Annotation Scheme for Discourse Relations in *Paraphrase* Corpus

Sasha Calhoun

June 5, 2005

1 Overview

The aim of this work is to build on the annotation of information structure in the Switchboard dialogues done as part of the *Paraphrase* Link project (Nissim, Dingare, Carletta & Steedman 2004). The *Switchboard* corpus is a collection of spontaneous telephone conversations between speakers of American English on pre-arranged topics. While the *Paraphrase* project was entirely text-based, in this project we are interested in the relationship between information structure and prosody. In the *Paraphrase* project, NPs were annotated according to one dimension of information structure, *information status*: NPs were assigned the categories of *old*, *mediated* or *new*. Here, we want to annotate discourse relations that speakers use to mark certain entities as salient; where that salience is based on the entity being differentiated from other related entities suggested by the discourse. Individual words will be marked according to one of five discourse relations that trigger this salience, or *background* if it is not triggered.

2 Starting the Tool

Go to the directory KontrastTool (from your home directory type `cd KontrastTool`). Type `sh kontrastcoder.sh`. The tool and a window with dialogue numbers will pop up. Click on the desired dialogue.

3 Markables

The set of markable words have been obtained by selecting all nouns, verbs, adjectives, adverbs and demonstrative pronouns (i.e. words with the POS tags CD, JJ, JJR, JJS, MD, NN, NNS, NNP, NNPS, PRP, PRP$, RB, RBR, RBS, VB, VBD, VBG, VBN, VBP, VBZ, WP, WRB, and DT where the word is ‘this’, ‘that’, ‘these’ or ‘those’) in full sentences.
The tool should start by highlighting the first word to be annotated, and then jump automatically to the next markable word once the current one has been marked. In addition you can use the forward and back buttons to move between markable words.

You will see that NPs that contain more than one markable word are marked at the level of the individual words as well as the whole NP. This is because in some cases the discourse relations apply more naturally to entire NPs, not just individual words. For instance, (1) is contrastive between the words movie and book, not good movie and book (see below for category definitions).

(NB - All examples are from the Switchboard corpus. Parentheses indicate prosodic boundaries, words in SMALL CAPS are pitch accented and words in bold indicate the markables, and italics the triggers, being discussed.)

(1) ( TWO thousand and ONE was a good MOVIE ) (IF you had read the BOOK)

whereas in (2) the NP purple buddy, and not purple and/or buddy separately most naturally is an instance of an answer. In cases like this mark the appropriate category when the NP is highlighted, not the individual word. Either leave the individual word unmarked or annotate it status-uncoded. Note that to highlight the NP and not individual words you need to click on the bracket marking the beginning of the NP.

(2) (I’ll TELL you another PLANT) (that’s PURPLY) ... (it’s REALLY a PRETTY PLANT) (it’s SOLD down here as) (PURPLE BUDDY)

If you are unsure, listen carefully to see whether you think the speaker intends to pick out one word or include the whole NP in the discourse relation. If the discourse relation appears to apply to a whole syntactic unit other than an NP, e.g. VP or AdvP, only mark the word or words you think the speaker makes central to the discourse relation.

4 Exclusions

In some rare cases you may wish to exclude certain markables:

4.1 Comprehension

If you cannot fully understand the dialogue after listening and reading the transcript (this should be very rare), it should not be annotated. Mark each word in the section you cannot understand as noncomp.
4.2 Non Applicability

If you think the markable word is being used in an entirely idiomatic or formulaic way, it should not be included. Or if a word sounds like a hesitation or a false start, so that it’s unclear how it was intended to be used, it should not be included. In each case, mark the affected word as nonapplic, e.g.

- “in [fact]”
- “of [course]”
- “[you] [know] [like]”
- “[it] [is]- [it] [has]- [I] [mean]...”

**Idiomatic Phrases** Both the pronoun and verb in phrases like “I mean”, “you know” and the like should be considered as non-applicable. However this category should be used sparingly, i.e. only for highly formulaic usages where the words themselves offer very little clue to the meaning.

**False Starts and Hesitations** If a speaker starts a clause, and then interrupts themselves and starts again, the point up to the beginning of the repair should be marked nonapplic, e.g. in “i- i- it- it is a purply blue” i- i- it should be marked nonapplic.

5 Trigger Links

In all categories except for answer, other and background, the word or NP that motivated the category assignment needs to be marked. To create a new link select the markable and click on the add new link button. Then highlight the trigger. If the trigger is not a markable word (e.g. because it is in a short sentence fragment), do not mark the link but type ‘non-markable trigger’ and then the trigger word in the Notes field.

If you want to delete a link, go into the links window, select the little folder relative to the link you want to delete, and click on the delete link button.

