The Mechanism of *Pro*-drop: Quantifying The Discourse Support For *Pro*-drop Across Languages Based on Statistical Models

by

Shulin Zhang

(Under the Direction of John T. Hale)

Abstract

This dissertation delves into the phenomenon of *pro*-drop, where pronouns can be grammatically omitted in certain languages. The research focuses on three key questions: (1) How does discourse coherence support *pro*-drop languages, and how can we quantify discourse coherence? (2) What is the order of importance among linguistic factors relevant to *pro*-drop? (3) If an AI model is trained to predict *pro*-drop based on language dependency structures, what factors are crucial cross-linguistically? These questions are addressed in three experiments.

Experiment 1 quantifies verb usage continuity to compare character salience in *pro*-drop and non-*pro*-drop cases across Chinese (CN), Brazilian Portuguese (BP), and Spanish (ES). Results indicate that *pro*-drop cases necessitate higher character salience in verb usage continuity, highlighting the role of discourse coherence. Differences exist among languages, with radical *pro*-drop languages demanding more coherence than partial *pro*-drop and consistent *pro*-drop languages.

Experiment 2 employs Binomial Logistic Regression and Random Forests to model *pro*-drop based on syntactic, semantic, morphological, and logical features. Findings reveal the significance of character consistency between main-embedded and current-previous clauses in all languages (Chinese, Brazilian Portuguese, and Spanish). ES and BP exhibit richer verbal morphology, potentially contributing to a higher model fit than Chinese. Sentential discourse relation features, such as coordinate structures, encourage *pro*-drop in CN and ES.

Experiment 3 employs Graph Attention Networks (GATs) to classify story characters using dependency structure graphs from CN, BP, and ES discourse material. Results align with key elements in pronoun resolution, highlighting the importance of subjects, main verbs, and objects. Language-specific importance emerges, with adverbial expressions (*e.g.* '就'(jiu4, means "just")) and auxiliary (*e.g.* '了'(leo, means "already"), '着'(zheo, means "going on")) crucial in CN, while determiner and case play vital roles in BP and ES.

INDEX WORDS: [*pro*-drop, zero pronoun, null subject, discourse coherence, character salience, Chinese, Brazilian Portuguese, Spanish, word embedding, Binomial Logistic Regression, Random Forest, Graph Neural Networks]

The Mechanism of *Pro*-drop: Quantifying The Discourse Support For *Pro*-drop Across Languages Based on Statistical Models

by

Shulin Zhang

B.E. in Electronic Engineering, Dalian University of Technology, China, 2014

M.S. in Applied Psychology, Southwest University, China, 2016

M.S. in Artificial Intelligence, University of Georgia, US, 2020

A Dissertation Submitted to the Graduate Faculty of the
University of Georgia in Partial Fulfillment of the Requirements for the Degree.

Doctor of Philosophy

Athens, Georgia

2024

©2024
Shulin Zhang
All Rights Reserved

The Mechanism of *Pro*-drop: Quantifying The Discourse Support For *Pro*-drop Across Languages Based on Statistical Models

by

Shulin Zhang

Major Professor: John T. Hale

Committee: Chad Howe

Vera Lee-Schoenfeld

Electronic Version Approved:

Ron Walcott

Dean of the Graduate School

The University of Georgia

May 2024

Acknowledgments

The five-year study at UGA's Linguistics program has been a priceless growing experience for me, and so many people have helped and supported me. First, I would like to thank my advisor Professor John Hale, for his guidance during my PhD study. He shares knowledge with his students and teaches us how to become a researcher. I appreciate Professor Hale's patience with my process of exploration in my research and his unconditional support of my interests in research projects. I would like to thank my committee members, Professor Vera Lee-Schoenfeld and Professor Chad Howe. Their advice and support for my research time are precious and always helpful.

I want to thank Dr. Jixing Li, for sharing the fMRI data during my first qualification study and kindly sharing experience and knowledge with me. For my dissertation study, I would like to thank my friends and collaborators Jean Costa-Silva and Cole Bryant, for their dedicated help in annotating the Brazilian Portuguese and Spanish language materials.

I want to thank all the professors in the linguistics department, and thank Professor Margaret Renwick, Professor Pilar Chamorro, Professor Jon Forrest, and Professor Mark Wenthe, for giving great lectures and guidance, I learned so much from you. Everyone I met in the UGA linguistics department is making it a wonderful place to study, grow, and explore. They are curious and respectful of different cultures, languages, and people from different backgrounds, and they hold the highest belief and dedication in the humanity research field, I feel honored to study here.

Last but not least, I want to thank my parents, family, and friends for their dedicated support for my achievement. I sincerely appreciate Preston and Sam, your support means so much to me. My friends, Weiwen Xu, Zenan Li, Xunan Shen, Shuang Yang, Yawei Shen, Chenxiao Li, Hui Du, you all are the best company and have made our time spent together so precious.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 What Is *Pro*-drop

Pro-drop is a phenomenon commonly seen in languages such as Chinese, Spanish, Portuguese, and Italian. In *pro*-drop languages, pronouns, or named entities can be omitted whereas native speakers can understand the dropped entities based on context information or world knowledge. The dropped pronouns are often referred to as zero pronouns in previous studies. The main property of zero pronouns is that they are not overtly pronounced, and we are intrigued to understand "what makes them understandable", and that motivates this study to explore the mechanism of *pro*-drop with a focus on null subjects.

As shown in the following examples, "[ ]" indicates that there is a *pro*-drop happening in the clause, and the entity being dropped is shown in the row below. In the Spanish example (1), the verb *vio* is a form of *ver*('see') and bears the ϕ-features [3rd, sing, past]. In the Brazilian Portuguese example (2), the verb *falou* is in a form of *falar*('speak') and bears the ϕ-features [3rd, sing, past]. In the Italian example (3), the verb *vide* is in a form of *vedere*('see') and bears the ϕ-features [3rd, sing, past]. These three languages, Spanish, Brazilian Portuguese, and Italian, are known to have comparably richer subject-verb agreement information to indicate "who has been dropped". These languages have different levels of morphological agreement (mentioned as "verb syncretism level" in the following chapters when referring to the level of morphological feature richness in the verb), and this agreement level plays a role in languages' *pro*-drop behavior. On the other hand, as shown in the Chinese example (4), the verb "kan"(看) itself does not carry a person or tense or singular number feature, and the dropped subject can be inferred from the

discourse context. Chinese is therefore known as a radical *pro*-drop language with no morpho-syntactic information or conjugation provided by the verb inflection to indicate the omitted subject. To resolve the omitted subject in a radical language, more context is required, and the so-called Topic Chain theory was brought up to describe this contextual support for *pro*-drop (W. Li, 2004; K. Sun, 2019) (see Section 2.1.3 for a review of the Topic Chain theory).

(1)	ES:	*[The fox]*	*saw the Little Prince.*
		[ ]	Vio al Principito.
		El zorro	vio al Principito.
(2)	BP:	*[The fox]*	*spoke to the little prince*
		[ ]	Falou com o Pequeno Principe.
		A raposa	falou com o Pequeno Principe.
(3)	IT:	*[The fox]*	*saw the Little Prince.*
		[ ]	vide il Piccolo Principe.
		La volpe	vide il Piccolo Principe.
(4)	CN:	*[I]*	*saw a fabulous painting.*
		[ ]	kan dao le yi fu jingcai de chahua.
		Wo	kan dao le yi fu jingcai de chahua.

Intuitively speaking, native speakers of a *pro*-drop language would drop a pronoun when they think it is "obvious" to be inferred from the context. There are previous theories proposing a rationale for *pro*-drop, such as Ariel's Accessibility Theory (Ariel, 2001), and Almor's Informational Load Hypothesis (ILH) (Almor, 1999). A more in-depth review of *pro*-drop is presented in Chapter 2.

1.2 The Issue: Quantifying Linguistic Factors of *Pro*-drop

The early studies on *pro*-drop focused on European languages with rich agreement levels, and concluded that *pro*-drop is supported by the agreement feature in the languages. However, as languages such as Chinese and Japanese caught researchers' attention, it was realized that the agreement factor cannot explain *pro*-drop in these languages without morphological variation on the verb, which means that agreement cannot explain the radical *pro*-drop phenomenon. Later on, semantic factors, such as the Topic Chain theory, were introduced to explain the radical and partial *pro*-drop phenomenon.

However, previous studies have not yet provided a **statistical** description of (1) the features that distinguish *pro*-drop and non-*pro*-drop scenarios; (2) quantifying the feature difference between radical

vs. partial *vs.* consistent *pro*-drop languages; (3) quantifying the level of importance for different linguistic factors that affect *pro*-drop; (4) understanding how *pro*-drop resolution in large language models can help us understand *pro*-drop. Therefore, the current studies focus on shedding light on these questions and provide a framework to bridge computational and theoretical ideas on the *pro*-drop phenomenon.

1.3 Road Map of The Studies: Studying *Pro*-drop with Computational Models

Although the study of *pro*-drop and its factors has a long tradition in linguistics, the lack of statistical description of this phenomenon has motivated this current series of studies: exploring the mechanism of *pro*-drop using mathematical feature representation and machine learning algorithms. This angle of computational linguistics can be meaningful in (1) understanding the *pro*-drop features in a quantitative way, and comparing the importance level quantitatively within each language; (2) making it possible to compare feature significance across languages, and visualize the typological categories of *pro*-drop (*i.e.* radical, partial, consistent, etc.); (3) proposing potential linguistic features to be adopted in NLP pronoun resolution tasks; (4) providing a framework for including more features in future studies for *pro*-drop and other theoretical concepts.

In the following chapters, a literature review on *pro*-drop is carried out in Chapter 2, and three experiments adopting computational methods are the basis for reports on the mechanism of *pro*-drop across Chinese, Brazilian Portuguese, and Spanish in Chapter 4, 5, 6. The discourse materials used in these experiments are Chinese, Brazilian Portuguese, and Spanish translations (xiaowangzi.org, 2021) of Saint-Exupéry's *The Little Prince*.

1.3.1 Experiment 1: Quantifying Verb Usage Continuity as Discourse Support for Omitted Pronouns

In Experiment 1 (Chapter 4), the factor of verb continuity in the discourse on *pro*-drop is explored in Chinese (CN), Brazilian Portuguese (BP), and Spanish (ES). Verb continuity describes how the usage of similar verbs can support *pro*-drop, and it is measured with a cosine similarity of the verbs' word embed-

dings (BERT and GloVe). The continuity level is compared between *pro*-drop and non-*pro*-drop cases. This experiment is inspired by the Topic Chain theory: Given that there is a "chain" existing throughout the discourse, how can we describe it mathematically? Since the subject is dropped for *pro*-drop cases, the "S-V-O" structure's "V" component becomes a promising element to support the chain. Therefore, the verb continuity factor was intended to create a mathematical representation to describe how the "chain" can be formed across the verbs. In the results, within each language, it is shown that the verb continuity can distinguish *pro*-drop and non-*pro*-drop cases in all three languages, and the *pro*-drop cases demand higher verb continuity than the non-*pro*-drop cases. Cross linguistically, the *pro*-drop cases' verb continuity levels among the three languages are: CN $>$ BP $>$ ES, which indicates the agreement (Agr) level plays a role. The higher the agreement level of a language the less it relies on other semantic factors such as verb continuity to resolve a *pro*-drop case.

