A Corpus-driven Analysis of the Do-Be Construction

Dan Flickinger and Thomas Wasow

1.1 Introduction

In the early 1980s, Ivan was invited to start a natural language project at Hewlett Packard Labs, just up the hill from the Stanford campus. At the time, with several collaborators, he was developing the theory of Generalized Phrase Structure Grammar (see Gazdar, et al, 1985), which seemed particularly well-suited for computational implementation. The linguists on the project (Ivan, Carl Pollard, Geoff Pullum, the two of us, and, later John Nerbonne) worked closely with a team of programmers (whose composition changed considerably over the years) to develop a system that could parse a broad range of English sentences and translate them into appropriate semantic representations. We found the problem of balancing computational efficiency with linguistic generality at once difficult and immensely productive. This interplay of computational and linguistic concerns led to so many modifications of our theoretical framework that, at one point, it was renamed Head-driven Phrase Structure Grammar (HPSG).

Ivan and Carl worked out the theory of HPSG in detail in two books (Pollard and Sag, 1987, 1994) and a number of articles. Dan, along with Derek Proudian, led the team that kept improving the implementation's coverage and efficiency. Almost a decade after its start, the HP natural language project was closed down.

In the mid-1990s, Ivan started the LinGO lab at Stanford's Center for the Study of Language and Information (CSLI), initially with fund-

The Core and the Periphery: Standing on the Shoulders of Ivan A. Sag. Philip Hofmeister and Elisabeth Norcliffe (eds.). Copyright © 2013, CSLI Publications.

ing from the German Verbmobil project. Dan was hired to oversee the development of a broad-coverage HPSG-based system for both parsing and generation of English. This system, known as the English Resource Grammar or ERG (see Flickinger 2000, 2011), is conceptually a direct descendant of the system built at HP Labs, but was built from scratch, incorporating many advances in theory and technology that have become available over the intervening decades. A number of students and postdoctoral researchers contributed to the development of the ERG over the years, notably (and alphabetically): Tim Baldwin, Emily Bender, Francis Bond, Ann Copestake, Rob Malouf, Stephan Oepen, and Susanne Riehemann.

In order to test the performance of the ERG, Dan periodically runs corpora through it. After years of development, the grammar now provides complete coverage of about 90% of the sentences in such corpora as the English Wikipedia and the Wall Street Journal, and is usually missing analyses for only one or two phrases within the sentences it cannot fully parse. Seeing what kind of sentences it fails to provide reasonable parses and semantic representations for provides both a metric of progress in the development of the system and pointers to where work is needed. Over the years, people affiliated with the LinGO project have met frequently (sometimes on a weekly basis, sometimes less regularly) to discuss how to deal with types of examples that the ERG was having trouble analyzing. This paper arose out of one of these meetings.

1.2 The *Do-Be* Construction

One relatively common type of sentence that the ERG was not successfully analyzing is exemplified in (1).[1]

(1) a. what you have to do is get ready
b. all the government does is send out checks
c. the thing I'm doing is trying to learn from my mistakes
d. the best one can do is compare one risk to the next
e. the least we should do is make it as much fun as possible

These are examples of what we will refer to as the do-be construction or DBC. The salient characteristics of DBC are the following:

[1] Except where noted, examples are taken from the Corpus of Contemporary American English (http://corpus.byu.edu/coca/). We have edited many of the examples, to remove material irrelevant to the point we are making. Since some of the removed material originally preceded or followed what we present, we give our corpus examples without initial capitalization or final punctuation. Invented examples are capitalized and have final punctuation.

A. One of seven specific words (*what, all, thing(s), best, least, most,* and *worst*) appears near the beginning of the subject.
B. The subject contains a subordinate clause (a relative clause or possibly an embedded question) with a gap in it.
C. Embedded in the clause within the subject is some form of the verb *do.*
D. The main verb of DBC is a copula.
E. The copula is followed by a verb phrase.
F. The verb heading the post-copula VP is interpreted as having the same subject as the occurrence of *do* mentioned in C above.
G. The form of the verb heading the post-copula VP depends on the form of *do.*
H. The copula expresses identity; that is, the post-copula VP and the subject have the same denotation.

We will elaborate on each of these properties in section 4 below. Corpus searches revealed that not all of them are categorical constraints. In particular, we found some possible counterexamples to A and C. Some types of counterexamples are exceedingly infrequent and of arguable acceptability, so it seems reasonable to treat them as ungrammatical (perhaps performance errors). Others that are more common and less marginal might be instances of other, related constructions. We will deal with specific cases in our discussion of the properties.

The reader may wonder at this point why anyone not engaged in building a broad-coverage English language system should have any interest in a detailed examination of this particular type of sentence. We have three answers, one methodological, one theoretical, and one concerning the role of grammar in computational linguistics.

The first examples of DBC to attract our attention started with the word *what.* A bit of discussion led us to realize that the properties that made these *what* sentences hard to analyze were shared by examples starting with *all* or *the thing.* But it wasn't until we began examining corpus data that we realized that there were four other words that could be used in place of *thing.* This was just the first of many incorrect generalizations we arrived at about DBC on the basis of introspection about invented examples. For example, we believed at one time that the post-copula VP had to be in base form; but we were wrong. And introspection convinced us that the gap in the clause within the subject could not be the subject of that clause (that is, that the subject had to contain a non-subject relative clause), until we came across some counterexamples[2] in the data from the Corpus of Contemporary American

[2] See (8) below for two of them.

English (COCA). Corpus searches repeatedly revealed to us that the constraints on DBC are more intricate than we had earlier thought.

The lesson we draw from this is that, whenever possible, linguists should supplement their introspective and experimental data with naturally occurring usage data. We hasten to add that we are not arguing for exclusive reliance on corpora. Quite the contrary, it should be obvious that every type of linguistic data has its pros and cons. The best-supported theories are those that receive support from a variety of types of evidence.

The second conclusion we draw from our investigation of DBC is that the theory of grammar needs to have mechanisms for expressing the intricate complexes of constraints that define phenomena like DBC. In particular, mechanisms are needed for expressing properties that distinguish DBC from other sentence types, and the ways in which DBC are similar to other sentence types. Further, we identify several subtypes of DBC; a grammar of English must allow expression of the features that distinguish the subtypes, as well as their commonalities. In section 6 below, we demonstrate how a constraint-based grammar incorporating a lexical inheritance hierarchy can both capture the generalizations and make the distinctions that we note.

The analysis we present is formalized within the system currently implemented in the ERG. It is possible that a still higher level of generality could be achieved in a theory incorporating a formal notion of construction in the sense of Fillmore, Kay and O'Connor (1988), as long as it also included a type hierarchy allowing the expression of both overlaps and distinctions among construction types, as a means of minimizing redundancy while maintaining descriptive adequacy. The theory of Sign-Based Construction Grammar (SBCG) developed by Ivan Sag and collaborators (Sag 2010, Boas & Sag 2012) is just such a theory.

The third lesson we draw from our investigation of DBC is that grammar has a place in large-scale natural language processing systems. The intuitions that motivate A-H above are quite robust, and receive strong support from our corpus data. They exemplify the kind of interconnected morphological, syntactic, and semantic properties that languages have conventionalized. We doubt that a purely data-driven system, relying solely on co-occurrence statistics in corpora, could deal with the range of DBC examples our analysis handles, without simultaneously accepting many superficially similar strings that are not, in fact, well-formed English sentences. While we make no attempt to demonstrate that DBC requires a grammar, we hope our discussion puts the burden of proof on those who think grammar is unnecessary.

