Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 34–43
Marseille, 11–16 May 2020

NoEl: An Annotated Corpus for Noun Ellipsis in English

Payal Khullar, Kushal Majmundar, Manish Shrivastava
International Institute of Information Technology Hyderabad
Gachibowli, Hyderabad 500032
{payal.khullar@research., kushal.majmundar@students., m.shrivastava@}iiit.ac.in

Abstract

Ellipsis resolution has been identified as an important step to improve the accuracy of mainstream Natural Language Processing (NLP) tasks such as information retrieval, event extraction, dialog systems, etc. Previous computational work on ellipsis resolution has focused on one type of ellipsis, namely Verb Phrase Ellipsis (VPE) and a few other related phenomenon. We extend the study of ellipsis by presenting the No(oun)El(lipsis) corpus - an annotated corpus for noun ellipsis and closely related phenomenon using the first hundred movies of Cornell Movie Dialogs Dataset. The annotations are carried out in a standoff annotation scheme that encodes the position of the licensor, the antecedent boundary, and Part-of-Speech (POS) tags of the licensor and antecedent modifier. Our corpus has 946 instances of exophoric and endophoric noun ellipsis, making it the biggest resource of noun ellipsis in English, to the best of our knowledge. We present a statistical study of our corpus with novel insights on the distribution of noun ellipsis, its licensors and antecedents. Finally, we perform the tasks of detection and resolution of noun ellipsis with different classifiers trained on our corpus and report baseline results.

Keywords: noun ellipsis, ellipsis resolution, zero anaphora

1. Introduction

Ellipsis is a linguistic phenomenon whereby certain parts of a sentence are omitted or deleted, and have to be retrieved from discourse or real-world context. Being motivated by the principles of information efficiency encoding, information constancy, cognitive economy, prominence, cooperation in dialogue, relevance, and intentionality - ellipsis becomes an important tool for carrying out efficient, natural and engaging dialogue (Zhao, 2015). Ellipses occur pervasively in natural language, especially in conversational settings and world languages use some or the other form of eliding redundant information, making this phenomenon universal in nature and extremely important for linguistic research (Langacker, 1999). While human interlocutors effectively resolve and disambiguate any elided information in a sentence based on context and cognitive commonsense extension (Chen, 2016), ellipsis resolution is deemed as a hard task for Natural Language Processing (NLP) systems.

A type of ellipsis is noun ellipsis, where the head noun inside a Noun Phrase (NP) is elliptically omitted. For example, in the following conversation taken from the third movie (m2) of the Cornell Dialogs Dataset (Danescu-Niculescu-Mizil and Lee, 2011), the word *coffee* is elided in the third turn of the dialogue:

1. ⟨ L3315_m2_Jordy ⟩ Do you have **coffee**?
   ⟨ L3316_m2_Daphne ⟩ In the kitchen
   ⟨ L3316_m2_Jordy ⟩ I will make [$_{NP}$ some [e]] for us.

The first string in the angular brackets denotes the unique label given to a single turn in the dialogue in the corpus, m2 is the movie number (starting from m0) and *Jordy* is the name of the character contributing towards the dialogue. We write *e* inside square brackets to mark the ellipsis site, enclose the elliptical noun phrase inside square bracket with the subscript NP, and write the antecedent in bold font. We use the same convention to present examples of ellipsis from the movie dialogues throughout this paper.

Ellipses occur in the environment of certain syntactical structures or trigger words, also known as licensors of ellipses. In case of noun ellipsis in English, these trigger words are determiners and modifiers of the elided noun. In the example presented in (1), the quantifier *some* licenses the elided noun or target, which is recovered from the previous context or antecedent of the ellipsis, in this case it is *coffee*. Hence, recovering from the previous context, we get the full form *some coffee*.

Noun ellipsis is also referred to as nominal ellipsis, zero noun anaphora and Noun Phrase Ellipsis (NPE) in different textbooks. Some authors feel it is preferable to not use the name *Noun Phrase ellipsis* to refer to the phenomenon as it is not the whole NP that gets deleted. (Menzel, 2017). For the purpose of the current paper, we use the term noun ellipsis everywhere.

2. Previous Work

Ellipsis has been discussed fairly well in theoretical linguistics literature (Halliday and Hasan, 1976; Dalrymple et al., 1991; Lobeck, 1995; Lappin, 1996; Hobbs and Kehler, 1997; Hardt, 1999; Johnson, 2001; Merchant, 2004; Frazier, 2008; Chung et al., 2010; Merchant, 2010; Rouveret, 2012; Gunther, 2011; van Craenenbroeck and Merchant, 2013; Park, 2017), in cognitive sciences (Kim et al., 2019) and language acquisition studies (Hyams et al., 2017; Lindenbergh et al., 2015; Goksun et al., 2010; Wijnen et al., 2003). In recent years, ellipsis resolution has been identified as an important Natural Language Processing (NLP) task for improving accuracy of information retrieval, event extraction, dialogue systems, etc (Hansen and Sogaard, 2019; Dean et al., 2016). One of the earliest computational approaches on ellipsis resolution involved the detection of Verb Phrase Ellipsis (VPE) instances in the Penn Treebank using a syntactic pattern match (Hardt, 1992). Most of the work on ellipsis resolution since then has focused on VPE and related phenomenon such as gapping, sluicing and do-so anaphora, for instance, a transformation learning-based

approach to generated patterns for VPE resolution (Hardt, 1998), the domain independent detection and resolution of VPE using machine learning methods (Nielsen, 2003), automatically parsed text (Nielsen, 2004), sentence trimming methods (McShane et al., 2015), linguistic principles (McShane and Babkin, 2016), improved parsing techniques that encode elided material dependencies for reconstruction of sentences containing gapping (Schuster et al., 2018), etc. More recently, complex Neural Networks like Transformers and Multilayer Perceptrons (MLP) have been used to achieve promising results on both VPE detection and resolution tasks (Zhang et al., 2019). This has been possible because of the availability of linguistic resources on VPE such as the annotated corpus for the analysis of VPE in English (Bos and Spenader, 2011) with 487 cases of VPE plus related phenomenon and the corpus prepared by (Nielsen, 2005) containing 1500 cases of VPE from parts of Wall Street Journal (WSJ), British National Corpus (BNC) and Brown Corpus.