1.3.2 Experiment 2: Modeling Linguistic Factors for *Pro*-drop Using Binomial Logistic Regression and Random Forest Models

In Experiment 2 (Chapter 5), a broader set of linguistic features are examined as regards their effects on *pro*-drop. This examination applies two machine learning models: Binomial Logistic Regression and Random Forest. The factors included in this experiment are rooted in syntactic, semantic, morphological, and logical aspects. The Binomial Logistic Regression Model and the Random Forest model provide significant levels of the features on their capability to tell the difference cases between *pro*-drop and non-*pro*-drop throughout the discourse. In the results, it is shown that (1) verb consistency between sub-main (or embedded-matrix), current-previous clauses is significant in all three languages; (2) Spanish and Brazilian Portuguese have richer verb agreement to help recover the dropped pronoun which might indicate that these languages' models have a higher fit level than that of Chinese; (3) sentential discourse relation features such as coordinate structures can encourage *pro*-drop, and it was found significant in both CN and ES.

1.3.3 Experiment 3: Modeling *Pro*-drop Features with Attention Values in Graph Neural Networks

In Experiment 3 (Chapter 6), Graph Attention Networks (GATs) are trained to classify story characters when they are mentioned in the Subject position, and the training materials were built with dependency structure graphs from CN, BP, and ES discourse material. First, the results in GATs are consistent with the main elements in pronoun resolution: the GATs found the elements including the subject, the main verb, and the object take leading roles among all other dependency types. Second, the elements shown importance in CN are different from the ones in BP and ES: marker and auxiliary are more important in CN; determiner and case are more important in BP and ES. This is consistent with the innate language features.

1.3.4 Contribution and Significance for Linguistics

These series of experiments provided the rationale and proved the possibility of using machine learning models to study linguistic phenomena. It is possible to find out the consistency and difference between models and linguistic theories, even neurolinguistic studies.

Chapter 2

Literature Review of *Pro*-drop Studies

2.1 *Pro*-drop Studies in Various Subfields of Linguistics

In this chapter, a literature review for *pro*-drop studies is carried out to present previous studies on this topic. As such, various linguistic subfields' studies on *pro*-drop are reviewed in the following subsections. Section 2.3.1, 2.3.2, and 2.3.3 provide more details of the *pro*-drop studies in Chinese, Spanish, and Brazilian Portuguese, respectively, and these three languages are used as corpus materials in the experiments presented in the following chapters.

The phenomenon of *pro*-drop has been studied in many subfields of linguistics, including but not limited to Typology, Syntax, Semantics, Morphology, Language Acquisition, Neurolinguistics, and Historical Linguistics. In the following subsections, a brief overview will be given of how *pro*-drop is studied in these subfields. It should be noted that the studies mentioned in the subfields' sections are not clear-cut belonging to one area (*e.g.* the discussion of morphological complexity is also relevant to syntactic agreement level), and these subsection divisions are simply there to review the studies in an organized manner.

2.1.1 Typological Aspects of *Pro*-drop

Pro-drop is a widely recognized typological parameter among languages in linguistic universals and the Principles and Parameters theory (Geeslin, 1999). The *pro*-drop parameter describes languages with respect to whether they do (+) or do not (-) require an overt subject, and it is often referred to as "[+/- *pro*-drop]".

A typology of *pro*-drop languages. It should be noted that, on the sentential level, the *pro*-drop parameter can be considered binary, but on the language level it is not fully binary depending on the null subject frequency. For example, many East Asian languages, such as Mandarin Chinese, Japanese, and Korean, are known as radical *pro*-drop language, in which pronouns can be freely omitted in both subject and object positions given proper discourse context and there is no morphological change (*i.e.* inflection) on the verbs indicating the ϕ-features of the null subject (C.-T. J. Huang, 1989). On the other hand, languages such as Spanish and Italian have consistent null subjects, and the subjects can be generally dropped; and languages such as Finnish and Hebrew can drop some subjects, but not all. Therefore, a broader spectrum of *pro*-drop types can be represented as shown in Table 2.1 (P. Barbosa, 2011b; P. P. Barbosa, 2011; Biberauer et al., 2009). The *pro*-dropping behavior is related to many linguistic features, such as the level of agreement (see Table 2.2 as an example), which varies across languages.

Table 2.1: *pro*-drop / null-subject (NS) language types and examples

Type	**Features**	**Language Examples**
Non-pro-drop	Subjects cannot be dropped	English
Semi pro-drop	Languages that only have non-referential NSs. Expletive subjects can be dropped.	German, Dutch, Yiddish, Icelandic, Faroese, a range of creoles
Partial pro-drop	Languages that have agreement and referential NSs whose distribution is restricted. Some, but not all (e.g. 3rd person pronoun), referential subjects can be dropped.	Hebrew, Finnish, Marathi, Russian, colloquial Brazilian Portuguese.
Consistent pro-drop	Languages with rich subject agreement morphology. Referential subjects can be generally dropped under the appropriate discourse conditions.	Italian, Spanish, Portuguese, Hungarian, Greek, among many others
Radical pro-drop	Languages that lack agreement.Referential subjects and objects can be dropped in the absence of inflectional morphology. These languages have been described as topic-oriented languages and allow for any argument to be dropped not just subjects.	Chinese, Japanese, Korean

Table 2.2: Agreement features across languages (J. Zhang, 2016)

Group	***pro-drop languages***	***Agr***
A	Italian, Spanish etc.	++ Agr
B	German, Scandinavian, Modern Hebrew, Turkish, Esperanto, Occitan, Catalan, Portuguese, Romanian (except French), Croatian, Brazilian Portuguese, Finnish, Marathi etc.	+ Agr
C	Chinese, Korean, Japanese etc.	- Agr

Although the typological categories of *pro*-drop in null subject languages have been well established, there are arguments that these distinct types should be merged as a single parameter, "argument ellipsis", to describe the null subject languages (Duguine, 2014). However, these arguments do not counter the fact that the *pro*-drop frequencies and morphosyntactic features vary among languages. They just try to discuss methodological issues based on a unity hypothesis.

Zero pronoun's typology. As *pro*-drop happens, the dropped pronoun is often referred to as a *zero pronoun*. Zero pronouns have three types in general based on their relationship with the referents: anaphoric , cataphoric, and exophoric. "Anaphoric" zero pronouns (*i.e.* zero anaphora) refer upward to previously mentioned words; "Cataphoric" zero pronouns (*i.e.* zero cataphora) refer downward to subsequent words; "Exophoric" zero pronouns (*i.e.* zero exophora) refer to something outside the context. A subset of anaphoric elements, which typically occurs in embedded clauses referring to a co-referential subject in the matrix clause, is called logophoric zero pronouns (*i.e.* zero logophora). A pronoun being anaphoric, cataphoric, or exophoric can be either a zero pronoun or a non-zero pronoun, and these features describe their relationship with the referents in the discourse but not their *pro*-drop feature. If a zero pronoun and its referent are in the same sentence, it is "intrasentential"; if they are in different sentences, the zero pronoun is "intersentential".

2.1.2 Syntactic Aspects of *Pro*-drop

Within Generative Grammar, the *pro*-drop phenomenon started to gain attention as the Government and Binding (GB) Theory developed (Chomsky, 1981). GB theory investigates the nature and distribution of phonetically null but syntactically present categories, named Empty Categories (EC). Chomsky proposed the Extended Projection Principle (EPP) which indicates that all clauses must have a structural subject, and therefore triggers subject (NP-) movement from the verbal domain into the inflectional layer of

structure (Spec TP). All empty nominal categories, including NP-traces, *wh*-traces, PRO (the null subject of non-finite clauses), and pro, belong to the same EC.

Early observation of *pro*-drop mainly focused on languages with rich agreement inflection. Agreement, abbreviated as *Agr*, takes place when a word undergoes a change in form based on its relation to other words. The *Agr* parameter was considered to be the distinction between *pro*-drop and non-*pro*-drop languages. However, this was called into question as more analyses of *pro*-drop languages came out that lack agreement features, and it became clear that *Agr* was not a sufficient condition to predict *pro*-drop, and it was shown that an inherently unspecified nominal can not inherit its feature from *Agr* (C.-T. J. Huang, 1984). In the mid-90s, the Pronominal-*Agr* hypothesis was derived to claim that *pro* can be reduced either to a trace or to PRO.

As the *Agr* feature for *pro*-drop was challenged by radical *pro*-drop languages, the presence of "pro" in the subject position of finite clauses became feasible, despite the assumption that *Agr* was missing and no agreement morpheme was generated under *Infl*. To explain this limitation, attempts at an explanation have been made coming from different perspectives. Huang proposed a theory based on two parameters: zero topic and silent pronominal argument (C.-T. J. Huang, 1984). The zero topic parameter tries to explain the cases where null arguments are not bound by the closest c-commanding argument. Following Huang's theory, in Rizzi's work (Rizzi et al., 1986), a distinction is made between the syntactic process of licensing "pro" and the semantic process of identifying it. Licensing pertains to the syntactic realm, while identification involves the semantic aspect, where "pro" is recovered from prior discourse. Later studies adopted more semantic processes to explain features in radical *pro*-drop (see the next subsection for a review of the semantics of *pro*-drop) (Tomioka, 2003).

Tomioka's (2003) work raised various reactions in follow-up studies. For example, on the one hand, Neeleman and Szendrői (2007) interpreted Tomioka's main argument to be that "all languages that allow radical *pro*-drop allow (robust) bare NP arguments" and brought up counter-examples that allow *pro*-drop in the absence of agreement and require referential NPs to be accompanied by determiners. Neeleman and Szendrői consider *pro*-drop a spell-out phenomenon and claim that the implicit subject is a fully specified pronoun that either gets deleted during the Phonetic Form (PF) stage or does not receive a realization at PF. On the other hand, Barbosa (2019) extends Tomioka's theory and states that *pro* reduces to a bare

nominal in argument position or Null NP anaphora. Barbosa examines this reduced form of *pro* in radical, partial, and consistent *pro*-drop languages.

As mentioned above, early studies on *pro*-drop focused on the agreement level among languages (mainly European languages with rich morpho-syntactic agreement features), and as focus turned towards languages with fewer subject-verb agreement features, the contribution of semantic resources at the sentential and discourse level became more prominent in *pro*-drop language studies.

2.1.3 Semantic Aspects of *Pro*-drop

As mentioned in the previous subsection reviewing the syntactic studies of *pro*-drop, the *pro*-drop parameter originated from studying consistent *pro*-drop languages with rich inflections. Later on, when more attention was put on radical *pro*-drop languages, ideas that involved semantic factors were brought into the discussion since these languages lack rich agreement features. For example, Huang (1984) suggested including "zero topic" as a parameter, and Tomioka (2003) brought in semantic processes, property anaphora (type $< e, t >$), to demonstrate that the same semantic techniques employed for comprehending full noun phrases can also be applied to understand pronouns.

