Using Syntax to Resolve NPE in English

by

Payal Khullar, Allen Antony, Manish Shrivastava

in

Recent Advances in Natural Language Processing
(*RANLP-2019*)

Varna, Bulgaria

Report No: IIIT/TR/2019/-1

Centre for Language Technologies Research Centre
International Institute of Information Technology
Hyderabad - 500 032, INDIA
August 2019

Using Syntax to Resolve NPE in English

Payal Khullar   **Allen Antony**   **Manish Shrivastava**
Language Technologies Research Centre
International Institute of Information Technology, Hyderabad
Gachibowli, Hyderabad, Telangana-500032
{payal.khullar@research., allen.antony@research.,
m.shrivastava@}iiit.ac.in

Abstract

This paper describes a novel, syntax-based system for automatic detection and resolution of Noun Phrase Ellipsis (NPE) in English. The system takes in free input English text, detects the site of nominal elision, and if present, selects potential antecedent candidates. The rules are built using the syntactic information on ellipsis and its antecedent discussed in previous theoretical linguistics literature on NPE. We prepare a curated dataset of 337 sentences from well-known, reliable sources, containing positive and negative samples of NPE. We split this dataset into two parts, and use one part to refine our rules and the other to test the performance of our final system. We get an F1-score of 76.47% for detection and 70.27% for NPE resolution on the testset. To the best of our knowledge, ours is the first system that detects and resolves NPE in English. The curated dataset used for this task, albeit small, covers a wide variety of NPE cases and will be made public for future work.

1   Introduction

Nominal Ellipsis or Noun Phrase Ellipsis (NPE, henceforth) is a type of ellipsis in linguists wherein the sub-parts of a nominal projection are elided, with the remaining projection pronounced in the overt syntax. For example in the sentence presented in (1), the noun *chairs* is elided at the position marked by [e].

1. There are three **chairs** in the living room and two [e] in the hall.

The full meaning of such a sentence can only be understood when we reconstruct the meaning of the elided part from the antecedent, which can be present in the linguistic context as in (1) or has to be retrieved from real world knowledge as in (2), where *Mary's* actually means *Mary's place*.

2. We are all partying at Mary's [e].

All world languages use some or the other mechanism to elide redundant information and, hence, ellipses is fairly pervasive in natural language, more so in conversational settings (Langacker, 1999). While human interlocutors effectively resolve and disambiguate any elided information in a sentence based on context and cognitive commonsense extension (Chen, 2016), the realization of the complexity of processing of ellipsis becomes evident when it poses a serious challenge for computational systems involved in natural language understanding.

The approaches that handle ellipsis in both theoretical linguistics and NLP are largely classified as syntactic, semantic and pragmatic (Merchant, 2010). For the current paper, we present a system that automatically detects and resolves NPE in English using a syntax-driven approach.

2   Previous Work

The syntax and semantics of the ellipsis phenomenon has been thoroughly studied in theoretical linguistics (Halliday and Hasan, 1976; Dalrymple et al., 1991; Lobeck, 1995; Lappin, 1996; 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 linguistics (Kim et al., 2019), and in language acquisition studies (Hyams et al., 2017; Lindenbergh et al., 2015; Goksun et al., 2010; Wijnen et al., 2003). In the context of NLP, most of the work on handling ellipsis has been done on Verb Phrase Ellipsis (VPE) and related phenomenon, for instance the preparation of annotated corpus for analysis of VPE (Bos and Spenader, 2011), the detection of VPE in Penn treebank (Hardt, 1997), the domain independent detection and resolution

of VPE using machine learning (Nielsen, 2003) and parsed text (Nielsen, 2004), using linguistic principles (McShane and Babkin, 2016), with sentence trimming methods (McShane et al., 2015), reconstruction of sentences with gapping using improved parsing techniques that encode elided material dependencies (Schuster et al., 2018), etc.

There are no known systems that handle NPE detection and resolution in English. However, on a related linguistic phenomenon called one-anaphora or one-substitution, in which the elided noun is replaced by an overt pro-form, there is a thorough data-driven investigation (Gardiner, 2003) and machine-learning methods that use heuristics proposed in this study (Ng et al., 2005). Another phenomenon similar to NPE is zero-anaphora, which has been thoroughly studied in some pro drop languages such as Chinese (Yeh and Chen, 2019a,b) and Japanese (Iida et al., 2007; Asao et al., 2018; Chen, 2016). Zero-anaphora does not occur in English, although there is some evidence of the phenomenon being used to achieve certain interactional functions in ordinary conversational settings by English speakers (Oh, 2005). There are also proposed heuristics for determining antecedents of pronominal words (Lappin and Leass, 1994; Kennedy and Boguraev, 1996). In the present paper, we do not deal with one-anaphora, zero anaphora or pronominals, and restrict our focus to NPE.