1.3 Our Data

We extracted examples of DBC from the Corpus of Contemporary American English (COCA), a 425-million word web-based database, with a suite of user-friendly search tools, compiled by Mark Davies. COCA is tagged for part of speech, but not parsed.

Our basic search pattern was: [vd*] [vb*] [v*], which, in the notation of COCA, means a sequence consisting of any form of the verb *do*, immediately followed by any form of the verb *be*, immediately followed by any form of any verb. This pattern produced 9914 hits. COCA displays the hits by the exact word string matching the search pattern, listing the number of hits for that string; we downloaded all strings with at least two hits. This gave us a collection of 7100 examples. These were culled and coded by Rebecca Greene.[3]

6471 of these 7100 examples (91%) are at least arguably instances of DBC – a surprisingly high success rate, given the simplicity of our search pattern. A few dozen of these are questionable, either in terms of well-formedness or in terms of their status as instances of DBC. These 6471 serve as the primary database for our discussion below. In addition, we conducted searches with *to* between [vb*] (the copula) and [v*] (the post-copula verb), with adverbs between each of the adjacent pairs of verbs in our pattern, and with two-word prepositional phrases between [vd*] (*do*) and [vb*]. These searches yielded a few thousand more hits, but we did not study them systematically. We do, however, make use of some examples taken from these additional searches below.[4]

[3] Specifically, she coded for: whether the subordinate clause in the subject is introduced by *that*, what form of *do* appears in that clause, and what the form of the main verb of the post-copula VP is. She also noted any properties of individual examples that she thought might be of interest.

[4] Two other searches, which we will not discuss in detail, involved looking at the frequency of DBC broken down by genre, and by history. The genre breakdown revealed that DBC is used predominantly in speech (though with *all* it is also used frequently in fiction), supporting our intuition that DBC is a relatively informal construction. The historical breakdown (based on the Corpus of Historical American English, http://corpus.byu.edu/coha/, a 400-million word corpus of sources from 1810 to 2009) showed that DBC first started appearing around 1820 or 1830, remained quite infrequent for a century, and then gained frequency for the following half century, before leveling off. This pattern is nearly identical across the three main subtypes of DBC, namely, those with *what*, those with *all*, and those with *thing*. This provides some support for our claim that the three types constitute a single construction.

1.4 Characteristics of the *Do-Be* Construction

1.4.1 Seven Words

In the vast majority of our DBC examples, the head of the subject is one of the three words *what, all*, or *thing(s).* The clear exceptions to this generalization are of two types: cases in which other words appear in place of these three, and cases in which the word in question is not the syntactic head of the subject. Cases of the latter type are exemplified in (2).

(2) a. one of the first things John and Glenn do is take a look around the city
b. a big part of what we do is give them the wind field on the entire storm

But in these and all other cases of this kind, what precedes the word in question serves to quantify or partition the noun phrase headed by that word. This is true even in cases like (2a), in which the singular form of the copula indicates that *one*, not *things*, is the syntactic head of the subject. Arguably, then, *what* and *thing* can be regarded as the semantic heads of the subject NPs in which they appear.[5]

As noted earlier, there are four other words that occur as the head of the subject: *best, least, most*, and *worst.* These are quite rare, accounting for only 131 of the examples in our database: 64 with *least*, 57 with *best*, 6 with *most*, and 4 with *worst.* These examples seem somewhat formulaic: the word in question is never preceded by anything but *the* and (occasionally) *very*, and the following relative clause almost always contains a modal, usually *can* or *could.* We will have little to say about examples with one of these four words as head of the subject, because we believe that they behave very much like those with *thing.*

The fact that it is just these seven words that have to appear in the head of the subject of DBC examples is no doubt in part due to the semantics and pragmatics of the construction. Following a reviewer suggestion, we checked COCA to see what nouns most frequently head objects of the verb *do*. We found that *thing* is by far the most frequent, with *business, research, work, job, harm*, and *damage* all occurring at least 400 times as the head noun of the object of some form of *do*. We then searched for strings where one of these nouns heads the subject of a sentence that otherwise looks like an instance of DBC. We found only two:

(3) a. the only business I did was to talk to the secretary of the Treasury

[5] Our data contain no such examples with *all.*

b. part of the work we do is to raise the public's awareness of women farmworkers' realities

It appears from the rarity of such examples that the very restricted set of nouns that appears as the head of the subject in DBC examples is in part conventionalized – a fact about English that must be learned. Notice that substituting some other word – even a near synonym or, in the case of *thing*, another semantically light noun, usually leads to unacceptability.[6]

(4) a. what/*whatever I do is give everyone a fair chance
b. now what/*how we do is follow the other approach
c. all/*everything a person tries to do is survive
d. all/*the sum total you've got to do is bury your partner
e. the first thing/*act he does is put his arm around Larry
f. one thing/*step we don't want to do is reduce the size of the force

1.4.2 The Subordinate Clause

We will treat the subordinate clause in the subject of DBC uniformly as a relative clause (RC). Faraci (1970) and others have argued that the examples starting with *what* are embedded questions, but Akmajian (1979) and others have argued that they are relative clauses. Without going into the arguments on the two sides of this debate (see den Dikken, 2001, for a summary), we will adopt the RC analysis because the subordinate clauses in the other types of DBC (that is, those with *all, thing*, etc.) are clearly RCs.

Although relative clauses are normally optional modifiers, the RC in DBC is obligatory, as shown in (5), where the asterisk outside the

[6] The following six examples from our database are prima facie counterexamples to this claim:
(i) everything he's doing is trying to help somebody else
(ii) part of this procedure that we're doing is making sure that we don't go near those pressure barriers
(iii) the only compliance work he's doing is making sure my lunch is still hot when it gets here
(iv) part of the work that I do is going to different hospitals in Denver
(v) the only way we could do is hit them with the pipeline of money going to all these organizations
(vi) the third occupation that I want to do is be an aeronautical engineer
However, (i)-(iv) all contain post-copula phrases that could be analyzed as gerundive NPs, and hence not instances of DBC. (v) and (vi) strike us as unacceptable. All six of these examples come from spoken data, and hence could well be speech errors. Since their status as counterexamples is questionable and they constitute less than 0.1% of our database, we feel comfortable with our claim that DBC sentences require one of the seven words we cite.

parentheses means that the sentence is unacceptable without the parenthetical material.

(5) a. all *(you have to do) is follow the road
b. the best *(you can do) is observe
c. the first thing ?*(we do) is make up a new name
d. what *(I think that I do) is write an adventure novel

The RC is usually finite, but with *thing* as the head, it may be infinitival:

(6) the smart thing to do was leave her there

Such cases are relatively rare; only 93 of our 1119 examples with *thing* have infinitival RCs. In these cases, the understood subject of *do* is contextually determined, unless the infinitival relative is a *for-to* clause:

(7) the best thing for you to do is learn how to be patient

When the RC is finite, the gap is almost always something other than the subject of the RC. Only five cases of RCs with a subject gap occur in our database (three of them beginning *all that was left to do*), though it is our impression that they are more common among examples in which the post-copula VP is infinitival (that is, begins with *to*).

(8) a. all that was left to do was get the teams lined up
b. what needs to be done is simply achieve literacy

Among the finite RCs with non-subject gaps, the occurrence of a relativizer *that* is extremely rare: our database contains only 106 examples. When the subject is headed by *what*, it is categorically impossible to have *that* as a relativizer:

(9) what (*that) the Judiciary committee did was say, no

This is not surprising, since neither free relatives nor embedded questions ever take *that* as a relativizer or complementizer. But even if we exclude *what*-DBC cases, *that* occurs as a relativizer in less than 3% of our examples. By comparison, 43% of the non-subject RCs in the Switchboard corpus have *that* as a relativizer (see Wasow, et al, 2011).