Topic chain and *pro*-drop. Topic chain is an essential concept that was brought up to explain radical *pro*-drop. A topic chain is a chain of clauses sharing an identical topic that occurs overtly once in one of the clauses, and its boundary may cross several sentences and even paragraphs (W. Li, 2004). The topic chain can integrate information from multiple clauses (K. Sun, 2019), which makes long-distance coreference between zero pronouns and overt noun phrases possible. We can understand coreference resolution as searching for an appropriate antecedent in a topic chain.

Previous studies have discussed the role of the topic chain in supporting *pro*-drop in Chinese (see Section 2.3.1 for more studies reviewed). Besides radical *pro*-drop languages, Frascarelli and Jiménez-Fernández's (2019) study explored the role of topic chain and silent topic in partial *pro*-drop languages. In their study, the extent of partiality is ascribed to an interface condition that merges information-structure requisites with the phonological visibility of explicit duplicates within topic chains. The characteristics of partial pro-drop can consequently be clarified by invoking distinct syntactic conditions, including the inclination towards explicit minimal links and the responsiveness to islands.

The perception of the boundary of a topic chain can be subjective. Frascarelli and Casentini (2019) provided an in-depth theoretical review of the topic criterion and the formation of topic chains. The topic criterion is based on the analysis of spoken corpora, and states that "there is evidence that the interpretation of a referential pro depends on a matching relation with a specific type of topic: the so-called Aboutness-shift (A)-Topic." There is also the Interface Root Restriction (Bianchi & Frascarelli, 2010), and it states that an A-topic is restricted to root clauses that are endowed with illocutionary force. Therefore, in their criterion, a topic chain can only be started from a root (or root-like) clause.

Verbs' semantic features and *pro*-drop. For consistent *pro*-drop languages, there are studies carried out to explore the semantic factors from the aspect of verb features. For example, Herbeck (2021) examined the effect of the perspectival factor on *pro*-drop in Spanish. The factors include the usage of speaker and addressee pronouns in conjunction with cognitive verbs.

2.1.4 Morphological Aspects of *Pro*-drop

Although agreement level seems relevant to *pro*-drop in some early studies, it is not a robust feature to predict a language as *pro*-drop (Müller, 2007). Koeneman and Zeijlstra (2019) provided an in-depth review of the relationship between morphology and *pro*-drop, and compared verb agreement levels in the languages including Italian, English, Danish, Icelandic, and Dutch (see Table 2.3). On the one hand, they pointed out that high verb agreement does not necessarily lead to *pro*-drop for a language. For example, as shown in the table, although Icelandic has fairly high agreement richness, it is not a *pro*-drop language. On the other hand, a low verb agreement level does not guarantee that a language is radical *pro*-drop. For example, there are languages that do not have any person/number marking on the verb, such as mainland Scandinavian languages (Norwegian, Danish, and Swedish), as well as Afrikaans, but these languages are not radical *pro*-drop.

Table 2.3: Verb forms example in Italian, English, and Danish from Koeneman and Zeijlstra (2019) Table (2) and (3).

<table>
<tr><th></th><th>Italian</th><th>English</th><th>Danish</th><th>Icelandic</th><th>Dutch</th></tr>
<tr><td>Inf.</td><td>parl-are
(“to speak”)</td><td>speak</td><td>hor-er
(“to hear”)</td><td>heyra
(“to hear”)</td><td>horen
(“to hear”)</td></tr>
<tr><td>1SG</td><td>parl-o</td><td rowspan="2">speak</td><td rowspan="6">hør-er</td><td>heyr-i</td><td>hoor</td></tr>
<tr><td>2SG</td><td>parl-i</td><td rowspan="2">heyr-ir</td><td rowspan="2">hoor-t</td></tr>
<tr><td>3SG</td><td>parl-a</td><td>speak-s</td></tr>
<tr><td>1PL</td><td>parl-iamo</td><td rowspan="3">speak</td><td>heyr-jum</td><td rowspan="3">hoor-en</td></tr>
<tr><td>2PL</td><td>parl-ate</td><td>heyr-ið</td></tr>
<tr><td>3PL</td><td>parl-ano</td><td>heyr-a</td></tr>
</table>

Koeneman and Zeijlstra (2019) examine three assumptions on the relationship between agreement and *pro*-drop: “(i) the link between pro-drop and agreement is misdirected altogether; (ii) there are two types of pro-drop, agreement-based and non-agreement-based, and they have nothing to do with one another; (iii) there are two types of pro-drop, agreement-based and non-agreement-based, but they are related.” The difference between arguments (ii) and (iii) is whether agreement-based and non-agreement-based *pro*-drop languages are related or not. The evidence for (ii) is that radical *pro*-drop language like Chinese also allows null objects, and it relies on contextual information to recover the null object, and the consistent *pro*-drop language would not allow null object due to the lack of verb-object agreement. With this typological difference taken into consideration, the null object feature sets apart the types of *pro*-drop languages. Argument (iii) leads to the old typological issue brought about by using the agreement feature to categorize *pro*-drop languages, and we know that this cannot explain all the languages.

Since agreement itself is not able to provide a full-coverage typological classification for *pro*-drop subcategories, Koeneman and Zeijlstra (2019) bring in the “*agglutinative pronouns*” feature as discussed earlier by Neeleman and Szendröi (2008). Neeleman and Szendröi (2008) tackle the question why radical *pro*-drop is sensitive to the morphology of pronouns, focusing on the *pro*-drop phenomenon in radical languages including Japanese and Chinese. Their analysis is based on the idea that radical *pro*-drop is zero spell-out of regular pronouns. The proposal is that case is not overtly realized and that the spell-out rules for pronominal stems target a category lower than KP (Case Phrase).

Koeneman and Zeijlstra (2019) summarize the features of different types of *pro*-drop languages, showing how these subcategories of *pro*-drop languages rely on morphological support:

> Discourse (or radical) pro-drop requires agglutinative nominal morphology visible in pronouns, consistent (or agreement-based) null subjects require rich agreement under some definition, partial pro-drop languages require a morphological resemblance between subject agreement affixes and pronouns, whereas expletive and/or generic null subjects may require a morphological uniform paradigm.

Since the traditional notions of morphological paradigms (*i.e.* counting distinct forms and setting up a minimal threshold of relevant distinctions) to determine a language being *pro*-drop is not good enough, Müller (2007) introduces a concept of morphological richness based on impoverishment operations. His theory states that a language can have null subject *pro*-drop only if its system of verb inflection does not involve an impoverishment operation applying to an associated T item in the numeration, and it is expected that languages with subject pro-drop either do not involve any syncretism in their system of verb inflection at all or involve only instances of syncretism which are not system-defining.

2.1.5 Neurolinguistic Studies on *Pro*-drop

In our previous study (S. Zhang, Li, Yang, et al., 2022), we reported an fMRI study of language comprehension processes in zero pronoun resolution and explored how *pro*-drop is processed in the human brain. The fMRI data come from Chinese speakers who listened to an audiobook. We conducted both univariate GLM and multivariate pattern analysis (MVPA) on these data. We found increased left Temporal Lobe activity for zero pronouns compared to overt subjects, suggesting additional effort in searching for an antecedent during zero pronoun resolution. MVPA further revealed that the intended referent of a zero pronoun can be decoded in the Parahippocampal Gyrus and the Precuneus shortly after its presentation. This highlights the role of memory and discourse-level processing in resolving referential expressions, including unspoken ones, in naturalistic language comprehension.

More studies have been done on non-*pro*-drop reference processing (*i.e.* pronoun resolution) in the human brain. There are some fMRI and MEG studies on referential processing in general (Brodbeck et al., 2016; Brodbeck & Pylkkänen, 2017; Hammer et al., 2007; J. Li et al., 2021; Matchin et al., 2014; Nieuwland et al., 2007; Santi & Grodzinsky, 2012). Previous studies have adopted different task manipulations, making it unclear whether they tapped the same cognitive processes. For example, Nieuwland et al. (2007) compared the BOLD responses when participants read sentences containing a referentially failing

pronoun (*e.g.*, "Rose told Emily that *he* had a positive attitude towards life.") or a coherent pronoun (e.g. "Ronald told Emily that *he* had a positive attitude towards life."). Nieuwland et al. showed that referentially failing pronouns were associated with increased activation in the medial parietal regions and bilateral inferior parietal regions, possibly reflecting morphosyntactic processing. Hammer et al. (2007) manipulated syntactic gender matching between the antecedent and pronouns using German sentences and found that gender incongruency elicited the bilateral Inferior Frontal Gyrus (IFG), the left Medial Frontal Gyrus (MFG), and the bilateral Supramarginal/Angular Gyrus compared to congruent pronoun-antecedent pairs. Hammer, Jansma, et al. (2011) further investigated possible interactions between gender and distance between the antecedents and the pronoun. The results identified a fronto-temporal network including the bilateral IFG, the Superior Temporal Gyrus (STG), and the posterior Middle Temporal Gyrus (pMTG) for long-distance conditions, with the pMTG additionally driven by syntactic gender violation. These authors suggested that the temporal regions are sensitive to morphosyntactic information of the antecedents since the long distance between the antecedent and the pronoun increased the overall syntactic complexity of the sentence. Matchin et al. (2014) also examined the effect of distance but with the backward anaphora/filler-gap dependencies contrast. Matchin and colleagues observed specific activity in the bilateral Anterior Temporal Lobes, the bilateral Angular Gyrus (AGs), and the left Precuneus activity during the processing of backward anaphora compared to *wh*-fillers. Santi and Grodzinsky (2012) compared null pronouns, a parasitic-gap and a *wh*-trace in English sentences such as "[Which paper] did the tired student submit [*wh*-trace] after reviewing [parasitic gap/it]?". The results identified increased activity in the right Middle Frontal Gyrus (MFG), the left Ventral Precentral Sulcus, and the Left Supramarginal Gyrus for pronouns compared to parasitic gaps. An activation likelihood estimate meta-analysis on 16 fMRI studies on pronoun resolution suggested that the processing of pronouns demonstrated functional convergence in the left posterior middle and superior temporal gyri, possibly indicating the involvement of these areas in the retrieval, prediction, and integration processes related to pronoun processing (El Ouardi et al., 2023).

In addition to morphosyntactic manipulations, Brodbeck and Pylkkänen (2017) and Brodbeck et al. (2016) used a visual world paradigm in magnetoencephalography (MEG) and found medial parietal activity in cases of successful reference resolution. More relevant to the current study is J. Li et al. (2021)'s study on third-person pronoun processing using the same naturalistic listening paradigm. In both fMRI

and MEG, Li et al. found that the left middle temporal gyrus (LMTG) is consistently activated for third-person pronoun processing in both English and Chinese. Yet they also found additional medial parietal activity from the MEG data, consistent with Brodbeck and Pylkkänen (2017), Brodbeck et al. (2016), and Nieuwland et al. (2007).