3   Task Description

Resolution of ellipsis comprises two tasks - detection of the elided material and antecedent selection. In some cases, reference resolution might also be necessary (Liu et al., 2016; Nielsen, 2003). For example, in (3), the common sense interpretation is that Sam loves his girlfriend. But it could also lead to a sloppy reading where it means Sam loves John's girlfriend.

  3. John **loves his girlfriend**. Sam does [e] too.

Note that (3) presents an example of VPE as the verb along with its predicate are elided. In this paper, we focus only on the first two tasks, i.e. detection of NPE and antecedent selection.

4   Dataset Preparation

There are no dedicated linguistic resources or datasets for the analysis of NPE in English. However, there are many well-known corpora that contain annotated instances of NPE. One such resource is the Universal Dependency (UD) tree-

No.	Syntactic Category	Examples
	Can License NPE	
1.	Cardinal Numbers	I read three **chapters** from this book and Mary read [$_{NP}$ four [e]].
2.	Ordinal Numbers	Mary got first **position** in the university and John got [$_{NP}$ second [e]].
3.	Demonstrative Determiners (Plural)	Of all the **candidates** that applied for the job, [$_{NP}$ these [e]] got selected.
4.	Quantifiers (Not all)	Some **students** love physics and [$_{NP}$ some [e]] don't.
5.	Superlative Adjectives	He is the funniest **guy** here. And also [$_{NP}$ the weirdest [e]].
6.	Noun Possessives	That big **car** standing over there is [$_{NP}$ Joey's [e]].
7.	Pronoun Possessives	John is reading my **book** and I am reading [$_{NP}$ his [e]].
8.	Interrogative Determiners	I don't know which **pages** to read and [$_{NP}$ which [e]] to ignore.
	Cannot License NPE	
1.	Adjectives	* I have a big **house** and she has [$_{NP}$ a small [e]].
2.	Demonstrative Determiners (Singular)	* That **pen** belongs to Mary and [$_{NP}$ this [e]] belongs to John.
3.	Articles	* I really liked that **house** and I know you liked [$_{NP}$ the [e]] too.
4.	Quantifiers (Not all)	* There are 50 **students** in my class and [$_{NP}$ every [e]] went to see the movie.

Table 1: Syntactic categories that can and cannot license NPE in English, with examples for each category.

bank (Silveira et al., 2014) for English that contains example sentences for different types of ellipsis such as VPE, NPE, etc. The UD treebank marks NPE by raising the dependents of the elided noun to the position of head in cases where the dependents are overtly marked. Through a simple search for noun dependents that are given the status of noun heads, we get a total of 146 cases of NPE in 120 sentences from the UD treebank. There is another comparatively small corpus called the ParCorFull: a Parallel Corpus Annotated with Full Coreference (Lapshinova-Koltunski et al., 2018) that is dedicated to anaphora. This corpus targets anaphora, but deals partly with NPE cases as well, marking them with a nom-ellipsis tag. A simple search for this tag gives us 5 sentences containing 5 NPE cases. We also pick a total of 80 sentences containing 83 cases of NPE from linguistic textbooks on ellipsis (Lobeck, 1995; Saito et al., 2008; Menzel, 2017; Kim et al., 2019; Corver and van Koppen, 2011) to cover even the infrequently occurring cases. Finally, we randomly pick 132 sentences that do not contain NPE from UD treebank, ParCOrFull and the same linguistic textbooks. Some of these negative samples of NPE contain sentences with ellipsis other than NPE, such as VPE.

In total, we curate a small dataset of 337 sentences, of which 205 sentences have 234 instances of NPE (some sentences contain more than one instance of NPE) and the remaining 132 sentences without NPE. This dataset, albeit small, covers a wide variety of the cases of NPE discussed in the ellipsis literature and will be useful for future work. Since the focus of this paper is on presenting a system for detection and resolution of ellipsis, we do not undertake the formidable task of preparation of annotation guidelines and performing annotation to prepare a Gold dataset in this paper. However, this could be an important future work considering the limited available resources for the analysis of NPE.

To fine tune our rules, we only need positive samples of NPE. However, for testing, it is important to use both positive and negative samples. We split our dataset into two parts. For fine tuning the rules, we randomly pick 140 out of the 205 positive sentences (roughly 70%), containing a total of 158 NPE instances. The remaining 65 sentences (roughly 30%) contain 76 NPE instances and are included in the testset along with the 132

sentences without NPE. Hence, our testset has 76 positive and 132 negative samples.

