ELLIPSIS DETECTION

Team LingoTribe:
Harshita Sharma
Souvik Banerjee

WHAT IS ELLIPSIS?

Omission from a clause of one or more words that are nevertheless understood in the context of the remaining elements.

Examples:
I heard Mary's dog, and you heard Bill's.

John can play the guitar; Mary can too.

WHY IS IT IMPORTANT?

- She won't laugh but he will.

Manual:
vah nahi hansegi par vah -*(inflection is fused in the word)* ☒

Google Translator:
vah nahin hansegee lekin vah ☒

WHY IS IT IMPORTANT?

- John has one hat, and Bill has five.

Manual:

jon ke paas ek topee hai, aur bil ke pass paanch hain. ☑

Google Translator:

jon ke paas ek topee hai, aur bil mein paanch hain. ☒

WHY IS IT IMPORTANT?

- When Susan brings her dog, Sam brings his [dog] too.

Manual:

jab susan apne kutte ko laati hai, Sam bhi uske/apne ko lata hai. ☒

jab susan apna kutta laati hai, Sam bhi apna lata hai.

Google Translator:

jab susaan apane kutte ko laata hai, sam use bhee laata hai. ☒

WHY IS IT IMPORTANT?

- Some school kids like syntax, and some don't.

Manual:

Kuch school ke bachchon ko vakya rachna pasand hai aur kuch ko nahi ☑

Google Translator:

kuchh skoolee bachchon ko vaaky rachana pasand hai, aur kuchh nahin ☒

NOUN PHRASE ELLIPSIS

A mechanism that elides, or appears to elide, part of a noun phrase that can be recovered from context.

Examples:

- I read three chapters from this book and Mary read four [e].
- Some students love physics and some [e] don't.

PROBLEM DESCRIPTION

The problem of NPE detection can be broken down into two parts:

- Dataset preparation
- Ellipsis detection: Trigger and the elided phrase detection

Constraint: Sentence boundary - NPE detection and antecedent detection is done within the sentence boundary.

PREPARATION OF DATA

- No dedicated resources.
- Data collected from:
  - Research papers
  - UD tree bank
  - Linguistics textbooks explaining ellipsis

DETECTION

An instance of NPE contains 2 parts:

1. Licensor: indicate presence of NPE - determiners and modifiers of elided noun.
    a. There are three chairs in the living room and **two** [e] in the hall.
    b. Some students love physics and **some** [e] don't.
2. Antecedent: missing phrase/word

LICENSORS

- can only **belong to certain syntactic categories**. These include:
  - cardinal numbers: *John has one hat, and Bill has five*.
  - ordinal numbers: *Mary got first position and John got second*.
  - plural demonstrative determiners: *Of all the candidates that applied for the job, these got selected.*
  - possessives: That big car standing over there is Joey's.
  - adjectives: *He is the funniest guy here. And also the weirdest*.
  - quantifiers: *Some students love physics and some don't.*
  - interrogative determiners: *I don't know which pages to read and which to ignore.*

NPE DETECTION

STEP 1: Look for noun-modifiers and determiners (from the licensor categories)

STEP 2: Check if they show NPE by applying rules.

RULES

Rule 1: Check for **ordinal numbers**, if ordinal numbers are followed by punctuation or prepositions or has no noun phrase in the next three words, then it is chosen as a licensor. Example - second [position]

Rule 2: Previous rule is applied for **cardinal numbers, demonstrative determiners, possessive pronouns and quantifiers**. Example - four [chapters] , these [candidates], some [students]

RULES

Rule 3: Check for **interrogative determiners**, if they are followed by punctuation or the previous word is a noun modifier and next three words don't add up to a noun phrase, it is chosen. Example - which [pages]

Rule 4: Check for **superlative adjectives**, if they are preceded by a determiner and no noun phrase exists, it is chosen. Example - the weirdest [guy]

RULES

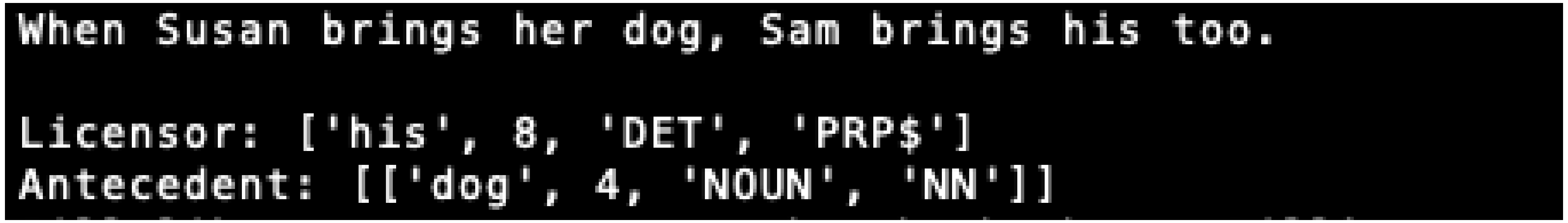
Rule 5: Check if the selected noun modifier is immediately **followed by a verb or auxiliary verb** as that would indicate the end of the given noun phrase immediately after the noun modifier. Ex - Of all the candidates that applied for the job, these got selected.