Previous Event-related Potentials (ERP) studies found that the initial activation and subsequent discourse-level integration of referents can be dissociated with event-related EEG activity, and are associated with respectively theta- and gamma-band activity (Coopmans & Nieuwland, 2020), and the anaphoric reference is related to P600 effect (Heinat & Klingvall, 2020). Another ERP study found that referential consistency results in a gamma-band increase to the posterior parietal cortex around 400–600 ms after anaphor onset and to the frontotemporal cortex around 500–1000 ms (Nieuwland & Martin, 2017), and new referents in a discourse elicited a subsequent frontal positivity (Nieuwland et al., 2019). In processing subject anaphors, the N400 is an index of reference predictability rather than a marker of the fit between antecedent salience and reference form, and frontal negativity marks referential ambiguity elicited by conjoined phrases (Almor et al., 2017). While searching for antecedents in the pronoun resolution, a larger differential distance between the pronoun and the antecedent elicited N400 and P600 effect and indicated a syntactic reanalysis corresponding to the antecedent searching task (Hammer et al., 2008), and larger brain activation in language neural networks are detected in inferior frontal and temporal regions with fMRI (Hammer, Jansma, et al., 2011). Long-distance dependency and gap filling were found related to activation in the supplementary motor cortex, left supramarginal cortex, precuneus, and anterior/dorsal cingulate (Piñango et al., 2016).

2.1.6 *Pro*-drop in Language Acquisition

There are studies on *pro*-drop in the field of language acquisition from multiple perspectives, such as the process of learning an L2 non-*pro*-drop language for L1 *pro*-drop speakers (D. Liu, 2008), learning an L2 *pro*-drop language as L1 non-*pro*-drop speakers, and acquiring L1 *pro*-drop languages.

Studies on L1 *pro*-drop learning L2 non-*pro*-drop. D. Liu (2008) reviewed the *pro*-drop resetting in ESL (English as a second language) learners, and mentioned that speakers of the Romance pro-drop languages show significantly more difficulty than speakers of the Asian pro-drop languages in learning

non-null subjects in English. Ynoa (2020) analyzed the *pro*-drop errors in L2 English by L1 Spanish speaker, and their study showed that L2 English by L1 Spanish speakers had a higher error rate of *pro*-drop with the 3rd person neuter singular, especially in nominative position, than other pronoun cases.

Studies on L1 non-*pro*-drop learning *pro*-drop. Liceras (1989) reviewed the studies on L2 acquisition of *pro*-drop and investigated French and English speakers learning Spanish in a classroom setting. Their study suggests that while the interpretation of sentences may be affected by language distance, it might not directly impact the resetting of the parameter. Yamada and Miyamoto (2017) compared the European *pro*-drop L1 (Spanish) and non-*pro*-drop L1 learning Japanese. Their study found that, on the one hand, L1 Spanish learning Japanese at the initial stage would allow a lenient interpretation of null arguments; On the other hand, L1 non-*pro*-drop European learners for Spanish did not allow a lenient interpretation with null arguments even at an advanced level. Yamada and Miyamoto (2017) argued that the outcomes for the pro-drop learners stem from positive transfer from their first language, Spanish, which permits a lenient interpretation with null subjects in specific well-defined contexts. White (1985) studies Spanish and French speakers' opinions on null subject English materials, and found that Spanish L1 speakers have higher acceptance of the null subject materials and showed improvement with increasing levels of proficiency, whereas French L1 speakers do not have these effects. White (1985) proposed that the *pro*-drop parameter, as a Universal Grammar parameter, plays a role when activated in L1 but not L2.

2.1.7 *Pro*-drop in Natural Language Processing Studies

Zero pronoun resolution is a key part of natural language processing (NLP) tasks such as machine translation and question-answering systems, and it involves sub-tasks including zero pronoun detection and coreference resolution. In earlier studies, researchers had proposed theories such as Centering Theory (Kameyama et al., 1993; Walker, Walker, et al., 1998) and Hierarchical Network of Concepts (Z. Huang, 1997) to represent the progress of zero pronoun resolution, and these theories had been applied in computational methods as well (Yeh & Chen, 2003; Yun & JIN, 2012).

On the one hand, zero pronoun resolution computational models before the 2010s were equipped with linguistic features such as syntactic (Iida et al., 2006, 2007; Iida et al., 2015; Yeh & Chen, 2007) or dependency (L. Zhang, Shen, et al., 2020) parsing results, coreference chain (Kong et al., 2019), case frames

(Kawahara & Kurohashi, 2004; Yamashiro et al., 2018), verb features (Yoshida & Nagata, 2009), discourse structure (Cheng et al., 2017), discourse focus (Rao et al., 2015). It shall be noted that these studies were aimed at obtaining higher accuracy performance for zero pronoun resolution, but their employment of various linguistic features motivates us to examine and improve the theories for zero pronoun (*i.e.* what features can improve the accuracy). On the other hand, since around 2015, with increasing computational resources, larger corpora, and surging popularity of artificial neural networks, researchers have explored methods such as Markov Chain, Deep Reinforcement Learning, Transformer, and Data Augmentation, and have made acknowledgable progress (S. Chen et al., 2021; Kim et al., 2021; Konno et al., 2020; S.-j. Sun, 2022; Tong et al., 2019; L. Wang et al., 2017; J. Yang et al., 2020; Q. Yin et al., 2017). There is no doubt that these works made impressive contributions to improve zero pronoun resolution performance from the perspective of the engineering goal, but their lack of using overt linguistic features made them hardly able to provide insights for zero pronoun mechanisms.

2.2 *Pro*-drop Theories

In the previous section, the studies on *pro*-drop are reviewed according to their subfields in linguistics in a developmental view, and we can see that many theories have been developed and extended.

In the first stage, the study of *pro*-drop originally started in the Principles and Parameter theory within the Universal Grammar tradition. Noam Chomsky's Universal Grammar proposes that there is a set of grammatical structures shared by all human languages, and pro-drop is considered to be one of these universal features. However, the specific manifestation of pro-drop can vary across languages. Developed within the framework of Generative Grammar, the Principles and Parameter Theory suggests that pro-drop is a parameter that languages can either have or lack. Languages with the *pro*-drop parameter (or null subject parameter) "on" allow subject pronouns to be dropped, while those with the parameter "off" require overt subject pronouns. As early studies focused on languages with rich subject-verb agreement, the Rich Agreement Theory posits that languages with rich verb agreement systems are more likely to allow pro-drop. When verb conjugations carry enough information about the subject, it becomes unnecessary to include an overt pronoun.

In the second stage, as research had been expanded to radical *pro*-drop languages, the Functional and Discourse-Related Theories took semantic, pragmatic, and discoursal factors into consideration to explain *pro*-drop in radical *pro*-drop and partial *pro*-drop phenomena. These theories emphasize the functional or discourse-related motivations for *pro*-drop. In languages allowing *pro*-drop, the choice to include or omit the subject pronoun may be influenced by pragmatic or discourse factors. The Topic Chain theory is representative of these theories and states the discourse factor in *pro*-drop languages.

In the third stage, the principle of economy of expression suggests that languages tend to evolve to be efficient in communication. *Pro*-drop can be seen as a strategy for minimizing redundancy in languages with rich verbal inflectional systems. As computational linguistics develops, more theories have been brought up concerning *pro*-drop resolution including Centering Theory, Accessibility Theory, Information Load Hypothesis, and Hierarchical Network of Concepts. Some other theories, including the Parallelism Hypothesis, Minimal Chain Principle, and Focus Antecedent Hypothesis, also combine various syntactic and semantic traditions to explain *pro*-drop. In the following review, these representative theories will be introduced.

Accessibility Theory was brought up by Ariel (1990, 2001). This universal theory assumes a logically prior distinction between identifiable/Given entities (coded as definite) and nonidentifiable/Given entities (coded as indefinite). Accessibility theory aims to explain the selection and interpretation of all definite referring expressions. Unlike assuming a fundamental distinction between first versus subsequent mentions, the theory offers a unified account for both referential expressions (*e.g.* proper names) commonly employed for initial mentions in discourse and anaphoric expressions (*e.g.* pronouns) typically used for subsequent mentions. Ariel provided an "accessibility marking scale" to describe the entities' mental accessibility level:

> Full name + modifier > full name > long definite description > short definite description > last name > first name > distal demonstrative + modifier > proximate demonstrative + modifier > distal demonstrative + NP > proximate demonstrative + NP > distal demonstrative (-NP) > proximate demonstrative (-NP) > stressed pronouns + gesture > stressed pronoun > unstressed pronoun > cliticized pronoun > verbal person agreement markers > zero

Ariel (2001) brought up that the accessibility marking scale is guided by three coding principles: informativity, rigidity, and attenuation, and also affected by distance (recency, number of clauses) between a previous and current mention of an entity. Ariel points out that "the prediction is that accessibility marker selection is determined by weighing together a whole complex of accessibility factors, which together determine what the degree of accessibility of a given discourse entity is at the current stage of the discourse", and this indicates that the factors used to predict the accessibility of a zero pronoun are not fixed and depend on the "current" state of the discourse. The universal feature of Accessibility theory and this "current" state feature can be seen in both this dissertation's Experiments 1 and 2 results (see Chapter 4 and Chapter 5).

Informational Load Hypothesis (ILH) was proposed by Almor (1999) to explain the psychological processes that underline the anaphora expression, and it has also been applied in later psychological and neuro-linguistic studies on referential entities' processing mechanism (Almor et al., 2017; Almor et al., 2007; Boiteau et al., 2014). The theory suggests that the processing of NP anaphors is an optimization process, emphasizing that the cost of processing, measured by the activation of semantic information, should serve a specific discourse function. This function may involve identifying the antecedent, introducing novel information, or a combination of both. This Information Load Hypothesis is similar to the Gricean Maxim of Quantity (Grice, n.d.), which states that "speakers should make their contribution as informative as required but not more than required, or in other words that speakers should use the least complex linguistic form that is sufficiently informative for their communicative purpose". The main statement of the Information Load Hypothesis is that

> Because of constraints imposed by the underlying architecture of the psychological mechanisms involved in processing anaphoric expressions, anaphor use can be generally described by the maxim of quantity with the following two additions. First, complexity is expressed by the measure of informational load, a notion that expresses the constraints on the simultaneous storage and processing of information in verbal working memory. Second, the information conveyed by an anaphoric expression consists of information that is required for identifying the antecedent and information that is included as new information about the referent.

The Information Load Hypothesis points out the essential role of the verbal working memory, which is considered important in the processing and distribution of anaphors. This supports our factor construction in Experiment 1, which explores the role of verb usage continuity in the discourse that supports null subject resolution.