5 System Overview

Our system is divided into two parts. An input sentence is fed into the NPE detection system that decides whether an NPE is present or not. If an NPE is detected, it sends the sentence to the Resolution system where a potential antecedent is selected. The output of the complete system is either a decision that there is no ellipsis present in the sentence or an ellipsis site marked along with its antecedent.

5.1 NPE Detection

For the task of NPE detection, we exploit a very useful syntactic feature, which is the presence of overt remnants of the noun phrase at the ellipsis site. In case of NPE in English, these trigger words are often determiners and modifiers of the elided noun. These are also known as licensors of ellipsis. In the examples presented in (1) and (2), the licensors of elided noun are the cardinal number *two* and possessive proper noun *Mary's* respectively. We use these remnants or noun modifiers present at the ellipsis site as cues to locate the elided noun.

An interesting feature about these license of NPE in English is that they can only belong to certain syntactic categories. These include cardinal and ordinal numbers, plural demonstrative determiners, possessives, adjectives, quantifiers, and a certain types of determiners. Table 1 provides examples from each of these categories, along with examples of syntactic categories that cannot license NPEs in English. Hence, the idea is to use the syntactic environment of the nominal ellipsis site to perform detection. Linguistically, our approach is similar to detection of VPE by using auxiliary and modal verbs as cues (McShane and Babkin, 2016), as VPE in English are licensed by auxiliary and modal verbs.

5.1.1 Look for Pre-Modifiers and Determiners

NPE detection is carried out in two steps. In the first step, the input sentence is parsed using the state-of-the-art spaCy parser (Honnibal and Johnson, 2015) and using the Part-of-Speech (POS) tags, we check for the presence of nominal modifiers and determiners from the aforementioned syntactic categories that can potentially license an

Figure 1: F1-score corresponding to different window sizes for searching nouns.
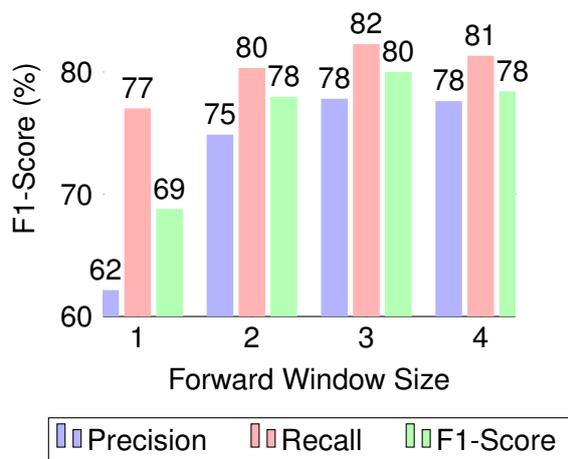
NPE. If the system detects POS tags corresponding to any such category, it proceeds to the second step.

5.1.2 Filter Using Syntactic Features

In this step, the system decides if the selected noun modifier is a licensor of an elided noun or not. This decision is taken using the following syntactic features:

(a) Search for Noun Heads

This simple feature looks for a noun word after the selected noun modifier. We check nouns in a forward search window of 3 words. This window is forward because in English noun heads follow their dependents. If there is no noun present in the next 3 words, the system marks the modifier in question as a licensor. The optimum size of the window as 3 is obtained after experimentation with different window sizes on the 100 sentences from the curated dataset. Figure 1 presents the results of experiments done with different forward window sizes searching for noun heads after noun dependents.

(b) Check for Noun Modifiers as Verbal Arguments

The feature looks for verbs with the selected noun modifiers as the main argument. This is because, many times, spaCy raises a noun modifier to the position of a head in the absence of its noun head, which confirms the presence of an elided noun.

(c) Check for Punctuation

We first check for a simple feature that checks for noun modifiers close to punctuation marks. Since a punctuation can indicate a sentential or phrasal break, this could indicate the absence of a noun head for the given noun modifier in the sentence or noun phrase respectively.

(d) Check for Prepositions

We check if the selected noun modifier is immediately followed by a preposition as that would indicate the beginning of a new (prepositional) phrase and imply that the noun modifier in the given noun phrase does not have a noun head overtly present.

(e) Check for Verbs and Auxiliaries

We also 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.

5.2 NPE Resolution

Ellipses can be resolved textually when their antecedents are present in the same text as in (1). Such cases of ellipsis are called *endophoric*. However, 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 pencils in a shop and uttering a sentence such as (4).

