On the Syntax and Processing of \textit{Wh}-questions in Spanish

Grant Goodall  

\textit{University of California, San Diego}

It is often stated that minimalism has a “therapeutic effect” on syntactic theory. One sense in which this is true is that minimalism forces us to examine analyses that are suspiciously complex and stipulative and ask whether the apparent complexity and stipulation can’t be avoided by forcing ourselves to limit our analyses to a severely restricted set of formal mechanisms in the syntax.

It is partly in this spirit that I approach one of the most widely studied aspects of Spanish and other Romance languages: the inversion that is found in \textit{wh}-questions. The basic facts are given in (1) and (2).

\begin{enumerate}
\item[(1)a.] \textbf{Juan} compró una botella de vino.
\item[(1)b.] Compró \textbf{Juan} una botella de vino.
\end{enumerate}

\begin{enumerate}
\item[(2)a.] *Qué \textbf{Juan} compró?
\item[(2)b.] Qué compró \textbf{Juan}?
\end{enumerate}

‘Juan bought a bottle of wine.’

‘What did Juan buy?’

Subjects are generally allowed either pre- or postverbally, as in (1), but in \textit{wh}-questions, the preverbal subject is generally impossible (with certain exceptions to be discussed below), as in (2).


\*

Thanks to Robert Klunder, Maria Polinsky, and audiences at the University of Arizona and WCCFL 23 for many useful comments. Thanks as well to colleagues and students at the University of Texas at El Paso for their invaluable help with the experimental work reported on here. All errors are my own.

I will propose here that the reason (2a) has been so troublesome is that in the end this may not be a strictly syntactic phenomenon, at least not in the way that we’re used to thinking. I will show that the contrast seen in (2) gives evidence of being highly sensitive to well-known working memory effects, and that if we pursue the possible influence of constraints on working memory on the contrast in (2), we are able to account for many basic properties of this construction, the very properties that syntacticians have been struggling with for decades. If this is correct, we are able to leave the basic computational properties of the syntax in a much more minimalist state, without the undesirable stipulations. More concretely, the syntax in the traditional sense will not need to rule out (2a), since working memory will do this for us.

1. Filler-gap structures and working memory: basic assumptions

A well-known finding from the processing literature is that filler-gap structures (such as wh-questions) strain working memory capacity, because the filler (wh-phrase) must be held in working memory until it can be assigned to a gap. Moreover, the filler’s level of activation in working memory declines continuously once it is first processed, and when this activation level gets low, assigning the filler to a gap becomes more difficult. Gap-assignment occurs when processing the head that subcategorizes for the gap (Pickering & Barry 1991, Hawkins 1999).

Given these assumptions, it is clear that (2a) is more difficult for working memory than (2b). In (2a), qué ‘what’ must be held in working memory longer than in (2b), because the subcategorizing head compró ‘bought’ is processed later in (2a) than in (2b). This means that the activation level of qué will be lower in (2a) at this point, which in turn means that it will be more difficult to assign it to a gap.

2. Evidence that (2a) is excluded because of working memory constraints

The fact that (2a) is expected to strain working memory more than (2b) does not in itself tell us that this can account for (2a)’s unacceptability, but such a conclusion would gain plausibility if factors which are known to affect the ability to hold fillers in working memory were to also affect the acceptability of (2a). This appears to be the case, as will be shown in the sections below.
2.1 Intervening DPs

Given the basic properties of filler-gap processing that we saw in section 1, it stands to reason that the longer the subject DP which intervenes between the filler and the gap (i.e. between the wh-phrase and the subcategorizing head), the more difficult the sentence should be for working memory. In addition, since a D-linked or “referential” DP has a high processing load (deVicenzi 1991, Kluender 1998), a D-linked DP that intervenes between the filler and the gap will reduce the resources available for holding the filler in working memory, thus causing its activation level to decrease more quickly than usual. This accounts for the contrast in (3), where the less D-linked subject in (3a) permits wh-extraction more readily than the subject in (3b).

(3)a. That’s the article that we need to find someone who understands.
   b. That’s the article that we need to find the reviewer who understands.

This same effect seems to be evident in the paradigm in (2). To see this, consider the sentences in (4)-(6), where the ratings are from an experiment with 23 subjects judging sentences on a scale from 1 (“very bad”) to 5 (“very good”).

(4) Qué tú leíste en la biblioteca?
   ‘What did you read in the library?’
   Mean rating: 2.174

(5) Qué el niño leyó en la biblioteca?
   ‘What did the boy read in the library?’
   Mean rating: 1.913

(6) Qué los amigos de tu hermana leyeron en la biblioteca?
   ‘What did the friends of your sister read in the library?’
   Mean rating: 1.833

Notice that as the intervening subject gets longer and more D-linked, the acceptability decreases.