Centering Theory is a theory of discourse coherence and salience, which is measured by tracking the attentional state of the speaker within a local discourse. The Centering Model based on this theory is a discourse model focused on the conversants' center of attention, examining the interplay among attentional state, inferential complexity, and the form of referring expressions (Grosz et al., 1995; Joshi & Weinstein, 1981; Poesio et al., 2004; Walker, Walker, et al., 1998; Xiao, 2021). Centering Theory has been applied in areas such as (1) explaining the mechanism of anaphora and coreference resolution (Chai & Strube, 2022; Manjuan & Ping, 2010) and zero pronoun resolution (Aroonmanakun, 2000; Kong et al., 2009; Yeh & Chen, 2003) using empirical methods, (2) quantifying the discourse local coherence level (Grosz et al., 1995; Jeon & Strube, 2020; Rus & Niraula, 2012), and (3) identifying the topic in the discourse (Yeh & Chen, 2004).

Poesio et al. (2004) summarized the main claims of Centering Theory as follows:

> Centering is simultaneously a theory of discourse coherence and of discourse salience. As a theory of coherence, it attempts to characterize entity-coherent discourses: discourses that are considered coherent because of the way discourse entities are introduced and discussed. At the same time, centering is also intended to be a theory of salience: that is, it attempts to predict which entities will be most salient at any given time.

The main idea of discourse coherence and salience in the Centering Theory inspired the parameter setup in Experiment 1, which is based on local coherence in the discourse. Since the null subjects do not have pronounced entities, we measure the verb continuity and examine the boundary of the "localness", to see how the range of the discourse segments affects the results.

As reviewed by Xiao (2021), Centering Theory offers a comprehensive framework for understanding the structure of discourse:

> The fundamental assumption of centering model as far as the focus of attention within a discourse is concerned is that the basic structure of an utterance in discourse often singles out

an entity to be called "center", which an utterance most centrally concerns. This centered entity is further developed by two lines of research work. To be specific, two discourse models, that is, (i) discourse focusing model and (ii) discourse centering model, are put forth by the two trains of thoughts.

In Biezma (2014)'s review on the Multiple Focus Strategies in *pro*-drop languages, it was mentioned that "the resolution of ellipsis is affected by parallelism (Carlson, 2013), focus structure (Clifton, 1998), and general processing constrains favoring 'less effort' structures (De Vincenzi, 1991b)". In the following paragraphs, we will briefly review the theories behind these three factors:

Parallelism Hypothesis explains the phenomenon where different forms of parallelism among conjuncts in coordinate structures aid the processor, making the processing of the second conjunct easier when it exhibits some form of parallelism with the first (Frazier et al., 2000; Frazier et al., 1984). The two main rules of the Parallelism Hypothesis are:

a. The most parallel analysis of a conjoined structure is preferred.

b. An analysis is parallel if featurally similar DPs in distinct conjuncts end up with similar syntactic roles (theta-roles and grammatical functions).

The Parallelism Hypothesis has been applied in explaining coreference (Hall & Yoshida, 2021), ellipsis, and gapping structures (Carlson, 2001). Hall and Yoshida (2021) summarized that the structures reflecting the parallelism effect including (1) Animacy parallel: When there was parallelism in animacy across the object DPs in the conjuncts, the reading speed for the second conjunct was faster compared to cases where animacy was not parallel; (2) Active vs. Passive voice parallel; (3) Sentential vs. DP objects parallel; (4) Non-shifted vs. shifted heavy NPs parallel; (5) Thematic role parallel. Carlson (2001) explored the parallelism effect in processing gapping structures, which can be compared with our *pro*-drop analyses as well. In their study, they proposed the inference based on the Parallelism Hypothesis that "the interpretation of a potentially gapping sentence would depend on the featural similarity of DPs in the different conjuncts."

This Parallelism Hypothesis is applied in Experiment 2, and the factors are generated based on the hypothesis, the features are included such as the subjunctive or conjunctive feature in the "Word-level morphological features", character consistency between main/embedded previous/current clauses, morphological feature consistency between main/embedded previous/current clauses.

Minimal Chain Principle was brought up by De Vincenzi (1991a). The Minimal Chain Principle (MCP) operates as a foundational concept in language processing, placing a strong emphasis on optimizing efficiency. At its core, the MCP represents the cognitive tendency of the processor to steer away from unnecessarily intricate linguistic structures. As the processor follows the MCP, its primary goal is to reduce complexity, prioritizing streamlined and accessible representations of language. However, it is essential to recognize that instances of complexity can arise, particularly when faced with grammatical and processing constraints.

"Minimal Chain Principle: Avoid postulating unnecessary chain members at Surface Structure, but do not delay required chain members."

Focus Antecedent Hypothesis describes the relation between focus structure and the remnant in ellipsis, and it was stated by Clifton (1998):

"Focus Antecedent Hypothesis: The antecedent of the remnant is preferentially focused."

As outlined by the Focus Alternatives Hypothesis (FAH), the process of resolving ellipsis involves the consistent search for a focused element in the antecedent clause to function as the counterpart of the remaining segment. This constraint proves beneficial to us, offering a valuable tool for utilizing ellipsis resolution as a diagnostic technique to discern the placement of focus within the discourse.

Principle of Economy of Expression The principle of economy in linguistics is a guiding concept that suggests linguistic structures tend to be organized in a way that minimizes unnecessary complexity and resources. This principle is often associated with the broader idea of linguistic economy, which encompasses various linguistic theories and approaches. Some relevant sub-principles relevant to the Principle of Economy, including the Gricean Maxims, the Minimalist Program, and the Principle of Least Effort. In *pro*-drop relevant studies, Speas (1995) proposed that "null arguments occur wherever general principles of economy permit them to occur", and disagreed with the *pro*-drop playing a role as a parameter in licensing. A language acquisition study in early childhood English claimed that the economy-based *pro*-drop theory can explain the empty subjects produced by young children acquiring English (Roeper & Rohrbacher, 2000).

2.3 *Pro*-drop Studies in CN, BP, ES

2.3.1 Chinese *Pro*-drop Studies

Chinese, known as a Radical *pro*-drop language, does not have a system of overt morpho-syntactic agreement (C.-T. J. Huang, 1989) so that the reference of a null subject or object cannot be inferred from the morphological forms of the verbs. Compared to other *pro*-drop languages such as Spanish and Italian, Chinese lacks morphological markers for person or gender information. This information is not the only basis on which people can recover co-reference relationships. To characterize the panoply of information that does make such resolution possible, the concept of a topic chain has been advanced (Pu, 2019a; Shi, 1989; Tsao, 1977).

As shown in example (1), sentences with omitted pronouns are ungrammatical. However, the Chinese examples in (2) show that the omitted pronouns in Chinese (marked with *e*) can show up at subject positions and/or object positions (C.-T. J. Huang, 1989).

(1) Speaker A: Did John see Bill yesterday?
Speaker B:
a. Yes, he saw him.
b. *Yes, *e* saw him.
c. *Yes, he saw *e*.
d. *Yes, *e* saw *e*.
e. *Yes, I guess *e* saw *e*.
f. *Yes, John said *e* saw *e*.

(2) Speaker A: Zhangsan kanjian Lisi le ma?
Zhangsan see Lisi LE Q?
"Did Zhangsan see Lisi?"

Speaker B:
a. Ta kanjian ta le.
He see he LE.
"He saw him."
b. *e* kanjian ta le
"[He] saw him."
c. Ta kanjian *e* le.
"He saw [him]."
d. *e* kanjian *e* le.
"[He] saw [him]."
e. Wo cai *e* kanjian *e* le.
I guess see LE.
"I guess [he] saw [him]."
f. Zhangsan shuo *e* kanjian *e* le.
Zhangsan say see LE.
"Zhangsan said that [he] saw [him]."

Syntax studies. In the 1980s, early studies of zero pronouns in Chinese followed the tradition of Chomsky's Government and Binding Theory, and considered zero pronouns as Empty Categories (ECs) (C.-T. J. Huang, 1984). In C.-T. James. Huang's seminal study (C.-T. J. Huang, 1984), he noted the maximal freedom of zero pronoun usage in Chinese and the asymmetry between zero subject and zero object. He classified Empty Categories (ECs) into four categories: zero subject (PRO) in the tenseless clause, zero subject (pro) in the tensed clause, zero object (pro), and zero topic. Among these four categories, only zero object, he claimed, was prohibited in Chinese, and Huang explained this asymmetry was affected by the interaction of the Generalized Control Rule (GCR) and Disjoint Reference (DJR). He considered object-zero anaphors in Chinese as variables instead of pronominal. Furthermore, Huang was trying to tackle the zero pronoun behavior in Chinese from the distinguishment between "discourse-oriented *vs.* sentence-oriented" parameter. Chinese, as a discourse-oriented language, he mentioned, has a discourse grammar of "Topic NP Deletion", which can act as a topic-chain interpretation rule. Apart from this discourse rule, Chinese also has the "topic-prominent" feature, in contrast to the "sentence-prominent" in English (C. N. Li, 1976), and this feature makes Chinese rely less on structural subjects. Importantly, C.-T. James Huang concluded that "the distribution of a zero subject pronoun is closely tied to the presence or absence of a potential antecedent rich enough in content (agreement or actual NP)", and the motivation

of this study is consistent with his conclusion: To explore what exactly is the "rich" content that makes zero pronoun understandable.

Xu's study of zero pronouns argued against treating zero objects as variables (Liejiong, 1986), as C.-T. James Huang had proposed. While discussing features of embedded zero objects, Xu mentioned that "whether a reading with an intrasentential antecedent is available or preferable largely depends on how easily the action-patient relation can be established". Even though this observation was made on zero objects, it indicates the rationality of considering semantic roles when trying to examine whether an antecedent is a good candidate for a zero pronoun.

Some scattered syntactical features of zero pronouns were also observed in previous studies. For example, Li & Thompson pointed out that subjects of NPs that follow right after prepositions "跟(gen1), 给(gei3), 把(ba3)" can not be zero form (C. N. Li & Thompson, 1989). A review article (Xian & Keliang, 2009) mentioned studies published in Chinese had findings including: Zero pronouns usually act as subjects; Intrasentential zero pronouns rely on the main verbs and syntactical structures of the sentences, whereas intersentential zero pronouns are affected by syntactical structures and the way of statement; Zero pronouns are more often found in clauses with related words. These findings show that the behavior of *pro*-drop is affected by contextual features.

Pragmatics studies. In the 1990s, Y. Huang (Y. Huang, 1994) claimed that the wide use of zero pronouns in Chinese contradicts the predictions by the *pro*-drop/null subject parameter in the Government and Binding Theory, and the typology of Chomsky's Empty Categories is not suitable to analyze zero pronouns in Chinese. Instead, Y. Huang developed a pragmatic theory within the neo-Gricean framework of conversational implicature. This theory states that two neo-Gricean principles (*i.e.* M[anner]- and I[formativeness]-principles) and their interactions determine the usage of anaphora in Chinese. The so-called M- and I-principles follow the tradition of Levinson's I[formativeness]- ("*say as little as necessary*"), and M[anner]- ("*do not use a prolix or marked expression without reason*"), and Q[uanty]- ("*do not say less than it is required*") principles (Levinson, 1987).