We believe that the rarity of *that* relativizers in DBC is related to the obligatoriness of the RC. Wasow, et al (2011) argue that the relativizer *that* occurs at a lower rate when the RC it introduces is more predictable (see also Jaeger, 2006, for a more thorough corpus study). This observation can be explained in functional terms: a uniform rate of information transfer enhances communicative efficiency (Levy & Jaeger, 2007, Jaeger 2010, Jaeger 2011). Since the RC in the subject of DBC is obligatory, it is highly predictable, so including *that* would usually lead to a trough in information density. Thus, the low rate of *that* in these examples supports Wasow and Jaeger's earlier work, as well as the theory of uniform information density.

1.4.3 The Verb *do*

The examples that got us started on our investigation of DBC all contained forms of the verb *do*, and *do* appears in our search pattern for finding instances of the construction. The results of our searches using that pattern consequently all contained instances of the verb *do* within the RC in the subject. We initially tested for the obligatoriness of *do* in the construction by substituting other verbs in selected examples from our corpus searches and consulting our own intuitions about them. This led us to conclude that a form of *do* was indeed obligatory. It seemed to us that no other verb could be substituted – not even another semantically light verb or one identical to the verb following the copula.

(10) a. what we do/*make is create Frankensteins
b. the first thing we do/?*try is make up a new name
c. all they did/*took was take a daily walk
d. the least you could do/*be is be a little grateful

However, a corpus search substituting other verbs in the search pattern revealed that the situation is not so straightforward. We conducted searches replacing *do* in the original pattern with *make, try, take*, and *be*. In all cases, the number of hits was vastly smaller than with *do* (by at least an order of magnitude), and most of them differed in many other ways from our DBC examples.

For instance, when the verb *make* was substituted in the search pattern, the subject in almost all of the examples was headed by a noun (like *mistake, adjustment*, or *decision*) that occurs often in collocation with *make*, and almost all of the post-copula VPs were headed by *-ing* forms or infinitives. Similarly, when the verb was *take*, the subjects were headed by nouns like *step, risk*, and *action*; and, again, the post-copula VPs were mostly headed by *-ing* forms or infinitives. Typical examples are given in (11).

(11) a. the biggest mistake she made was doing this in Columbus, Ohio, instead of Aruba
b. the best decision I could make is to go back and finish my education
c The biggest risk I've ever taken is going on American Idol and trying to be myself
d. the first step we must take is to improve our knowledge of the subject

Substituting *try* in the search pattern resulted in only a handful of hits similar to DBC, again with *-ing* forms or infinitives heading the post-copula VP.

Substituting *be* for *do* our search pattern yielded more hits, but the vast majority of these are cases where the main copula is predicative (e.g. *what the alternative might be is known only to the eye of faith*), the post-copula phrase is arguably adjectival (*all I wanted to be is loved*), or examples we feel comfortable treating as ungrammatical speech or transcription errors (e.g. **what we are is are looking at a new electorate*).

The examples with post-copula VPs headed by *-ing* forms or infinitives, while still far rarer when the subject lacks an occurrence of *do*, are too numerous to ignore. Although we do not have a definitive analysis of such cases, we think it is no coincidence that these are the only two forms that may head a VP functioning as the subject of a clause:

(12) a. Going on American Idol made me famous.
b. To improve our knowledge of the subject would help us understand the problem.
c. *Went on American Idol made me famous.
d. *Improve our knowledge of the subject would help us understand the problem.

Arguably, then, VPs headed by these two verb forms may be treated like noun phrases in certain contexts, including the position following a copula expressing identity. If so, these apparent counterexamples to the claim that *do* is obligatory in the subject of the DBC might be instances of another related construction. We leave this as an open question for now.

Our searches for instances of DBC with other verbs substituting for *do* turned up a very small number of additional potential counterexamples:

(13) a. all it takes is run my finger down the page
b. what I think the process should be is withdraw, as you pointed out, one piece of it
c. what you don't want to be is be part of that conducting line from the line down to the ground
d. what we think the right process should be is let the Congress set the priority

Given the rarity and (in our opinion) marginality of these examples, we will tentatively stick with our claim that some form of *do* is obligatory in DBC. In our concluding section, we return to the issue of the status of strong statistical generalizations that are not categorical.

The *do* in the subject of DBC sentences is not the auxiliary *do*, as evidenced by examples like the following:

(14) a. the only thing that the yellow bulb didn't *(do) was keep the bugs away
b. what it does do is give him more energy

The obligatory occurrence of *do* may be deeply embedded within the relative clause.

(15) a. all she had ever meant to do was make her mother proud of her
b. one thing I never have trouble getting her to do is eat
c. what I'm really working to try to do is get people past the pro- and anti-war arguments

Although *do* is rarely embedded more than one or two levels down in our corpus data, it is not hard to construct acceptable sentences in which it is more deeply embedded:

(16) What I instructed you to tell Pat to ask Sandy to try to be sure to do was get to the meeting on time.

We conclude that *do* may be arbitrarily deeply embedded in the RC. However, the intervening clauses and verb phrases must be complements. It is not acceptable for *do* to be buried somewhere in an adjunct:

(17) *What I will sleep until you do is wake me up.

In most cases, *do* is the last word of the subject, but modifiers are possible between *do* and the copula that serves as the main verb of DBC.

(18) a. what I suggest we do instead is put on our wigs
b. all he wanted to do in July was make the team
c. the most interesting thing we do in bed is eat cereal

The material intervening between *do* and *be* can be arbitrarily long, as is easily seen by adding more modifiers in (18c) – e.g. *The most interesting thing we do in bed on summer mornings when we have enough time and don't have to worry about cleaning up is eat cereal.*

The obligatory *do* in the subject RC can be in any of its inflectional forms: present, past, base, infinitive, present participle, past participle, or passive participle.

(19) a. what this bill **does** is allow you to sue your employer
b. all I **did** was sit around
c. the least we should **do** is make it as much fun as possible
d. the last thing she wants to **do** is get up early and make the call

e. all this is **doing** is helping parents raise their kids
f. what he has **done** is set his case in writing
g. what has to be **done** is to ask the health industry to sacrifice

1.4.4 The Copula

The copula in DBC is usually finite, but it may be preceded by auxiliary verbs that impose another form on it.

(20) a. the best thing to do would be bring in the bloodhounds
b. what Pat Buchanan has done has been complain about George Bush

Again, no synonym or near synonym can be substituted.

(21) a. what I do is/*equals set out the seven steps for thinking it through
b. the thing you always have to do is/*constitutes prove to people you belong there

We confirmed this through a corpus search in which we replaced the copula in the search pattern by an arbitrary verb.

The crucial property of the DBC copula is that it expresses identity. This is what Akmajian (1979) and others have called the specificational interpretation of the copula, and still others have called equative. This is to be distinguished from the predicational use of *be*. Sentences that look superficially very similar can differ with respect to which of these interpretations of the copula makes sense, as in (22).

(22) a. what we really need to be doing is building up the regular army
b. What we really need to be doing is being discussed in the Pentagon.

Some examples are ambiguous between the two interpretations:

(23) The thing we have been doing is keeping the police chief awake at night.

