Lexical Noun Phrases in Dialog

Prajnya Satish

LING 5430: Semantics and Pragmatics

May 8, 2018

1 Introduction

In dialog, the use of pronouns to refer to subjects is more common than the use of lexical noun phrases. Lexical NPs are mainly used to introduce the subjects as the topic under discussion as in the case of sentence focus arguments. Participants rarely use the lexical noun phrase to access the referent once the referent has been evoked. There are cases however when an already introduced subject is referred as a lexical noun phrase rather than as the pronoun even though the sentence structure does not require it.

In this paper, I plan to study the reasons for participants in dialog to switch to lexical noun phrases for pronominal phrases. I also intend to find if there is any universality in this case and try to predict if there is a minimum number of turns before a subject is reintroduced.

2 Method

2.1 Data

The switchboard corpus was used for this task. The corpus is a collection of telephone conversations between two anonymous callers (A and B). The corpus is divided into a single transcript for each dialog between caller A and B. There are a total of 2866 such transcripts. The callers discuss a wide range of topics. This is a rich annotated dataset with several features such as part of speech tags and tree structure for each utterance and information about the caller such as their location, age, gender, dialect and also the prompt given to the speaker.

The features that were particularly useful for this project were the part of speech tags and the parse trees. The lexical noun phrases (proper nouns) are tagged with an NNP tag according to the Penn Treebank P.O.S tags.

2.2 Procedure

1. The lexical noun phrases were identified using the NNP tag on the corresponding to the word.
2. The line number or the utterance in which the lexical noun phrase occurred was recorded.
3. If there were multiple occurrences of the same NP in the transcript, the distance between the successive occurrences was computed.
4. A histogram plot of the distance between successive occurrences and the frequency of such distances was computed.
5. Following observations from the histogram, the transcripts were viewed to investigate the distances and the possible reasons for such distance.

2.3 Results

The histogram plot of the distance between successive occurrence of lexical NPs is shown in Figure 1. The plot has the distance between two utterances on the x-axis and the number of ties this distance occurs on the y-axis.

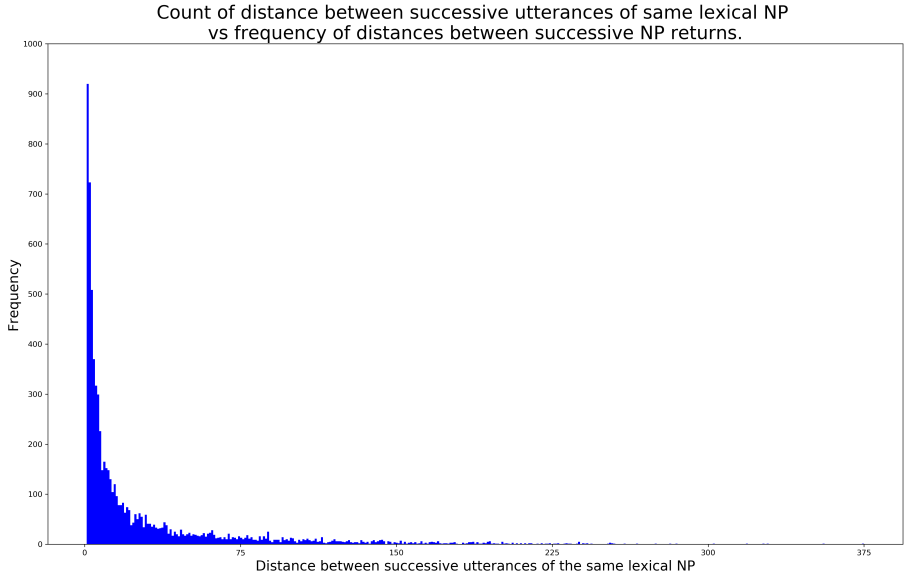

Figure 1: Histogram plot of the distance between successive occurrence of NPs.

From Figure 1, it is clear that a majority of lexical NPs occur within the first three turns of the dialog. In fact, about 13.06% of the re-occurrences are just within the first turn. The $50th$ *percentile* of the distances is 9, which shows that most NPs occur within nine turns of their last utterance.

Stats	Values
Mean	26.58
Median	9.0
Std Dev	42.19
50th %ile	9

Figure 2: Stats derived from the histogram.

The mean distance between the utterances is 26.58. There are some transcripts for dialogs with more than 300 turns. This shows that distance between the re-occurrence is pretty low even for long conversations.

Because of the nature of the data, I will first discuss the lexical NPs that occur within 9 utterances of each other and then discuss the NPs for all other utterance distances.

