CrossMark

English Language and Linguistics 19.3: 477–503. © Cambridge University Press 2015
doi:10.1017/S1360674315000180

The structural non-integration of *wh*-clefts[1]

WOJCIECH GUZ
John Paul II Catholic University of Lublin
(Received 23 September 2014; revised 24 February 2015)

Weinert & Miller (1996) suggest that English *wh*-clefts are a heterogeneous class in that they can have varied degrees of structural integration. Many such constructions depart structurally from the canonical *wh*-cleft which consists of a *wh*-clause, the copula and a focus constituent, and in which all the three elements are brought together into a fully integrated utterance. In the types of *wh*-clefts displaying looser structure, their lack of syntactic integration has so far been related to such linguistic features as (a) omission of the copula, (b) non-canonical copular complementation, e.g. independent main clauses instead of standard infinitival phrases appearing in the focus constituent, (c) lack of a clearly identifiable copular complement, (d) the focusing effect of the *wh*-clause extending over several clauses (Weinert & Miller 1996; Koops & Ross-Hagebaum 2008; Hopper & Thompson 2008; Callies 2012). Although the disintegrating effect of these features has been observed, the extent of the phenomenon in modern English has not been properly established and other non-integration features have not been investigated. Therefore, the purpose of this study is to use corpus data to identify and examine such non-integration features and to investigate the extent – expressed quantitatively – to which these features are found in *wh*-clefts with the verb *do* in the *wh*-clause. The article also points out the formulaic status of those *wh*-clauses which become disconnected from their focus phrases.

1 Introduction

The starting point for this paper is Weinert & Miller's (1996) suggestion that English *wh*-clefts are a heterogeneous class in that they can have varied degrees of structural integration; see the examples below (from Weinert & Miller 1996: 177):

(1) now **what you want to do is curve** round that wood
(2) **what you're going to do is you're going to continue** your downward line for about another inch

(1) is a classic *wh*-cleft with a cleft constituent or focus phrase (*curve round that wood*) that complements the *wh*-clause (*what you want to do*). The cleft constituent contains a bare verbal stem (*curve*) and has a form appropriate to its syntactic function, i.e. it is a phrasal complement of the *wh*-clause. Thus it is fully integrated. Also, prosodically, there is typically no pause between the two clauses, and the entire utterance has a single pitch pattern.[2]

In contrast, example (2) departs from this classic pattern in that the cleft constituent (*you're going to continue...*) is less integrated into the *wh*-cleft. Firstly, Weinert & Miller

[1] I would like to thank two anonymous reviewers of *ELL* for their most valuable comments on the first draft of this article. Thanks are also due to Piotr Pęzik for the opportunity to access the BNC audio files.
[2] This is a typical prosodic pattern. However, see section 5.3 for evidence of pausing in *wh*-clefts with infintival complements.

https://doi.org/10.1017/S1360674315000180 Published online by Cambridge University Press

report pauses and pitch markers separating the two parts of the sentence; therefore, utterances such as (2) have a bipartite prosodic structure.[3] Secondly, the cleft constituent in (2) is an independent clause acting as the copula complement. The independent clause – complete with a subject and finite verb – contrasts with the copula complements of classic *wh*-clefts, which, for the verb *do*, include bare infinitives, *to*-infinitives and *-ing* forms. Compare the following three options (after Quirk *et al*. 1985: 1388):

(3) **What he does is spoil** the whole thing.
(4) **What he does is to spoil** the whole thing.
(5) **What he's doing is spoiling** the whole thing.

Beyond the three types of complements in (3)–(5), Quirk *et al.* (1985: 1388) make an allowance for the occasional occurrence of matching verb forms similar to that in (5) but extended to perfective aspect, e.g. *What he's done is spoilt whole thing*, although the authors mark this type with doubtful acceptability and suggest that the structure may be seen as an ellipsis of an appositional construction (*What he's done is ((this): he's) spoilt whole thing*).

Reference grammars of present-day English vary in their acknowledgement of the type of *wh*-clefts in (2), in which the copula complement is a main clause. For example, Huddleston & Pullum (2002: 1422) briefly note that, with *do* in the *wh*-clause, main clauses are possible in informal style. Biber *et al.* (1999) ignore this possibility altogether. Carter & McCarthy (2006: 475d) say that main clauses can be used as copula complements when the focus of *wh*-clefts is a whole clause, as in (6), or even a longer stretch of discourse, possibly stretching over several sentences, as in (7) below (from Carter & McCarthy):

(6) **What I'll do is** I'll take all your details from you.
(7) And **what happened was** I'd arrived so late that I'd missed the tour round the school which I'd had previously when I'd been up. And I got there the time that I was due to see the headmistress. So I went straight to see the headmistress, went in and apologised for being late.

A similar explanation of multi-clause complementation seen in (7) can be extended to apply to *wh*-clefts with *do*, as confirmed by Koops & Ross-Hagebaum (2008), who cite examples such as (8) below:

(8) So **what a penguin will do, is** it'll go back, it'll squeeze the gland, it'll get some oil on its beak, and then, ... it's rubbing ... that waterproofing all over its feathers.

Following Lambrecht (2001) and Koops & Ross-Hagebaum (2008) we will refer to the variant of *wh*-clefts represented in (2), (6), (7) and (8) as AMALGAM *wh*-clefts. These, as stated above, exhibit a looser integration compared with standard *wh*-clefts. Additionally, the former represent a spectrum of degrees of integration. Weinert & Miller (1996) offer the following example of a loosely structured cleft:

[3] Collins (2006: 1713) reports that 94.6 per cent of his *wh*-clefts are realized as two or more tone units.

https://doi.org/10.1017/S1360674315000180 Published online by Cambridge University Press

(9) OK, well **what you've got to do is** moving across to the left, **you should be** about three inches from the bottom of the page

On a scale of integration, (9) ranks lower than (2), (6) and (8) (and, by implication, than (1)). Note that in (2) the focus clause mirrors the syntax of the *wh*-clause (tense, aspect) and repeats some lexical material (*you're going to*). This is typical of amalgam *wh*-clefts (see also Koops & Ross-Hagebaum 2008). Such parallelism and repetition establishes a link between the clauses that binds them together, although, at the same time, the separateness of the main clause is still distinct. In contrast, (9) has less of such parallelism and repetition: cf. *you've got to* vs *you should be.*

Let us represent the cline of integration of a series of *wh*-clefts derived from a simple non-cleft sentence:

(10) John lost his keys. (non-cleft monobloc)
(11) What John did was (to) lose his keys. (classic cleft)
(12) What John did was he lost his keys. (amalgam; second clause with subject NP, aspect and tense)

The loss of integration from (11) to (12) is further emphasized by the fact that the tense of the copula – the fulcrum of the construction – is to some extent independent of the tense of the *wh*-clause. Although the present/past tense specification of the copula tends to match that of the *wh*-clause verb (see (11)), as Ross (2000: 8) notes, a present tense copula seems to always be possible, even in past tense environments:

(13) What John did **is** (to) lose his keys. / What John did **is** he lost his keys.

What is more, in amalgam *wh*-clefts in spontaneous speech, the copula may be omitted:

(14) What John did, he lost his keys.

We note these variations in tense mismatch and ellipsis as another indication of the broken structure of clefts. And this inherent non-integration is indeed a matter of degree: note that the two variants in (13), with their present/past tense mismatch, are less integrated than (11) and (12), which consistently use the past tense. Additionally, sentence (14) is even less integrated because the link between the two parts of the cleft – the copula – is omitted. This cline of integration is particularly visible in *wh*-clefts with *do* in the *wh*-clause, and it is only these that will be the object of analysis in this article.

In summary, *wh*-clefts represent a spectrum of degrees of structural integration associated with syntactic, prosodic and/or lexical cohesion. Still, Weinert & Miller (1996) make this observation in reference to limited data and without considering the scale of the problem. Other studies have not rectified the situation, mostly due to limited data, or the fact that the authors either did not pursue the quantitative dimension of the issue at all or merely considered the relative proportion of integrated and unintegrated clefts, without zooming in on specific non-integration features (see section 3). Since a more thorough analysis of the problem has not been offered, the purpose of this study is to use a comparatively large collection of naturally occurring *wh*-clefts to address the following questions:

https://doi.org/10.1017/S1360674315000180 Published online by Cambridge University Press

1. What is the scale on which the non-integration of *wh*-clefts can be mapped?
2. What types of non-integration features/effects are at work?
3. How frequent are those non-integration effects?