6 Discourse Relations Marking

We wish to identify cases where speakers make a word or NP salient with an implication that this salience is in comparison or contrast to other related words or NPs explicitly or implicitly evoked in the context. We define several categories of discourse relations which bring about this salience.
For each selected word that has not been marked noncomp or nonapplic, you should decide whether you think it fits into one of these categories by listening to and reading enough of the surrounding context to be able to understand the sentence within the context of the conversation. Mark the word as the first category that applies. If you think a word or NP does convey this differentiated salience, but it does not fit into any of the given categories, mark it as other (use this sparingly). The five categories of discourse relations, plus other, form a complement to the category background, which should be used when no differentiated salience is induced. Note that a word can be both a trigger for another category, and be marked in its own right (see extract at the end for examples). If you cannot decide on a category, then you can mark the word status-uncoded, this should be used very rarely, however.

correction The speaker’s intent was to correct or clarify another word or NP just used by them or the other speaker. Mark the correction or clarification as correction and the word or NP being corrected as the trigger. For example, in (3) the speaker wishes to clarify whether her interlocutor really meant “hyacinths” as opposed to any other bulbs.

(3) (A)... it was a hyacinth have you ever seen those? Oh they are pretty in the Spring but the leaves i do not like them, you know once...
(B)(now are you sure they’re HYACINTHS) (because that is a BULB)

contrastive The speaker intends to contrast the word with a previous one which was (a) a current topic; (b) semantically related to the contrastive word, such that you can naturally think of a plausible set to which they would both belong (you do not need to identify this set). The set may be fairly abstract but the intended contrast must be clear. In (4), B contrasts recycling in her town “San Antonio”, with A’s town “Garland”, from the set places where the speakers live. In cases like this mark San Antonio as contrastive and then link to the trigger Garland.

(4) (A) I live in Garland, and we’re just beginning to build a real big recycling center that recycles everything imaginable...
(B) (YEAH there’s been) (NO emphasis on recycling at ALL) (in San ANTONIO)

In cases like (5) the speaker highlights both the trigger and the referent in making the contrast. In such cases, where both trigger and referent are said by the same speaker and the trigger is highlighted, mark both words as contrastive, and create a link with the second word the referent and the first the trigger.

(5) (I have got SOME in the BACKYARD that) ( bloomed BLUE) ( Which I WOULD have liked those in the FRONT) (because they match my PORCH)

subset The speaker highlights one word which is (a) a current topic, (b) a member of a more general set that is mentioned in the surrounding context. Again, the set may be fairly abstract but
the intended set-subset relationship must be clear. In (6), the speaker introduces the general set “three day cares”, and then gives a fact about each. Mark two in Lewisville, one in Irving and the second one and then three day cares as the trigger of each.

(6) (THIS woman owns THREE day cares) (TWO in Lewisville) (and ONE in Irving) (and she had to open the SECOND one up) in Lewisville (because her WAITING list was) just like you like (a YEAR old)

As in the contrastive case, if both the trigger and the referent are said by the same speaker and the trigger’s role as a superset is highlighted, mark both trigger and referent as subset, but only make one link between them. If the trigger is not highlighted or is said by the other speaker, do not mark it as subset.

**adverbial** The speaker uses a focus-sensitive adverb, e.g. only, even, always, especially, just, also and too to highlight that word, and not another in a plausible set that you can think of. The set can be fairly abstract but the intended invocation of a set of alternatives must be clear, i.e. you do not need to mark every use of one of these adverbs. Mark the focussed word as adverbial and the adverb as the trigger. Again, if the adverb itself is highlighted, mark it as adverbial also. In (7), B didn’t even like the “previews” of ‘The Hard Way’, let alone the movie.

(7) (A) I like Michael J Fox, though I thought he was crummy in ‘The Hard Way’.
(B) (I didn’t even like) (the PREVIEWS on that)

**answer** The word or NP, and no other, fills an open proposition set up in the context by either speaker. It must be so that it would have made sense if they had only said that word or phrase. In (8), A sets up the “bloom” she can’t identify, and B answers “lily”. You do not need to mark the trigger.

(8) (A) Well everybody down here calls these flags... they get just one bloom... I’m not sure what they are called but ... they come in all different colours the blooms are on some of them is yellow, purple, white just all different colours
(B) (I’m going to BET you) (that is a LILY)

**other** The speaker highlights a word or NP, implying that it is salient in opposition to related words or concepts explicitly or implicitly evoked by the context. This category should be used sparingly when the criteria for the previous types of discourse relations cannot be met and the category background is not applicable. For example, in (9) the speaker clearly wishes to highlight that it was Christmas Eve, and not any other day, that they’d forgotten, so it is not background. However, it is not contrastive, as there is no explicit trigger, nor is there an explicit superset of all days, so it is not a subset. There is no focus-sensitive adverb for adverbial, nor an open proposition set up for answer.
(9) (When I was a little KID) (I saw ‘the INCREDIBLE JOURNEY’) (on CHRISTMAS EVE) (and it was SO GOOD) (that I had FORGOTTEN) (it was CHRISTMAS EVE)

background The category of background is best described as the opposite of the six categories above. It is used for the words that the speaker either does not wish to make salient, or does make salient, but there is no implication of related alternatives. This could be because the word relates back to what has just been said, has already been mentioned. For example, in [9], “it was” in “it was so good” relates back to “the Incredible Journey”, and “I had” to “when I was a little kid”, a clue that they are background. Or it could be that the speaker is introducing a completely new proposition, and so there is no implication of contextual alternatives, e.g. in [10], “rats in the attic” is highlighted, but the speaker doesn’t wish to differentiate them from other things in the attic, or other places rats might be, etc.