The properties of DBC we have enumerated hold only of sentences with this identity interpretation of the copula. For example, in (23), *thing* can be replaced by *dance* and *doing* can be replaced by *planning*, but only if *keeping the police chief awake at night* is predicated of the subject NP, not if it is equated with it.

1.4.5 The Post-Copula VP

Almost all of the examples in our database have non-finite VPs following the copula; that is, the head verb is in base, infinitive, present participle, or past participle form. We did, however, find eleven examples in which the verb is finite, two of which are given in (24).

(24) a. the first thing I did was went to the public library
b. what it does is takes the candidate off the ballot

Although such examples are exceedingly rare and arguably less than fully acceptable, we see no compelling reason to rule them out, so we treat them as grammatical.

If the form of *do* in the subject is a present participle, then the verb heading the post-copula VP must be a present participle (and conversely). If *do* is any other form, the head of the VP may be in base form, infinitive form, or the same form as *do*. This is illustrated in (25).

(25) a. all he's doing is talking about sport
b. *All he's doing is (to) talk about sport.
c. what Microsoft does is make the computing experience easier
d. What Microsoft does is makes/to make/*made/*making the computing experience easier.
e. the first thing he did was take the players' names off the backs of the jerseys
f. The first thing he did was took/to take/*takes/*taken/*taking the players' names off the backs of the jerseys.
g. what they have done is take it on an interim basis
h. What they have done is taken/to take/*took/*taking it on an interim basis.

The understood subject of the post-copula VP is the same as the subject of *do*. This is evident from the interpretation of DBC examples, and is also demonstrated by bound anaphors, when they are in the VP.

(26) a. the first thing people did was bless themselves
b. what Steve and Marlene have to do is eliminate themselves
c. all they do is tell each other how great they are
d. all we could do is make our way down through the cedar trees

Despite the intuition that the verb *do* is in some sense standing in for the post-copula VP, that VP may express a stative predicate, as exemplified by the line from an old Bob Dylan song, *All I really want to do is, baby, be friends with you.* (27) gives more examples from COCA.

(27) a. what I think we would all like to be able to do is know that the energy choices we're making as a nation are good for the environment
b. the smarter thing to do is realize that blueberries have a nice taste
c. all it has to do is be attached
d. all you will do is be sick until you go home
e. the best Smits can do is be tall, and at 7'4, he's got that covered

Interestingly, the post-copula VP may contain negative polarity items (NPIs), if *do* in the subject is in an environment where NPIs are licensed, as (28) shows.

(28) a. what we don't do is add any taxes
b. the only thing he won't do is serve you dinner on anything resembling disposable paper.
c. what he can't do is invade his neighbors anymore
d. another thing she does not have to do is decide any cases of real significance
e. what they were unable to do was provide any connection

In similar-looking sentences with a predicative interpretation of the post-copula VP, no such licensing is possible:

(29) *What we aren't doing is upsetting to anyone.

Similar contrasts have been known for many years. The examples in (30) are from den Dikken (2005), who cites Ross (1972), Akmajian (1970), and Halvorsen (1978) in this connection.

(30) a. What I have never noticed is any sign of dissatisfaction.
b. *What I never noticed was noticed by any of us.

We assume that NPI licensing is determined by the semantics (see Ladusaw, 1979, *inter alia*), so that the facts in (28) will not need any special stipulations, if the semantics of DBC is correctly analyzed.

1.4.6 Inversion

Another interesting property of DBC that we discovered when we examined our corpus data is that many of DBC examples cannot naturally be converted into yes-no questions through inversion of the copula. For example, inverting the examples in (1) yields (31).

(31) a. ?*Is what you have to do get ready?
b. *Is all the government does send out checks?
c. ?*Is the thing I'm doing trying to learn from my mistakes?
d. *Is the best one can do compare one risk to the next?
e. *Is the least we should do make it as much fun as possible?

In our corpus data, we found two examples in which speakers tried to invert DBC sentences but produced ungrammatical results – in one case by doubling *is*, and in the other by using *is* where the copula should have been *be*.

(32) a. *is what you do is take the other guy's best issue
b. *wouldn't all they need to do is shut off your power

The analysis we have implemented in the ERG does not allow the copula in DBC to be inverted. It seems plausible, however, that what makes examples like (31) sound so bad is a processing fact, not a grammatical one[7]. This assumption receives some support from the fact that inversion becomes more acceptable if the post-copula VP is introduced by *to*, as in *Wouldn't all they need to do be to shut off your power?*.

1.5 Ross's Proposal

Ross (2000, 2008), following earlier suggestions[8], proposes that pseudoclefts be derived via a deletion rule from sentences which underlyingly have a full clause following the copula. For example, under such an analysis, (33a) would be derived from (33b) via deletion of the second occurrence of *we need to.*

(33) a. what we need to do is invest in our future
b. What we need to do is we need to invest in our future.

Ross makes no attempt to state the relevant deletion rule precisely, nor does he say what its domain of applicability is. Thus, we can only conjecture whether he would advocate a similar treatment of the other subtypes of DBC. Still, this approach has several appealing features.

Perhaps the most attractive property of the deletion analysis is that the type of structure it posits as the source of pseudoclefts does in fact occur with some frequency, at least in speech. Moreover, the subject-sharing between *do* and the post-copula VP automatically follows from this analysis, as does the possibility of licensing negative polarity items in the VP with negation (or another licenser) in the subject RC.

The form dependencies between VP and *do* are a bit trickier to handle under Ross's analysis. Ross assumes that the verbal morphology is added to verbs via a transformation (affix-hopping of some sort), which may apply either before or after the deletion transformation. But this would not cover the biconditional dependency with present participles. That is, it would not be able to rule out (34a) without similarly ruling out (34b).

(34) a. *What we were doing was keep his organs alive.
b. what he had done was find a way to alter the cellular process

[7] Of course, saying that (31) and (32) are due to processing problems would not constitute an analysis. As Mary Dalrymple (p.c.) has pointed out to us, any processing account would have to explain why sentences with predicational uses of the copula sound so much better when inverted:

(i) ?*Is what the White House is doing trying to go on offense?

(ii) Is what the White House is doing causing Congress to worry?

[8] Akmajian (1970) cites an unpublished 1968 paper by Bach and Peters as the source of the idea.

Moreover, it is by no means clear how the deletion rule would be restricted to examples with one of our seven words in the subject and *do* in the subject RC, when the post-copula constituent is a VP. That is, the deletion rule would have to be constrained to avoid the generation of examples like (35), repeated from (3) and (10) above.

(35) a. *The sum total you've got to do is bury your partner.
b. *The first act he does is put his arm around Larry.
c. *What we make is create Frankensteins.
d. ?*The first thing we try is make up a new name.

In short, while a few of our observations about the constraints on DBC are automatic consequences of the deletion analysis, others would remain unaccounted for.

In any event, a deletion analysis is not an option for us, since the ERG has no deletion mechanism. So we will not consider it further.

1.6 Analysis and Implementation

We have implemented one analysis of this construction, extending the ERG, the existing broad-coverage HPSG-based computational system cited above. Our analysis establishes a semantic dependency between a construction-specific verb *do* within the subject NP of a clause and a construction-specific verb *be* which has this NP as its subject. In addition to the new lexical entries for these two verbs, we define entries for the small class of nouns that can serve as the head of the subject NP, organized as a small type hierarchy to capture the commonalities among these seven words. With these entries in place, the existing standard inventory of feature principles and syntactic rules in the ERG will suffice to characterize the range of idiosyncratic constraints on grammaticality we have identified for this construction.