In other words, the objectives of this article are (i) to map types of *wh*-clefts on a scale of decreasing integration, (ii) to find, if possible, other (non-)integration features not reported previously, and (iii) to quantify the (non-)integration phenomena. These questions are addressed in the main part of the study – section 5 – in which we illustrate, discuss and quantify the non-integration effects observed in corpus data. These effects will include those reported in the literature, but we also add two new features pertinent to the discussion. In section 5, the following will be considered in turn:

1. mismatched copula tense (new)
2. prosodic separation (with emphasis on pausing)
3. omission of the copula
4. omission of subject pronouns (new)
5. non-coreferential subject NPs
6. tense–aspect–mood mismatches
7. multi-clause focus constituents, delayed focus constituents, and lack of overt focus constituents

In the course of the discussion, the article also points out the formulaic status of those *wh*-clauses which become disconnected from their focus phrases. However, before we turn to the analysis proper in section 5, sections 2 and 3 will review the literature with regard to what has already been said about the non-integration of *wh*-clefts. Section 4 will present the corpora and data underlying the present study as well as explain the method of data retrieval.

2 Previous research: types of non-integration features

The lack of integration of some *wh*-clefts has been observed to manifest itself in several different ways. Firstly, main clauses have been observed in the position of copular complement, i.e. in amalgam *wh*-clefts (Weinert & Miller 1996; Hopper & Thompson 2008; Koops & Ross-Hagebaum 2008; Miller 2011: 32–9; Callies 2012). All these authors note the matching syntax (tense and aspect) across the two clauses and repetition of lexical and syntactic material (identical or coreferential subject NPs, repetition of auxiliaries; see example (15)), but also indicate lexical and syntactic mismatches (non-coreferential subject NPs, see (16), and mismatching tense/aspect properties; see (17)). Koops & Ross-Hagebaum (2008: 463) conclude that 'speakers are not strictly bound by the syntactic form of the *wh*-clause in constructing the FP [focus phrase]'.

(15) **What you're doin' is** you're drawin' a line. (Weinert & Miller 1996: 177)
(16) **What we would do is**, they would take care of the account maintenance. (Koops & Ross-Hagebaum 2008: 463)
(17) Basically **what I'm going to do is** I want them to reply. (Callies 2012: 12)

https://doi.org/10.1017/S1360674315000180 Published online by Cambridge University Press

Another feature of syntactic non-integration pointed out is the omission of the copula,[4] as in (18) below. Koops & Ross-Hagebaum (2008: 461) report that copula omission occurs almost exclusively in cases of amalgam syntax. To illustrate a less frequent variant, Hopper & Thompson (2008: 102) report the copula-less cleft with an infinitival focus constituent in (19) below:

(18) **what we'do** when I was a kid in the sounds **ø** we used to sometimes catch fish ... (Miller 2011: 33).
(19) **What you ought do** though Mar, **ø** cook all the fish.

Collins (1991: 44) regards instances such as (19) as 'system-deviant, or at least untypical constructions', i.e. incomplete versions of the full-fledged *wh*-cleft construction.[5] Hopper & Thompson (2008: 102) argue against such a position and suggest that many such fragmented *wh*-clefts of spontaneous conversation should not be seen as entering into bi-clausal constructions at all. Instead, they are 'syntactically separate entities' serving specific discourse and interactional goals (see below).

The next non-integration effect noted in the literature is the focusing effect of the *wh*-clause extending over several clauses. Koops & Ross-Hagebaum (2008: 463) cite the example in (20):

(20) **What we'll do is**, John will write up a draft and you and I will send him our comments.

In (20) the pointing-forward function of the *wh*-clause extends over all of the following utterance, which in some cases may be 'an indeterminate stretch of discourse' (Hopper & Thompson (2008: 110) (see also examples (7) and (8) above). Another possibility, and another type of disintegration, is that there is a lack of a clearly identifiable copular complement, as illustrated in (21) and (22) below.

(21) **What I thought I'd do** Chairman: as you all know, the most important issue at the moment is the poll-tax. (Weinert & Miller 1996: 196)
(22) **What we do**, then that's... that's where the ferrier comes in. (Hopper & Thompson 2008: 103)

Hopper & Thompson (2008) comment on such disjointedness and lexico-syntactic mismatches by stating that in most *wh*-clefts in everyday conversation the two clauses are separate entities. The second part is typically not a clause but an indeterminate stretch of discourse and the function of the *wh*-clause is to anticipate (or 'project') that stretch of discourse and to frame it in terms of such categories as EVENT (for clefts with *happen*; *What happened*...), ACTION (clefts with *do*; *What I do*...), and PARAPHRASE (clefts with *say* and *mean*, *What I'm saying*...).

[4] When the tense of the copula does not correspond to that of the *wh*-clause, as argued above, it may be seen as another manifestation of disintegrated syntax, although this particular feature has not been identified in the literature as a non-integration effect.

[5] Collins makes this comment in reference to infinitival *wh*-clefts with copula omission. His (1991) and (2006) analyses do not seem to cover amalgam *wh*-clefts at all.

https://doi.org/10.1017/S1360674315000180 Published online by Cambridge University Press

Similarly, with regard non-integration and discourse functions, Miller (2011: 34) sees *wh*-clauses such as *What I thought I'd do* in (21) as 'complex highlighting discourse particles that do not seem to be part of a *wh*-cleft complex'. Koops & Ross-Hagebaum (2008: 463) argue that the discourse function of the *wh*-clause follows from the fact that the focusing effect of the *wh*-clause may extend over several clauses or even independent sentences. Therefore, according to the authors, the discourse function of the *wh*-clause in amalgams is that of *discourse topic*, as opposed to *sentence topic* (2008: 469).

Hopper & Thompson (2008) conclude that non-canonical conversational *wh*-clefts should not be viewed as a degenerate variety of the canonical biclausal *wh*-cleft construction. Instead, 'linguists would be better advised to consider the written pseudocleft [i.e. *wh*-cleft] construction a normativized version of the *what*-fragment found pervasively in interaction'. Weinert & Miller (1996: 196), Miller and Weinert (1998: 293) and Callies (2012: 12) seem to agree in that they suggest that the classical integrated pattern may be a special development of written English.

The non-integration of some *wh*-clefts has also been observed in their prosody (Collins 1991, 2006; Weinert & Miller 1996; Miller 2011). The great majority (Collins reports 94.6 per cent) are separated intonationally by tone unit boundaries. Weinert & Miller (1996: 196) report that in their data some *wh*-clauses have a faster tempo and undergo phonetic reductions, reflecting the secondary status of whatever information the *wh*-clause carries. Miller & Weinert (1998: 291–3) report that some of their clefts are separated by pitch and rhythm, with *wh*-clefts being pronounced more rapidly and at a lower pitch than the following clause. Miller (2011: 38–9) also reports that some *wh*-clauses are separated by a pause from the following clause. With reference to amalgams, but not classic clefts, Weinert & Miller (1996) report pauses separating the two parts of amalgam *wh*-cleft.[6]

To summarize, six types of structural non-integration have been noted: (i) main clauses as copular complements; (ii) lexico-grammatical mismatches across the two parts of a cleft; (iii) copula omission; (iv) the focus of the *wh*-clause extending over multiple clauses; (v) lack of a clearly identifiable copular complement; and (vi) prosodic separation. These phenomena have been reported in previous research, but the quantitative dimension of the matter has not been addressed in detail. The observations that have been made in the literature are the focus of the next section.

3 Previous research: quantitative evidence

Weinert & Miller (1996) have little to say about the scale of the non-integration of *wh*-clefts given their limited database: the authors admit that clefts in general are relatively

[6] Weinert & Miller (1996: 177) say that in a classic *wh*-cleft there is no pause separating the clauses and that the clefted constituent is fully integrated into the construction. However, see section 5.3 for evidence of pausing in *wh*-clefts with infinitival focus constituents.

https://doi.org/10.1017/S1360674315000180 Published online by Cambridge University Press

rare in their data and *wh*-clefts are particularly rare: 22 occurrences in task-oriented dialogues and 31 in spontaneous conversation. Based on this database, they conclude that infinitival complements are more typical of written language (1996: 177) or 'a special development of written English' (1996: 196). At the same time, the authors find that most of the *wh*-clefts in their dialogic data do not fit the classical integrated pattern (1996: 196) and state that the cleft clause + main clause structure is 'regular and frequent ... in spontaneous speech' (1996: 178). In a footnote (1996: 178), the authors note the continuum of degrees of non-integration exhibited by *wh*-clefts. Miller & Weinert (1998) analyse the same collection of *wh*-clefts and report that '[a]lmost all the *wh*-clefts in the Map Task dialogues and the spontaneous conversations are unintegrated' (1998: 74).

