), manito(3), mano(11), manta(1), manzanilla(1), merienda(1), mesita(1), mitad(2), muevas(2), musiquita(3), naricina(1), naricita(1), nariz(1), novia(1), oreja(1), orejina(1), paciencia(1), papita(2), parte(2), patita(1), pelicula(5), pelota(2), piel(1), pierna(2), pieza(1), plusvalia(1), posicion(1), puerta(1), pulsera(1), resolucion(1), revista(4), rota(1), sabes(1), señora(4), sensacion(1), silla(2), tapa(2), tarta(2), tenemos(1), tienda(8), trata(1), ultima(1), uña(2), vaca(2), vana(1), vida(1), voz(3), zapatilla(1)

*le __ a* (169 tokens, 42 types)
aviso(1), cantamos(1), cogiste(1), cuentas(2), cuentes(3), da(1), dabas(1), das(1), decia(1), decias(1), dice(3), dices(10), dices(9, digas(1), dijiste(8), doy(1), ensenias(1), gu(1), gusta(3), gustaban(1), gustan(1), hace(1), haces(1), hiciste(3), iba(1), ibas(3), pasa(4), pasaba(2), paso(2), pediste(1), pegas(1), perdio(2), pone(1), quieres(1), regalamos(1), toca(5), va(14), vamos(16), van(1), vas(60), ve(1), voy(5)

*un __ de* (122 tokens, 37 types)
amigo(1), bastoncillo(1), beso(1), bibi(1), bocata(1), cachito(1), cartel(1), censo(1), chanchito(1), cuento(1), daikiri(1), dibujito(2), galito(1), gnomo(1), hueso(2), huevo(1), juego(2), librito(2), lunar(1), oco(1), osito(1), paquete(2), par(2), pedacito(4), pellejito(1), pendiente(2), poco(29), poquitito(1), poquito(45), pupuyu(1), rojo(1), solomillo(1), sorbito(3), tipo(1), trozo(1), vaso(1), zapato(2)

## English end-frames

*that*__. (378 tokens, 158 types)
afterwards(1), airplane(2), alright(3), animal(4), apart(1), awful(1), baby(3), be(1), becca(1), better(3), bibbie(1), blanket(2), block(1), blouse(1), blue(1), book(5), bottle(3), box(3), boy(2), button(3), cake(1), called(15), card(2), carton(1), cereal(1), chair(1), chirp(1), clay(1), cookie(1), corner(2), cute(1), d(1), daddy(1), darling(1), dog(3), doggy(2), dolly(3), door(1), easterbunny(2), either(1), eve(9), first(1), fit(3), flower(1), foot(2), for(2), frasers(1), frosting(1), fun(1), funny(3), game(2), girl(2), go(3), good(3), gun(1), hand(1), hard(1), hat(1), help(1), here(2), hole(3), home(3), honey(6), horse(2), hot(2), house(3), hurt(2), hurts(1), in(1), ink(1), is(18), it(4), jar(1), jumped(1), kangaroo(1), kitty(1), knife(1), lamb(1), leila(1), letter(1), lion(1), long(1), makes(1), man(4), many(2), mean(1), means(2), mess(1), milk(1), mom(2), naomi(2), necessary(1), needs(1), nina(3), noise(6), nomi(24), off(2), okay(1), on(3), one(25), page(1), painful(1), papa(1), paper(2), part(3), pen(2), pencil(1), picture(13), piece(2), pillow(1), plate(2), please(1), pool(1), pot(1), present(2), pretty(1), puppet(1), puzzle(1), rabbit(2), racketyboom(2), rain(1), red(1), right(8), round(1), sarah(2), seat(1), shoe(2), side(2), song(1), sounds(1), soup(1), space(1), spell(1), spells(2), spoon(1), squeezes(1), stick(1), stool(1), story(2), stuff(1), sweetie(3), then(1), there(3), thing(2), tonight(1), too(2), tower(1), toy(1), train(2), tree(1), trip(1), two(1), valentine(2), water(1), way(14), what(1), window(2), you(1)