3 Discussion

Distance 1: For words with re-occurrence distance equal to 1, a few rules can be applied for the lexical NP to reoccur:

1. In most cases, the first utterance is part of a question or statement and the second utterance is an affirmation of what is heard by the other speaker. It is mostly just the single NP in the second utterance. An example is shown below in Figure 3 . This was found to be the case in about 30% of the cases.

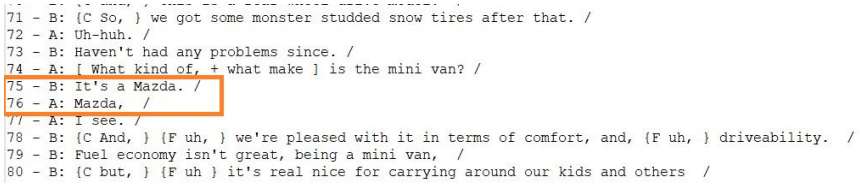

Figure 3: Example with distance = 1 between NPs.

2. Another frequent occurrence is when the current speaker continues on the the next turn too and repeats certain portions of the previous utterance.

3. **Disagreements:** When the responder disagrees with the initiator, the responder often uses the full lexical NP in the subject. For example in Figure 4. . If the NP is longer than a

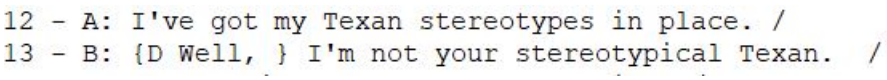

Figure 4: Distance = 1 between NPs and disagreement from the responder.

single word, sometimes only one NP word in the whole phrase is repeated in the succeeding utterance.

4. **Assessments:** A full NP is used when the utterance in the second part is an assessment. This is more frequent in cases of negative assessments. For example,

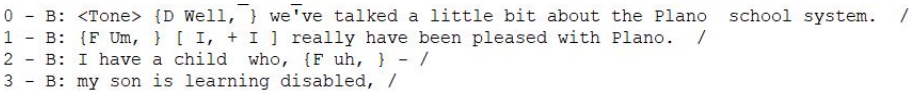

Figure 5: Distance = 1 between NPs and assessment in the utterance.

5. A full NP is also used when a category label can be used to refer to the pronoun. In this dataset, teams are referred to with their full NPs rather than pronouns. It is also the case with names of cities where the name is used to refer to the frame, such as when the name of the city is used to refer to the people in the city, the weather, the culture, etc.

Distances 2 - 9: If the current speaker does not select the next speaker and another speaker does not self-apply, then the current speaker can take the turn[3]. With this rule applied, the current speaker tends to take up the next 2 or 3 turns. In these cases, the current speaker may bring up the same lexical NP in the new utterance too. This is the only case that differs from the above set of rules for distance 1 samples.

Distances greater than 9: The five rules listed above can be consistently applied to transcripts with distances greater than 9, if the sequence has not been closed yet. A sequence is understood to be closed when utterances such as OK, bye, etc are used. There are two situations that cause a full NP to be reused as the subject:

1. **Return Pop** - When an adjacency pair ties to a pair other than the immediately preceding one, the structure is called a return pop. [1](Fox 1987, p. 27) When a full NP is used in a return pop structure, it means that the previous NP is "popped over" and closed off. The current NP is used as the referent in all further utterances. Figure 6 is an example of this. In this dialog, speaker A talks about moving to Texas in utterance 4. It is only in utterance 68, that speaker B returns back to the announcement made and further inquires about it. It is a classic case of return pop where the sequence is closed off and popped over at 68 and a new sequence is opened. Figure 7 is another case of return pop which occurs after 260 lines. Clearly, there is no limit to the minimum number of utterances between the two lexical NPs in this case.

2. Pronouns can be easily used if the lexical or grammatical structure of the sentence allow the listener to resolve the anaphora. If not, the speaker tends to repeat part of the lexicon

```
15 - A: in New York we had to pay  a state and a city, {F uh, } tax.  /
16 - A: {C And } it was really a great relief when I got to Texas  that I
didn't have to pay that city income tax, though. /
17 - B: Yeah,  /
18 - B: makes quite a bit of difference in [ your, + {F uh, } {D well } your ]
income level. /
19 - A: Right. /
20 - B: {D You know, } {C because } other states quite a bi-,  a chunk out
for, {F uh, } city  and  state taxes  /
21 - B: {C so } you're, {F uh, } looking at your yearly salary, it makes quite
a difference. /

(45 lines ...)


66 - B: that's true. /
67 - B: Although [ I wi-, + ]  I wish, {D you know, } overall they'd try the
lottery again.  /
68 - B: When did you first come to Texas?  <Dishes>. /
```