1. Subjects were initially given an instruction sheet and 4 practice sentences, followed by group discussion of the task and their ratings of the practice sentences. They were then given 10 test sentences. Each sentence was presented with an appropriate context, and forward/backward presentation was balanced across subjects. Subjects were encouraged to compare the test sentences. All subjects were native speakers of Spanish (native bilinguals were excluded).
2.2 Dialect variation

As has been widely reported, overt subject pronouns in Caribbean Spanish do not have the same discourse status as in standard Spanish (Contreras 1989, Toribio 2000). In particular, they seem to lack the contrastive or emphatic status that they have in other dialects. As a result, they should then present a lighter processing load than in standard Spanish, and we thus predict that an intervening subject pronoun in Caribbean Spanish should cause relatively little disruption to processing in a filler-gap structure. This prediction appears to be true, in that sentences such as (4) above are reportedly acceptable in Caribbean Spanish (Ordóñez and Olarrea 2001).

2.3 D-linked wh-phrases

As we have seen so far, the more D-linked an element is, the heavier its processing load. In the case of a D-linked intervening DP, this means that its own processing demands will begin to deplete resources, making it difficult to hold the filler in working memory until it can be assigned to a gap. In the case of a D-linked wh-phrase, on the other hand, its high processing load (and consequently high activation level in working memory) means that it will take longer to drop to the critically low activation level that would make gap-assignment difficult and is thus able to tolerate a longer wait until the subcategorizing head (and thus the gap) is processed. The effects of this may be seen in (7), where the D-linked head of the relative clause in (7a) allows for a long wait for the gap, while the wh-word in (7b) tolerates this wait much less well.

(7)a. That's the article that we need to find someone who can understand.
   b. What do you need to find someone who can understand?

   (Kluender 1998, see also Cinque 1990, Chung 1994)

The same effects are observed in the paradigm in (2). A D-linked wh-phrase, as in (8), is able to tolerate the delay in gap-assignment caused by the intervening subject better than a non-D-linked wh-phrase, as in (9).²

(8) Cuáles de esos libros Ana leyó?
   ‘Which of those books did Ana read?’

   Mean rating: 3.885

(9) Qué Ana leyó?
   ‘What did Ana read?’

   Mean rating: 2.192

². The experimental procedure here is the same as in footnote 1, but N=26.
Notice, however, that (8) is still less acceptable than (10), where the postverbal position of the subject means that there is no delay at all in locating the subcategorizing head and gap.

(10) Qué leyó Juan en la biblioteca?  
‘What did Juan read in the library?’  

Mean rating: 4.739

This three-way contrast among (8), (9), and (10) is just what our constraints on working memory would predict: (9) should be the most difficult case for assigning the filler to a gap, (8) should be somewhat better, and (10) should be (nearly) perfect.

2.4 Relative clauses

Against this backdrop, the behavior of subjects in relative clauses now makes sense. The head of a relative clause is clearly D-linked/‘referential’ (e.g. Chung 1994), so we expect it to be able to tolerate a longer wait for the subcategorizing head. In particular, it should be able to tolerate a preverbal subject, and indeed this is the case.

(11)a. el libro que Ana compró  
b. el libro que compró Ana  
‘the book that Ana read’

We would expect (11a) to be roughly on a par with (8), and this seems to be true (I do not have data indicating whether there is a subtle contrast between (11a) and (11b) such as that between (8) and (10)).

2.5 Lack of matrix vs. embedded contrast

Another clear prediction of this analysis is that there should be no contrast between matrix and embedded clauses with respect to the paradigm in (2). If (2a) is ruled out because the intervening subject strains the processor’s ability to keep the filler in working memory until it can be assigned to a gap, this same effect should be observed when the filler-gap dependency is located within an embedded clause, and indeed it is, as has been widely noted:

(12)a. *No sé [qué Juan compró]  
b. No sé [qué compró Juan]  
‘I don’t know what John bought.’

The problem of holding the filler in working memory across an intervening
subject is essentially the same in both (2a) and (12a).

2.6 Main verbs vs. auxiliaries

So far we have examined properties of the filler (the wh-phrase) and of the intervening subject, and we have seen that as we manipulate these in ways that should place a greater or lesser strain on working memory, the acceptability of the sentence decreases or increases correspondingly. Let us now turn to the verb, which plays an important role in our analysis because it is the head that subcategorizes for the filler and thus allows for assignment to a gap. This should hold true only for main verbs, however. An auxiliary verb will be useless in helping assign the filler to a gap, since the auxiliary provides no relevant subcategorization information. As expected, then, an auxiliary that is adjacent to the wh-phrase, as in (13a), does not help acceptability.