ANTECEDENT

To resolve NPE, we look for antecedents - Match POS tags of the licensor with other noun modifiers. If a POS tag matching the licensor of the NPE is found in the sentence, the system outputs the noun that the modifier with the same tag modifies as the antecedent of the NPE. If there are more than one such modifiers found, the system selects the one nearest to the NPE as generally has a role to play in anaphora and coreference resolution tasks.

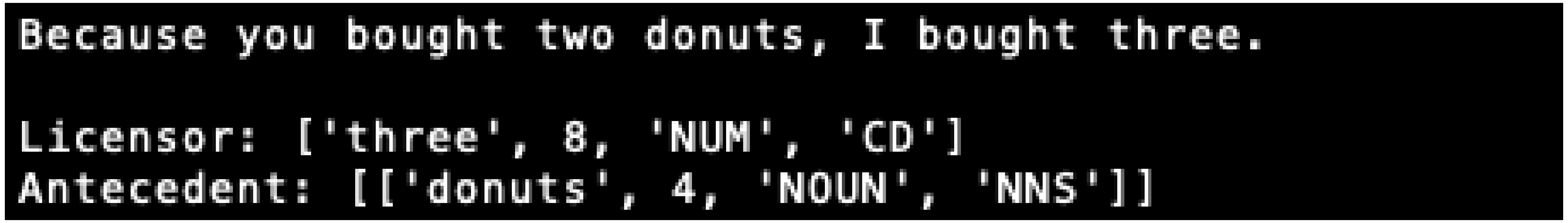
I don't know *which* pages to read and **which** to ignore.

OUTPUT



```
When Susan brings her dog, Sam brings his too.

Licensor: ['his', 8, 'DET', 'PRP$']
Antecedent: [['dog', 4, 'NOUN', 'NN']]
```

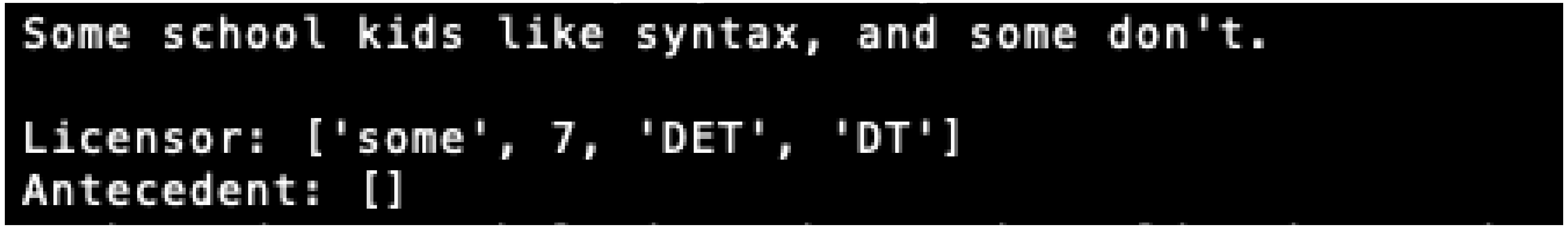


```
Because you bought two donuts, I bought three.

Licensor: ['three', 8, 'NUM', 'CD']
Antecedent: [['donuts', 4, 'NOUN', 'NNS']]
```

ERROR ANALYSIS

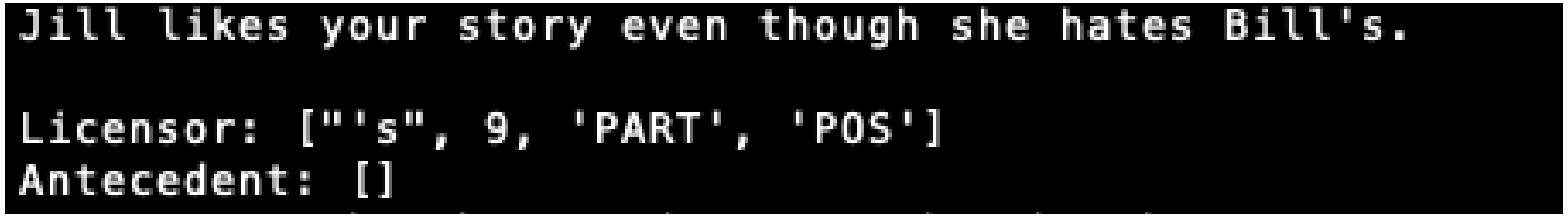
1. One-anaphora errors
2. Parsing fails



```
Some school kids like syntax, and some don't.

Licensor: ['some', 7, 'DET', 'DT']
Antecedent: []
```