Coming to computational work on noun ellipsis, there is a rule based system that detects noun ellipsis using syntactic constraints on licensors of ellipsis and resolves them by matching Part-of-Speech (POS) tag similarity between the licensor of ellipsis and the modifier of the antecedent, and fine tunes these syntactic rules on a small curated dataset that contains 234 instances of noun ellipsis along with some negative samples (Khullar et al., 2019). Although this dataset curates examples of noun ellipsis from Universal Dependency (UD) treebank (Silveira et al., 2014) and ParCorFull: a Parallel Corpus Annotated with Full Coreference (Lapshinova-Koltunski et al., 2018) also, a majority of these examples are actually from linguistic textbooks which may not fully represent the real world occurrence of this phenomenon. There is another corpus called the GECCo (German-English Contrasts in Cohesion) Corpus (Menzel and Lapshinova-Koltunski, 2014) that contains manual annotations on nominal, verbal and clausal ellipsis presented as cohesive device in a total of fourteen written and spoken registers of English and German. The corpus is not publicly available and the number of annotated noun ellipses are not enough for training machine learning models.

In this paper, we create the No(oun)EL(lipsis) corpus - a gold-standard annotated corpus containing 946 instances of noun ellipsis in the movie dialogues of the Cornell Movie Dialog Corpus (Danescu-Niculescu-Mizil and Lee, 2011) using a stand-off annotation scheme that does not modify the original corpus text. Using these annotations, we present statistical insights on noun ellipsis and baseline results on the detection and resolution tasks by training a simple classifier. This corpus will open up further avenues for computational work on noun ellipsis.

3. Scope of Annotation

This paper focuses on annotating noun ellipsis in English. There are two linguistic phenomena related to noun ellipsis that deserve a mention. One of them is subject ellipsis, where a reference position of the subject of a clause is filled with a morphologically unrealized form. For example, the null element in the second clause in (2).

2. John came early and ∅ ate all the snacks.

In pro-drop languages, the subject of the main sentence can also be dropped. However, in English, the subject, for instance, John in (2) needs to be overtly present. Subject ellipsis has been studied in detail in Chinese (Yeh and Chen, 2019a; Yeh and Chen, 2019b) and Japanese (Iida et al., 2007; Asao et al., 2018; Chen, 2016). There is some evidence of the phenomenon being used to achieve certain interactional functions in ordinary conversational settings by English speakers. For example, English speakers sometimes delete subjects in informal speech or conversational settings as in (3), although the sentence is ungrammatical (Oh, 2005).

3. *?* ∅ Told you so.

The second related phenomenon is one-anaphora, in which the elided noun is replaced by a phonologically overt pro form inside the noun phrase. There have been studies on one-anaphora in English, for instance a data-driven investigation of one-anaphora (Gardiner, 2003) and machine-learning methods for detection and resolution of one-anaphora that use Gardiner's heuristics (Gardiner, 2003), corpus study on identity of sense anaphoric relationships (Recasens et al., 2016). For the purpose of this corpus, we do not annotate subject ellipsis and one-anaphora, and focus on noun ellipsis only.

4. Dataset and Tools

We annotate the first 100 movies of the Cornell Movie Dialogs dataset (Danescu-Niculescu-Mizil and Lee, 2011) that accounts to 6,72,024 words out of 41,79,920, which is roughly 16.08% of the total corpus. The choice of the dataset is governed by a fundamental property of ellipsis being more frequent in conversational settings (Langacker, 1999). This corpus contains a large metadata-rich collection of fictional conversations extracted from raw scripts of a total of 617 movies. We take the first 100 movies, labelled m0 to m99, from the corpus for our task. We use Brat (Stenetorp et al., 2012) as the annotation tool for two reasons. It is a free web-based tool that helps annotate things or entities and relationships between entities. Secondly, it allows for text span annotations and marking multiple attributes of an entity, which are useful features for our task. Finally, the annotations are created in a stand-off scheme where the annotations and references to the original texts are collected in a separate file. This type of annotation does not modify the raw files of the corpus and is completely independent from further processing such as tokenization, chunking, parsing or other computational tasks.

5. Annotation Guidelines

Our annotation guidelines are created following the guidelines used to manually annotate nominal ellipsis as cohesive device in the GECCo Corpus (Menzel, 2017), albeit with a few necessary modifications as the purpose of building this linguistic resource is not to study cohesion in discourse but to analyse noun ellipses and provide a means to computationally handle it in text. The annotation guidelines are described in detail to ensure consistency throughout the annotation procedure.

5.1. Marking Ellipses

Since it is not possible to select and mark ellipses as such as they do not appear overtly in the sentence, we focus on the remnants of ellipses present at the ellipses site, also known as licensors of ellipses. In case of noun ellipsis in English, these remnants are typically determiners and modifiers to the elided noun head. Hence, we annotate these noun phrase remnants that license noun ellipsis. For example, in the sentence presented in (4), we annotate the cardinal number *two*, which is the licensor of the elided noun *pens*.

4. You gave me three pens, but I asked for [$_{NP}$ two [e]].

5.2. Noun Ellipsis Cases

Ellipses is a fuzzy topic in linguistics, and what constitutes as ellipsis and what does not often is a result of different perspectives on the phenomenon. Hence, it is important to decide which cases of noun ellipsis are to be incorporated in the corpus. We try to include a wide variety of cases, with an emphasis on those that could possibly improve performance of NLP systems dealing with ellipsis.

(a) We mark cases of noun ellipsis where the head noun in a noun phrase gets elided in identity with an antecedent to avoid redundancy, in a different clause such as in (4) or in a different sentence such as in (1). These are the atypical noun ellipsis cases for which there is a consensus in the ellipses literature.