4. Give me three [e].

Using visual context, the shopkeeper can easily resolve the ellipsis in this sentence as *three pencils*. Such cases of ellipses are called *exophoric* or situational as they need situational context to resolve. Since we are only limited to text processing at this stage in the current paper, we only focus on resolving NPE that have textual antecedents.

5.2.1 Ellipsis-Antecedent Environment

It is shown that clauses that are linked by an ellipsis-antecedent relation often have similar syntactic structure and priming effects (Xiang et al., 2014). Further, there is evidence that parallelism in discourse can be applied to resolve possible readings for VPE and possibly other ellipsis and

reference phenomenon (Hobbs and Kehler, 1997). This becomes our motivation to resolve NPE.

(a) Match POS tags of the Licensor with other Noun Modifiers

The syntactic environment of the an NPE comprises the remnants left in the noun phrase. One simple way to see structural parallelism between the syntactic environment of antecedent and that of ellipsis to locate antecedents is through matching the syntactic category information. From the detection task, we already have the POS tag information of the licensor of the NPE. In the first step of antecedent selection, the system checks for other noun phrases in the sentence that contain the modifiers with the same POS tag as that of the licensor of the NPE.

(b) Select Antecedent

If a POS tag matching the licensor of the NPE (detected in the NPE detection task) 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 distance generally has a role to play in anaphora and coreference resolution tasks (Lappin, 1996).

6 Results

Simply looking for nouns in the context of the noun modifiers gives a poor F1-score of 64.20%. Addition of only the feature that checks for noun modifiers raised as verbal arguments results into an increased in F1-score by 10.25%. Addition of only the punctuation feature after auxiliary and modal verbs resulted into an increase in accuracy by 3% for VPE detection task (Nielsen, 2004). However, in our task, this resulted into a drop in accuracy by 0.05%. Hence, we excluded this feature from the final system. Addition of only the feature that checks for prepositions immediately following the noun modifier results into an

increase in F1-score by 13.73%. Finally, addition of only the feature that checks for verbs and auxiliaries that immediately follow the noun modifier gives an increase in F1-score by 11.46%. These features are independent of each other and do not follow any hierarchy.

The final system together with only the significantly important features is tested on the testset containing 76 positive and 132 negative samples of NPE. The detection system is able to correctly detect 65 instances of NPE out of 76. It also rightly predicts 113 out of 132 negative samples as not containing any NPE. It fails to detect 11 positive cases and falsely detects 29 others. This gives us a final precision of 69.15%, a recall of 85.53% and an F1-score 76.47%. Out of the 65 NPE cases detected by the system, 41 have a textual antecedent and the remaining 24 are exophoric and need extra-linguistic context to resolve. Our system is able to select a potential antecedent for 37 of these from the text, of which 32 are correct predictions. The system fails to select any antecedent for the 9 cases. This gives us a final precision of 78.79%, a recall of 63.41% and an F1-score 70.27%. See table 2 for precision, recall and F1-score values for the NPE detection and resolution tasks on the testset.

One of the main reasons of the low accuracy of our system is wrong POS tags generated for sentences with missing or incomplete information as in the case of ellipses (Menzel, 2017). Secondly, although, licensors of NPE and modifiers of the antecedent indeed show similarity in terms of syntactic category information, this might not always be the case.

5. The books were new, and all six [e] were on syntax.

For example in (5), the NPE licensor is a cardinal number, but antecedent *books* has the definitive article in its Noun Phrase.

7 Conclusion & Future Work

This paper described a syntax-based system for automatic detection of NPE in English. The sys-

Task	Positive Samples	Negative Samples	Precision	Recall	F1-Score
Detection	76	132	69.15%	85.53%	76.47%
Resolution	65	29	78.79%	63.41%	70.27%

Table 2: Performance of NPE detection and resolution systems on the testset.

tem takes in free English text and exploits syntactic constraints on the licensors of NPE to mark the site of ellipsis and syntactic parallelism between antecedent-ellipsis syntactic environments to select potential antecedents. Evaluated on a testset containing both positive and negative NPE samples, the system achieves an F1-score of 76.47% on the detection task and 70.27% on the resolution task. Although these numbers are not high, they can be useful as baselines for future work in this direction. NLP research on ellipsis and NPE in particular suffers from a scarcity of resources. While a rule-based system such as ours does not need sizable data for training, with more language resources available in future, machine learning methods can also be used.

