A Clarification Request Markup Scheme for the BNC

Matthew Purver

February 19, 2002

Abstract

This paper describes the markup scheme used in annotating a sub-corpus of the British National Corpus (BNC) with information about clarification requests.

1 Introduction

Clarification requests (CRs) are common in human conversation. They can take various *forms* and can be intended by the speaker making the request (the CR *initiator*) to request various types of clarification information (i.e. they can have various *readings*), but have in common the fact that they are in a sense utterance-anaphoric – they refer to the content or form of a previous utterance that has failed to be fully comprehended by the initiator.

As part of an attempt to exhaustively categorise CR forms and readings, a sub-corpus of the British National Corpus (BNC) – (see Burnard, 2000) – has been used to examine the nature of CRs naturally occurring in dialogue. Once a set of possible CR forms and readings had been constructed, a markup scheme was constructed and the corpus marked up accordingly. The markup scheme also allowed CRs to be tagged with information concerning the source of the clarification (i.e. the sentence being clarified). This document describes the markup scheme and the tagging process.

The resulting tagged corpus has allowed us to determine a set of and to extract information about the relative frequencies of these forms and readings. Details of the results are given in (Purver et al., 2001).

1.1 Aims and Procedure

Our intention was to investigate the forms and readings for CRs that are present in a corpus of dialogue. For this purpose we used the BNC, which contains a 10 million word sub-corpus of English dialogue transcripts. For this experiment, a sub-portion of the dialogue transcripts was used consisting of c. 150,000 words. To maintain a spread across dialogue domain, region, speaker age etc., this sub-portion was created by taking a 200-speaker-turn section from 59 transcripts.

All CRs within this sub-corpus were identified and tagged, using the markup scheme and decision process described in 4 and 4.1 below. At time of writing, this process has been performed by one expert user and repeated by one naive user. Reliability seems reasonable, with kappa figures (see Carletta, 1996) varying between 75% and 90% (see Purver et al., 2002, for details).

Initial identification of CRs was performed using SCoRE (Purver, 2001), a search engine developed specifically for this purpose (in particular, to allow searches for repeated words between speaker turns, and to display dialogue in an intuitive manner). However, in order to ensure that all clarificational phenomena were captured, the final search and markup were performed manually.

2 Clarification Forms

In this section we list the CR forms identified, and illustrate them with examples from the BNC.

2.1 Non-Reprise Clarifications

In this form, the nature of the clarifying information being requested by the CR initiator is spelt out for the addressee. Utterances of this type thus often contain phrases such as *"do you mean..."*, *"did you say..."*, *"what is..."*,

"what does... mean", as can be seen in examples (1) and (2).

(1)[1]

Cassie:	You did get off with him?
Catherine:	Twice, but it was totally non-existent kissing so
Cassie:	**What do you mean?**
Catherine:	I was sort of falling asleep.

(2)[2]

Leon:	Erm, your orgy is a food orgy.
Unknown:	**What did you say?**
Leon:	Your type of orgy is a food orgy.

2.2 Literal Reprise Sentences

Speakers can form a CR by echoing or repeating a sentence from the source utterance in full, as shown in example (3).

(3)[3]

Orgady:	I spoke to him on Wednesday, I phoned him.
Obina:	**You phoned him?**
Orgady:	Phoned him.

The sentence echoed need not consitute the complete source utterance, but can be a subpart of it as long as it is a complete sentence that could stand in its own right – as is the case in example (3) above and example (4) below:

(4)[4]

Unknown:	Stay in the car `<pause>` stay with the car and put these `<pause>` erm, motorists lights on.
Heather:	**Stay in the car?**
Unknown:	Mm.

Repeats need not be verbatim, due to the possible presence of phenomena such as anaphora and VP ellipsis (see example (5)), as well as changes in

[1]BNC file KP4, sentences 521–524
[2]BNC file KPL, sentences 524–526
[3]BNC file KPW, sentences 463–465
[4]BNC file KNF, sentences 728–730

indexicals as already shown in example (3) above.

(5)[5]

Anon 5	Oh he's started this other job
Margaret	**Oh he's started it?**
Anon 5	Well, he he <pause> he works like the clappers he does!

2.3 Reprise Fragments

This form involves echoing or reprising a fragment (not a full sentence) of a previous utterance.

(6)[6]

Lara:	There's only two people in the class.
Matthew:	**Two people?**
Unknown:	For cookery, yeah.