(b) We mark cases of noun ellipsis that occur in a phrase after a linking verb, such as in a sentence with a predicate phrase. For such sentences, the predicate is annotated as the licensor of the noun ellipsis. For example, in the sentence in (5), the predicate *Ireland's best* has an elided head noun *cathedral* in the predicate.

5. This **castle** is [$_{NP}$ Ireland's best [e]].

This type of noun ellipsis involving nominal predicate can also be analyzed as clausal ellipses. But since it involves the elision of a head noun in the noun phrase, we include them in our annotated corpus.

(c) We mark cases of noun ellipsis that are locally bound and refer to antecedents within the same phrase. For example, the deletion of the noun head *minute* after *two* in the second conjunct of a coordinated noun phrase in the sentence presented in (6).

6. I will just take a **minute** or [$_{NP}$ two[e]].

This type of noun head deletion in coordinated noun phrase is not the same as in (7), where the conjunction coordinates two modifiers of the same noun head, without any deletion.

7. I want [$_{NP}$ a red and yellow hat].

(d) Not all ellipses can be recovered or inferred from a co-text. It is also possible that the antecedent of a given ellipsis is present outside the given text. For example, consider a speaker pointing towards roses in a shop and uttering a sentence such as in (8).

8. I will take two [e].

Using visual context, the shopkeeper can easily resolve the ellipsis in this sentence as *two roses*. Such cases of ellipses are *exophoric*, and can be resolved from extra linguistic, situational context using the knowledge of the grammar of the language. We mark cases of NPE such as in (8), which can be resolved from situational context.

(e) We mark conventionalised or lexicalised cases of noun ellipsis (Agel, 1991). For example, *Sam's* refers to Sam's place of stay in (9).

9. Let's party at [$_{NP}$ Sam's [e]] this Friday.

Similarly, cardinal noun modifiers in certain contexts modify elided nouns by convention. For example in (10), the word *two* actually means *two people* and in (11), the word *five* means *five minutes*.

10. We want to book ten days in Hilton for [$_{NP}$ two [e]]!
11. I will be there in [$_{NP}$ five [e]].

These noun ellipses can be resolved from the knowledge of idiomatic usage of the language. Since such cases clearly have a missing noun head, we mark their occurrences in the text.

(f) We mark nominalised pronouns as they are closely related to noun ellipsis. For example, the nominalised pronoun *mine* in (12) that is resolved as *my car* from the previous context. Although it is problematic to call such cases as noun ellipsis, they can be resolved using the same pipeline as noun ellipsis, if we treat them as possessive pronoun modifiers of the elided noun.

12. I drove my friend's **car** today. [$_{NP}$ Mine/ my [e]] was at the workshop.

Such cases pose a challenge during annotation as they require text modification. In fact, they are marked problematic in the GECCo corpus. Since we follow a stand-off annotation scheme that does not change the original text, marking these cases is easy in our corpus.

5.3. Antecedents

In this section, we explain the decisions we make while selecting the antecedent of a detected noun ellipsis from the text.

(a) It is possible that the ellipsis and antecedent do not match in the number feature. For example, in (13), the antecedent to the ellipsis marked at site [e] is plural *cars*, however the omitted structure is singular.

13. John wanted to get his own car. After looking at all the **cars** in the showroom, he decided to buy [$_{NP}$ Alice's [e]].

We mark antecedents that show non-identity in the number feature with the ellipsis as they provide the lexico-grammatical content necessary for the resolution of the noun ellipsis.

(b) We select the maximal antecedent boundary. For example, in (14), the elided noun can be resolved from just the antecedent *love*, however, the logical full form is *some of my undying love*.

14. **My undying love**. Have [$_{NP}$ some [e]].

For such cases, the resolution to the head word of the antecedent phrase may also be acceptable. However, for our task, we select the complete antecedent to get maximum information from the text.

(c) While selecting the antecedent boundary, we do not include punctuation marks or spaces at its beginning or end, unless they are a part of quoted speech. This is to ensure that we only select useful information on ellipsis resolution from the text.

15. He said he found a thick **book**, which has a brown cover. I think it should be [$_{NP}$ Alice's [e]].

Hence, in (15), we select *a thick book* as the antecedent of the ellipsis, and do not include the comma after it.

5.4. Categories and Labels

Apart from marking ellipses and selecting their antecedents, we also add labels to mark certain information to the licensors of the noun ellipsis and modifiers of the antecedent.

(a) We categorize noun ellipsis on the basis of how they can be resolved. If the resolution of the noun ellipsis is present in the text as in (1), (4), (5) and (6), we add the label *endophoric* to the licensor of noun ellipsis. This label corresponds to the coherent and the clause-internal cases of noun ellipsis presented in the GECCo Corpus, i.e. those with a resolution in the same clause (weakly cohesive) and those with a resolution in a different clause (strongly cohesive) respectively.

If the noun ellipsis cannot be resolved textually as in the sentences presented in (8), (9), (10) and (11), we add the label *exophoric* to the licensor of noun ellipsis. This label corresponds to the *non-coherent* noun ellipsis cases in the GECCo corpus, i.e. the ones that do not contribute towards textual coherence.

(b) In English, words with only certain grammatical categories can license noun ellipsis. These typically include cardinal and ordinal numbers, quantifiers, adjectives, classifier nouns, indefinite pronouns and possessives. We manually label every noun ellipsis licensor with its grammatical category. We do this because the grammatical category information of licensor of an ellipsis can serve as an important cue for its detection, for example using auxiliary verbs and modals in case of VPE detection (McShane and Babkin, 2016).

We use the CLAWS5 Part-of-Speech Tagger for English to semi-automatically add POS tags to every licensor of the noun ellipsis. Since most parsers give erroneous output sometimes for sentences containing ellipsis (Menzel, 2017), we verify the tags manually before annotation. We choose the CLAWS5 tagset over other tagsets such as the Penn Treebank tagset (Marcus et al., 1994) because it is more fine grained and lets us assign distinct tags for different grammatical categories of licensors that we wish to mark. For instance, articles, demonstrative determiners and general determiners are all represented by the DT (determiner) tag in the Penn Treebank tagset, but CLAWS5 tagset has separate tags for these. This distinction is potentially very useful for the noun ellipsis detection and resolution as we show in the section on corpus summary.