Koops & Ross-Hagebaum (2008) study spoken American English against a database that is notably larger than that of Weinert & Miller's (256 *wh*-clefts). They note that in their data amalgam *wh*-clefts are especially common with the verb *do* as the *wh*-clause predicate. In fact, with the verb *do* in this position, they observe the amalgam type in the majority of cases (53 per cent). After this statement on the frequency of occurrence of amalgam vs standard *wh*-clefts, the authors zoom in on the factors predicting the variation between amalgam and standard *wh*-clefts. Firstly, the authors find that amalgams are more frequently used in informal than in formal spoken discourse (spoken academic English in their case). The second predictor is the length of the focus phrase expressed in the number of syllables in the main clause repeated after the *wh*-clause. The authors find that the amalgam construction is strongly favoured when only one syllable (typically a subject pronoun) is repeated, as in *What **she** did is **she** called me*. However, they find no such effect in the standard variant. Thirdly, the standard *wh*-cleft is favoured when the focus constituent contains a single clause, or two at the most. The amalgam shows no such skewing and also occurs with multiple-clause focus constituents.

Hopper & Thompson (2008) make numerous insightful remarks on non-integration phenomena (see section 2) although the quantification they offer is somewhat vague. The authors report that '*often* ... the *wh*-clause is followed by a whole finite clause' and that '*quite often* the follow-up to the *wh*-clause extends indefinitely far into the subsequent interaction' (2008: 102) (emphasis added). Further, they note that '*by no means all* of the *wh*-clauses are followed up by anything resembling a *be*-clause' and that '*most* pseudoclefts in everyday conversation are not syntactically biclausal' (2008: 102).

Callies (2012: 15) reports in spoken English 'a large number of structural types that are difficult to classify in terms of "canonical", bi-clausal cleft-structure' and the majority of these are reportedly of the amalgam type featuring identical or co-referential subject NPs and matching syntax. Callies' discussion of (non)-integrated *wh*-clefts with *do* in the *wh*-clause is based on 1.25 million words of spoken language (from three spoken corpora). The author confirms Koops & Ross-Hagebaum's (2008) finding that the non-integrated type is in fact the most frequent variant in spoken *wh*-clefts (in both BrE and AmE according to Callies 2012: 12). Specifically, in 500,000

https://doi.org/10.1017/S1360674315000180 Published online by Cambridge University Press

words of a spoken section of the BNC, the frequencies of occurrence per 100,000 words are the following: 6.4 for the non-integrated pattern, 6.0 for standard *wh*-clefts with bare infinitives, and 5.8 for standard *wh*-clefts with *to*-infinitives. In the *Santa Barbara Corpus of Spoken American English* (250,000 words), the non-integrated type is especially frequent compared to the standard *wh*-cleft (respectively 17.6 and 6.4 per 100,000 words) (Callies 2012: 9–12). Judging by corpus size and the recorded frequencies, the total of spoken *wh*-clefts in Callies' study is approximately 150, excluding his analysis of written *wh*-clefts, in which he is not concerned with the integrated/unintegrated contrast.

4 The study: corpora and data

In the remainder of this article we will focus on the results of the present study with a view to answering the research questions posed in the introduction. This section presents the corpora and data underlying the analysis.

The primary source of data is the *Corpus of Contemporary American English* (COCA; Davies 2008–). The COCA is a balanced corpus of 464 million words, of which 95 million represent spoken language in the form of unscripted speech (mostly conversation but also monologue and interviews) from TV and radio programmes. Exceptionally, in section 5.3, where analysis of audio data is necessary, the data come from the 10-million spoken subcorpus of *British National Corpus* (BNC; Davies 2004–).

The data used are *wh*-clefts with different forms of *do* in the *wh*-clause. The reasons for this decision are the following. Firstly, *do*, *mean* and *happen* are overwhelmingly the most frequent verbs in *wh*-clefts, especially *do* (Weinert & Miller 1996: 194, Hopper & Thompson 2008: 103). Secondly, the amalgam type is far more common with *do* than with any other verb (Koops & Ross-Hagebaum 2008: 464) and has higher acceptability ratings in combination with this verb (Ross 2000: 385). Thirdly, the verb *do* has a number of variant complementation patterns in the focus clause (bare infinitive, *to*-infinitive, main clause) thus displaying varied degrees of syntactic integration.

The target structures were retrieved by searches of word strings of the form: *what* + NP + (optional intevening words) + *do/does/done/did/doing*. Individual hits were inspected to ensure that only relevant occurrences were considered. The total of the body of data used is 8,857 *wh*-clefts from the COCA and 277 from the BNC (the latter used only in section 5.3). All the examples cited below come from the COCA unless otherwise stated. Individual examples are identified according to subcorpus (spoken/written) and text-type (e.g. unscripted conversation).

Below we will consider *wh*-clefts with respect to their structural non-integration (syntactic and prosodic). Simultaneously we will distinguish between two broad types: standard clefts, i.e. infinitival and gerundive on the one hand, and amalgams on the other. The following features will be discussed in turn: mismatched copula tense, prosodic separation (with emphasis on pausing), omission of the copula, omission

https://doi.org/10.1017/S1360674315000180 Published online by Cambridge University Press

of subject pronouns, non-coreferential subject NPs, tense–aspect–mood mismatches, multi-clause focus constituents, delayed focus constituents, and lack of overt focus constituents.

5 Results and discussion

5.1 Infinitival, gerundive and clausal complements in spoken and written language

We will start by considering the general distribution of *wh*-clefts across written and spoken language. This is because the occurrence of integrated vs non-integrated structures has been closely associated with spoken language and especially with spontaneous unplanned speech. Figure 1 represents the distribution of *wh*-clefts with *do* in the *wh*-clause.

In the COCA, *wh*-clefts are over thirteen times more frequent in spoken than in written language, the spoken data being unscripted conversation of TV and radio programmes. Biber *et al*. (1999: 961) report, in a similar vein, that *wh*-clefts are most common in conversation. The authors attribute this proliferation of *wh*-clefts in conversation to the fact that the *wh*-clause often has low information content, as in phrases such as *What this means...* or *What I'd like to do...*, which can be used in the real time of spontaneous speech as a springboard to start an utterance (1999: 963). Others point out other rhetorical functions that, again, may be especially relevant in spoken interactional language: to buy time and hold the floor while the speaker formulates his point, or to give the appearance of fluency by filling in what otherwise would be awkward silences (Hopper & Thompson 2008). Consider example (23), in which one possibility is that the *wh*-cleft, straightforwardly, foregrounds the focus constituent, but the second possibility is that the cleft construction, together with the filling and hedging expressions *well*, *essentially* and *I think*, allows the speaker time to think and plan what to say.

(23) Well, there's all sorts of speculation about what they're going to do, but essentially I think **what they want to do** is put power back into the hands of prosecutors (spoken, unscripted conversation)

The low information content of *wh*-clauses is most visible in amalgams with *do*, in which there is much repetition of lexical and grammatical material; typically, the *wh*-clause contributes very little information.