Y. Huang examined his pragmatic theory on discourse anaphora in his study. He stated that the discourse anaphora relies on the assumption made by the speaker about "how the hearer will recognize the intended referent", and the two conflicting principles Q- and I- are constantly competing to reach a

compromise in Chinese. This statement implies that, in our case, there should be the **right amount of information** in the discourse for the hearer to resolve a zero pronoun. This "right amount of information" in the discourse context, inspired by the Theory of Relevancy (Sperber & Wilson, 1986), contributes to the **Discourse Relevance**, and makes it possible to recover both the implicit and explicit content of an utterance.

Discourse studies. Many Chinese narrative discourse analytical studies have been done to explore the features of zero pronouns, and the concept of "**topic-continuity**" was highlighted among them. In 1979, Li & Thompson carried out a discourse study on two Chinese novels "水浒传 Shui3 Hu3 Zhuan4"("All men are brothers") and "儒林外史 Ru2 Lin2 Wai4 Shi3"("Romance of Confucian Scholars") (C. N. Li & Thompson, 1979), and suggested that the occurrences of zero pronouns are not controlled by structural factors, but by pragmatic features. They observed that the most frequent type of zero pronouns in Chinese discourse is the "topic chain", which was defined as: "the topic established in the first clause serves as the referent of the unrealized topics in the chain of clauses following it". Chen reported referents with higher continuity in Chinese discourse tend to be encoded with pronouns or zero pronouns (P. Chen, 1986).

W.D.Li examined the functional and distributional characteristics of topic-chain/sentence-initial zero pronouns (W. Li, 2004), and claimed that chain initial zero pronouns show overall lower occurrences compared to non-initial ones, and they play the role of providing background information in the discourse.

Pu's studies provided insights into the features of zero pronoun and topic-chain (Pu, 2019b; Pu & Pu, 2014). The Topic Chain Principle she brought up highlights the function of topic chain in the discourse for continuity and coherence: "topic chain encodes a referent that is cognitively most accessible at the moment of discourse production, as enhanced by maximum discourse coherence of topic continuity and thematic coherence". Pu's theory provides a more dynamic way of viewing zero pronoun processing in the discourse, and it shows how zero pronoun relies on information offered "at the moment". Consistent with this idea, our study explores this cognitive-functional perspective by quantifying the "information coherence at the moment", and examines its consistency with *pro*-drop behavior in the discourse.

2.3.2 Spanish *Pro*-drop Studies

Spanish is a consistent *pro*-drop language, and *pro*-drop occurrence in Spanish have the following scenarios (Pešková, 2013): (1) *pro*-drop must happen when it is impersonal and generic structures; (2) *pro*-drop is optional in some personal sentences; (3) *pro*-drop is required in some personal sentences.

Previous studies have explored the omission and expression of pronominal subjects in Spanish, including the subject dropping rates and the factors that affect *pro*-drop. For the dropping rates in Spanish, previous studies had explored *pro*-drop or null subject rates in Spanish, and showed that the *pro*-drop rates vary across Spanish spoken in different locations and are different among pronoun types. As shown in Table 2.4, Otheguy et al. (2007) investigated the overt subject rates in Spanish spoken in different locations. As shown in Table 2.5, the overt subject rates vary across pronoun types and differ between Peninsular and Porteño Spanish. Rello and Ilisei (2009) created a zero pronoun corpora ("Z-corpora") and reported the *pro*-drop distribution in different genres, which include legal, encyclopedic, and instructional. The *pro*-drop rate reported by Rello and Ilisei (2009) averaged among the genres is 54%. They reported the *pro*-drop rate based on clause types, and the results indicated that the *pro*-drop rate is higher in relative clauses than in simple clauses, and the *pro*-drop rate is the highest for subordinate clauses than the other types, which include main clauses, coordinate clauses, and juxtaposed clauses (see Table 2.6 for the results Rello and Ilisei (2009) reported).

Table 2.4: The overt subject rates in Spanish spoken in different locations (Otheguy et al., 2007).

Location	**Rate of overtly realized PS (%)**
Mexico	19
Colombia	24
Ecuador	27
Cuba	33
Puerto Rico	35
Dominican Republic	41

Table 2.5: Overt pronoun rates in Peninsular and Porteño Spanish in a study by Soares da Silva, summarized in Pešková (2013).

Grammatical Persons		**Peninsular (%)**	**Porteño Spanish (%)**
1st Person	yo ('I')	35	37
	nosotros ('we/us')	11	38
2nd Person	tú/vos ('you', SG)	22	22
	usted ('you', SG)	31	40
	ustedes ('you', PL)	33	37
3rd Person	él/ella ('he/she')	12	19
	ellos/ellas ('they')	9	23

Table 2.6: The *pro*-drop rate among different types of clauses reported by Rello and Ilisei (2009).

Genre	**Clause Type pro-drop Rate**					
	Simple	*Relative*	*Main*	*Subordinate*	*Coordinate*	*Juxtaposed*
Legal	0.29	53.22	5.56	69.88	19.01	5.56
Encyclopediac	5.7	39.6	17.95	70.66	7.12	4.27
instructional	15.47	40.11	29.51	50.43	10.03	10.03

As for the factors that affect *pro*-drop in Spanish, Pešková (2013)'s production experiment showed that the expression of pronominal subjects in Spanish is optional, and examined the linguistic internal factors (*i.e.* structural factors of a language or dialect), which includes grammatical person, morphological and contextual ambiguity, verb semantics, clause type, and switch reference. Their experiment results indicated that the usage of pronominal subjects is related to these factors with the ranking as *grammatical persons* > *verb semantics* > *(syntactic) clause type* > *(semantic) sentence type*. Pešková (2013) reviewed earlier studies on the null and overt subject in Spanish. On the one hand, traditional Hispanic grammarians in the past century held the view that the null subject in Spanish is due to *verb affixes* (*e.g. pro hablo* 'I speak'), and these studies had also focused on the morphosyntactic and licensing conditions for pro and its overt counterpart. These early studies explain non-*pro*-drop using the reasons including ambiguity resolution, contrast, and emphasis. On the other hand, descriptive non-empirical research based on corpus studies claimed that the expression or omission of subject pronouns is due to internal/structural factors.

Table 2.7: Internal factors for overt or null subject in Spanish explored by Pešková (2013) using statistical correlation analyses.

Factor	**Example**
grammatical person	o 'I', vos 'you-SG (familiar)', él/ella 'he/she', usted 'you-SG (formal)', nosotros 'we', ustedes 'you-PL', él/ella 'they'
verb semantics	epistemic verbs vs. perceptive verbs
type of sentence	declarative, absolute interrogative, wh-interrogative
type of clause according to its structural complexity	matrix clause with or without subordinate clause, subordinate clause

Biezma (2014) carried out a syntactic analysis on the role of *pro*-drop and non-*pro*-drop in Spanish, and the study claimed silent *pro* subjects are preferred in Spanish and non-*pro*-drop in the subject position in Spanish are marked as focused. In their syntactical experiment, they (1) explored structural preferences for resolving ellipsis in bare argument ellipsis and replacives, investigating the influence of syntax and the information-structure status of the subject; (2) contrast ellipsis resolution with antecedents containing overt DP subjects against those with pro subjects. Biezma (2014) also reviewed the previous theory on *pro*-drop including Parallelism Hypothesis (Carlson, 2001, 2013), Minimal Chain Principle (De Vincenzi, 1991b), and Focus Antecedent Hypothesis (Clifton, 1998), and these theories are further introduced in Section 2.2.

Herbeck (2021) examined perspectival factors' effect on *pro*-drop in Spanish, and the factors include the usage of speaker and addressee pronouns in conjunction with cognitive verbs. Their findings revealed that the role of cognitive verbs, pronoun type, and polarity have a preference between overt and zero pronouns, for example, their data showed that: Explicit pronouns are more prevalent when associated with the 1SG person than with the 2SG person; The higher occurrence of overt pronouns with positive forms, as opposed to negative forms.

Gelormini-Lezama and Almor (2011) studied the roles of referential entities including repeated names, overt pronouns, and null pronouns based on the information load hypothesis (Almor, 1999), which explains anaphor processing as reflecting a balance between function and processing cost. Gelormini-Lezama and Almor (2011)'s experiment measured the participants' reading speed to examine the processing of dis-

Graph 1: Null subjects across time in Brazilian Portuguese

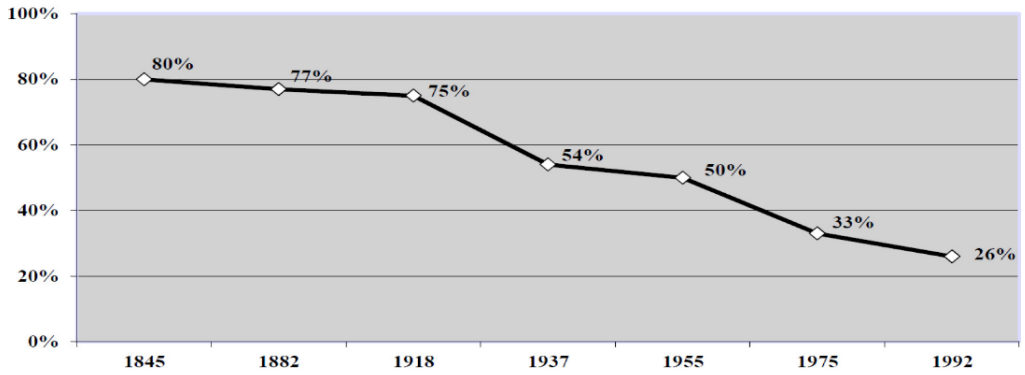


Figure 2.1: The null subject rate in Brazilian Portuguese reported in M. E. L. Duarte and KATO (2017).

courses with repeated names, overt pronouns, and null pronouns. Their results showed that processing penalties are incurred when utilizing both repeated names and overt pronouns in comparison to null pronouns for syntactically salient antecedents. Furthermore, they showed that repeated names are read slower than null pronouns when their antecedent is syntactically salient, and this is similar to the way in English and Chinese.

2.3.3 Brazilian Portuguese *Pro*-drop Studies

Brazilian Portuguese is in a grammatical change from *pro*-drop language to partial *pro*-drop language (I. Duarte & Silva, 2016; M. E. L. Duarte & Marins, 2021), and shows increased usage of overt subject. This changing process is affected by multiple factors including the verb agreement syncretism level change throughout history. The verb syncretism level tends to be higher across time (see Table 2.8), and this change leads to the null subject rate in BP decreasing across time (see Figure 2.1). In European Portuguese, the null subject is the normal case and it is unmarked. As for Brazilian Portuguese, the null subject is marked and is a less common case restricted to some specific contexts.