(c) We also add POS tags to the modifier(s) of the antecedent of the noun ellipsis using the CLAWS5 tagger. For example, for the antecedent *pens* to the noun ellipsis in the sentence in (4), we add the label CRD (Cardinal Numeral) to the modifier *three* of the antecedent. Clauses that are linked by an ellipsis-antecedent relation often have similar syntactic structure and priming effects (Xiang et al., 2014). Hence, marking grammatical category information of the licensors of noun ellipsis and modifiers of the antecedent might be useful for resolution of noun ellipsis in English.

(d) If there are multiple modifiers present at the ellipsis site, we add POS tags to all of them separately. For example, in (16), we add POS tags to both *my* and first. We explain in the corpus summary section how this information might be useful.

16. You are [$_{NP}$ my first [e]].

(e) It is possible that the entities to which we want to add POS tags are compound words or multiwords. For example, in (17), the numeral is actually two words combined with a hyphen. In English, adjectives can also occur as compound words, however, cases of them licensing an noun ellipsis are rare (Menzel, 2017).

17. I gave the shopkeeper fifty **dollars** and he returned [$_{NP}$ twenty-five [e]].

We assign a single POS tag to *twenty-five* in (17) instead of two same tags to *twenty* and *five* because not only is it redundant to do so, but we also do not want to make modifications to the text including splitting a word and removing punctuation.

5.5. Not Marked

(a) We do not mark cases of nominalised adjectives as noun ellipsis. The word *poor* in the example presented in (18) refers to the generic notion of *poor people*.

18. We should help *the poor*.

An argument of such cases as exophoric noun ellipsis (Halliday and Hasan, 1976), with adjectives as modifying an elided silent noun head (Gunther, 2011) is also available in literature. However, following (Menzel, 2017), we treat such adjectives as nouns only and do not mark them.

(b) We do not mark deletions of nouns arising due to incomplete sentences, dialogue breaks or pauses, which are common in movie dialogues. For example, the noun deletion after *big* in the sentence (19) is ungrammatical and does not fall into the definition of the phenomenon.

19. I would like to eat [$_{NP}$ a big...]. Wait, what is that?

(c) We do not annotate ellipsis occurring in a language other than English in the movie dialogues.

6. Annotation Format

A single annotation comprises either 2 or 5 lines, depending upon whether the noun ellipsis is exophoric or endophoric respectively. Let us take an example sentence to explain the annotation format.

20. ⟨ L937_m0_Kat ⟩ How'd you get a **tux** at the last minute?
⟨L936_m0_Patrick ⟩ It's [$_{NP}$ Scurvy's [e]].

The sentence in (20) from the first movie of the Cornell Dialogue dataset has an endophoric noun ellipsis after *Scurvy's* which can be resolved from the previous context *tux*. The annotation for this sentence is presented below:

T18 Licensor 39630 39638 Scurvy's
A18 Scurvy_NP0 's_POS Endophoric
T19 Antecedent 39755 39758 tux
A19 a_AT0
R9 Resolution Arg1:T18 Arg2:T19

The first line begins with the string "T" that implies it is an entity. The number next to it represents the count of the entity in the movie file, starting from 1. This is followed by the label of the entity, i.e. Licensor and the index position of the character at the starting and end of the word. This line ends with the text of the licensor in the original movie file. The second line with the label A is for attributes of the licensor, i.e. the POS tags for the licensor and the category of the noun ellipsis. For exophoric noun ellipsis, there are only these two lines. For endophoric noun ellipsis as in (18), the third line contains the index of the antecedent and its text value. The fourth line contains the POS tags for modifiers of the antecedent. The fifth and last line links the licensor with the antecedent.

6.1. Inter-annotator Agreement

Annotation is carried out manually. Three annotators who are linguists by training and proficient in the language perform the task of hand annotation. All of them work independently on all sentences of the hundred movies in the dataset. For each sentence, the annotation is either a two or four-step procedure. The first task is that of detection of the noun ellipsis in a given sentence. For positive cases, the annotators select the type of ellipsis and the grammatical category of the licensors from a list of options. For endophoric noun ellipsis cases, they move to the next step of marking the antecedent of the noun ellipsis. The final step is to mark the grammatical category of the modifier of the selected antecedent.

After the detection of an instance of noun ellipsis, the annotators perform multiple annotations on a single sentence. To get a better picture of disagreements and error types, we calculate the inter-annotator agreement on five tasks - the detection of noun ellipsis cases in the text, identifying the type of the noun ellipsis, adding a POS tag to the licensor of the noun ellipsis, the selection of antecedent boundary from the text in those cases where the noun ellipsis could be resolved, and the selection of the POS tag of the modifier of the antecedent. We use the Fleiss's Kappa coefficient to calculate the inter-annotator agreement between multiple annotators.

For the number of noun ellipsis detected, we get Fleiss's Kappa coefficient of 0.92, for the labels on the licensor of noun ellipsis, we get 0.85, for the type of noun ellipsis, we get 0.92, for antecedent boundary selection, we get 0.88 and for the label on the modifier of the antecedent, we get 0.87. These numbers confirm reliability of our annotations. All the disagreements are finally resolved at the end of the task by discussion among the three annotators and the agreed-upon noun ellipsis cases are included in the final corpus.

7. Corpus Summary

In this section, we present a statistical summary of our corpus. The observations will be useful for an empirical analysis of the noun ellipsis phenomenon in English.

(a) We find a total of 946 cases of noun ellipsis in the first hundred movies of the Cornell Movie Dialogs dataset. Out of these, 438 are endophoric and 508 are exophoric. The higher count of exophoric noun ellipsis in the dataset is predictable because many anaphoric expressions in movie dialogues can be resolved from visual scene context. Another reason for this could be the fact that the dialogues in the Cornell Movie Dialogs dataset are often only parts or fragments of a larger conversation. It is possible that the ellipsis may be resolved in discourse, but the information is not present in the provided text.