(24) And **what they could have done** is they could have given up to him. (spoken, unscripted conversation)
(25) Usually **what these people want to do** is they want to get rid of their pain. (spoken, unscripted conversation)

Spontaneous spoken English is thus the typical domain of *wh*-clefts. Having said that, below we will investigate the COCA's written and spoken *wh*-clefts in turn with respect to the degree of their integration.

https://doi.org/10.1017/S1360674315000180 Published online by Cambridge University Press

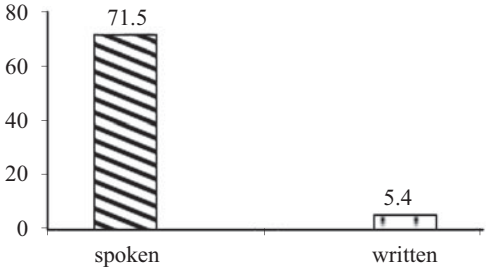


Figure 1. *wh*-clefts in COCA (per 1 million)

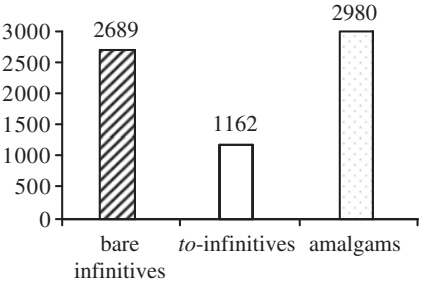


Figure 2. *wh*-clefts in spoken English

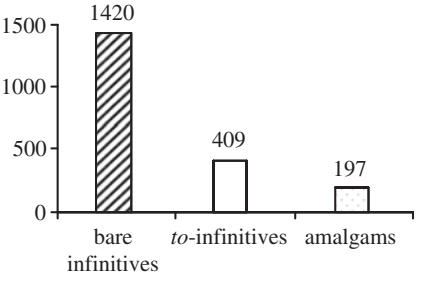


Figure 3. *wh*-clefts in written English

In the spoken data, 43.5 per cent of all *wh*-clefts are of the amalgam type, the rest being accounted for by standard clefts (figure 2). Considering that infinitival cleft clauses come in two varieties – bare and *to*-infinitives – the amalgam variant is the most frequent complement type of the three options available. In written language, on the other hand, 90 per cent of occurrences are of the canonical type, with *to*-infinitives being markedly less common than bare infinitives (figure 3).

We will now examine specific non-integration effects. Firstly, we will look into infinitival clauses, considered fully integrated syntactically, to find out whether this standard complement type displays any signs of non-integration. Individual examples will be identified according to text type.

https://doi.org/10.1017/S1360674315000180 Published online by Cambridge University Press

5.2 Mismatched copula tense

One feature of a minor lack of syntactic integration is when the tense of the copula does not match that of the *wh*-clause. This is possible when the present tense copula is used in a past tense environment:

(26) I think sexuality is confusing and complex, and **what this process did for me is** to open the door to be able to discuss my own sexuality. (written, newspaper)
(27) they found out that **what we were doing is** actually using the picture to protect that information (spoken, unscripted conversation)

Tables 1 and 2 represent the frequency of occurrence of copula tense mismatch with bare- and *to*-infinitives in the COCA. The pattern is more common with the bare infinitive than with the *to*-infinitive. With bare infinitives, when the choice between the present and past copula is available, i.e. when the *wh*-clause is in the past tense, the present copula is used in 15 per cent of cases. With *to*-infinitives, the present copula is used in 9.5 per cent of cases.

The figures also show that the inconsistent tense use is typically to be found in spontaneous spoken English. For example, for bare infinitives, 20 per cent of spoken *wh*-clefts are of the mismatching type as opposed to 7 per cent in written language. More to the point, on closer inspection, it becomes clear that, with only two exceptions, all the occurrences in the written medium are in fact quoted speech (in the press) or intended to mimic speech (in fiction). Clearly, unplanned speech is the domain of this non-integration effect. Note that this is in line with the view that speech in general,

Table 1. *Consistency in tense use (bare infinitives)*

Bare infinitives						
			Spoken		Written	
	Raw counts	%	raw	%	raw	%
Consistent (past+past)	631	85%	326	79.2%	305	92.2%
Inconsistent (past+present)	112	15%	86	20.8%	26	7.8%

Table 2. *Consistency in tense use (*to*-infinitives)*

to-infinitives						
			Spoken		Written	
	Raw counts	%	raw	%	raw	%
Consistent (past+past)	392	90.5%	258	89%	134	93.7%
Inconsistent (past+present)	41	9.5%	32	11%	9	6.2%

https://doi.org/10.1017/S1360674315000180 Published online by Cambridge University Press

Table 3. *Consistency in tense use (amalgams)*

Amalgams						
			Spoken		Written	
	Raw counts	%	raw	%	raw	%
Consistent (past+past)	372	55.9%	363	56%	9	53%
Inconsistent (past+present)	293	44%	285	43.9%	8	47%

in comparison to written language, is marked by unintegrated syntax (e.g. Miller & Weinert 1998; Hopper & Thompson 2008).

The copula has a loose syntactic connection with the cleft structure. It is a connecting point, but also a breaking point for the two clauses. At the same time, the two parts on both sides of the copula have a loose connection with each other, and are to some extent independent (especially in amalgams). Consider example (28), in which the speaker reports a tornado attack – an account of a past event, i.e. when the past tense is normally used. Note that the finite verbs on both sides of the copula are consistently in the past tense, where they need to be in the past tense, but not the copula.

(28) And **what we did is get** in the center of the house which was in the baby's room in the closet. We put some mattresses against the windows and we had a mattress against the door and that's when we heard the big roar and it just took the roof. (spoken, news report monologue)

Although the pattern of tense mismatch is relatively rare with infinitives, it shows that *wh*-clefts with infinitival complements may display a degree of non-integration. This is rather unsurprising if one sees the copula as an element of the construction that is to some extent independent of the rest of the construction. The purpose of the copula is to link the *wh*-clause and the cleft constituent, but the tense of the copula may vary between past and present. The mismatched copula tense is another non-integration feature of *wh*-clefts, although it has not been recognized as such in the previous literature.

Below we carry out an analogical analysis of copula tense use in amalgam *wh*-clefts. Table 3 represents their quantitative distribution. As can be seen in table 3, inconsistent copula tense use is far more common in amalgam *wh*-clefts (in 44 per cent of cases) and especially in the COCA's unscripted conversation, as can be seen in the recorded raw frequencies: only eight instances were found in written texts, all of them, once again, representing quoted speech. The greater frequency of mismatched copula tense in amalgams confirms that the connection of the copula to the cleft construction is looser than in clefts with infinitival complements. Consider examples (29) and (30) below. Both are similar accounts of sequences of past actions but the present tense copula alternates with the past tense form.

(29) And **what I did is** I ran downstairs, and I woke up the girls, and I said, you know, that this has happened, and they called him. (spoken, unscripted conversation)

https://doi.org/10.1017/S1360674315000180 Published online by Cambridge University Press

(30) So **what I did was** I went on the Internet, I made a twitter just as a sounding board and it worked. (spoken, unscripted conversation)

Because in amalgams there is a finite verb in the post-copula clause, tense mismatches are likely to be more frequent, as in (31):

(31) **What they did** right **is they allow** a much wider variety of payment models, (written, newspaper)

In other cases, tense mismatches may be triggered by local factors, e.g. by the proceeding discourse, as below:

(32) A: Madonna hired Michael Jackson's publicist.
B: No, **what Madonna did is she has** a licensing company now that's going to handle all of her merchandise (spoken, unscripted conversation)

In (32), it is possible that the tense mismatch occurs because the tense of *what Madonna did* mirrors that of the immediately preceding statement (*Madonna hired*...). Speaker B swiftly moves on to correct speaker A's statement and the resulting combination of the past and present tenses is non-canonical, although perfectly suitable in its context.

To summarize this section, we have seen that the mismatched tense of the copula is another feature contributing to the non-integration of *wh*-clefts. It is especially common in spoken amalgam *wh*-clefts, but also observable with bare and *to*-infinitives, which offers grounds on which one may see even the standard *wh*-cleft structure as not fully integrated.

5.3 The prosodic separation of the clauses

The loose structure of the *wh*-cleft construction can also be seen in the frequent prosodic separation of the two clauses in speech. Collins (2006) notes that 94.6 per cent of his *wh*-clefts are intonationally divided into two or more tone units. Weinert & Miller (1996) report pauses separating the two parts of the cleft although they offer no quantification of this observation. This section outlines an analysis of audio data with a view to establishing the extent to which *wh*-clefts are separated by unfilled (silent) pauses. We will consider the location and frequency of occurrence of such pauses as well as their duration. As the COCA corpus does not allow auditory analysis, the spoken subcorpus of the BNC has been used to retrieve relevant audio files (Pęzik 2015–). Silent pauses preceding or following the copula were identified and their duration was measured with PRAAT (Boersma 2001). For this analysis, a silence equal to or longer than 0.2 seconds is counted as a pause; in choosing this cut-off point, we follow the common standard adopted elsewhere (e.g. Kendall 2013: 63). Occurrences of clefts with obvious disfluencies such as false starts or self-repairs were excluded as they would have been relevant in pausing.