1.6.1 Lexical Entries

The lexical entry for this construction-specific verb *do* is unusual in that its single NP complement is obligatorily extracted, establishing one end of an unbounded syntactic dependency. Similar lexically licensed gaps are also found in the entries for adjectives such as *easy*, which select for an infinitival complement with an obligatory NP gap: *This problem is easy to solve.* The standard principles which account for the propagation of ordinary gaps (Lexical Slash Amalgamation and Head Feature Principles) ensure this NP gap will still be visible at the top of the relative clause which is sister to the head noun of the subject NP.

The feature structure for the lexical entry for the DBC *do* is shown in Fig. 1, following the HPSG formalism of Pollard and Sag (1994, Ch. 9), where signs include constraints on both syntactic and semantic (SYNSEM, here SSM) properties, grouped into LOC(al) and NONLOC(al) features, the latter expressing constraints on long-distance dependencies such as in filler-gap phrases, wh-questions, and relative clauses. The local syntactic properties of a sign are grouped in the feature CAT(egory) and the semantic ones in CONT(ent), with the CAT features including both HEAD and VAL(ence) features. These groupings enable simple expression of general feature principles governing the propagation of constraints within phrases. Note that the valence features SUBJ(ect), COMP(lement)S, and SP(ecifie)R are (possibly empty) lists whose elements express the subcategorization requirements of the sign, with each element of these lists being again a *synsem* constraining both the syntax and semantics of these dependents. Modification in HPSG is treated as a dependency where the modifier constrains what it will modify, via the MOD attribute (here treated as a HEAD feature), whose value is again a *synsem* list, typically either empty or singleton.

The entry for *do* in Fig. 1, then, expresses (i) the requirement for an ordinary subject NP: the value of the valence feature SUBJ, a singleton list whose HEAD value is *noun* and whose SPR feature is saturated (an empty list); (ii) no overt complements: an empty list as the value of COMPS; and (iii) an unusual lexically-introduced non-local dependency: a singleton list value for the SLASH feature, requiring a nominal phrase of a particular kind to be satisfied by a SLASH-discharging construction elsewhere in the clause containing this verb. The semantic index of this nominal phrase gap (the value of the feature IND) is constrained to have the idiosyncratic SORT value *do-event*, discussed below.

$$
\begin{bmatrix}
\text{ORTH} \ \langle \textit{do} \rangle & \\
\text{SSM} & \begin{bmatrix}
\text{LOC} & \begin{bmatrix}
\text{CAT} & \begin{bmatrix}
\text{HEAD} & \textit{verb} \\
\text{VAL} & \begin{bmatrix}
\text{SUBJ} & \left\langle \begin{bmatrix} \text{LOC} & \begin{bmatrix} \text{CAT} & \begin{bmatrix} \text{HEAD} & \textit{noun} \\ \text{VAL} & \begin{bmatrix} \text{SPR} \ \langle\ \rangle \end{bmatrix} \end{bmatrix} \\ \text{CONT} & \begin{bmatrix} \text{IND} \ \boxed{1} \end{bmatrix} \end{bmatrix} \end{bmatrix} \right\rangle \\
\text{COMPS} & \langle\ \rangle
\end{bmatrix}
\end{bmatrix} \\
\text{CONT} & \begin{bmatrix} \text{RELS} & \left\langle \begin{bmatrix} \text{PRED} & \textit{\_do\_v\_be} \\ \text{ARG1} & \boxed{1} \\ \text{ARG2} & \boxed{2} \end{bmatrix} \right\rangle \end{bmatrix}
\end{bmatrix} \\
\text{NONLOC} & \begin{bmatrix} \text{SLASH} & \left\langle \begin{bmatrix} \text{CAT} & \begin{bmatrix} \text{HEAD} & \textit{noun} \\ \text{VAL} & \begin{bmatrix} \text{SPR} \ \langle\ \rangle \end{bmatrix} \end{bmatrix} \\ \text{CONT} & \begin{bmatrix} \text{IND} \ \boxed{2} \begin{bmatrix} \text{SORT} \ \textit{do-event} \end{bmatrix} \\ \text{XARG} \ \boxed{1} \end{bmatrix} \end{bmatrix} \right\rangle \end{bmatrix}
\end{bmatrix}
\end{bmatrix}
$$

FIGURE 1 Feature structure for DBC verb *do*

Since the relative clause containing this verb *do* is obligatory in DBC, we cannot analyze it as a modifier of the head noun (assuming with Pollard & Sag that modification is by definition optional), so the small class of nouns that can head the subject NP is defined to instead select the relative clause as an obligatory complement. The head noun preserves the necessary information provided by the gap as part of the noun's semantic index, and propagates it to the top of the subject NP via the standard principles of semantic composition, so this information is visible when the full NP is combined as the subject of the *be* VP.

The seven lexical entries that can serve as the head of the subject NP in DBC have some properties in common, but also exhibit idiosyncrasies, which we express in a small hierarchy of lexical types, described in a following section. They all select for an obligatory complement which has the form of a relative clause, and they all preserve three pieces of information provided by the gap within the relative clause: the semantic index of the gap itself, the semantic index of the syntactic subject of the verb *do* within the relative clause, and (most surprisingly) the inflectional form of the verb *do*. These head nouns differ in whether the relative clause can be infinitival or finite, and whether a relative pronoun is possible, so appropriate subtypes of this noun class are defined that capture these differences.

We show in Fig. 2 the feature structure of the lexical entry for one of these head nouns, *thing*. Note that this entry selects for an obligatory clausal complement (verbal head and saturated valence lists), which is a relative clause (non-empty MOD value) constrained to modify a noun with the DBC-specific SORT value *do-event*. The only such clauses admitted by the grammar are those which contain the lexical entry for the verb *do* shown above. The entry for *thing* also identifies its own semantic XARG (external argument – its subject) value with that of its relative clause complement. It is via this identity that the relevant constraints are linked between internal properties of the relative clause: the semantic index of the subject of *do* in the relative clause, and the inflectional form of *do*.

The lexical entry for the DBC verb *be*, whose feature structure is shown in Fig. 3, is unusual in at least two respects: (i) it requires a subject NP which is headed by one of the small class of seven permissible nouns (ensured by the SORT value of the subject's semantic index), and (ii) it requires a complement VP whose inflectional form and controlled semantic argument are both constrained by properties of that subject NP (via the XARG value of the subject NP, and the IFORM value of its index). However, the complement VP does not itself contain any construction-specific elements; it must only conform to the constraints imposed by the other elements of the construction.

$$\begin{bmatrix} \text{ORTH} < \textit{thing} > & \\ \text{SSM} & \begin{bmatrix} \text{LOC} & \begin{bmatrix} \text{CAT} & \begin{bmatrix} \text{HEAD } \textit{noun} & \\ \text{VAL} & \begin{bmatrix} \text{CMP} < & \begin{bmatrix} \textit{synsem} & \\ \text{LOC} & \begin{bmatrix} \text{CAT} & \begin{bmatrix} \text{HD} & \begin{bmatrix} \textit{verb} & \\ \text{MOD} < & \begin{bmatrix} \textit{synsem} \\ \text{...IND } \boxed{1} \begin{bmatrix} \text{SORT} \\ \textit{do-event} \end{bmatrix} \\ \text{...XARG } \boxed{2} \end{bmatrix} > \\ \text{INV -} & \\ \text{FRM } \textit{fin-or-inf} & \end{bmatrix} \\ \text{VAL} & \begin{bmatrix} \text{SUBJ} < > \\ \text{COMPS} < > \end{bmatrix} \end{bmatrix} \\ \text{CONT} & \begin{bmatrix} \text{XARG } \boxed{1} \end{bmatrix} \end{bmatrix} \\ \text{OPT -} & \end{bmatrix} > \\ \text{SPR} < & \begin{bmatrix} \textit{synsem} & \\ \text{LOC} & \begin{bmatrix} \text{CAT} \begin{bmatrix} \text{HEAD } \textit{det} \end{bmatrix} \end{bmatrix} \end{bmatrix} > \end{bmatrix} \end{bmatrix} \\ \text{CONT} & \begin{bmatrix} \text{IND } \boxed{1} & \\ \text{XARG } \boxed{2} & \\ \text{RELS} & < \begin{bmatrix} \text{PRED} & \textit{\_thing\_n} \\ \text{ARG0} & \boxed{1} \end{bmatrix} > \end{bmatrix} \end{bmatrix} \\ \text{NONLOC} & \begin{bmatrix} \text{SLASH} < > \end{bmatrix} \end{bmatrix} \end{bmatrix}$$