(b) The total number of tokens in the selected part of the corpus are 6,72,024. This means the frequency of noun ellipsis is 14.08 per 10,000 tokens. This is significantly higher than the frequency of noun ellipsis in the GECCo Corpus (211 cohesive plus 125 non-cohesive nominal ellipsis in 4,08,016 tokens or 5.17 + 3.06 = 8.23 in 10,000 tokens).

Ellipsis is a fundamental property of conversational speech (Langacker, 1999). The frequency of ellipsis is expected to be higher in a corpus such as the Cornell Movie Dialogs dataset that comprises of informal

Syntactic Type of the Licensor	**Example Sentence from Cornell Movie Dialogs Corpus** (Danescu-Niculescu-Mizil and Lee, 2011)
Cardinal Numbers	⟨*L17395_m11_Carlton*⟩ We got three **MiGs** running around and [$_{NP}$ six [e]] on the way.
Ordinal Numbers	⟨*L3432_m2_Jordy*⟩ You'll be [$_{NP}$ my first [e]].
Possessives	⟨*L378_m0_Michael*⟩ I think I speak correctly when I say Cameron's **love** is pure. Purer than say –[$_{NP}$ Joey Dorsey's [e]].
Demonstrative Determiners	⟨*L31008_m16_Schikanader*⟩ Oh, and a few **trick animals**. You'd have to use [$_{NP}$ those [e]].
Quantifiers	⟨*L30280_m16_Mozart*⟩ There are just as many **notes**, Majesty, as required. Neither [$_{NP}$ more [e]] nor [$_{NP}$ less [e]].
Adjectives	⟨*L12128_m10_Lariviere*⟩ We're looking for the funniest **costume**! And [$_{NP}$ the scariest [e]]. And [$_{NP}$ the most imaginative [e]].
General Determiners	⟨*L162657_m49_Mrs.Bruce*⟩ It's dirty laundry for one **thing** and for [$_{NP}$ another [e]], you still haven't worn the clothes I bought you.
Interrogative Determiners	⟨*L17195_m11_GeneralNorthwood*⟩ Like any good poker player, they are checking over their hand seeing which **cards** to play and [$_{NP}$ which [e]] to discard.

Table 1: The first column shows different syntactic categories that can license a noun ellipsis in English. The second column has an example sentence of each licenser type from the Cornell Movie Dialogs corpus. Dialogue information is given inside the angular brackets. The site of ellipsis represented by [e] and the antecedents of the ellipsis are marked in bold.

conversations, as compared to GECCo that contains both written and spoken registers of English. The frequency of nominal ellipsis in only the spoken registers of GECCo is 6.98 per 10,000 which is still very low as compared to the frequency of noun ellipsis in the Cornell Dialogs dataset. This is because the spoken registers of GECCo do not represent informal, everyday speech. Out of the three spoken registers included for ellipsis annotation, the *Academic* one comprises monologic, planned texts and the *Forum* one is a heterogeneous mix of written and spoken forms. The *Interview* is the only one with actual conversational text.

(b) Sentence wise, annotated corpus has 946 noun ellipsis in 49,804 sentences, which makes the frequency 1.99%. As compared to the reported VPE frequency in a previous annotated VPE corpus (487 VPE + 67 related cases in 53,561 sentences of the Penn Treebank, making it roughly 1.0%) (Bos and Spenader, 2011), we find the frequency of noun ellipsis to be much higher. Although these counts are not directly comparable as the choice of datasets used are different, this finding is in line with the distribution of nominal and verbal/clausal ellipsis in the GECCo corpus, where the reported frequency of nominal ellipsis is higher than that of verbal/clausal ellipsis (211 cohesive plus 125 non-cohesive nominal ellipsis versus 186 cohesive plus 47 non-cohesive verbal/clausal ellipsis in 4,08,016 tokens). Results on the annotation of VPE cases in the Cornell Movie Dialogs dataset could further help in confirming these findings.