(13)a. *A quién había la madre de Juan visto? 
   who had the mother of Juan seen

b. A quién había visto la madre de Juan?
   who had seen the mother of Juan
   ‘Who had Juan’s mother seen?’ (Ordóñez 1997)

Acceptability increases dramatically, though, when the subcategorizing verb is close to the wh-phrase (as close as the syntax will allow), as in (13b). As Ordóñez (1997) points out, it is unlikely that (13a) can be explained away by a constraint prohibiting separation of the auxiliary and main verb, since in some other circumstances, such separation is possible.

2.7 Arguments vs. adjuncts

Even with main verbs, there is much variability in the degree to which the verb subcategorizes for the filler. In our paradigm case in (2), there is a clear subcategorization relation between the filler and the verb, so anything that disrupts the processor’s ability to hold the filler in working memory (such as the intervening subject) has a clear effect on its ability to assign the filler to a gap. When the filler is a clear adjunct, however, and by definition does not have a subcategorization relation with the verb, an intervening subject should have no effect. Finally, when the filler is a locative or temporal adjunct, there is a less direct relation with the verb, since these adjuncts are often taken to be arguments of a higher functional head such as Event, and the verb is the main overt indicator of the clause’s event structure.

With regard to the disruptive effects of an intervening subject, then, we are able to make a clear prediction. Argument fillers should be most affected,
locative and temporal fillers should be somewhat less affected, and true adjunct fillers should not show signs of being affected at all. This prediction is borne out by the following data, from the experiment described in footnote 1 (see Baković 1998 for related results).

(14) **Qué Juan leyó en la biblioteca?**
    ‘**What** did Juan read in the library?’

    Mean rating: 2.130

(15) **A quién Maria vio en el parque?**
    ‘**Who** did Maria see in the park?’

    Mean rating: 2.478

(16) **Dónde Ana compró el periódico?**
    ‘**Where** did Ana buy the newspaper?’

    Mean rating: 2.957

(17) **Cuándo José escribió la carta?**
    ‘**When** did José write the letter?’

    Mean rating: 3.043

(18) **Por qué Miguel trabaja tanto?**
    ‘**Why** does Miguel work so much?’

    Mean rating: 4.783

In all five sentences there is an intervening subject, but this is most disruptive when the filler is an argument, as in (14) and (15) ((15) may be somewhat better than (14) because *a quién* ‘who’ is slightly D-marked). With the locative filler *dónde* ‘where’ and the temporal filler *cuándo* ‘when’, the disruption is lessened, and with the adjunct filler *por qué* ‘why’, there does not appear to be any disruption at all.

### 2.8 Extraction out of embedded clauses

Let us now turn to cases where the filler is in the matrix clause and the gap is in an embedded clause. Looking first at just the embedded clause, we would expect to see the by now familiar pattern in which an intervening (preverbal) subject decreases acceptability of the sentence. This expectation is borne out, but there is a mitigating factor: The decrease in the activation level of fillers in working memory is asymptotic (Baboyanshev & Gibson 1999), so by the time the embedded clause is processed, the difference between having or not having an intervening subject is not great. This effect may be observed in (19) and (20) (from the experiment described in footnote 2).

(19)a. **Qué cree Juan que **María** leyó en la escuela?**
    ‘What does Juan think that María read at the school?’

    Mean rating: 4.385

b. **Qué cree Juan que leyó María en la escuela?**
    ‘What does Juan think that María read at the school?’

    Mean rating: 4.577

(20)a. **Qué dices que tus papás compraron?**
    ‘What do you say that your parents bought?’

    Mean rating: 4.346
b. Qué dices que compraron tus papás?
‘What do you say that your parents bought?’

As can be seen, the intervening subject in the embedded clause (in the (a) sentences) does lead to degradation in acceptability in relation to the non-intervening subjects (in the (b) sentences), but the effect is not nearly as pronounced as in simple clauses (cf. (10) vs. (14)).