FIGURE 2 Feature structure for DBC noun *thing*

Beyond the construction-specific lexical entries for *do*, *be*, and the seven head nouns, we added to the grammar's type hierarchy a new type of semantic index for the gapped NP, one which includes the surprising attribute IFORM which records the inflectional form of the governing verb *do*. The value of this attribute can thus be propagated using standard feature principles up through the relative clause and then further to the full subject NP, so it is available to appropriately constrain the inflectional form of the VP complement of the verb *be*.

1.6.2 An Example

In Fig. 4 we show a phrase structure tree summarizing the syntactic structure assigned by the ERG to an example sentence containing this construction. In this analysis the V/NP node above *did* indicates the NP gap lexically introduced by this *did* (via its non-empty SLASH value), a gap propagated on the dominating VP/NP and S/NP nodes via the Head Feature Principle. The verb *did* also records its own subject index as the external argument (XARG) value of the gap's semantics, so this index will be preserved up through the subject NP, to remain accessible to the verb *be*, as elaborated below. The unary rule with the S/NP as its daughter is the normal rule for *that*-less relative clauses, where the semantic properties of the NP gap in the daughter (the INDEX and

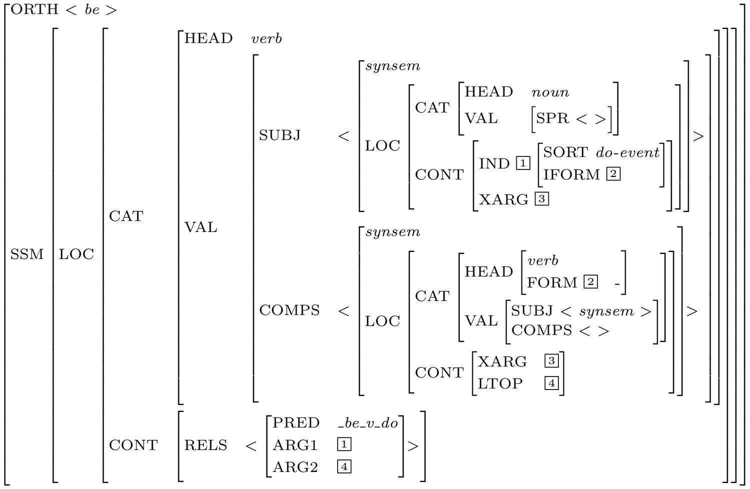

FIGURE 3 Feature structure for DBC verb *be*

the XARG values) are identified with those in the MOD value on the mother, and then unified with those of the head noun which is sister to the relative clause.

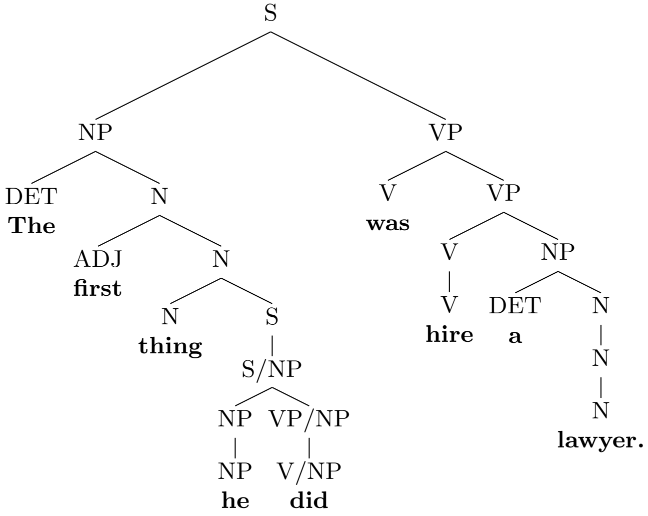

FIGURE 4 Syntactic structure for sentence exemplifying DBC

The head noun *thing* which selects this relative clause *he did* as a complement thus identifies its own INDEX and XARG values with those of the gap (the missing direct object of *did*); these values are then preserved as the noun is combined with the adjective *first* and the determiner *the*, since the modifier-head rule and the specifier-head rule both identify the mother's INDEX and XARG values with those of their syntactic head daughter. As a result of the interactions among the new lexical entries for *thing* and *do*, and the standard feature principles and syntactic constructions already present in the grammar, the subject NP *The first thing he did* has as its INDEX value the newly added semantic type which includes an attribute recording the inflectional form of *did*, and this subject NP also has as its XARG value the index of the subject of *did*.

The lexical entry for *was* here introduces a two-place semantic relation denoting something like identity, where the value of the first argument in the relation is identified with the subject's INDEX value, that of *thing*, and the second argument of the relation is identified with the INDEX of the VP complement of *was*. The entry for *was* also identifies the XARG value of its subject with that of its VP complement, so the semantic index of *he* is constrained to also be the subject index of *hire a lawyer*, as desired. Finally, the inflectional form of *did* (finite) is visible to *was*, since it is stored as a property of the index of the subject NP, so it can be used by *was* to constrain the inflectional form of the head of the complement VP of *was*, here the base form *hire* which is typical when the form of *do* is anything other than a present participle, as noted above.

1.6.3 Some further details of the analysis

The other possible head nouns in DBC have corresponding lexical entries, sharing some properties with the entry for *thing* (and the closely related but separate entry for *something*), but they impose additional restrictions on their dependents, including the complement relative clause. All of these others require the relative clause to be finite, with the relative pronoun *that* optional except with *what*, where it is not possible. The entries for the superlatives *best, worst, most*, and *least* share with *thing* the selection for a determiner as their specifier, but unlike *thing*, they do not permit adjectival modification.

In order to avoid spurious ambiguity that would result from the two distinct analyses of an NP such as *the first thing that he did* (one with the ordinary entries for *thing* and *do*, and one with the DBC-specific entries), all verbs except for the DBC-specific *be* exclude subjects whose INDEX value is the newly added type that the DBC-specific nouns introduce. Thus the grammar will not assign two analyses to a sentence such as *The first thing he did surprised us.*

Finally, we observe in the corpus some subject noun phrases whose syntactic heads are not one of the seven nouns we have discussed, where one of these nouns is embedded in a partitive NP, as in the example *One of the things I want to do is make a movie.* Since partitives in general need to preserve some properties of the semantic index of the embedded NP, such as number and semantic type (for example the temporal/non-temporal noun distinction), the definition of the new DBC-specific index type also specifies the semantic sort of the index to be a DBC-specific type, and it is this semantic sort which the DBC verb *be* requires of its subject NP's INDEX value.