In all, 277 spoken clefts were examined. In the majority the *wh*-clause constituted a separate tone unit, as reported by Collins (2006), and the tone unit boundary occurred at the copula. Of the 277 occurrences, 73 (26 per cent) were additionally separated

https://doi.org/10.1017/S1360674315000180 Published online by Cambridge University Press

Table 4. *Pauses in BNC*

	Pause ≥ 0.2s		Pause < 0.2s	
	raw	%	raw	%
Amalgams	36		95	
Bare infinitives	25		71	
to-infinitives	12		38	
Total	73/277	26.3%	204/277	73.6%

Table 5. *Length of pauses*

	73 spoken *wh*-clefts	
	raw	%
Pauses 0.2s – 0.3s	22	30.1%
Pauses 0.3s – 0.4s	14	19.1%
Pauses 0.4s – 0.5s	10	13.6%
Pauses > 0.5s	27	36.9%
Total	73	100%

by a pause of at least 0.2 seconds. These pauses typically occurred before the copula, with only nine cases (12 per cent) in which the pause occurred after the copula. Thus, examples (33–35) are prosodically typical of the 73 clefts with pauses (the numbers in brackets represent the location and duration of pauses):

(33) **What you can do** [0.38] **is** you could turn the graph that way up. (BNC, spoken, classroom)
(34) and **what I want to do** [0.22] **is** to look at it er over that particular period (BNC, spoken, a training course in a commercial context)
(35) now ideally **what I'd like to do** [0.80] **is** to have a chat with somebody, (BNC, spoken, private conversation)

All three types of copula complementation (bare infinitives, *to*-infinitives and amalgams) showed prosodic separation to approximately the same degree, i.e. 26 per cent of bare infinitives, 24 per cent of *to*-infinitives and 27 per cent of amalgams were found to be separated by silent pauses at the copula. Thus, in this respect, infinitival clefts can be said to be as (non-)integrated as amalgams. Table 4 shows the quantitative contribution of each complementation type to the total of the sample, i.e. 277 items.

As can be seen, *wh*-clefts are not nearly as frequently separated by a pause as they are by a tone unit boundary. However, on average, in one in every four cases, the *wh*-cleft structure is perceptibly divided by a pause. In fact, pauses longer than 0.5 seconds, i.e. far longer than the minimum threshold, were most frequent. The pauses were all the more noticeable given that the pause co-occurs with the tone unit boundary. Table 5 shows the durations of the pauses.

https://doi.org/10.1017/S1360674315000180 Published online by Cambridge University Press

Functionally, these pauses may be boundary markers used for segmentation purposes, and/or they may have a non-segmenting communicative function of drawing the listener's attention to the upcoming information (cf. Murkherjee's (2000) concept of anticipation pauses). In either case, the pausing produces prosodic segmentation that complements the separation by tone unit boundaries reported by Collins (2006), as well as the phonetic reduction and faster tempo of *wh*-clauses reported by Weinert and Miller (1996) (also observed in this study). In combination, all these prosodic phenomena render the cleft construction prosodically non-integrated to a considerable degree and also facilitate the use of the *wh*-clause as a formulaic chunk. In clefts such as those in (33)–(35), the *wh*-clauses are relatively fixed, formulaic chunks allowing limited modifications and with low information content which can be used for a variety of rhetoric functions (see Hopper & Thompson 2008). When such semi-fixed phrases are used for such purposes, they are likely to be prosodically separated from the focus phrase.

Another prosodic argument pointing to the formulaic status of the *wh*-clause is the fact that in 88 per cent of cases the copula is prosodically not part of the *wh*-clause, suggesting a strong tendency for the prosodic separation of the *wh*-clause and the copula, and indicating a formulaic template that consists of the *wh*-clause on its own. Interestingly, in two cases of a post-copula pause, there was an additional pre-copula pause, further indicating the tendency to separate the *wh*-clause and the copula. All in all, in only 7 cases (9 per cent) is the copula unseparated from the *wh*-clause by a pause. This is compatible with Miller (2011: 37–8), who observes that when the copula is prosodically part of the *wh*-clause (including a pause following the copula), as in (36), the prosodic characteristics of both pre- and post-copula clauses suggest an imperative clause analysis.

(36) Er, what you're going to do is go in between the camera shop and the lefthand side of the page (Miller 2011: 37)

5.4 Omission of the copula

Another non-integration feature, and a more radical one, is the omission of the copula, which results in a fragmented structure in which the variable contained in the *wh*-clause lacks an overt syntactic link to its value in the focus constituent. Such omission is typically seen in amalgams rather than standard *wh*-clefts:

(37) I mean, **what that ad did, it** got Gil Aust a lot of press coverage in Washington and other places (spoken, unscripted conversation)
(38) **What I would do with the glasses, I** would look to see the ridge detail (spoken, unscripted conversation)

Clefts with infinitival complements are also subject to copula omission, as in (39) from Hopper & Thompson (2008: 102):

(39) What you oughta do though, Mar, cook all the fish.

In (39), *cook all the fish* might also be interpreted as the speaker's directive given to Marilyn, thus inviting an interpretation under which the phrase is an imperative

https://doi.org/10.1017/S1360674315000180 Published online by Cambridge University Press

Table 6. *Copula omission in COCA*

	Amalgams		Infinitives	
	raw	%	raw	%
With copula	2,804	88.2%	5,680	99.8%
Without copula	373	11.7%	11	0.2%

clause. In this way some copula-less infinitival complements have an ambiguous status (see Weinert & Miller 1996: 195; Miller & Weinert 1998: 120–3). However, in other examples, including (40) and (41), the imperative reading is unlikely and the utterances are best seen as copula-less *wh*-clefts.

(40) Dealers aren't in the business to do social work. **What they want to do, get** you to come in. (spoken, news report monologue)
(41) First, **what we have to do, curb** our spending where we can. (spoken, unscripted conversation)

Admittedly, copula omission occurs only sporadically with infinitival complements, as evidenced by only eleven such instances in the COCA. Amalgams, on the other hand, allow the omission more freely (373 instances) and by virtue of this fact, once again, exhibit a less integrated structure. Table 6 represents the extent to which copula omission is observed with both types of complementation. As before, occurrences found in written language constitute quotations from unplanned speech (in the press) or mimic its style (in fiction).

5.5 Omission of subject pronouns

Biber *et al*. (1999: 963) note that 'cleft constructions are syntactically integrated, making use of function words to build a close-knit structure'. However, in amalgam *wh*-clefts, this syntactic integration built with, for example, pronominalization can be disrupted by the ellipsis of a subject pronoun in the focus phrase, as a result of which finite clauses lack subject NPs.

(42) Clearly, **what television does is** instantly **immerses** an entire continental nation in a common, single experience, (spoken, unscripted conversation)
(43) So **what the prosecutor did is tipped** his hand that he's going to investigate this. (spoken, unscripted conversation)

In other cases, the pronoun is ellipted along with other function words such as auxiliaries (sentence (44)) and the negative *not*, as in (45):

(44) **What we ought to have done is deployed** troops from the area [...] (spoken, unscripted conversation)
(45) **What he hasn't done is adopted** the rhetoric of his predecessor. (spoken, unscripted conversation)

https://doi.org/10.1017/S1360674315000180 Published online by Cambridge University Press

In terms of the ellipted material, the surface forms of these constructions resemble those of canonical *wh*-clefts in which the same function words are omitted and the infinitive is all that is left of the verb group. The difference in surface forms is that finites replace non-finites, e.g. *what we ought to have done is **deploy** troops* vs *what we ought to have done is **deployed** troops*. In this sense then, ellipsis of pronouns improves the integration of amalgams, although the pattern is non-canonical and less integrated than classic *wh*-clefts. It may even be a source of confusion, as when this type of clefts are interpreted predicationally rather than specificationally. Consider (46):

(46) **What he should have done was telegraphed** at the beginning, (spoken, unscripted conversation)