With regard to the matrix clause in this context, we might expect that the position of the subject would have little effect, since the matrix verb does not subcategorize for the filler. However, the matrix verb does subcategorize for the clause which contains the gap, and we know from evidence in other domains that the relation between the matrix verb and the embedded clause plays a crucial role in filler-gap structures (e.g. Chung’s (1994) demonstration that in Chamorro wh-agreement, the matrix verb in environments like this agrees with the embedded clause that contains the gap, not with the gap itself). It is thus plausible that the matrix verb plays a major role in allowing the processor to assign a gap to the filler. Given this, it follows that processing will be greatly facilitated by having the verb adjacent to the filler, and hampered by allowing other material (such as a preverbal subject) to intervene between the filler and the verb. The facts seem to come out as expected, as shown in (21) and (22), where we see that an intervening subject leads to severe degradation.

(21) Qué cree Juan que María leyó en la escuela?
‘What does Juan think that María read at the school?’

(22) *Qué Juan cree que María leyó en la escuela?
‘What does Juan think that María read at the school?’

Quantitative data are not available, but impressionistically, (22) is as bad as (9).

A related phenomenon is observed with extraction of por qué ‘why’. When it is extracted out of an embedded clause, a matrix preverbal subject is prohibited, as shown in (23).

(23) *Por qué Juan dice que beberá cerveza? (from Ausín and Martí 2001)
‘Why does Juan say that he’ll drink beer?’

The same explanation given for (22) suffices to rule out this reading of (23) as well, since the matrix verb subcategorizes for the clause containing the gap. (23) could also be the result of movement of por qué from the matrix clause, though, and under this reading, (23) is perfect. Again, this is what we would predict, since under this reading the matrix verb does not subcategorize for the gap or anything containing it.
2.9 Satiation

As a further piece of evidence supporting the idea that sentences like (2a) are excluded because of working memory constraints, consider the phenomenon of satiation, in which unacceptable sentences increase in acceptability after repeated exposure. Snyder (2000) has shown that this phenomenon can be induced experimentally, and that certain sentence types are susceptible to it and others are not. He suggests that those that are susceptible might owe their unacceptability to reasons of processing. If this is correct, then it could provide a new tool for discerning the source of unacceptability in particular sentence types.

I explore this idea in an experiment reported in Goodall (2004), and I will briefly present here those aspects of that study that are relevant for our present purposes. After receiving instructions and a practice set of 4 sentences, 59 native speakers of Spanish were presented with 5 blocks of sentences, where each sentence was preceded by a context (“the situation”) and followed by the question Does this sound good? Possible answers were yes or no. Each block consisted of 4 grammatical and 6 ungrammatical sentences in random order, and the same sentence types were repeated (with varied lexical items) in each block. Each subject was thus presented with 5 tokens of each sentence type over the course of the experiment. Forward/backward presentation of the total set of 50 sentences was balanced across subjects.

Crucially for our purposes, one of the ungrammatical sentence types in the set was a wh-question with an intervening subject, as in (24a) (and similar sentences with other lexical items).

\[(24)a. *Qué Juan compró en la tienda?\]
\[b. *A quién habló José con Irma después de ver?\]

(24b) is given as an example of one of the other ungrammatical sentence types (an adjunct island violation).

Subjects were classified as “No → Yes switches” for a given sentence type when they gave one of the following response sequences for the five presentations of that sentence type: YYYYY, NYYYY, NNNYY, NNNNY. Likewise, “Yes → No switches” were those who gave one of the following response sequences for a given sentence type: YNNNN, YYNNN, YYYNN, YYYYY.

For the two sentence types given in (24), the results are as follows:
This shows a clear satiation effect for sentences like (24a), but not for (24b), thus supporting the idea that the intervening subject in \textit{wh}-subjects causes a problem for the processor, not the syntax.

3. A comparison with English

We have now seen evidence of various types that capacity constraints on working memory could be responsible for some of the basic properties of \textit{wh}-questions in Spanish. The paradigm that we started with in (2) is of course very similar to the English pattern in (25); in both languages the subject must appear in a postverbal (or post-auxiliary) position.

\begin{table}
\begin{tabular}{|l|c|c|c|}
  \hline
  Sentences & \textit{No} $\rightarrow$ \textit{Yes} & \textit{Yes} $\rightarrow$ \textit{No} & \textit{p} \\
  switches & switches & \\
  \hline
  (24a) & 12 & 1 & 0.00171 \\
  (24b) & 1 & 1 & 0.75 \\
  \hline
\end{tabular}
\end{table}

It is thus tempting to speculate that both (2a) and (25a) should be ruled out in the same way, and indeed, this has at times been a common assumption in the literature. We shall now see, however, that the paradigms in (2) and (25) differ in a number of significant ways which suggest that (25a) is unacceptable for primarily syntactic reasons.