By making some additions to the existing lexical type hierarchy to accommodate the seven DBC nouns, as sketched in Fig. 5, and defining DBC-specific variants of the transitive verb *do* and the identity-copula *be*, we can account for virtually all of the examples of DBC that we found in the corpus, without adding any new syntactic rules or feature principles.

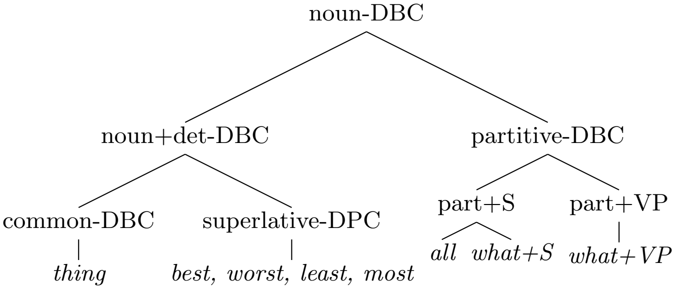

FIGURE 5 Hierarchy of lexical types for DBC nouns

Of the 6331 well-formed sentences in our corpus (setting aside all examples with typos or other syntactic errors), the ERG (version 1111) successfully produces DBC analyses for 6283, leaving 48 sentences – less than one percent of our corpus –unanalyzed or only parsed without an occurrence of a DBC construction. The DBC parsing results are summarized by sentence length in the following table:

Table 1 Parse results on the well-formed examples in the DBC corpus

Sentence length (in tokens)	number of sentences	lexical items	total parses	% parsed coverage
40–45	1	346.0	1	100.0
30–35	7	259.6	6	85.7
25–30	174	271.6	172	98.9
20–25	921	218.8	915	99.3
15–20	2469	180.8	2445	99.0
10–15	2040	136.5	2027	99.4
5–10	719	96.9	717	99.7
Total	6331	165.2	6283	99.2

Here, the corpus examples are aggregated by the number of tokens per sentence, with the columns showing, in turn, the number of sentences in each aggregate, the number of lexical items considered by the parser per sentence (including inflected variants), the number of successfully parsed sentences[9], and these results expressed as percentages.

Of the 48 unparsed grammatical DBC examples in our corpus, 19 fail to parse due to unrelated shortcomings of the ERG, such as a robust treatment of inserted quoted expressions as in this example:

(36) all he had to do was remember those mandatory NASA Thou shalt not! lectures to squelch it

More interestingly, the remaining 27 unparsed sentences reveal a handful of inadequacies in our current implementation that are related to the DBC analysis. First, since we treat the relative clause within the subject NP as a complement of the head noun, the grammar should allow at least some optional modifiers to intervene, as seen in the four corpus sentences below, but it does not, since it is not yet clear how freely such modifiers can appear.

(37) a. the last thing in the world you want to do is come in and get together
b. the last thing on earth I would do is kill the remaining dull tool that I had
c. last thing in the world I want to do is agree with Scott
d. the only thing left for you to do is go out

The second handful of problematic examples present the subject phrase of DBC as something other than an NP headed by one of the seven words identified in Section 4 above or a partitive noun such as *one* or *some*. While it remains our hypothesis that one of these seven

[9] A successful parse means that the sentence was fully analyzed using the ERG and at least one of the parses included the correct DBC analysis, though this analysis may not have been ranked as the most likely reading.

nouns always serves as the semantic head of the subject phrase, this notion will need some further refinement in the implementation in order to admit the following three corpus examples, where the subject in the first two is syntactically a prepositional phrase headed by *among*, and in the latter two, it is an NP headed either by the common noun *part* (modified by the adjective *big* and with the overt determiner *a*), or by the word *something* again with an overt determiner *the*.

(38) a. among the things we've got to do is see to it that there are options
b. among the things bishops and other ecclesiastical teachers do is teach
c. a big part of what we do is give them the wind field on the entire storm
d. the something that we should be doing is supporting the SADC process

The other 19 corpus examples that remain unanalyzed by our implementation are all sentences which we are inclined to treat as disfluencies, where a full clause is followed by what would be the finite verb phrase of ordinary DBCs. We present a few typical examples:

(39) a. that's all you can do is be depressed
b. that's what we've been trying to do is get the money back
c. it's what we did is put in a very strict set of preconditions for them

In the process of adding DBC to the ERG, one other variation on the theme emerged in the corpus, and is now accommodated in the implementation. While the complement of the head word in the subject NP of a DBC is almost always a (finite or infinitival) relative clause, here is one example where what follows the head word *what* is a finite VP containing the required *do* (a variant not possible with *thing* or *all*):

(40) what's left for us to do is achieve

In the implementation, given the hierarchy of lexical types to capture the generalizations that hold for DBC, and also the fine-grained idiosyncrasies that we have described and exemplified, this additional variant is implemented by adding a second lexical entry for the DBC *what* which selects for a finite VP complement of the right kind, in place of the relative clause that is required by all of the other DBC head noun lexical entries.

1.7 Conclusion

The *do-be* construction presents interesting challenges to syntacticians considering the kinds of theoretical machinery required to provide comprehensive descriptions of natural languages. The constraints on DBC detailed in section 4, involving syntax, semantics, and morphology, are just the sort of complex interdependencies that have led some linguists to propose including a notion of construction in the theory of grammar (e.g. Boas & Sag, 2012, Croft, 2001, Fillmore, et al, 1988, Goldberg, 1995). Two widely cited examples of constructions in English are the so-called *let alone* construction (Filmore, et al, 1988) and the *what's X doing Y* construction (Kay & Fillmore, 1999). One major motivation for dealing with these in constructional terms is that they are subject to a special array of morphosyntactic constraints (Kay & Fillmore, 1999; 4). As we have argued above, DBC has a remarkably rich array of such constraints.

In particular, the dependencies between the post-copula VP and the verb *do* in the subject involve elements that do not bear any kind of dominance or command relationship to one another, and this poses a challenge to any grammatical theory incorporating strong locality conditions. Our ERG implementation, however, shows that the standard feature-passing mechanisms of HPSG, together with a well-developed type hierarchy, are sufficient to handle the facts of DBC quite efficiently. Whether a fully constructional account would be preferable is a question we leave for future research.

Some might question the status of DBC as a distinct construction, given its shared properties with some other constructions (notably pseudoclefts), and the fact that our corpus studies found possible counterexamples to some of the properties we had used to characterize it.

The relationship to other constructions should come as no surprise. As Fillmore, Kay, Sag, and other construction grammarians have argued, languages have families of related constructions that can be represented with an inheritance hierarchy of construction types. Thus, for example, it would be natural to posit at least three subtypes of DBC, depending on whether the head of the subject is *what*, *all*, or one of the other five words. And sentences with specificational *be* might well constitute a construction of which DBC is a subtype.

The question of the status of DBC in light of the apparently noncategorical nature of the lexical constraints that characterize it is more difficult. Our corpus studies revealed that there are sentences very much like DBC but with nouns other than one of our seven heading the subject, and/or with another verb in place of *do* in the relative clause. But the number of examples we found that shared all of the properties A-H dwarfs the number of similar examples differing in these ways.

This sort of statistical convergence would be accepted as clear evidence of a real phenomenon in any of the social sciences, but most grammatical research has made categorical claims. This has been facilitated by the standard methodology of generative grammarians, which consists of testing hypotheses by inventing example sentences and introspecting on the sentences' acceptability. As the field moves from thought experiments to corpus studies and controlled experiments as its sources of data, the insistence on exceptionless generalizations becomes increasingly untenable. Clearly, this raises important and controversial methodological issues that are beyond the scope of this paper. We maintain, however, that the strong quantitative trends we found justify the claim that DBC is a distinct construction.