Sentence (46) is ambiguous between a predicational reading ('someone telegraphed what he should have done') and a specificational reading ('what he should have done is telegraph'). The finite verb *telegraphed* is responsible for the ambiguity as it makes possible, together with *was*, a passive voice reading. The use of the infinitive would resolve the ambiguity and improve the syntactic integrity of the sentence. The use of finite clauses in examples (42)–(46) brings these constructions closer to the amalgam type rather than standard *wh*-clefts. Consider the typology suggested below ((47) and (49) are invented examples modelled on (46)):

(47) **What he should have done was** he should have telegraphed at the beginning, (unellipted amalgam)
(48) **What he should have done was** ~~he should have~~ telegraphed at the beginning, (ellipted amalgam)
(49) **What he should have done was** telegraph at the beginning, (standard *wh*-cleft)

Amalgam *wh*-clefts are exceptional in that they allow finite clauses to complement *do* as the verb of the *wh*-clause. These finite clauses can be complete with subject NPs and copy lexical material from the *wh*-clause, as in (47), or they may involve ellipsis, as in (48). In terms of structural integration, we will claim that example (48) has a more integrated structure than (47), and the reasons are the following. Firstly, the copula cannot be as easily omitted as in the case of unellipted amalgams such as (47) – no such cases have been found in the COCA. Secondly, in the 163 occurrences of finite clauses with pronoun ellipsis there are no commas separating the two clauses, which is often the case in transcribing amalgams in corpus data and which is quite universal when the copula is omitted. Although this is not a reliable diagnostic, the comma seems to reflect the perceived non-integration of amalgams. We take the absence of the comma in clefts with pronoun ellipsis to be an indication of better integration, compared to unellipted amalgams. Thirdly, the ellipsis that occurs in examples such as (48) eliminates repetition of material, thus creating a more coherent and close-knit structure. In conclusion, we propose the following scale of decreasing syntactic integration:

standard *wh*-clefts > ellipted amalgams > unellipted amalgams

https://doi.org/10.1017/S1360674315000180 Published online by Cambridge University Press

Table 7. *Pronoun omission in COCA*

	Amalgam *wh*-clefts	
	raw	%
Overt pronoun	3,014	94.8%
Pronoun omission	163	5.1%

Of course, the scale is not complete in detail as specific representatives of one and the same type may still vary as to specific non-integration features. For example, within the ellipted amalgam type, the tense of the copula may or may not match the tense of the *wh*-clause, thus maintaining or undermining integrity, e.g.

(50) **What I did is took** that photographic evidence, color photographs, and shipped it to seven different experts (spoken, unscripted conversation) [past-present mismatch = integrity]
(51) **What I did was took** that photographic evidence, color photographs, and shipped it to seven different experts (modified) [tense concord = higher integrity]

Table 7 shows the extent to which subject pronoun omission is present in amalgam *wh*-clefts in the COCA.[7] As can be seen, subject pronouns are dropped in 5 per cent of eligible cases.

5.6 Non-coreferential subject NPs

Another non-integration effect in amalgam clefts is the non-coreferentiality of subject NPs, typically pronouns. Canonically, the two subjects should be identical, as in (52), or coreferential, as in (53) and (54):

(52) **what he began to do was he began** to separate himself from everybody (spoken, interview)
(53) And **what South Africa did was they learned** from that experience. (spoken, unscripted conversation)
(54) because **what mathematicians do is we** look at the world and we see and we identify patterns (spoken, unscripted conversation)

The coreference of the two subjects brings the two clauses together, thus improving sentence integrity. On the other hand, lack of coreference undermines integrity, as in (55) and (56):

(55) A: How did you guys pull this off? How did you get ants on a space shuttle?

[7] The full extent of pronoun omission is difficult to ascertain. Ellipted pronouns can only be detected in 3rd person singular present tense forms and perfective forms; other persons and tense forms obscure it (e.g. *what he does is goes* vs *what I do is go*). The ellipsis is also obscured by identical perfective forms of irregular verbs. In table 7, only explicit examples of omission are counted.

https://doi.org/10.1017/S1360674315000180 Published online by Cambridge University Press

B: Well, actually, **what we did is it started** as a normal high school science project. (spoken, unscripted conversation)

(56) It used to be you'd say something to management and they'd say "I'm the boss. You're the worker." Now **what they do is we sit** down and talk about the problem, (written, newspaper)

In (55) and (56), although the changes in subject NPs weaken the integration of the clefts themselves, they may be seen as perfectly fitting in the context of the preceding discourse. In (55) *we* in B's answer is used in response to *you guys* in A's question. Speaker B begins with the kind of *wh*-clause (*what we did*) that is a sensible response to A's question (*How did you...?)* but then switches to a non-coreferential subject to accommodate the focus phrase. Similarly, in (56) the speaker switches from *they* (i.e. management) to the more inclusive *we*, i.e. management and the worker. The implication is that it is the management's new policy now to solve problems together ('what they do now is include the worker in the discussions').

Similar examples of splitting groups of participants are illustrated in (57), in which *they* includes *the father*, and in (58), in which *we* includes *I*.

(57) So **what they did was, the boy's father testified** about how much the 9-year-old missed his mom. (spoken, unscripted conversation)

(58) But **what I do do is we have** wonderful discussions about princesses that really do exist. (spoken, unscripted conversation)

Still, in other cases, the use of non-coreferential subjects does not seem to be dictated by the surrounding discourse. It is in these cases that the lack of integration is most detectable. Consider the fragmentation and incoherence brought about by the lack of coreferential subjects in examples (59) and (60).

(59) A: Were not critical of the plans or the employers or unions that have negotiated them.
B: No, you're – no, you are critical – you – you are critical of taxing them. You are critical of the bill for wanting to tax them.
A: **What the report does is, they look at the taxes.** And there are two parts to the tax. One is an excise tax, a new health care tax that would be on pharmaceutical price – costs on devices on health insurance. (spoken, unscripted conversation)

(60) It is an organization that has tentacles into a lot of places. And so **what they're doing in Pakistan, we are working** with the Pakistani government. That's why you have military raids in the frontier region in Pakistan. (spoken, unscripted conversation)

In a number of *wh*-clefts in the COCA, non-coreferential subjects are used to produce an interesting effect that is not present in examples (54)–(60) above. In these clefts the NPs represent two distinct entities (X and Y) that enter a kind of cause-and-effect relationship in which the constructions may be paraphrased as 'what X did is cause Y to do sth'. Examples (61) and (62) are two cases in point.

(61) So what these two **science papers did was I realized** how to make that happen. (spoken, unscripted conversation)

(62) **What the Internet has done is, we have given up** being the purveyor of top-secret information. (written, newspaper)

https://doi.org/10.1017/S1360674315000180 Published online by Cambridge University Press

Table 8. *Non-coreferential NPs in COCA*

	Amalgams	
	raw	%
Identical or coreferential NPs	2,927	97.1%
Non-coreferential NPs	87	2.8%

In clefts such as above, the use of non-coreferential subjects does not generate confusion, incoherence or reference problems, as is the case in (59) and (60). The use of non-coreferential subjects makes perfect sense as the two NPs refer to distinct entities which are interrelated in a way that is predictable: 'X causes Y to do sth'; here, the two examples can be paraphrased as 'these two science papers made me realize ...' and 'the Internet has caused us to give up ...'. In terms of logic then, the type of *do*-clefts represented by (61) and (62) is clearer and more transparent than examples (59) and (60); structurally, however, (61) and (62) are still non-canonical and marked by a degree of non-integration. This type is especially frequent with *what it does* as the *wh*-clause (see (63) and (64)) and also includes focus constituents introduced by *that*-complementizer (65).

(63) Well, in reality this experiments doesn't measure neutrino mass directly, **what it does is we are able to** distinguish between two types of neutrinos, (spoken, unscripted conversation)
(64) Now, **what it does is the enzymes** in the bacteria **breaks down** (spoken, unscripted conversation)
(65) if the police are seen as unfair in their policing, what **that will do is that people will take** the law into their own hands (spoken, news report monologue)

What it/that/this does is especially convenient in referring to something mentioned previously and indicating the relevance it has for something else. This suggests that *what it does* and similar low-content *wh*-clauses can be used hollistically in discourse as chunks. Additional support comes from the fact that the chunk is often not complemented by a focus phrase until later in the following sentence, as in (66), or not overtly complemented at all, as in (67). But even then one should not see it as an incomplete cleft which has been abandoned mid-sentence, but as one that has a complement derivable from context. We will return to this point in section 5.8.