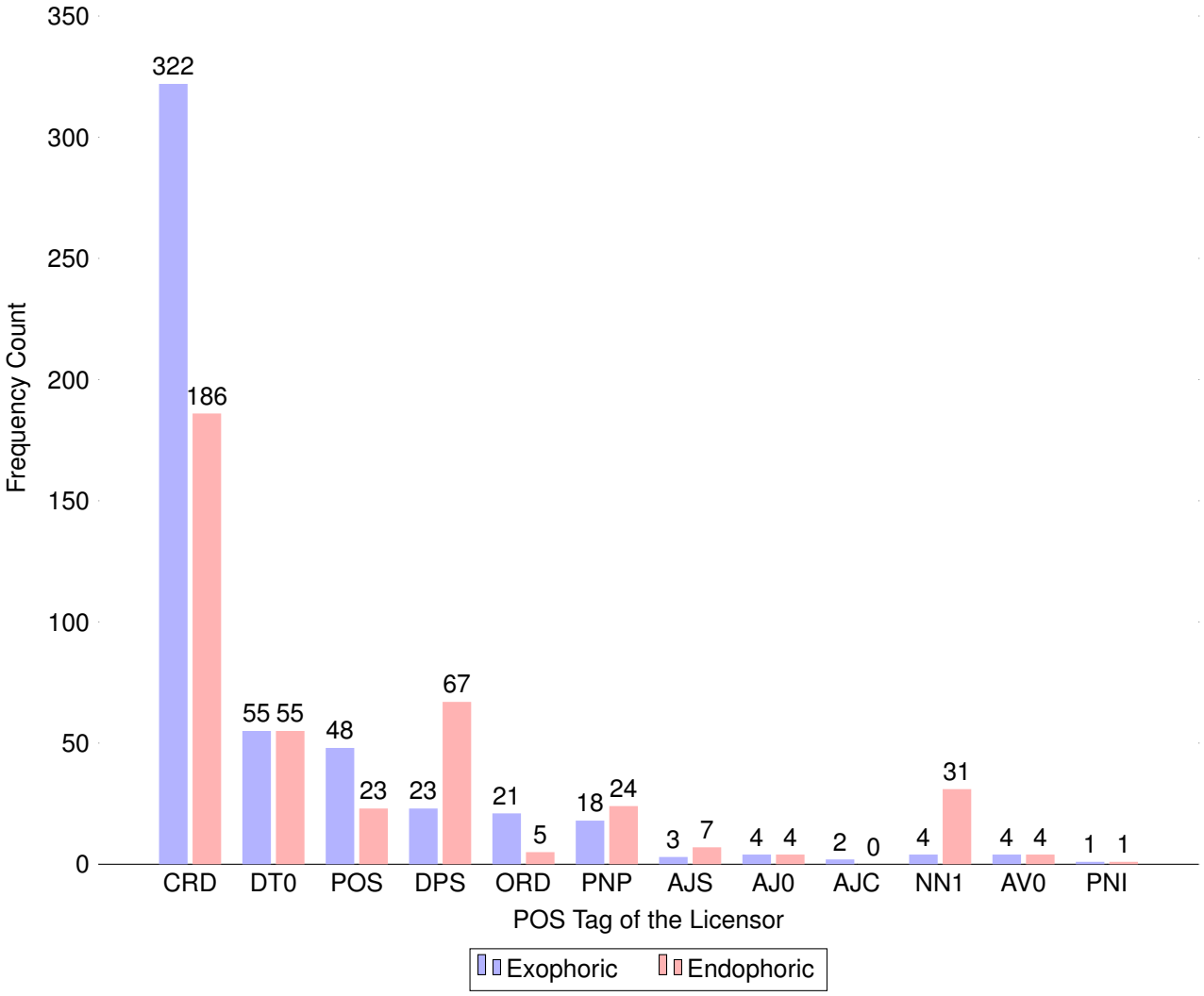
(c) There are various syntactical studies cited earlier that discuss the possible syntactic categories of the licensors of noun ellipsis. However, empirical studies on the distribution of these categories in the language are rare because of the lack of a sizeable dataset to make generalisations. In our annotated corpus, cardinal numbers (CRD tag) are the highest in number to license noun ellipsis, followed by possessives (DPS, POS and PNP tags) and determiners (DTO and DTQ tags). These are, in fact, the syntactic categories with most discussed examples in ellipsis textbooks. Noun ellipsis followed by adverbs and adjectives are rarest in the corpus. Out of the 20 adjective licensors, 12 are in the superlative and comparative forms. This can be accounted by the fact that adjectives in English often need one-substitution to license a noun ellipsis (a phenomenon known as one-anaphora, as discussed previously). This can be seen from the sentence in (21) which is ungrammatical as opposed to the sentence with one-substitution in (22).

21. *John has a big **house** but I have [$_{NP}$ a small [e]].
22. John has a big **house** but I have a small one.

A bar graph depicting the frequency distribution of endophoric and exophoric noun ellipses followed by licensors with different POS tags in our annotated corpus is presented in the figure 1. Exophoric noun ellipses are higher than endophoric noun ellipses for licensors with all POS tags, except for the possessive determiner (DPS) tag. This is because as compared to numerals and quantifiers, it is uncommon to use a possessive determiner in a situational context.

(f) We report an interesting parallel between syntactic type of the licensor and the modifier of the antecedent

Figure 1: Frequency distribution of licensors in the annotated corpus.



of the ellipsis. The licensor and the modifier of the antecedent of the noun ellipsis very often have the same POS tag. For example, there are 186 noun ellipses in our corpus that have licensors with the CRD POS tag. The modifier of the antecedent of 71 of these also has a CRD POS tag, which is 38.17 % of the times. Table 2 presents a list of licensors and antecedent modifier pairs that exhibit this POS tag similarity very frequently in our corpus. This finding is in line with the parallelism theory in discourse (Hobbs and Kehler, 1997) that can be applied to resolve possible readings ellipsis and reference phenomenon. In our case, the parallel between the licensor of an ellipsis and its antecedent could be useful in the search for the antecedent of the ellipsis in the text.

Licensor Tag	**Tag of the Corresponding Antecedent Modifier (Most Frequent)**	**Frequency %**
CRD	CRD	38.17
NN1	NN1	74.19
DPS	DPS	53.73
POS	POS	52.17
AJ0	AJ0	50.00
AJS	AJS	42.85
DTQ	DTQ	100.00

Table 2: First column shows Part-of-Speech (POS) tag of the licensor of the noun ellipsis and the second column shows the most frequent POS tag of the corresponding modifier of the antecedent of the ellipsis.

8. Corpus Utility for Machine Learning

Our corpus is useful for doing a statistical study of noun ellipsis as presented in the previous section. The size of our corpus also allows us to perform experiments with machine learning models on the tasks of ellipsis detection and resolution. In this section, we demonstrate the utility of our corpus for machine learning.

Task	Averaged Results ML Model	Precision	Recall	F1-Score
Noun Ellipsis Detection	Naive Bayes	0.6217	0.8376	0.7137
	Linear SVM	0.6407	0.8587	**0.7339**
	RBF SVM	0.6054	**0.9045**	0.7253
	Nearest Neighbors	**0.7369**	0.5949	0.6583
	Random Forest	0.1750	0.3500	0.2333
Noun Ellipsis Resolution	Naive Bayes	0.6096	0.6008	0.6052
	Linear SVM	**0.6213**	0.4258	0.5053
	RBF SVM	0.6007	0.9858	0.7465
	Nearest Neighbors	0.6061	0.3418	0.4371
	Random Forest	0.6000	**0.9989**	**0.7497**

Table 3: Results on the Precision and Recall and F1-Score values for noun ellipsis detection and resolution task with different classifiers trained on our annotated corpus. The parameter values are set to default and the results are averaged over 20 iterations to get unbiased results. The values in bold indicate best performance.