First, recall that the claim that an intervening subject is disallowed in Spanish \textit{wh}-questions because of working memory constraints was supported by the fact that acceptability varies depending on the length and D-linking of the subject. English \textit{wh}-questions do not seem to be susceptible to this type of variation:

\begin{table}
\begin{tabular}{|l|}
  \hline
  (25)a. *What \textbf{John} will buy? \\
  b. What will \textbf{John} buy? \\
  \hline
\end{tabular}
\end{table}

\begin{table}
\begin{tabular}{|l|}
  \hline
  (26)a. *What John will buy? \\
  b. *What the friends of your sister will buy? \\
  \hline
\end{tabular}
\end{table}

The intervening subject in (26b) is longer than its counterpart in (26a), but it is not clear that there is a perceptible difference in acceptability between the two.

Second, in Spanish we saw that the burden on working memory caused by an intervening subject can be alleviated somewhat if the filler is D-linked. If the correct explanation for English (25a) were that the intervening subject
places too much of a load on working memory, we would then expect that it too would improve with a D-linked filler, but this does not seem to be true:

(27) *Which of those books John will buy?

(25a) and (27) seem to be roughly equally unacceptable.

Third, the processing of a filler-gap structure should be essentially the same in either a matrix or an embedded clause, and we saw that in Spanish this could explain why there is no matrix/embedded contrast with regard to the paradigm in (2). In English, on the other hand, there is a contrast of this type:

(28a) *What John will buy?
   b. I wonder what John will buy.

The fact that this contrast obtains argues against ruling out (28a) through working memory limitations.

Fourth, under the account proposed here for Spanish, argument fillers need to be assigned to a gap as soon as possible, and this is why the subcategorizing head (the verb) prefers to be adjacent to the filler. Adjunct fillers, on the other hand, do not have this need, so they are not disturbed by an intervening subject. In English, this contrast between argument and adjunct fillers does not obtain:

(29a) *What John will buy?
   b. *Why John will buy that?

It thus appears that some other factor is at work which prevents John from intervening in (29).

Finally, we saw earlier that the fact that Spanish sentences like (2a) are susceptible to satiation suggested that these sentences are excluded for processing reasons. English sentences like (25a) are not susceptible to satiation, however. This is shown by an experiment following the same protocol as in 2.9 above with 45 subjects, all native speakers of English. Among the sentences presented to the subjects were those of the types shown in (30).

(30a) *What John will buy at the store?
   b. *Who did Alice speak with Tim after seeing?

(30a) is a wh-question with an intervening subject, and (30b) is an adjunct island violation.

The results of the experiment for the two sentence types in (30) are given
As can be seen, there is no satiation effect with either (30a) or (30b).

### 4. More on the difference between Spanish and English

Our results so far suggest that Spanish (2a) is out because of capacity constraints on working memory that make it difficult for the filler to be assigned to a gap, while the superficially similar English (25a) is out for syntactic reasons (presumably *will* must move to C). If this is correct, then how is English (25b) possible? Why does this not violate working memory constraints?

The answer, I believe, lies in the fact that Spanish, unlike English, regularly has either a preverbal or a postverbal subject position available, and as would be expected, the discourse roles of the two positions differ, with the preverbal position receiving the more specific discourse role (e.g. Casielles-Suárez 1999, Goodall 2001, Rizzi 2004). If, as seems plausible, this special discourse role of the preverbal subject results in a significantly heavier processing burden, the difference between Spanish (2a) and English (25a) becomes clear. In Spanish, the increased processing load of the intervening (preverbal) subject means that it is more difficult to hold the filler in working memory until it can be assigned to a gap. In English, on the other hand, the intervening subject does not have any special discourse role and thus does not have an increased processing load. This in turn means that it does not pose any special difficulty for holding the filler in working memory until it can be assigned to a gap. This view is supported by the facts from Caribbean Spanish that we saw in section 2.2. In that dialect, preverbal subjects are more English-like in their discourse role, with the result that they do not disrupt processing of the filler-gap dependency.

### 5. Conclusion

As we have now seen, some very simple properties of working memory seem to be able to account for many characteristics of *wh*-movement in Spanish, including many that would otherwise be puzzling. This opens up the possibility that the syntax of Spanish *wh*-questions could be reduced to its bare
essentials: raising of V to T, possible raising of the subject to \( \text{SPEC} \) of T (or to a higher position, as many have proposed), and \( \text{wh} \)-movement to \( \text{SPEC} \) of C. The syntax would then not need to say anything about the constraints on the interaction of the \( \text{wh} \)-phrase and the preverbal subject; these would be handled by the independently needed constraints on working memory.

References