3. Possessive marker error

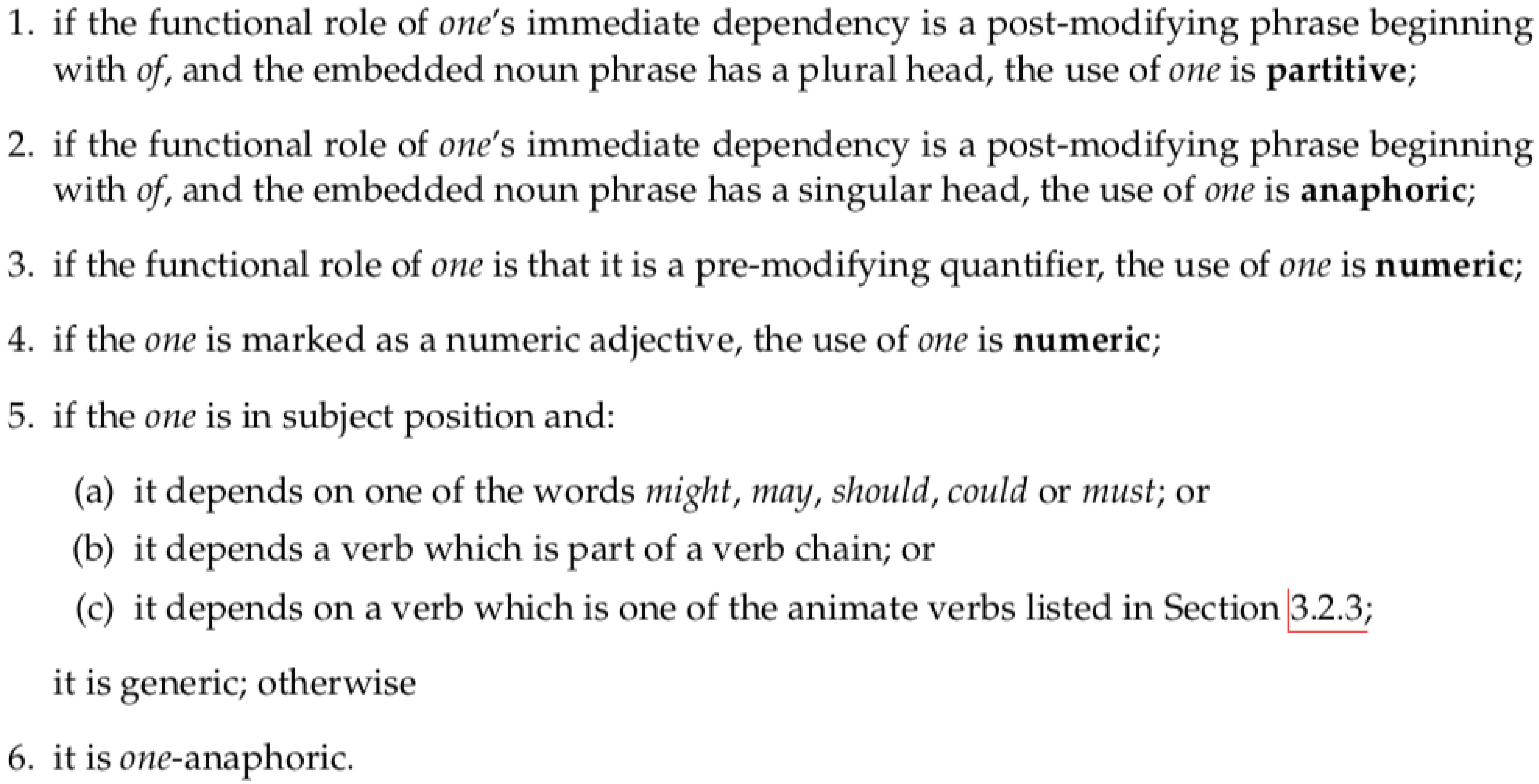


```
Jill likes your story even though she hates Bill's.

Licensor: ["'s", 9, 'PART', 'POS']
Antecedent: []
```

ERROR ANALYSIS

1. One-anaphora errors :



1. if the functional role of *one*'s immediate dependency is a post-modifying phrase beginning with *of*, and the embedded noun phrase has a plural head, the use of *one* is **partitive**;
2. if the functional role of *one*'s immediate dependency is a post-modifying phrase beginning with *of*, and the embedded noun phrase has a singular head, the use of *one* is **anaphoric**;
3. if the functional role of *one* is that it is a pre-modifying quantifier, the use of *one* is **numeric**;
4. if the *one* is marked as a numeric adjective, the use of *one* is **numeric**;
5. if the *one* is in subject position and:
   (a) it depends on one of the words *might*, *may*, *should*, *could* or *must*; or
   (b) it depends a verb which is part of a verb chain; or
   (c) it depends on a verb which is one of the animate verbs listed in Section 3.2.3;

   it is generic; otherwise
6. it is *one*-anaphoric.