8.1. Noun Ellipsis Detection

For a given input sentence, we first get all its trigrams and map each of them to 0 or 1, depending upon whether the center element of the trigram is a noun ellipsis licensor or not. For simplicity, we only take the previous and following words to capture the context. Noun ellipsis detection can be posed as classification problem, where the classifier has to predict whether the given trigram has a licensor in its centre or not. We take all the 946 instances of noun ellipsis from our annotated corpus and do a standard train-test split of 80-20. For each trigram, we take the three constituting word tokens along with their POS tags and generate a Sentence2Vec embedding for these using inbuilt Word2Vec of Gensim and Fast Sentence Embeddings using Smoothened Inverse Frequency. We use sklearn (Pedregosa et al., 2011) to train different machine learning models on the obtained vectors and average the results over 20 iterations to eliminate bias in the results. We set the parameters to their default values for all the models. The precision, recall and F1-Score values with Naive Bayes, Linear Support Vector Machine (SVM), Radial Basis Function (RBF) SVM, Nearest Neighbours and Random Forest classifiers are presented in 3. The performance of Naive Bayes, Linear SVM and RBF SVM classifiers is comparable, with Linear SVM getting the highest F1-score on the detection task. Given that we use classifiers with simple syntactic feature vectors and default parameter values, these results are promising.

8.2. Noun Ellipsis Resolution

For the noun ellipsis resolution task, we only take the sentences containing endophoric noun ellipsis from our corpus, i.e. the ones that have a textual resolution. For each of these sentences, we make a triad of [licensor, antecedent candidate, all tokens in the sentence] and map it to 0 or 1 depending upon whether the antecedent candidate corresponds to an actual resolution for the given ellipsis or not. Since there are 438 instances of endophoric noun ellipsis in our corpus, we get 438 such triads. We convert these into 438*600 dimensional sentence embeddings using inbuilt Word2Vec of Gensim and Fast Sentence Embeddings using Smoothened Inverse Frequency. Each of these 100 dimension represents sentence embeddings of each element and POS tag of the triad. We take the words other than antecedent candidates in the sentence as the negative class samples. Our negative samples are inevitably a lot more than the positive ones. We take a 80-20 split on the positive class and a 10-90 split on the negative class to capture the positive class properly. We pose Noun ellipsis resolution as a classification problem where for a given triad of [Licensor, Antecedent, all tokens in a Sentence], the classifier has to predict whether the antecedent candidate is the resolution of the ellipsis (licensed by the licensor) or not. The precision, recall and F1-Score values with Naive Bayes, Gaussian Process, Linear Support Vector Machine (SVM), RBF SVM and Nearest Neighbours are presented in 3. The performance of RBF SVM and Random Forest classifiers is comparable, with the latter getting a higher F1-score on the resolution task. Again, given that we use classifiers with simple syntactic feature vectors and default parameter values, these results are promising.

9. Conclusion

We present the NoEl Corpus, a gold-standard corpus of noun ellipsis in English by hand annotating the first hundred movies of the Cornell Movie Dialogs dataset. We mark a total of 946 instances of exophoric and endophoric noun ellipsis in the selected corpus, which makes it the biggest annotated corpus for noun ellipsis in English. Ellipsis is a fuzzy concept in linguistics and what constitutes as ellipsis and what does not often changes with different perspectives on the analysis of the phenomenon. We try to incorporate various cases of noun ellipsis discussed in literature in our corpus. For each identified case of noun ellipsis, we mark the licensor of noun ellipsis along with information such as the POS tag of the licensor, the type of the noun ellipsis, the antecedents when present textually and the syntactic information of the modifiers of the antecedents. We present a statistical summary of noun ellipsis in our corpus and present results on noun ellipsis detection and resolution task by training various classifiers on our corpus using sklearn. This corpus will be beneficial for theoretical work on noun ellipsis in linguistics and for improving the performance of NLP systems that handle ellipsis.

10. Bibliographical References

Agel, V. (1991). Lexikalische ellipsen fragen und vorschlÃ¤ge. *Zeitschrift fÃ¼r Germanistische Linguistik*, 19.

Asao, Y., Iida, R., and Torisawa, K. (2018). Annotating zero anaphora for question answering. In *LREC*.

Bos, J. and Spenader, J. (2011). An annotated corpus for the analysis of vp ellipsis. *Language Resources and Evaluation*, 45(4):463â494.

Chen, W. (2016). The motivation of ellipsis. *Theory and Practice in Language Studies*, 6(11):2134–2139.

Chung, S., Ladusaw, W., and McCloskey, J. (2010). Sluicing (:) between structure and inference. *In Representing language: Essays in honor of Judith Aissen*.

Dalrymple, M., Shieber, S. M., and Pereira, F. C. N. (1991). Ellipsis and higher-order unification. *Linguistics and Philosophy*, 14(4):399–452.

Danescu-Niculescu-Mizil, C. and Lee, L. (2011). Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics, ACL 2011*.

Dean, K. K., Cheung, J. C. K., and Precup, D. (2016). Verb phrase ellipsis resolution using discriminative and margin-infused algorithms. *In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, page 1734â1743.

Frazier, L. (2008). Processing ellipsis: A processing solution to the undergeneration problem? In *Proceedings of the 26th West Coast Conference on Formal Linguistics*.

Gardiner, M. (2003). *Identifying and resolving one-anaphora*. Department of Computing, Division of ICS, Macquarie University.

Goksun, T., Roeper, T. W., Hirsh-Pasek, K., and Golinkoff, R. M. (2010). From nounphrase ellipsis to verbphrase ellipsis: The acquisition path from context to abstract reconstruction.

Gunther, C. (2011). Noun ellipsis in english: adjectival modifiers and the role of context. *The structure of the noun phrase in English: synchronic and diachronic explorations*, 15(2):279–301.

Halliday, M. A. K. and Hasan, R. (1976). Cohesion in english. page 76.

Hansen, V. P. B. and Sogaard, A. (2019). What do you mean âwhy?â: Resolving sluices in conversations.