References

Yoshihiko Asao, Ryu Iida, and Kentaro Torisawa. 2018. Annotating zero anaphora for question answering. In *LREC*.

Johan Bos and Jennifer Spenader. 2011. An annotated corpus for the analysis of vp ellipsis. *Language Resources and Evaluation*, 45(4):463494.

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

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

Norbert Corver and Marjo van Koppen. 2011. Np-ellipsis with adjectival remnants: A microcomparative perspective. *Natural Language and Linguistic Theory*, 29.

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

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

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

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

Tilbe Goksun, Tom W. Roeper, Kathy Hirsh-Pasek, and Roberta Michnick Golinkoff. 2010. From noun-phrase ellipsis to verbphrase ellipsis: The acquisition path from context to abstract reconstruction.

Christine Gunther. 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.

Michael Alexander Kirkwood Halliday and Ruqaiya Hasan. 1976. Cohesion in english. page 76.

Daniel Hardt. 1997. An empirical approach to vp ellipsis. *Computational Linguistics*, 23(4):525541.

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

Jerry R. Hobbs and Andrew Kehler. 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.

Matthew Honnibal and Mark Johnson. 2015. An improved non-monotonic transition system for dependency parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1373–1378, Lisbon, Portugal. Association for Computational Linguistics.

Nina Hyams, Victoria Mateu, and Lauren Winans. 2017. Ellipsis meets wh-movement: sluicing in early grammar.

Ryu Iida, Kentaro Inui, and Yuji Matsumoto. 2007. Zero-anaphora resolution by learning rich syntactic pattern features. *ACM Trans. Asian Lang. Inf. Process.*, 6.

Kyle Johnson. 2001. What vp ellipsis can do, and what it cant, but not why. pages 439–479.

Christopher Kennedy and Branimir Boguraev. 1996. Anaphora for everyone: pronominal anaphora resolution without a parser. In *In Proceedings of COLING*, page 113118.

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

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

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

Shalom Lappin and Herbert Leass. 1994. A syntactically based algorithm for pronominal anaphora resolution. *Computational Linguistics*, 20:535561.

Ekaterina Lapshinova-Koltunski, Christian Hardmeier, and Pauline Krielke. 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*.

Charlotte Lindenbergh, Angeliek van Hout, and Bart Hollebrandse. 2015. Extending ellipsis research: The acquisition of sluicing in dutch. *BUCLD 39 Online Proceedings Supplement*, 39.

Zhengzhong Liu, Edgar Gonzalez, and Dan Gillick. 2016. Verb phrase ellipsis detection using automatically parsed text. pages 32–40.

Anne Lobeck. 1995. *Functional Heads, Licensing, and Identification*. Oxford University Press.

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

Marjorie McShane, Sergei Nirenburg, and Petr Babkin. 2015. Sentence trimming in service of verb phrase ellipsis resolution. In *EAPCogSci*.

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

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

Jason Merchant. 2010. *Three Kinds of Ellipsis: Syntactic, Semantic, Pragmatic?*

Hwee Tou Ng, Yu Zhou, Robert Dale, and Mary Gardiner. 2005. A machine learning approach to identification and resolution of one-anaphora. pages 1105–1110.

Leif Arda Nielsen. 2003. Using machine learning techniques for vpe detection.

Leif Arda Nielsen. 2004. Verb phrase ellipsis detection using automatically parsed text.

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

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

Alain Rouveret. 2012. Vp ellipsis, phases and the syntax of morphology. *Natural Language & Linguistic Theory*, 30(3):897963.

Mamoru Saito, Jonah Lin, and Keiko Murasugi. 2008. Nominal-ellipsis and the structure of noun phrases in chinese and japanese.

Sebastian Schuster, Joakim Nivre, and Christopher D. Manning. 2018. Sentences with gapping: Parsing and reconstructing elided predicates. *ArXiv e-prints*.

Natalia Silveira, Timothy Dozat, Marie-Catherine de Marneffe, Samuel Bowman, Miriam Connor, John Bauer, and Christopher D. Manning. 2014. A gold standard dependency corpus for English. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*.

Frank Wijnen, Tom W. Roeper, and Hiske van der Meulen. 2003. Discourse binding: Does it begin with nominal ellipsis?

Ming Xiang, Julian Grove, and Jason Merchant. 2014. Ellipsis sites induce structural priming effects.

Ching-Long Yeh and Yi-Chun Chen. 2019a. Using zero anaphora resolution to improve text categorization.

Ching-Long Yeh and Yi-Jun Chen. 2019b. An empirical study of zero anaphora resolution in chinese based on centering model.