The constraints on DBC pose a challenge to developers of natural language processing systems trying to achieve broad coverage of English. The construction is common enough that such systems need to be able to handle it. The intricacy of the constraints on it, together with the unbounded character of the dependencies, suggest that it might create problems for systems relying entirely on statistical cooccurrence patterns. We have demonstrated that one grammar-based system can be readily extended to deal with DBC. We suggest that it may provide a good test of the versatility of any type of English-language NLP system.

Finally, as noted above, there are methodological lessons to be derived from our investigation into DBC. We relied crucially on both usage data (in the form of corpus examples) and invented examples, as well as introspective judgments about both. We contend that it is important for grammarians seeking to achieve descriptive adequacy to make use of all of these. We also made crucial use of a computational implementation to test our analysis. Languages are extremely complex systems, with so many interacting parts that it is typically difficult or impossible for people to keep track of all of the empirical consequences of any change to a grammar. Moreover, studies of different phenomena often make slightly different theoretical assumptions. Putting everything together into one computational system forces consistency in the analyses of distinct phenomena, and it makes possible reliable testing of what one's grammar predicts. In short, theoretical linguistics, at its best, should make use of a wide range of available methods for testing hypotheses.

Many open questions remain regarding DBC. How is it related to other similar constructions? In particular, DBC examples with *what* heading the subject are a species of pseudoclefts. But most DBC examples are not pseudoclefts, and most pseudoclefts are not instances of DBC. So how should the relationship between the two constructions be characterized? More generally, what properties of DBC follow from the fact that the copula in it receives a specificational interpretation? To what extent can the differences among the subtypes of DBC be explained semantically? When there are multiple choices for the form of the post-copula verb (that is, in all cases except where the form of *do* in the subject relative clause is present participle), what factors influence which one the speakers choose? How much variation is there in the use of DBC, both historically and across varieties of contemporary English? What other languages have similar constructions? We leave these and other questions for future research. We hope that our informal description of the construction and our description of the implemented analysis will be useful to those who investigate such questions.

Acknowledgments

We are grateful to Mark Davies for helpful advice regarding our corpus searches. Useful comments on earlier versions of this material were provided by Philip Hofmeister and an ostensibly anonymous reviewer (initials TFJ), as well as members of the audience at a talk in the Linguistics Department at Oxford University. Special thanks go to Rebecca Greene, not only for her careful and patient coding, but also for useful insights about the data she was annotating. Of course, we take full responsibility for any shortcomings in this paper.

References

Akmajian, Adrian (1970) Aspects of the Grammar of Focus in English, MIT doctoral dissertation, published in 1979 by Garland Publishing's Outstanding Dissertations in Linguistics series.

Boas, Hans and Ivan A. Sag, eds. (2012) Sign-Based Construction Grammar. Stanford: CSLI Publications.

Croft, William (2001) Radical Construction Grammar: Syntactic Theory in Typological Perspective. Oxford: Oxford University Press.

den Dikken, Marcel (2005) Specificational copular sentences and pseudoclefts, in Martin Everaert & Henk van Riemsdijk, eds, The Blackwell Companion to Syntax. Vol. IV, Chapter 61. Hoboken, NJ: Wiley.

Faraci, Robert (1970) On the Deep Question of Pseudo-clefts. Unpublished paper. MIT.

Fillmore, Charles, Paul Kay, and Mary Catherine O'Connor (1988) Regularity and idiomaticity in grammatical constructions: the case of let alone. Language, 64, 3, 501538.

Flickinger, Dan (2000) On Building a More Efficient Grammar by Exploiting Types. Natural Language Engineering 6:1, 15-28.

Flickinger, Dan (2011) Accuracy vs. Robustness in Grammar Engineering. In E.M. Bender and J.E. Arnold (eds) Language from a Cognitive Perspective: Grammar Usage, and Processing, pp. 31-50. CSLI Publications, Stanford.

Gazdar, Gerald, Ewan Klein, Geoffrey K. Pullum, and Ivan A. Sag (1985). Generalized Phrase Structure Grammar. Cambridge, MA: Harvard University Press.

Goldberg, Adele (1995) Constructions: A Construction Grammar Approach to Argument Structure. University of Chicago Press.

Halvorsen, Kris (1978) The Syntax and Semantics of Cleft Constructions. Texas Linguistic Forum 11. Austin: The Univ. of Texas Linguistics Dept.

Jaeger, T. Florian (2006) Redundancy and Syntactic Reduction in Spontaneous Speech. Stanford University PhD dissertation.

Jaeger, T. Florian (2010). Redundancy and Reduction: Speakers Manage Syntactic Information Density. Cognitive Psychology, 61(1), 23-62.

Jaeger, T. Florian (2011) Corpus-based Research on Language Production: Information Density and Reducible Subject Relatives. In Bender, E. M. and Arnold, J. E. (eds): Language from a Cognitive Perspective: Grammar, Usage, and Processing. Studies in honor of Tom Wasow, 161-197. Stanford: CSLI Publications.

Kay, Paul and Charles Fillmore (1999) Grammatical Constructions and Linguistic Generalizations: The What's X doing Y? Construction. Language 75, 1,1-33.

Ladusaw, William (1979) Polarity Sensitivity as Inherent Scope Relations. Austin: University of Texas PhD dissertation, publishing in 1980 by Garland Publishing's Outstanding Dissertations in Linguistics series.

Levy, Roger and Jaeger, T. Florian (2007) Speakers optimize information density through syntactic reduction. In B. Schlkopf, J. Platt, and T. Hoffman (Eds.), Advances in Neural Information Processing Systems (NIPS) 19, 849-856. Cambridge, MA: MIT Press.

Mikkelsen, Line (2011). Copular Clauses. In K. von Heusinger, C. Maienborn, and P. Portner (eds) Semantics: An International Handbook of Natural Language Meaning. Berlin: de Gruyter.

Pollard, Carl and Ivan A. Sag (1987) Information-based syntax and semantics. Volume 1. Fundamentals. CLSI Lecture Notes 13. Stanford: CSLI Publications.

Pollard, Carl and Ivan A. Sag (1994) Head-driven Phrase Structure Grammar. Chicago: University of Chicago Press and Stanford: CSLI Publications.

Ross, John R. (1972) Act. In G. Harman & D. Davidson, eds, Semantics of

Natural Language, pp. 70-126. Dordrecht: Reidel.

Ross, John R. (2000) The Frozenness of Pseudoclefts Towards an Inequality-based Syntax. In A. Okrent and J. P. Boyle (eds.), Proceedings of the Thirty- Sixth Regional Meeting of the Chicago Linguistic Society, pp. 385-426. Chicago Linguistic Society, University of Chicago.

Ross, John R. (2008) An Automodular Perspective on the Frozenness of Pseudoclefts. Paper presented at the Pragmatics, Grammatical Interfaces, and Jerry Sadock Conference. University of Chicago.

Sag, Ivan A. (2010) English Filler-Gap Constructions. Language 86.3: 486-545.

Sag, Ivan A. (2012) Sign-Based Construction Grammar: An Informal Synopsis. In Boas & Sag, eds, (2012).

Wasow, Thomas, T. Florian Jaeger, and David Orr (2011). Lexical Variation in Relativizer Frequency. In H. Simon & H. Wiese, eds, Expecting the Unexpected: Exceptions in Grammar. Berlin: de Gruyter.