Hardt, D. (1992). An algorithm for vp ellipsis. page 9â14.

Hardt, D. (1998). Improving ellipsis resolution with transformation-based learning. 1998.

Hardt, D. (1999). Dynamic interpretation of verb phrase ellipsis. *Linguistics and Philosophy*, 22(2):187–221.

Hobbs, J. R. and Kehler, A. (1997). A theory of parallelism and the case of vp ellipsis. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, ACL '98/EACL '98, pages 394–401, Stroudsburg, PA, USA. Association for Computational Linguistics.

Hyams, N., Mateu, V., and Winans, L. (2017). Ellipsis meets wh-movement: sluicing in early grammar.

Iida, R., Inui, K., and Matsumoto, Y. (2007). Zero-anaphora resolution by learning rich syntactic pattern features. *ACM Trans. Asian Lang. Inf. Process.*, 6, 12.

Johnson, K. (2001). What vp ellipsis can do, and what it canât, but not why. pages 439–479.

Khullar, P., Anthony, A., and Shrivastava, M. (2019). Using syntax to resolve npe in english. In *Proceedings of Recent Advances in Natural Language Processing*, pages 535–541.

Kim, N., Brehm, L., and Yoshida, M. (2019). The online processing of noun phrase ellipsis and mechanisms of antecedent retrieval. *Language, Cognition and Neuroscience*, 34(2):190–213.

Langacker, R. W. (1999). *Grammar and Conceptualization (Cognitive Linguistics Research)*, volume 14. Mouton de Gruyter, New York.

Lappin, S. (1996). The interpretatin of ellipsis. In Shalom Lappin, editor, *The Handbook of Contemporary Semantic Theory*, pages 145–176. Blackwell.

Lapshinova-Koltunski, E., Hardmeier, C., and Krielke, P. (2018). Parcorfull: a parallel corpus annotated with full coreference. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018.*

Lindenbergh, C., van Hout, A., and Hollebrandse, B. (2015). Extending ellipsis research: The acquisition of sluicing in dutch. *BUCLD 39 Online Proceedings Supplement*, 39.

Lobeck, A. (1995). *Functional Heads, Licensing, and Identification*. Oxford University Press.

Marcus, M., Kim, G., Marcinkiewicz, M. A., MacIntyre, R., Bies, A., Ferguson, M., Katz, K., and Schasberger, B. (1994). The penn treebank: Annotating predicate argument structure. In *Proceedings of the Workshop on Human Language Technology*, HLT '94, pages 114–119, Stroudsburg, PA, USA. Association for Computational Linguistics.

McShane, M. and Babkin, P. (2016). Detection and resolution of verb phrase ellipsis. *Linguistic Issues in Language Technology*, 13(1).

McShane, M., Nirenburg, S., and Babkin, P. (2015). Sentence trimming in service of verb phrase ellipsis resolution. In *EAPCogSci*.

Menzel, K. and Lapshinova-Koltunski, E. (2014). Kontrastive analyse deutscher und englischer kohÃ¤sionsmittel in verschiedenen diskurstypen. *tekst i dyskurs - Text und Diskurs. Zeitschrift der Abteilung fÃ¼r germanistische Sprachwissenschaft des Germanistischen Instituts Warschau*.

Menzel, K. (2017). *Understanding English-German contrasts: a corpus-based comparative analysis of ellipses as cohesive devices*. Ph.D. thesis, Universitat des SaarÂ¨ landes, Saarbrucken.

Merchant, J. (2004). Fragments and ellipsis. *Linguistics and Philosophy*, 27(6):661–738.

Merchant, J. (2010). *Three Kinds of Ellipsis: Syntactic, Semantic, Pragmatic?*

Nielsen, L. A. (2003). Using machine learning techniques for vpe detection. 1.

Nielsen, L. A. (2004). Verb phrase ellipsis detection using automatically parsed text. 01.

Nielsen, L. A. (2005). *A corpus-based study of verb phrase ellipsis identification and resolution*. Kingâs College London.

Oh, S.-Y. (2005). English zero anaphora as an interactional resource. *Research on Language and Social Interaction*, 38(3):267–302.

Park, D. (2017). *When does ellipsis occur, and what is elided?* PhD dissertation, University of Maryland.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Recasens, M., Hu, Z., and Rhinehart, O. (2016). Sense anaphoric pronouns: Am i one? pages 1–6, 01.

Rouveret, A. (2012). Vp ellipsis, phases and the syntax of morphology. *Natural Language & Linguistic Theory*, 30(3):897â963.

Schuster, S., Nivre, J., and Manning, C. D. (2018). Sentences with gapping: Parsing and reconstructing elided predicates. *ArXiv e-prints*.

Silveira, N., Dozat, T., de Marneffe, M.-C., Bowman, S., Connor, M., Bauer, J., and Manning, C. D. (2014). A gold standard dependency corpus for English. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*.

Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., and Tsujii, J. (2012). Brat: A web-based tool for nlp-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '12, pages 102–107, Stroudsburg, PA, USA. Association for Computational Linguistics.

van Craenenbroeck, J. and Merchant, J. (2013). Ellipsis phenomena. In *The Cambridge Handbook of Generative Syntax*, pages 701–745. Cambridge University Press.

Wijnen, F., Roeper, T. W., and van der Meulen, H. (2003). Discourse binding: Does it begin with nominal ellipsis?

Xiang, M., Grove, J., and Merchant, J. (2014). Ellipsis sites induce structural priming effects.

Yeh, C.-L. and Chen, Y.-C. (2019a). Using zero anaphora resolution to improve text categorization. 03.

Yeh, C.-L. and Chen, Y.-J. (2019b). An empirical study of zero anaphora resolution in chinese based on centering model. 03.

Zhang, W.-N., Zhang, Y., Liu, Y., Di, D., and Liu1, T. (2019). A neural network approach to verb phrase ellipsis resolution.

Zhao, G. (2015). The motivation of ellipsis. *Theory and Practice in Language Studies*, 5(6):1275–1279, June.