It is possible that this fragment consitutes the entire source utterance (see example (7)) – what is important is that it cannot stand alone as a sentence in its own right.

(7)[7]

Geoffrey:	we spent it walking.
	Ten weeks.
DV:	**Ten weeks.**
	And where did you stay?

The fragment in the CR need not be identical to the fragment in the source utterance, as long as it is clarifying it – see example (8).

(8)[8]

Catherine:	And Sharon.
Unknown:	Oh.
Unknown:	**His girlfriend?**
Catherine:	Yes.

A similar form was also identified in which the bare fragment is preceded

[5]BNC file KST, sentences 455–457
[6]BNC file KPP, sentences 352–354
[7]BNC file KRG, sentencens 1361–1364
[8]BNC file KP5, sentences 640–643

by a wh-question word:

(9)[9]

Ben:	No, ever, everything we say she laughs at.
Frances:	**Who Emma?**
Ben:	Oh yeah.

As these examples appeared to be interchangeable with the plain fragment alternative (in example (9), *"Emma?"*), they were not distinguished from fragments in our classification scheme.

2.4 Wh-Substituted Reprise Sentences

Like the literal reprise, this form involves echoing a previous utterance in full, but the specific part being clarified is substituted for a wh-phrase (e.g. *what*, *who*, *which X*), as illustrated by example (10).

(10)[10]

Unknown:	He's anal retentive, that's what it is.
Kath:	**He's what?**
Unknown:	Anal retentive.

2.5 Reprise Sluices

This form is an elliptical version of the wh-substituted reprise sentence, where only the wh-phrase is used:

(11)[11]

Sarah:	Leon, Leon, sorry she's taken.
Leon:	**Who?**
Sarah:	Cath Long, she's spoken for.

There may be a continuum of forms between *wh-substituted reprise sentences* and *reprise sluices*. Consider the following exchange (12):

(12)[12]

Richard:	I'm opening my own business so I need a lot of money
Anon 5:	**Opening what?**

This form seems to fall between the full wh-substituted reprise sentence

[9]BNC file KSW, sentences 698–700
[10]BNC file KPH, sentences 412–414
[11]BNC file KPL, sentences 347–349
[12]BNC file KSV, sentences 363–364

"You're opening (your own) what?" and the simple reprise sluice *"(Your own) what?"*. The actual form employed in this case appears closer to the sluice and was classified as such.

2.6 Gaps

The *gap* form differs from the reprise forms described above in that it does not involve a reprise component *corresponding to* the component being clarified. Instead, it consists of a reprise of (a part of) the utterance *immediately preceding* this component – see example (13).

(13)[13]

Laura:	Can I have some toast please?
Jan:	**Some?**
Laura:	Toast

2.7 Gap Fillers

The *filler* form is used by a speaker to fill a gap left by a previous incomplete utterance. Its use therefore appears to be restricted to such contexts, either because a previous speaker has left an utterance "hanging" (as in example (14)) or because the CR initiator interrupts.

(14)[14]

Sandy:	if, if you try and do enchiladas or
Katriane:	Mhm.
Sandy:	erm
Katriane:	**Tacos?**
Sandy:	tacos.

2.8 Conventional

A *conventional* form is available which appears to indicate a complete breakdown in communication. This takes a number of seemingly conventionalised

[13]BNC file KD7, sentences 392–394

[14]BNC file KPJ, sentences 555–559

forms such as *"What?"*, *"Pardon?"*, *"Sorry?"*, *"Eh?"*:

(15)[15]

Anon 2:	Gone to the cinema tonight or summat.
Kitty:	**Eh?**
Anon 2:	Gone to the cinema

3 Clarification Readings

This section presents the CR readings that have been identified, together with examples.

3.1 Clausal

The *clausal* reading takes as the basis for its content the *content of the conversational move* made by the utterance being clarified.

This reading corresponds roughly to *"Are you asking/asserting that X?"*, or *"For which X are you asking/asserting that X?"*. It follows that the source utterance must have been partially grounded by the CR initiator, at least to the extent of understanding the move being made.

(16)[16]

Clare:	Right, hold on a moment please `<pause>` Sarah `[last name]` for you.
Gary:	**For me?**
Clare:	Yes.

(17)[17]

Peggy:	He would bet it.
Arthur:	**Who?**
June:	My dad.