(66) Well, high definition television, **what it does is,** let's say you're watching a sporting event. You can actually see the individual blades of grass. You can see the pinholes in the jersey. (spoken, unscripted conversation)
(67) If you listen to the television commentators, the talking heads, all of us, it sounds like what he should do is suddenly compromise, retreat, water down his agenda. I think that would be the worst thing he could possibly do. **But what you can do**, you know, Bill Clinton did a very good job in thwarting the Republican majority in Congress for six

https://doi.org/10.1017/S1360674315000180 Published online by Cambridge University Press

years, and so it's a question whether George W. Bush is going to play hardball or softball, and I don't think the jury is spoken to that one yet. (spoken, unscripted conversation)

Table 8 shows the frequency with which non-coreferential subjects occur in amalgams. As can be seen, they occur in 2.8 per cent of eligible cases.

5.7 Tense–aspect–mood mismatches

As was pointed out above, amalgam *wh*-clefts often exhibit a fair amount of repetition between the *wh*-clause and the cleft constituent, as seen in (68) and (69) below, where some lexical material is repeated (*I'm trying to, love to*) and the syntactic features of tense, aspect and mood (TAM) are consistently the same (present tense, progressive aspect).

(68) **What I'm trying to do here is I'm trying to** get around the corner (spoken, unscripted conversation)
(69) Typically, **what people love to do is they love** to have these big dishes. (spoken, unscripted conversation)

Likewise, negation is repeated in both clauses:

(70) But **what we're not doing is we're not** going to have a big government takeover (spoken, unscripted conversation)
(71) **What this bureau doesn't do well is it doesn't** account for its own blind (spoken, unscripted conversation)

However, as pointed out in section 2, in amalgam *wh*-clefts, the cleft constituent does not necessarily mirror the syntax of the *wh*-clause resulting in mismatches of TAM values. This is illustrated in (72):

(72) Well, **what I wanna do is, I don't wanna have a debate where** it simply degenerates. (spoken, unscripted conversation)

What is striking in (72) is the conjunction of two clauses – one positive (*what I wanna do*), the other negative (*I don't wanna have*) – joined by the copula. In such examples, it is especially visible that *wh*-clauses can be distinctly formulaic and detached from their *be*-complements.

In other such cases, the negative clause is merely an introduction to the genuine *be*-complement which appears later in the text. In example (73), the complement for the *wh*-clause appears in the next sentence, thus marking the separation of the two clauses especially heavily. As has been noted elsewhere (Carter & McCarthy 2006; Hopper & Thompson 2008; Koops & Ross-Hagebaum 2008) this spread of the projectibility of the *wh*-clause over several clauses is rather common in amalgam *wh*-clefts. We will return to this point in the next section.

(73) Yes, it sounds a little oversensitive to me. My – **what I was doing**, I didn't speak about anybody by name. **I talked about members** of Congress. (spoken, unscripted conversation)

https://doi.org/10.1017/S1360674315000180 Published online by Cambridge University Press

Table 9. *TAM inconsistencies in COCA*

	Amalgams	
	raw	%
Consistent TAM values	2,985	94%
Inconsistent TAM values	193	6%

Sentence (73) is additionally marked for lack of syntactic integrity in that the use of TAM values is inconsistent across the relevant clauses. The Past Continuous tense in the *wh*-clause is mismatched with the Past Simple tense in both clauses following it. Other types of TAM inconsistencies are exemplified below:

(74) (Present Simple+Past Simple) **What it also does is, it helped** move what Keynes called the animal spirits of the market. (spoken, unscripted conversation)
(75) (Present Continuous+*like to*) Michael, **what these people are doing is they like to take** the word 'Democrat' and make it synonymous with the word 'liberal'. (spoken, unscripted conversation)
(76) (Present Perfect+Present Continuous) I think **what Jim has done is that he is looking** at it through the lens of today (written, newspaper)
(77) (Past Simple+Present Simple) So **what we did in this one is** we go back in time to 1969 in the second act (spoken, unscripted conversation)
(78) (Present Simple+*will*) **what this does is he'll have to** have a roll call vote (spoken, unscripted conversation)
(79) (Present Simple+*can*) **What you do, you can** marinate it. (spoken, unscripted conversation)
(80) (*have to*+*need*) Now **what we have to do, we need to** tidy up a little bit (spoken, unscripted conversation)
(81) (*could*+*would*) **what it could do, it would** put one image over another, it would remove enough image so you could be right there. (spoken, unscripted conversation)
(82) (*like to*+*could*) **what I like to do, you could** add a little bit of parmesan cheese (spoken, unscripted conversation)
(83) (*would like to*+*will*) **What I'd like to do, we** will just rejoin the Barry Gray show (spoken, unscripted conversation)
(84) (*going to*+*like to*) But **what we're going to do, I** like to use my hands when I cook (spoken, unscripted conversation)
(85) (*would*+*might*) So, **what we'd do is we** might show that information to your friends a little bit (spoken, unscripted conversation)

Of course, in some of the sentences above, several non-integration effects occur simultaneously, e.g. in example (82) the copula is dropped, there is a mismatch in lexical material, and the subject NPs are non-coreferential. The sentence is also of the amalgam type, which is less integrated than the infinitival type. As a result of this combination of features, the cleft is heavily disjointed and the *wh*-clause itself (*what I like to do*) seems to be more of an extra comment on the focus clause rather than an integral part of the assertion. Table 9 shows the extent to which TAM mismatches occur in the COCA.

https://doi.org/10.1017/S1360674315000180 Published online by Cambridge University Press

5.8 Multi-clause focus constituents, delayed focus constituents, and lack of overt focus constituents

As noted above, in amalgam *wh*-clefts the focusing effect over the post-copula constituent may be placed over longer stretches of text: several clauses or even sentences, as in example (86) below:

(86) I didn't use any scientific method to determine it. **What I did was I took the circle and I divided it in half and then I divided it in half again to a quarter. And I just cut out a piece.** And that's how I did it. (spoken, unscripted conversation)

This results in a looser link between *wh*-clauses and their focus constituents as in many cases, as in (87) below, it is difficult to ascertain the exact length of text over which the focus is placed.[8] Instead, in (87), *what I did* acts as a formulaic chunk introducing the speaker's account of what happened. Thus the phrase *what I did* is functionally similar to *what I mean* or *what I'm saying*.

(87) And so, basically **what I did is** I couldn't live with the pain any longer, I mean, I had to start giving up sports. I mean, I love to play basketball, play tennis, I couldn't do that any longer because every time I played, my knee – it would take me two weeks to recover from it. (spoken, unscripted conversation)

Extra material between the *wh*-clause and the focus constituent may also occur, which further exacerbates the fragmentation effect of such constructions. Consider examples (88) and (89), in which *wh*-clauses and focus constituents (in bold) are separated by parenthetical information. Note that in these examples, precisely because of the intervening material, infinitival complementation in focus constituents is highly unlikely.

(88) And **what they did was,** within three months of our going into Iraq while our troops were in harm's way, **they attacked the commander in chief** as a liar, a fraud. (spoken, unscripted conversation)

(89) **What they do is**, we have a form in the back of the catalog and **they can fill out an absentee bid form**. (spoken, unscripted conversation)

In example (90) below, the separation of the two elements is quite extreme. In the excerpt, the speakers discuss research into using bees to detect explosives and the pronoun *they* in the *wh*-clause refers to bees; the focus constituent complementing the *wh*-clause appears much later in the next sentence (in bold).

(90) A: What we're doing is we're training bees much like you would train a dog, to search out a certain odor.
B: Right.
A: And **what they do is** you take a feeder that is doped with sugar water, which the bees like...
B: Mm-hmm.

[8] For this reason, frequency data are not given in this section.

https://doi.org/10.1017/S1360674315000180 Published online by Cambridge University Press

A: and then you, basically, sprinkle, or you put some of the odor that you want them to search for near that feeder. Then **as the bees come and feed, they smell the odor and then they begin to associate the odor with a food source.** (spoken, unscripted conversation)

In other cases, focus constituents are not overtly present and need to be deduced from context. In most cases the deduction is fairly straightforward, as in (91) and (92), although the lack of an overt complementation for *wh*-clauses (along with copula omission) makes for a strong sense of structural non-integration (translations into canonical *wh*-clefts are given in brackets).