*it*__. (223 tokens, 152 types)
again(35), all(4), alone(7), along(2), anymore(3), apart(3), around(4), away(16), awhile(2), back(12), back(7), becca(1), before(1), belong(4), belongs(1), better(2), big(1), black(1), blue(3), break(2), bright(1), broke(3), broken(2), called(6), carefully(1), clean(1), cold(3), comes(3), cromer(1), cute(4), did(2), didn't(1), do(2), does(5), doesn't(1), doing(2), down(11), dried(1), drip(1), drop(1), either(3), eve(8), fantastic(1), feel(2), fell(1), first(6), fits(3), fixed(1), flies(1), fly(1), from(2), fun(8), funny(3), go(19), goes(7), going(3), goldie(1), gone(1), good(13), green(1), hard(4), here(11), hit(1), hmm(1), honey(12), horse(1), hot(2), hurt(6), hurts(3), in(24), into(2), is(115), isn't(1), itches(1), later(5), linda(1), look(1), louder(1), melted(1), moves(1), need(1), nina(1), no(2), nomi(27), nomi's(1), normally(1), now(4), off(40), okay(1), on(23), open(2), out(11), out(14), please(1), popped(1), pretty(1), rachel(1), raining(3), rattles(1), red(2), rest(1), right(2), rolled(1), rubs(1), sad(1), say(2), scary(1), shut(1), sideways(1), silly(1), sleeping(1), slowly(1), softly(1), somewhere(2), special(1), spin(1), stew(1), stop(1), sunny(1), sways(2), sweetheart(2), sweetie(1), swims(1), that(1), then(4), there(8), this(1), to(3), today(2), together(9), tomorrow(1), too(7), tore(1), untied(1), up(27), warm(1), was(5), wasn't(1), went(2), what(1), where(2), whistles(2), whole(1), will(3), with(2), work(1), works(2), would(1), yeah(4), yes(4), yet(3), yourself(4)

*is*__. (728 tokens, 120 types)
asleep(1), awake(1), baby(2), barking(1), better(1), bigger(2), billowing(1), black(1), blue(7), broken(6), called(7), clipclop(1), cold(3), colleen(1), coming(1), cool(2), cromer(2), crying(1), dancing(1), different(1), doing(1), dolly(3), drawing(1), eating(4), empty(2), eve(7), evecummings(1), fine(1), finished(1), fraser(2), frasers(1), fred(1), froggy(1), frosty(1), funny(1), furniture(1), good(1), grass(1), green(1), happening(1), hard(1), he(37), heavy(1), here(3), home(1), honey(4), hot(5), humm(1), hungry(1), indeed(2), inside(1), it(179), jackie(1), jenko(1), jim(1), jumping(1), kanga(1), left(1), lemon(2), light(1), like(1), little(1), locked(1), missing(1), mommys(1), next(1), nice(2), nina(2), nomi(9), now(1), nuzzling(1), ohio(3), ok(1), one(1), open(1), out(3), outside(1), painting(1), papa(5), racketyboom(1), raking(1), reading(1), ready(1), red(4), ricci(1), right(1), roo(1), running(1), saying(1), she(17), sleeping(6), smaller(1), snoopy(1), soap(1), soup(1), stuck(3), sugar(1), tea(1), that(212), there(4), this(80), time(1), timmy(1), tired(3), today(1), too(1), upsidedown(1), walking(1), weak(1), wearing(1), what(4), where(1), white(1), who(1), woopsie(1), working(1), yawning(1), yellow(4), yes(1), yours(1)

**Spanish end-frames**

*los* __. (369 tokens, 171 types)
abrimos(1), amiguetes(1), ángeles(1), animales(1), anises(1), autobuses(1), bambis(2), barcos(2), barriletes(3), bebes(3), bebitos(1), besos(1), bibis(2), bichos(1), bocatas(1), bracitos(3), brazos(2), caballitos(1), caballos(2), cables(1), cacharritos(1), cacharros(1), calcetines(3), carros(2), catalanes(1), cereales(1), chicos(1), chiquitos(1), ciervos(1), coches(2), colores(1), columpios(5), comen(1), como(1), conejitos(1), conejos(2), conoce(2), contamos(2), cordones(1), cristales(1), cucharros(1), cuento(1), cuentos(3), de(3), dedit(1), deditos(2), dedos(1), demás(13, descuidas(1), diablitos(1), días(2), dibujitos(1), dibujos(1), dientes(7), dos(9), elefantes(2), elotes(1), elotitos(1), enemigos(1), enseño(1), fantasmas(1), fideos(1), gallegos(2), gallos(1), ganó(1), gatos(1), guardamos(1), guardes(1), guardo(1), helados(1), hiciste(1), honguitos(6), huesos(1), jamases(1), juguetes(10), lavo(1), leotardos(2), libritos(5), mayores(1), meto(1), mezclamos(2), mickeys(1), mimitos(1), mocos(2), moquetes(1), moquitos(1), moros(2), mosqueteros(2), mosquitos(1), muebles(2), muñecos(3), muñequitos(4), nenes(12), nervios(2), niñitos(1), niños(15), números(3), ojinos(3), ojitos(5), ojos(8), ositos(2), osos(3), otros(3), pajarinos(2), pajaritos(1), palitos(1), palotes(1), pantalones(4), papás(3), papelitos(1), pasajeros(1), patitos(5),) patos(2), peces(3), pedacitos(1), pedales(2), pegotitos(1), pellejitos(1), pelos(1), pequeñitos(1), perdio(1), perritos(7), perros(1), pescaditos(4), petes(1), piececitos(1), pies(15), pifufos(5), pisos(1), pollitos(2), ponemos(1), pongas(1), poquís(1), portales(1), primos(1), que(1), quesitos(1), quiere(1), quieres(1), quillitos(2), recoges(1), regalamos(1), regaló(1), regalo(1), regalos(1), reyesmagos(7), sacó(2), saltitos(1), sapos(1), se(1), seco(1), semaforos(1), señores(4), silloncillos(2), techos(1), tiene(1), titos(1), toquen(1), tumbamos(1), tuyos(4), vagones(7), varones(1), vecinos(1), vera(1), ves(1), vestidos(1), videos(1), vio(1), viris(1), zapatitos(6), zapatos(8)