3.2 Constituent

Another possible reading is a *constituent* reading whereby the content of *a constituent* of the previous utterance is being clarified.

[15]BNC file KPK, sentences 580–582
[16]BNC file KSR, sentences 427–429
[17]BNC file KSS, sentences 681–683

This reading corresponds roughly to *"What/who is X?"* or *"What/who do you mean by X?"*.

(18)[18]

Frances:	She likes boys called `[...]` Bill Leigh `[last name]`, B J.
Ben:	**B J.**
Frances:	She, she's writing a note
Ben:	**B J?**
Frances:	you know Ash, B J

3.3 Lexical

Another possibility appears to be a *lexical* reading. This is closely related to the clausal reading, but is distinguished from it in that the *surface form* of the utterance is being clarified, rather than the content of the conversational move.

This reading therefore takes the form *"Did you utter X?"* or *"What did you utter?"*. The CR initiator is attempting to identify or confirm a word in the source utterance, rather than a part of the semantic content of the utterance.

This reading is usually conveyed by conventional forms (see example (15)).

3.4 Corrections

The correction reading appears be along the lines of *"Did you intend to utter X (instead of Y)?"*.

(19)[19]

Grace:	Yeah.
	`<laugh>` Fifteen for the first time round.
Anon 3:	**Third.**
Grace:	Third time round.

4 Markup Scheme

This section describes the markup scheme and tagging process.

[18]BNC file KSW, sentences 611–615

[19]BNC file KPE, sentences 327–330

A multi-layered approach was taken, along the lines of the DAMSL dialogue act markup scheme (Allen and Core, 1997) – this allowed sentences to be marked independently for three attributes: *form*, *reading* and *source*.

The *form* and *reading* attributes have finite sets of possible values. The possible values are as described in sections 2 and 3, plus an extra catch-all category *other* to deal with any otherwise uncategorisable phenomena.

The *source* attribute can take any numerical value and is used to specify the number of the sentence being clarified (according to the BNC sentence-numbering scheme).

4.1 Decision Process

Following the methods described by Allen and Core (1997), binary decision trees were designed to guide the classification process. The trees are designed so that a naive user can follow them. Trees are available for initial identification of a CR, for classification of CR form and for determination of CR source: they are given in the appendix section A.

4.1.1 Ambiguity of Reading

In the (common) case of ambiguity of reading, the response(s) of other dialogue participants were examined to determine which reading was chosen by them. The ensuing reaction of the CR initiator was then used to judge whether this interpretation was acceptable. If the CR initiator gave no reaction, the reading was assumed to have been acceptable. The following example (20) shows a case where the other participant's initial (clausal) reading was incorrect (the initiator is not satisfied), as a constituent reading was required. In such cases, both CRs were marked as constituent.

(20)[20]

George:	you always had er er say every foot he had with a piece of spunyarn in the wire
Anon 1:	**Spunyarn?**
George:	Spunyarn, yes
Anon 1:	**What's spunyarn?**
George:	Well that's like er tarred rope

In example (21), however, the other participant's clausal interpretation

[20]BNC file H5G, sentences 193–196

provokes no further reaction from the CR initiator, and is taken to be correct:

(21)[21]

Anon 1:	you see the behind of Taz
Selassie:	**Tazmania?**
Anon 1:	Yeah.
Selassie:	Oh this is so rubbish man.

To ensure that this process is used correctly, 10 turns before and after the sentence being tagged must be examined before the tagging decision is made. In order to facilitate this process in the case of CRs near the beginning or end of the 200-turn section being marked, an additional 10 turns of backward and forward context were shown (but not themselves marked up).

4.1.2 Ambiguity of Source

In the case of ambiguity as to which sentence was being clarified, the most recent one was taken as the source.

The BNC sentence numbering scheme does not assign numbers to sentences containing no transcribed words. Such sentences are common where recording quality was poor or the environment was noisy – these sentences are marked in the BNC as `<unclear>` and given no number. Of course, these sentences are often unclear to other conversational participants, and so often cause CRs (usually with a lexical reading). In these cases, sentence numbers were assigned during tagging. Non-integer numbers were used, with values chosen to be consistent with the BNC numbering of surrounding sentences. For example, in example (22), the unclear sentence was given the number 589.1, and the source of the CR in sentence 590 was tagged with this number.