PAPER PRESENTATION

VERB PHRASE ELLIPSIS DETECTION USING AUTOMATICALLY PARSED TEXT - LA NIELSEN

PROBLEM DESCRIPTION

- Detecting ellipsis occurrences: VPE + pseudo-gapping
- Do so/it/that and so doing anaphora are not handled, as their resolution is different from that of VPE

Example of VPE:
John can play the guitar; Mary can too.

PSEUDO-GAPPING

Gapping: occurs in the non-initial conjuncts of coordinate structures and **elides minimally a finite verb and further any non-finite verbs** that are present.

Examples:

- ○ John can play the guitar, and Mary can too. - VPE
- ○ John can play the guitar, and Mary the violin. - Gapping

PSEUDO-GAPPING

Pseudo-gapping: elides **most but not all of a non-finite verb phrase**; at least one part of the verb phrase remains, which is called the remnant.

Examples:

- He drinks milk more often than she does [drink milk]. - VP-ellipsis
- He drinks milk more often than he does [drink] water. - Pseudo-gapping

The paper can be divided into three parts:

First, **previous work** done on tagged corpora is summarised.
Then, **new work** on parsed corpora is presented, showing the **gains possible through sentence-level features**.
Finally, **experiments using unannotated data** that is parsed using an automatic parser are presented.

CORPUS DESCRIPTION

1. British National Corpus (BNC):
    1. Training: 370k words with 645 samples of VPE
    2. Development: 74k words with 200 samples of VPE
2. WSJ and Brown corpus:
    1. 540k words and contains 522 samples of VPE
    2. 140k words and contains 150 samples of VPE.

EXPERIMENTS USING THE PENN TREEBANK

1. Words and POS Tag
2. Close to Punctuations
3. Heuristic Baseline: It searches forwards within a short range of words, and if it encounters any other verbs, adjectives, nouns, prepositions, pronouns or numbers, classifies the auxiliary as not elliptical. It also does a short backwards search for verbs. The forward search looks 7 words ahead and the backwards search 3.

EXPERIMENTS USING THE PENN TREEBANK

4. Auxiliary-final VP: this feature checks if the final element in the VP is an auxiliary or negation. If so, no main verb can be present, as a main verb cannot be followed by an auxiliary or negation.
5. Empty VP

EXPERIMENTS USING AUTOMATICALLY PARSED DATA

Step 1: Use the BNC and Treebank, but strip POS and parse information, and parse them automatically using two different parsers.

Parsers' used:

1. Charniak's Parser + Johnson [empty categories]
2. RASP

EXPERIMENTS USING AUTOMATICALLY PARSED DATA

Step 1a: Parsing the treebank: Compared to results on the **original treebank with similar data, the results are low**, which is not surprising, given the **errors introduced by the parsing process.** It is noticeable that the addition of features has less effect; 0-6%.

Previous experiments on the Treebank give 82% F1, with the most informative feature, empty VP's, giving 70% F1.
Parsing the Treebank gives 67% F1 for both parsers. Charniak's parser combined with Johnson's algorithm generates the empty VP feature with 32% F1.

EXPERIMENTS USING AUTOMATICALLY PARSED DATA

Step 1b: Experiments using parsed versions of the BNC corpora show similar results to the original results but the **features generate only a 3% improvement**, suggesting that many of the *cases in the test set can be identified using similar contexts in the training data and the features do not add extra information*. The performance of the features remain **similar to those for the re-parsed treebank experiments [71% F1]**, except for empty VP, which is reduced to 25% F1., due to Charniak’s parser being trained on the Tree-bank only.

COMBINED TREEBANK AND BNC

Combining the re-parsed BNC and Treebank data gives a more **robust** training set of 1167 VPE’s and a development set of 350 VPE’s. The results show only a **2-3% improvement when the features are added**. Again, simple contextual information is successful in correctly identifying most of the VPE’s. It is also seen that the **increase in data size is not matched by a large increase in performance**. This may be because *simple cases are already handled, and for more complex cases the context size limits the usefulness of added data.*