Figure 6: Return Pop

```
8 - A: What area do you live in? /
9 - B: I live in Houston. /
10 - A: {F Oh, } okay.  /


(261 lines ...)


271 - A: that's true  /
272 - A: Houston, in itself, is a pretty tough place. /
```

Figure 7: Return Pop

to make it easy for the listener to resolve the pronoun. If both of these are not possible, then the speaker resort to using a full NP to refer to the subject. In Figure 8, for example, the topic of the sequence is North Carolina throughout. But it is not possible to have a lexico-grammatical structure that will allow North Carolina to be pronominalized. So, it has to be referred to as a full NP. It must also be noted that this example is not a return pop because a new sequence is not started at line 78. There is no indication of a sequence close off. .

```
21 - B: {D Well, } there was an article in our paper  Sunday that the average
person sentenced to seven to ten years in North Carolina,  [ the average + ]
serves six weeks <laughter>. /
22 - A: {D Now, } {D see, } [ that's a, +  {A you talk about, } that's a ]
crime right there. /
23 - B: {D Well, } {D see, } we don't have any  room in our prisons  /

(42 lines ...)


75 - A: {C and } you spend  [ thousands, +  hundreds of thousands ] of
dollars to, {F uh, } take care of them  and all, {D you know, }  /
76 - A: it's just unbelievable. /
77 - B: Yeah,  /
78 - B: you can appeal now in North Carolina  for almost a period of twenty
years. /
```

Figure 8: Full NP without return pop.

3. The last case is when a sequence is completely closed off by both the speakers and the subject is reintroduced into the consciousness of the listener. Consider Figure 9.. Here, a lexical NP can be reintroduced after any duration of time and number of turns. This would be like the first mention of the NP.

```
0 - A: Okay,  /
1 - A: yeah,  /
2 - A: Donna, {F um, } you said you have three children.  /
3 - A: What, {F uh, } kinds of activities do you spend with them?   /
4 - A: Is it  ((   )) ? /
5 - B: Okay,  /
6 - B: {F um, } we have an eleven year old, an eight year old, and a three
year old #  /
7 - B: {C and, }  {F um, } #
8 - A: # Uh-huh. # /
9 - B: <Sniffing> {F uh, } my husband, {A let's see, }  he coaches basketball
with Brian, {F um, }  /
10 - B: {C so } he spends time doing that with him, planning and, {F um, }
practicing outside. /
11 - A: Uh-huh. /
12 - B: {F Um, } he's a real sports kind of person,  /
13 - B: {C so } he plays tennis [ with, + with ] the kids.  /
14 - B: We all kind of play tennis as a family,  /
15 - B: {C except } the baby --
16 - A: Uh-huh. /

(41 lines ...)


57 - B: They go upstairs at nine,  /
58 - B: {C and, }  {F um, } now, this weekend we'll have a baby-sitter come  /
59 - B: {C and, }  {F uh, } my husband and I will go out to dinner and come
home [ and, + ] {F um, } probably not too late. /
60 - A: (( Uh-huh )) . /
61 - B: Maybe Brian will still be up, and spend some time with him.  /
```

Figure 9: Reintroduction of subject

4 Conclusion

In summary, there are multiple ways in which an already introduced lexical NP reoccurs in dialog. It could be based on structural rules like the return pop or lexico-grammatical rules in long distance utterances or other rules such as disagreements, assessments, affirmations or other such rules.

The cases discussed above are the most frequent and identifiable cases in which a lexical NP is used as referent to a subject. These cases however are not exhaustive. There are several situations where an NP is used by the speaker which I could not identify to fit into any of the categories discussed above nor could I come up with a new category. These utterances are few and sparse. A larger dataset and more time would help in coming up with more such structured categories for the re-occurrence of lexical NP but there will always be cases which cannot be fit into these categories.

No universal minimum distance for the occurrence of successive lexical NPs could be arrived at. it is clear however, that a majority of them occur within a 9 utterance limit.

References

[1] Barbara Fox. *Discourse structure and anaphora : written and conversational English.* Cambridge University Press, Cambridge England New York, 1993.

[2] Dan Jurafsky. *Speech and language processing : an introduction to natural language processing, computational linguistics, and speech recognition.* Pearson Prentice Hall, Upper Saddle River, N.J, 2009.

[3] Stephen Levinson. *Pragmatics.* Cambridge University Press, Cambridge England New York, 1983.

[4] Laura A Michaelis and Hartwell S Francis. Lexical subjects and the conflation strategy.

[4, 2]