(22)[22]

Peter: `<589>`	But he couldn't work out why I was in school?
Muhammad:	`<unclear>`
Peter: `<590>`	**What?**

4.2 Markup Details

Details of the markup tag syntax are shown below, together with an example of a marked-up CR in the SGML-format used in the corpus.

[21]BNC file KNV, sentences 548–551
[22]BNC file KPT, sentences 589–590

Attribute	Possible Values	
rform	non	Non-Reprise
	lit	Literal Reprise
	sub	Wh-Substituted Reprise
	slu	Reprise Sluice
	frg	Reprise Fragment
	gap	Gap
	fil	Gap Filler
	oth	Other
rread	cla	Clausal
	con	Constituent
	lex	Lexical
	cor	Correction
	oth	Other
rsource	-	(any sentence number)

Table 1: Clarification Request Markup Scheme

```
<u who=PS1BY>
<s n="363">
<w PNP>I<w VBB>'m <w VVG>opening <w DPS>my <w DT0>own <w NN1>business
<w AV0>so <w PNP>I <w VVB>need <w AT0>a <w NN1>lot <w PRF>of <w NN1>money
</u>
<u who=PS1K6>
<s n="364" rform="slu" rread="lex" rsource="363">
<w NN1-VVG>Opening <w DTQ>what<c PUN>?
</u>
```

Figure 1: Example CR (12) after markup

```
<!ATTLIST s
       id ID #IMPLIED
       n CDATA #IMPLIED
       p (Y | N) "N"
       rform (non | lit | sub | slu | frg | gap | fil | wot | oth) #IMPLIED
       rread (cla | con | lex | cor | oth) #IMPLIED
       rsource CDATA #IMPLIED
       TEIform CDATA "s" >
```