(91) And **what you also do,** I am for stimulating, I am for the fact that you have to give people incentives to make jobs in the economy. ('What you also do is stimulate people to make jobs') (spoken, unscripted conversation)

(92) **What we're going to do tonight**, I've got Lucianne Goldberg just itching to come out backstage. Of course, she was one of the original players in this drama. And also, Michael Moore, who comes at it from a different side. ('What we're going to do tonight is talk to Lucianne Goldberg and Michael Moore') (spoken, monologue)

In other cases, deducing the focus element can be slightly problematic, as in (93), where the speakers are talking about how long the economical recession will continue.

(93) A: Well, here's the thing, let's just – let's just assume that this is going to continue, OK?
B: Mm-hmm.
A: Because all indicators show that this will. **What we have to do is** we've already been in it for a while, we've made our adjustments, but the key point here is we've adjusted to a time that is very, very difficult, this is going to take about a year to go. None of us can call the bottom, none of us are going to be able to call the top, but this is going to be at least a year, we can all assume. (spoken, unscripted conversation)

In (93), it seems that the *wh*-clause (*What we have to do*) can optionally be complemented by three ideas which appear in the text, i.e. (a) 'assume that it is going to continue', (b) 'assume that it is going to be at least a year' or (c) 'make our adjustments', or by a conclusion that can be inferred from the speaker's words, i.e. 'What we have to do is brace ourselves for at least another year'.

6 Conclusion

The lack of structural integration exemplified in this study is a combination of the inherent bipartite structure of the *wh*-cleft construction and the natural fragmentation and parataxis of spontaneous speech. In amalgam *wh*-clefts, which are very common in spontaneous spoken English, the separation into two parts is particularly noticeable. All in all, we have proposed four broad types of constructions representing a scale of decreasing integration of *wh*-clefts:

https://doi.org/10.1017/S1360674315000180 Published online by Cambridge University Press

standard *wh*-clefts (*what he does is leave*)



ellipted amalgams (*what he does is leaves*)



unellipted amalgams (*what he does is he leaves*)



amalgams without an overt focus phrase (but derivable from context)

As we move down the scale, the structural integration of *wh*-clefts decreases. Additionally, this effect is reinforced by the occurrence of syntactic and prosodic non-integration phenomena which are relatively rare in standard *wh*-clefts and more frequent in amalgams. These phenomena can be prosodic or syntactic in nature. The former include the use of intonation patterns that separate the two clauses and pausing at the copula (before, which is more frequent, or after it). The latter include a variety of inconsistent syntax phenomena, two of which have not been considered in the previous literature, i.e. mismatched copula tense and the omission of subject pronouns. Another original contribution of this article is using a much larger sample to establish the extent to which individual non-integration features are observed. In the COCA, these features occur with the following frequency rates:

Standard *wh*-clefts:

- copula tense mismatch (15% for bare infinitives and 9.5% for *to*-infinitives)
- prosodic separation by a pause (bare infinitives: 26%; *to*-infinitives: 24%)
- copula omission (0.2%)

Amalgam *wh*-clefts:

- copula tense mismatch (44%)
- prosodic separation by a pause (27%)
- copula omission (11.7%)
- omission of subject pronouns (5.1%)
- non-coreferential subject NPs (2.8%)
- inconsistent TAM values (6%)
- multi-clause focus constituents, delayed focus constituents, lack of an overt focus constituent (frequency of occurrence is difficult to establish)

We have also seen that, when *wh*-clauses are isolated as a result of syntactic and prosodic fragmentation, they can act as semi-formulaic units projecting focus onto an indefinite extent of upcoming discourse. The common separation of the *wh*-clause from a single-clause focus phrase means that the *wh*-clause gains some independence,

https://doi.org/10.1017/S1360674315000180 Published online by Cambridge University Press

becomes disconnected, and thus may act as a pre-packaged chunk. In this respect then, *wh*-clauses with *do* may act much in the way that *wh*-clauses with *mean* and *say* do (in phrases such as *what I mean* and *what I'm saying*).

With regard to language use, the occurrence of the above features of structural non-integration is not necessarily an infelicitous departure from the compactness and integration of the standard *wh*-cleft construction in every case. Some occurrences seem perfectly justified and natural in appropriate context. For example, tense mismatches (example (32)) and the use of non-coreferential subject NPs (example (55)) may be dictated by preceding discourse. Likewise, the use of non-coreferential subject NPs is natural in contexts where a group of participants is split (example (57)), or in contexts where two referents are bound by a cause-and-effect relationship (e.g. *what it does is we no longer need to...*).

Author's address:

English Department
John Paul II Catholic University of Lublin
Al. Racławickie 14, 20-950 Lublin
Poland
wguz@o2.pl

References

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad & Edward Finegan. 1999. *Longman grammar of spoken and written English*. Harlow: Pearson Education.

Boersma, Paul. 2001. Praat, a system for doing phonetics by computer. *Glot International* 5(9/10), 341–5.

Callies, Marcus. 2012. The grammaticalization and pragmaticalization of cleft constructions in present-day English. In Sebastian Hoffmann, Paul Rayson & Geoffrey Leech (eds.), *Corpus linguistics: Looking back – moving forward*, 5–21.Amsterdam: Rodopi.

Carter, Ronald & Michael McCarthy. 2006. *Cambridge grammar of English CD-ROM: A comprehensive guide to spoken and written English usage*. Cambridge: Cambridge University Press.

Collins, Peter. 1991. *Cleft and pseudo-cleft constructions in English*. London: Routledge.

Collins, Peter. 2006. *It*-clefts and *wh*-clefts: Prosody and pragmatics. *Journal of Pragmatics* 38(10), 1706–20.

Davies, Mark. 2004–. *BYU-BNC*. (Based on the *British National Corpus* from Oxford University Press). Available online at http://corpus.byu.edu/bnc/.

Davies, Mark. 2008–. *The Corpus of Contemporary American English: 450 million words, 1990–present*. Available online at http://corpus.byu.edu/coca/.

Hopper, J. Paul & Sandra A. Thompson. 2008. Projectibility and clause combining in interaction. In Ritva Laury (ed.), *Crosslinguistic studies of clause combining: The multifunctionality of conjunctions*, 99–124. Amsterdam: John Benjamins.

Huddleston, Rodney & Geoffrey Pullum. 2002. *The Cambridge grammar of the English language*. Cambridge: Cambridge University Press.

Kendal, Tyler. 2013. *Speech rate, pause and sociolinguistic variation: Studies in corpus sociophonetics*. Basingstoke: Palgrave Macmillan.

https://doi.org/10.1017/S1360674315000180 Published online by Cambridge University Press

Koops, Christian & Sebastian Ross-Hagebaum. 2008. Information structure and discourse function of amalgam *wh*-clefts. *Proceedings of the 37th Annual Meeting of the Berkeley Linguistics Society*, 461–72.
Lambrecht, Knud. 2001. A framework for the analysis of cleft constructions. *Linguistics* 39(3), 463–516.
Miller, Jim. 2011. *A critical introduction to syntax*. London: Continuum.
Miller, Jim & Regina Weinert. 1998. *Spontaneous spoken language: Syntax and discourse*. Oxford: Clarendon Press.
Mukherjee, Joybrato. 2000. Speech is silver, but silence is golden: some remarks on the function(s) of pauses. *Anglia* 118(4), 571–84.
O'Keeffe, Anne, Michael McCarthy & Ronald Carter. 2007. *From corpus to classroom: Language use and language teaching*. Cambridge: Cambridge University Press.
Pęzik, Piotr. 2015–. *Spokes – a search and exploration service for conversational corpus data*. Available online at: http://pelcra.clarin-pl.eu/SpokesBNC
Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech & Jan Svartvik. 1985. *A comprehensive grammar of the English language*. London: Longman.
Ross, Haj. 2000. The frozenness of pseudoclefts – towards an inequality-based syntax. In Arika Okrent & John P. Boyle (eds.), *Proceedings of the 36th Meeting of the Chicago Linguistic Society*, 385–426. Chicago: University of Chicago Press.
Weinert, Regina & Jim Miller. 1996. Cleft constructions in spoken language. *Journal of Pragmatics* 25(2), 173–206.

https://doi.org/10.1017/S1360674315000180 Published online by Cambridge University Press