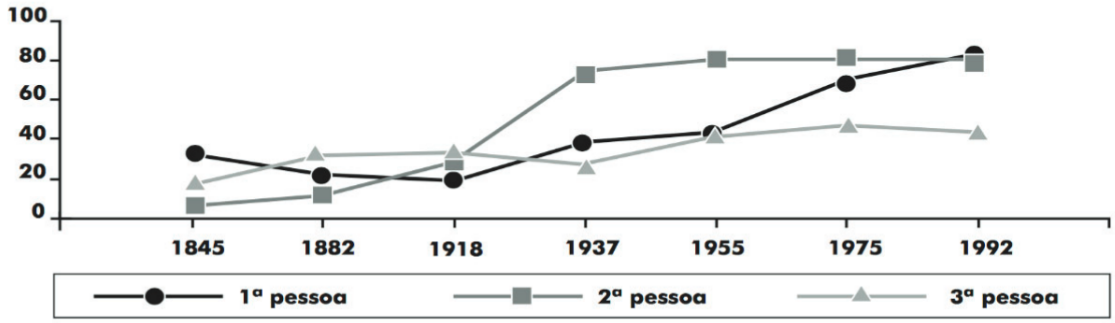


Figure 2.2: The overt subject rate among pronoun types in Brazilian Portuguese reported in M. E. L. Duarte (1996) and cited in Soares et al. (2020).

Table 2.8: An example of the present tense verb *falar*('to speak')'s syncretism level change from Ayres and de Ávila Othero (2021) Table 1. The inflectional paradigms in BP become impoverished.

	Paradigm 1	**Paradigm 2**	**Paradigm 3**
1st person singular	falo	falo	falo
2nd person singular	falas fala	fala	fala
3rd person singular	fala	fala	fala
1st person plural	falamos	falamos	(falamos) fala
2nd person plural	falais	falam	falam
3rd person plural	falam	falam	falam

The verb conjugations in BP make it possible for zero pronoun resolution in some cases, and the high verb syncretism demands more context input for the null subject to be resolvable. As shown in the following example from Ayres and de Ávila Othero (2021), the verb *adolo* (means 'love', 1st person singular) in the sentence (1) can indicate the null subject (*i.e.* 'I'), whereas the verb *adora* (means 'love', can be 2nd person singular, 3rd person singular, or 1st person plural) in the sentence (2) has a higher level of syncretism and leave the subject harder to resolve.

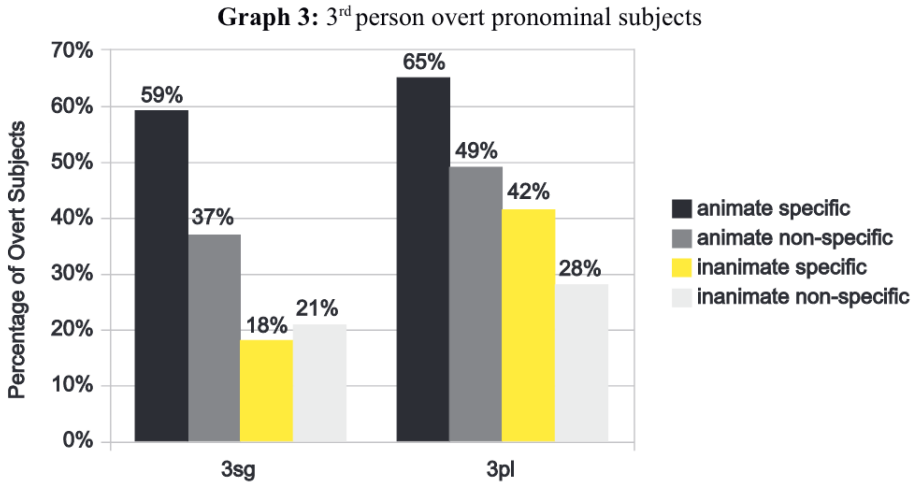


Figure 2.3: The over subject percentage among 3rd person compared based on Animacy ([animate]) and Specificity ([specific]) features in Brazilian Portuguese in Soares et al. (2020).

(1) Adoro pizza.
Love.1st.person.sg pizza
'I love pizza.'

(2) Adora pizza.
Love.{2nd.person.sg
/3rd.person.sg, pizza
/1st.person.pl}
'You/He/She/It/We love(s) pizza.'

Apart from the verb agreement factor, semantic features also play a role in BP being partial *pro*-drop. Previous studies pointed out that semantic features including [animacy] and [specificity] affect the *pro*-drop usage in BP. For example, Soares et al. (2020)'s corpus study showed that the [+human] feature encourages the use of overt pronouns (see Figure 2.3), and their judgment study showed that the acceptance for null subject is lower for inanimate (*i.e.* [-animacy]) cases, and the acceptance for the overt subject is higher for animate (*i.e.* [+animacy]) cases (see Figure 2.4).

Another factor is the word order change in BP, from VSO (V1) structure to SVO (V2) structure, which has also been brought up to play a role in the null subject usage rate decreasing in BP (Kato, 2000). As mentioned in Kato and Duarte (2018):

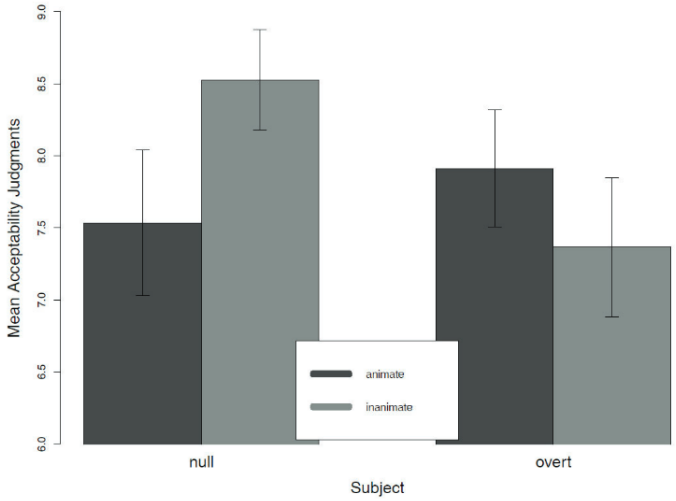


Figure 2.4: Acceptability results based on the animacy factor for null and overt subjects in Brazilian Portuguese in Soares et al. (2020).

> BP rejects the verb in sentence-initial position, filling it with an adjunct or a discursive element. The phenomenon is analyzed as distinct from the syntactic V2 structures, as any category can fill this position: a syntactic adjunct (whether a head or an XP) or a discursive element. Kato's hypothesis is that the constraint here is rhythmic or prosodic.

In a recent study, Ayres and de Ávila Othero (2021) reviewed these four main factors considered as relevant to null subject in BP: (1) verb agreement syncretism level; (2) semantic feature of the referent, which includes human, referential, and semantic gender (Cyrino et al., 2000); (3) linear order of the verb, whether it is the first element in its clause (Kato, 2000); (4) topic chain. Ayres and de Ávila Othero (2021) analyzed these four linguistic factors for *pro*-drop in Brazilian Portuguese with corpus analysis, and their study indicated that these four factors are both necessary and sufficient to understand and explain all null subject occurrences in any given corpus of contemporary spoken BP.

2.4 Summary

In this chapter, the previous studies on *pro*-drop have been reviewed from the perspective of various subfields in linguistics and cross-linguistically, looking at Chinese, Spanish, and Brazilian Portuguese.

The linguistic subfields' study review shows how *pro*-drop has been developed as a parameter in the early Generative Grammar framework. This early-stage study is also highly correlated with the morphological features in the language. Later on, as the *pro*-drop study paid more attention to the radical *pro*-drop languages, its theory was expanded by taking semantic and pragmatic factors into consideration and developing theories such as topic chain. The *pro*-drop study, in recent years, has also reached more applied linguistic areas, such as language acquisition on *pro*-drop acquisition, computational linguistics on zero pronoun resolution, corpus studies on *pro*-drop rate distribution, and neuro-linguistics on brain mechanism of the *pro*-drop resolution. We can see that these recent studies are using various methods to contribute to *pro*-drop theory construction. This is consistent with the current series of studies that are reported in the following chapters, and it is shown that cutting-edge machine learning and other statistical methods can shed light on our understanding of *pro*-drop cross-linguistically.

Table 2.9 provides an outline review of the studies that have been mentioned in Section 2.3.1, 2.3.2, 2.3.3. We can see that previous studies explored the factors that affect *pro*-drop in each language, and studied their *pro*-drop rate changes given the following factors (1) position in clause distribution and clause type; (2) historical change; (3) location change. These previous studies provide an inspiring framework for this current work, which focuses on exploring *pro*-drop factors using statistical modeling methods. Especially for Experiment 2 (see Chapter 5), the factors generated from syntactic, semantic, morphological, and discoursal/logic aspects are deeply rooted in these previous studies and theories. Experiment 1 (Chapter 4) and Experiment 3 (Chapter 6) take advantage of the development of Large Language Models (LLMs), which provide efficient high dimensional vector representation of words in each language trained on the deep learning models.

Table 2.9: A summary of *pro*-drop studies' outlines on CN, ES, and BP.

Language	pro-drop Type	pro-drop Previous Studies Summary
CN	*radical*	- null object - discourse-oriented topic drop - topic chain theory - pro-drop affected by contextual features - Manner and Informativeness theories - sentence-initial dominant distribution
ES	*consistent*	- pro-drop variation in different locations speaking ES - pro-drop factors: structural factors of a language or dialect), which includes grammatical person, morphological and contextual ambiguity, verb semantics, clause type, and switch reference - non-pro-drop leads to focus in ES - perspectival factors' effect - information load hypothesis
BP	*partial*	- pro-drop rate decrease throughout history - verb syncretism increase throughout history - null subject preferred for inanimated cases - VSO to SVO verb order transition - topic chain theory

Chapter 3

Background on Methodology

3.1 Introduction

In this dissertation, multiple natural language processing methods and machine learning models are adopted to explore the factors and mechanism behind *pro*-drop. These methods include (1) Word representation and similarity measurement using word embeddings; This task involves language model training and word embedding retrievement, and these methods are reviewed in Section 3.2. (2) Constituency tree and dependency tree parsing based on pre-trained language models, and these parsing methods are reviewed in Section 3.3. (3) Graph Neural Networks and Graph Attention Networks' training and interpretation, and it is introduced in Section 3.4. (4) Machine learning models including Binomial Logistic Regression and Random Forest, and these two models are introduced in Section 5.2.4 and Section 5.2.5. Apart from these methods mentioned above, there are other natural language processing tools are adopted in this dissertation, such as the lemmatizer (an NLP tool that obtains lemmas from the original words), the POS tagger (a Part-Of-Speech tagger that can provide the POS tag for words), and the morphological feature detector. These methods are more widely applied in the previous studies so there would not be further review on these topics.

3.2 Word Similarity Measurement

Measuring the word similarity is important in natural language processing tasks such as language modeling and information retrieval. It has been applied in semantic search, summarization, question answering, document classification, sentiment analysis, and knowledge graph construction (Farouk, 2019). This task appears in Experiment 1 in this dissertation (Chapter 4), and the word-level verb similarity is measured to reflect the discourse coherence level. This task is realized with word embedding, which is a high dimensional vector representation trained by language models. However, in the history of word similarity measurement, there are many other methods developed by previous researchers, and word embedding is one of them that has emerged recently. In this section, a review of word similarity measurement methods in NLP history is provided.