*te* __. (310 tokens, 134 types)
acalores(1), acercas(1), aclaras(2), acuerdas(39), advierto(1), ahogas(2), alejes(1), apetece(3), apuestas(2), atreves(1), ayudo(2), bajamos(1), bañamos(1), bañas(1), bañaste(2), busco(1), caes(4), caigan(1), caigas(1), cante(1), cantemos(1), cas(1), castiga(1), coja(1), conozco(1), costipas(1), cuente(2), cuento(1), da(2), decia(1), dejo(2), dice(4), dicen(2), diga(1), digo(13), dija(1), dijo(4), dio(1), diste(4), dolIa(1), doy(1), duele(3), eh(1), enfadabas(1), enfadas(3), enfadruscas(1), enseño(1), ensucias(1), entendí(1), enteras(1), entiende(1), entiendo(7), escapes(1), escondas(4), escuche(1), escucho(1), escurres(1), explico(1), expliques(1), gusta(27), gustan(4), gusto(3), hace(5), haga(1), hago(1), hizo(1), hundes(1), importa(2), lavan(1), levantas(1), levantaste(1), llamas(6), lleva(1), manchas(3), manches(2), mates(1), mato(1), moja(2), mojaste(1), mojes(1), molesta(2), moleste(1), muerda(1), ocurra(1), oí(2), oigo(1), oye(1), parece(11), parece(4), parezca(1), pasa(11), pasaba(1), paso(1), pega(1), pegas(1), peinamos(1), peino(2), perdono(1), pinchar(1), pone(3), pongo(1), preocupes(2), puso(1), qué(1), quemo(1), quiero(1), quito(1), quitó(1), rebelaste(1), ries(2), riño(1), sabes(2), sabrías(1), saco(1), sale(1), salió(1), seca(2), seco(1), seque(2), sequemos(1), sientas(e), sientes(1), tiraste(1), va(1), vas(2), vayas(1), ve(6), veamos(1), veía(1), ven(1), veo(3), vestimos(2), veyo(1), viste(2)

*está* __. (287 tokens, 158 types)
abajo(1), abierto(1), acá(1), acabando(2), adónde(1), agarrado(1), agusto(1), ahí(19), alto(1), angelines(1), apagada(1), apísima(1), apuntando(1), aquí(14), arregladita(1), arregladito(1), así(1), bailando(1), bañando(5), batido(1), bien(18), bonito(2), buenísimo(1), bueno(5), cabeza(1), calentito(1), caliente(6), cansada(1), cansadito(1), cariño(1), carlitos(2), carne(1), carnina(1), cerrada(1), cerrado(2), chica(1), chupando(1), colonia(1), columpiando(1), comiendo(1), cuqui(1), dentro(1), descalzada(1), descalzo(1), descansando(2), diciendo(3), disfrazado(1), distraído(1), donde(1), dormida(1), dormido(1), dura(1), durmiendo(2), duro(1), eh(3), en(1), enfermita(1), enseñasela(1), es(1), escuchando(1), esponja(1), esta(3), estornudando(1), fea(1), feya(1), fría(1), frio(1), fumanchu(1), grabando(1), gracias(1), gritando(2), grover(3), guapa(4), guapo(1), guardada(1), guardadita(1), gustando(1), hablando(1), haciendo(25), hay(1), hija(3), igual(1), irene(3), jugando(4), koki(3), la(1), ladrando(1), limpio(1), limpita(3), limpito(1), linda(1), llorando(5), lloviendo(4), luis(2), mal(1), mala(1), malito(1), mamá(2), mami(1), mano(1), maría(4), masticadin(1), mejor(1), merendando(1), mintiendo(1), mirando(1), mojado(1), nada(1), nena(2), niña(5), no(2), ocupado(1), oscuro(2), otra(2), paco(1), papá(5), pedrin(1), peque(1), pequeño(2), perrita(1), peterpan(3), pieza(1), prohibido(1), pulsera(1), qué(7), quedando(2), quien(1), quilli(1), quique(1), retorcido(1), revista(1), rica(4), rico(12),

rota(2), rotita(1), rotito(1), roto(2), sacando(1), santi(1), seco(1), semama(1), sentado(3), solito(2), sucio(2), suficiente(1), tampoco(1), tarde(4), terminamos(1), tito(1), tocando(1), todo(3), tolumpiando(1), triste(1), usado(1), vacilando(1), vale(1), verdad(2), vestida(1)