Figure 2: Excerpt from updated BNC Document Type Definition

A Decision Tree Details

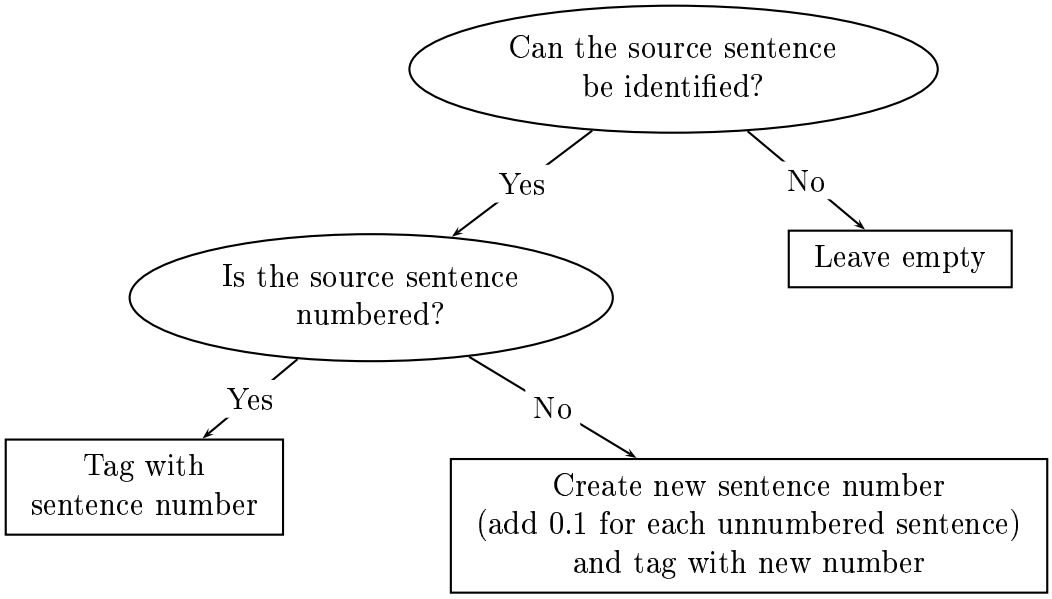

Figure 3: Decision Tree: CR Source

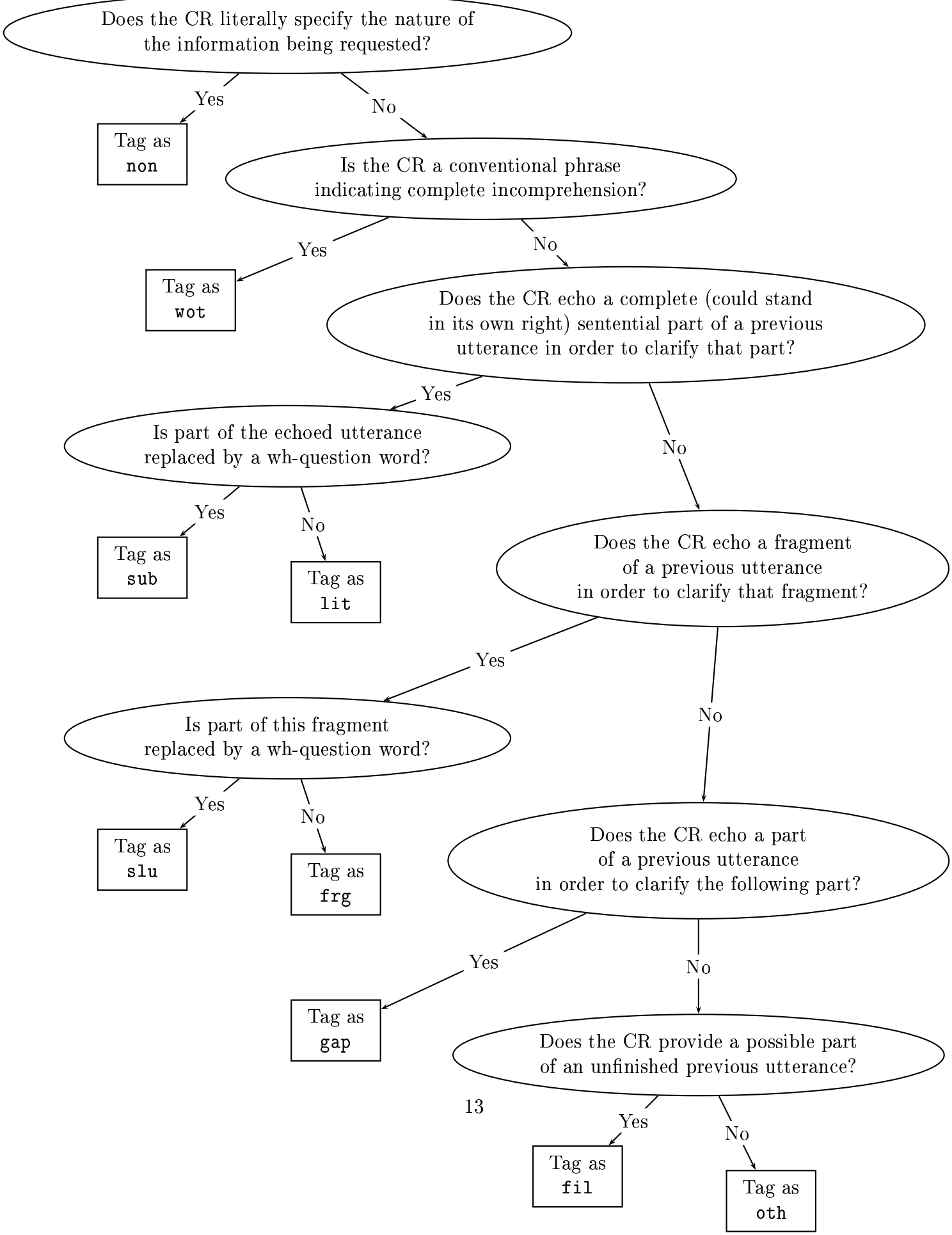

Figure 4: Decision Tree: CR Form

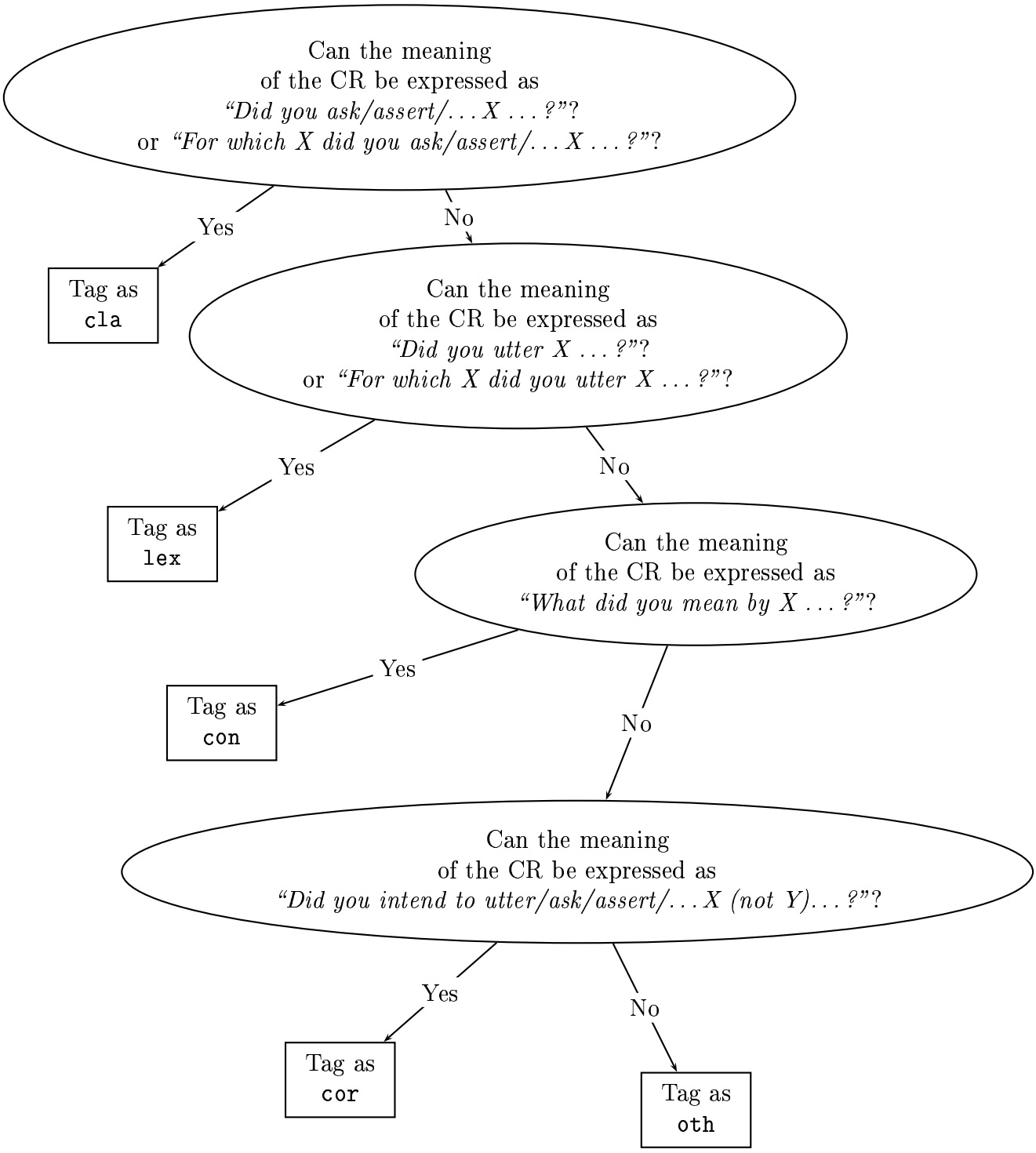

Figure 5: Decision Tree: CR Reading

References

James Allen and Mark Core. Draft of DAMSL: Dialog act markup in several layers, 1997. URL `http://www.cs.rochester.edu/research/cisd/resources/damsl/RevisedManual/`.

Lou Burnard. *Reference Guide for the British National Corpus (World Edition)*. Oxford University Computing Services, 2000. URL `ftp://sable.ox.ac.uk/pub/ota/BNC/`.

Jean Carletta. Assessing agreement on classification tasks: the kappa statistic. *Computational Linguistics*, 22(2):249–255, 1996.

Matthew Purver. SCoRE: A tool for searching the BNC. Technical Report TR-01-07, Department of Computer Science, King's College London, October 2001. URL `ftp://ftp.dcs.kcl.ac.uk/pub/tech-reports/tr01-07.ps.gz`.

Matthew Purver, Jonathan Ginzburg, and Patrick Healey. On the means for clarification in dialogue. In *Proceedings of the 2nd ACL SIGdial Workshop on Discourse and Dialogue*, pages 116–125. Association for Computational Linguistics, September 2001.

Matthew Purver, Jonathan Ginzburg, and Patrick Healey. On the means for clarification in dialogue. In R. Smith and J. van Kuppevelt, editors, *Current and New Directions in Discourse & Dialogue*. Kluwer Academic Publishers, 2002.