(10) (I was living ALONE) (at the TIME) (and it was LATE at NIGHT and) (SCARY and) (you start HEARING NOISES) (and there’s RATS in the ATTIC)

You should expect that almost every sentence will contain at least one word which is background, as well as some sentences which are entirely background. In all cases, mark background as a property of words, not NPs.

References


7 Example - Selected Sentences from sw2969

Below are selected example sentences. In the actual annotation task, the tool described above would enable transcribers to use as much of the context as they need to do the annotation, here I’ve given just enough to get the sense. If the trigger is in the context, this is marked, though the rest of the context is not annotated.

Words in square brackets are markable. If no category is given, it is background. If more than one word is in the brackets, this shows they should be grouped. Numbers at the beginning of each sentence in brackets refer to the approximate start time of the utterance in the sound file of the whole conversation, the number at the end is the msstate ID.

1. Context: (Talking about gardening)
   We bought a a large house that was completely unlandscaped. I mean it, it was probably-
probably was up to your shoulder in in thistles, that was all that was there and... (38.19) and [we] [hired] a [professional landscaper] other to [do] [it] (sw2969B-ms98-a-0012)

2. **Context:** (Speaker talking about being frustrated how much work her garden is)
Speaker: ...rip up rip out all the rest of these junipers or anything that requires uh more care than I am willing to give the kids and replace it with hollies.
Reply: uh how many kids do you have? (70.96) [just] adverbial/trigger [two] adverbial/contrast/trigger but [they] [are] [eighteen months] other [apart] [so] [it] [seems] like a [lot more] contrast (sw2969B-ms98-a-0021)

3. **Context:** (speaker is talking about something else (not gardening)
yeah it is over with... but back to gardening (99.73) [are] [you] subset [one] of [these] uh [Howard Garrett organic enthusiasts] contrast/trigger/subset/trigger or a [Neil Sperry nuke the lawn with chemicals] contrast/subset/trigger (sw2969B-ms98-a-0032)

4. **Context:** (The other woman is talking about flowers around where she is)
Reply: it was a [hyacinth] correction/trigger have you ever seen those
Speaker: oh yeah
Reply: oh they are pretty in the Spring but the leaves i do not like them (161.79) [all right] nonapplic now [are] [you] [sure] [they] [are] [hyacinths] correction because [that] [is] a [bulb] other (sw2969B-ms98-a-0047)

5. **Context:** (the speaker explains the problems she’s had with hyacinths)
Well let me tell you they do not come back very well at all. Maybe they will if up further north but down... and you are not that much further north. (191.22) [I] [have] [probably] [planted] oh around [here] um [maybe] [six] [hundred] [bulbs] (sw2969B-ms98-a-0053)

6. **Context:** (talking about her experience growing hyacinths compared to the other speaker, whose location was explicitly mentioned much further back)
And if they do come back next year they are going to look real puny (215.74) [about] the [only] adverbial/trigger [bloom bulb] adverbial/subject/trigger that [repeats] other [well] other in [my area] contrast [is] the uh uh [daffodil] subset and some of the [narcissus bulb] subset (w2969B-ms98-a-0059)

7. **Context:** (the speaker is trying to help the other woman to identify a type of flowering plant, which she thinks might be an iris.
Spkr: and what color?
Reply: they get uh just [one bloom] correction/trigger on the plant i do not i am not sure what they are called but it has got g[ay]- oh they come in all different colors the blooms are on some of them is yellow purple white just all different colors (260.27) and [you] [only] [get] [one] correction of [them] (sw2969B-ms98-a-0067)

8. **Context:** (the speaker gives her qualifications for guessing what the other woman’s plant is)
You know what I bet you they are... and I am a pretty good gardener because... I... (283.93) [I] [work] on a uh [landscape] [committee] for [our] [neighbourhood] (sw2969B-ms98-a-0070)

9. **Context:** (talking about a plant the other speaker can’t identify)
Spkr: and what does the foliage look like does it...? Reply: green spiky leaves now it looks almost like a hyacinth only a a lot higher
(331.59) oh oh oh oh [i] [bet] [you] [those] [are] [i] [bet] [you] what [those] [things] [are] um [is] a [Dutch iris]answer (sw2969B-ms98-a-0080)

10. Context: (The other woman asks the speaker what flowers she should put in her garden)

Speaker: Ooh, blue is hard you know... if you want do you want a perennial flower that will bloom all summer?

Reply: I bought well... I would like to have one that would bloom next year too that will come back up...

(484.69) okay [put] in a [perennial blue salvia]answer (sw2969B-ms98-a-0118)