As summarized by Farouk (2019), there are three main methods to measure words' similarity: (1) Corpus-based words similarity; (2) Knowledge-based words similarity; (3) String-based words similarity.

3.2.1 Corpus-based words similarity

Corpus-based words similarity is also mentioned as Distributional word similarity (Navigli & Martelli, 2019). The corpus-based word similarity measurement relies heavily on large-scale corpora. Large-scale corpora reflect how words behave in context, and reveal various relationships of the words. The words' co-occurrence in the corpora can reflect their similarity. This characteristic makes them an excellent resource for understanding word distributions, which are then used to infer semantic properties and evaluate the level of similarity between two words. As Navigli and Martelli (2019) summarized:

"The fundamental assumption behind distributional approaches is that the semantic properties of a given word w can be inferred from the contexts in which w appears. That is, the semantics of w is determined by all the other words which co-occur with it."

The representation forms of the corpus-based methods can be classified as Explicit Representation and Implicit or Latent Representation. For **Explicit Representation**, each dimension can be directly interpreted, such as when words or senses are employed as the meanings of the vector's dimensions. The feature vector for an Explicit Representation could be *binary values* (*i.e.* a vector of 0 or 1, such as the one-hot encoding [0,0,0,..., 1, ...,0]), or an *association and probabilistic measures* which offer a score or

probability indicating the likelihood that a particular word co-occurs with the target word. Some representative algorithms for Explicit Representation are Sørensen-Dice index, Jaccard Index (Grefenstette, 2012), Pointwise Mutual Information (Church & Hanks, 1990), Positive PMI, and Explicit Semantic Analysis (using TF-IDF). For **Implicit/Latent Representation**, it encodes linguistic information in a form that cannot be directly interpreted. Representative methods for Implicit/Latent Representation include Latent Semantic Analysis (LSA) and Word Embeddings.

LSA (Deerwester et al., 1990) can discover the hidden, or latent, relationships between words and concepts within a given body of text, and it allows for the exploration of semantic relationships between terms and documents. The steps of LSA are as follows: First, LSA begins by constructing a matrix that represents the relationships between words and documents. Each cell in the matrix contains a numerical value that reflects the frequency of a term in a document. Second, Singular Value Decomposition (SVD) is applied as a mathematical technique used to factorize the term-document matrix into three matrices: U, Σ (sigma), and V. These matrices capture the underlying structure and relationships in the data. Third, in Dimensionality Reduction, the Σ matrix contains singular values that represent the importance of different dimensions in the data. By keeping only the top singular values and their corresponding columns in U and V matrices, LSA reduces the dimensionality of the original matrix. This step helps in capturing the most significant patterns while discarding noise and less important information. The reduced matrices form a new space, often referred to as a "semantic space". In this space, terms and documents are represented as vectors, and the similarity between them is determined by the cosine of the angle between their respective vectors.

Word Embedding methods get more attention recently as neural networks and deep learning techniques develop. Word Embeddings represent words as vectors in a continuous vector space, and these methods capture semantic relationships between words based on their contextual usage. Early efforts in Word Embedding methods can date back to the late 1980s, and Rumelhart et al. (1986) proposed the "back-propagation" method for neural networks to obtain continuous vectors. In the early 2000s, Bengio et al. (2000) brought up the neural probabilistic language model, which can learn word representation and probability function simultaneously. In the early 2010s, Collobert et al. (2011) proposed a unified neural network that can learn. internal representations based on vast amounts of mostly unlabeled training data. Most recently, Word Embedding methods have made progress and the computational models

and capacity have improved, and some famous models are Word2Vec, fastText, GloVe, SensEmbed, AutoExtend, Contextual Word Embeddings (such as BERT and ELMo). In this dissertation, GloVe and BERT word embeddings are used in Experiment 1 (Chapter 4), and these models mentioned above will be introduced in the following paragraphs.

Word2Vec , developed by Mikolov et al. (2013), is a popular word embedding method that learns distributed representations of words. It is trained on large corpora and uses a two-layer neural network to create dense vector-based representations for words. Word2Vec includes two models: Continuous Bag of Words (CBOW) and Skip-Gram. The CBOW exploits the context to predict a target word, and the Skip-Gram uses a word to predict a target context word. Word2Vec preserves relationships between vectors such as analogy (*e.g.* London - UK + Italy should be very close to Rome).

fastText , developed by Joulin et al. (2017), integrates subword information and extends the idea of word embeddings to subword-level embeddings. It represents words as bags of character n-grams (*i.e.* substrings of length n), allowing it to capture morphological information such as prefixes and suffixes, and handle out-of-vocabulary words more effectively. Compared to Word2Vec, fastText developed in the sense that it can build vectors for misspellings or out-of-vocabulary words.

GloVe (Global Vectors), developed by Pennington et al. (2014), focuses on global statistical information and considers the entire corpus when learning word representations instead of focusing on local context like Word2Vec. GloVe conducts unsupervised learning that involves deriving latent word vector representations by starting with global word–word co-occurrence information. GloVe creates embeddings that capture the relationships between words based on their co-occurrence probabilities. GloVe constructs a word co-occurrence matrix based on the frequency of word pairs occurring together. The entries in this matrix represent the likelihood of words co-occurring, providing a global perspective on the relationships between words. The training objective of GloVe is to learn word vectors that can capture the relationships between words in terms of their co-occurrence probabilities. The model aims to minimize the difference between the dot product of word vectors and the logarithm of the observed word co-occurrence probabilities. During the training process, a least square regression model is computed to learn latent word vector

representations, to minimize a loss function that is influenced by the soft constraint mentioned earlier, applied to all pairs of words within the vocabulary.

SensEmbed (Sense Embeddings), developed by Iacobacci et al. (2015), employs a multifaceted approach, and involves the transformation of word embeddings to the sense level. This approach harnesses knowledge from an extensive semantic network to enhance the effectiveness of semantic similarity measurement.

AutoExtend , developed by Rothe and Schütze (2015), employs an autoencoder neural architecture to obtain latent, embedded representations of lexemes and synsets. In this structure, word embeddings serve as both input and output layers, while the hidden layer yields synset embeddings.

Contextual Word Embeddings leverage the word distribution to acquire latent encodings that capture occurrences of words within a specified context. Three representative Contextual Word Embeddings are BERT, ELMo, and GPT-2. **ELMo** (Embeddings from Language Models), developed by Peters et al. (2018), uses a bidirectional Long Short-Term Memory (LSTM) network to process input sequences. ELMo generates word embeddings at multiple layers of the bidirectional LSTM. Each layer captures different aspects of syntax and semantics. These layered representations are combined to form the final contextualized embeddings. ELMo's training task is to predict the next word in a sequence, considering the bidirectional context. **BERT** (Bidirectional Encoder Representations from Transformers), developed by Kenton and Toutanova (2019), considers the context from both directions and can capture the bidirectional dependencies of words in a sentence. BERT is built on a Transformer architecture, which uses self-attention mechanisms to weigh the importance of different words in a sequence based on their contextual relevance. The training task of BERT is to predict missing words in a sentence by considering the surrounding context. This process results in the creation of rich contextualized word embeddings. For application, BERT can be fine-tuned on specific downstream tasks, such as text classification, named entity recognition, and question answering. **GPT-2** (Generative Pre-trained Transformer 2), developed by Radford et al. (2019), is also a bidirectional transformer-based model and has 1.5 billion parameters. In recent years, many fine-tuned models have been built based on Contextual Word Embedding models such as BERT, ELMo, and the GPT models, and Ethayarajh (2019) deployed a comparison study on how different layers of embeddings retrieved from these models can represent the contextual information.

3.2.2 Knowledge-based words similarity

Knowledge-based words similarity is measured based on the knowledge stored in the Lexical Knowledge Base (LKB). Examples of popular LKBs are: WordNet (Fellbaum, 1998), Roget's thesaurus, Wikipedia, Wiktionary, and BabelNet (Navigli & Ponzetto, 2012).

The two main methods of knowledge-based word similarity measurement are (Navigli & Martelli, 2019):

> (1) The first method computes the semantic similarity between two given items i1 and i2 by inferring their semantic properties on the basis of structural information concerning i1 and i2 within a specific LKB. Early attempts at this method are The depth of a given concept (*i.e.*, synset) in the LKB taxonomy; The length of the shortest path between two concepts in the LKB; The Least Common Subsumer (LCS), that is, the lowest concept in the taxonomical hierarchy which is a common hypernym of two target concepts.
> (2) The second method performs the extraction and comparison of a vector representation of i1 and i2 obtained from the LKB. These methods include: A vector-based word and sense representation is obtained by exploiting the structural information of an LKB; The obtained vector representations are compared by applying a similarity measure.

3.3 Tree Parsing Via Pre-trained Language Models

Constituency tree and dependency tree are well-known structural representation tools in linguistics, and their automatic parsing technology is always an important topic in computational linguistics. These two parsing types are used in all three experiments in this dissertation (Chapter 4, 5, 6): In Experiment 1, the dependency structure in the discourse is used to locate the "subject-verb-object (agent-verb-patient)" structures; In Experiment 2, both constituency tree and dependency tree are used to generate the *pro*-drop relevant features that adopted as feature inputs for the statistical models; In Experiment 3, the dependency tree is used to generate graph representation for each clause in the discourse, and reflects the structural feature in the language. Therefore, these two parsing strategies play an essential role in the following chapters and will be reviewed in this chapter on how they are realized by pre-trained language models.

Generally speaking, dependency parsing is a specific aspect of syntactic analysis that concentrates on revealing the grammatical relationships between words within a sentence. The primary goal is to showcase how words in a sentence depend on or relate to each other. In the process, a tree structure, known as a dependency tree, is constructed to visually represent these dependencies, offering a clear illustration of the hierarchical structure of the sentence. This helps in understanding the syntactic connections and dependencies between words. On a broader scale, syntactic parsing is based on context-free grammar, and it encompasses various techniques that aim to uncover the overall syntactic structure of a sentence, going beyond just dependencies. It includes the identification of phrase boundaries, constituents, and adherence to grammatical rules.

In the broader context, both dependency parsing and syntactic parsing play essential roles in extracting meaning and insights from textual data. They provide the foundational understanding needed for various language-processing tasks. Some of the key concepts associated with these tasks include relation extraction (Y. Zhang et al., 2018), machine translation (K. Chen et al., 2018), and sentiment analysis(Poria et al., 2014; Vilares et al., 2017). Each of these concepts contributes to the overall process of understanding and extracting valuable information from natural language text.

3.3.1 Dependency Parser

A dependency parser analyzes the grammatical structure of a sentence, and it builds relationships between "head" words and words that modify those heads. The dependency parsing methods can be classified into two main groups, as summarized by Y. Zhang and Clark (2008):

(1) Transition-based: the outputs are built by explicit transition actions, such as "Shift" and "Reduce";
(2) Graph-based: it is dependency graphs (not transition actions) that the parsing model assigns scores to.

Anderson and Gómez-Rodrıguez (2020) and Choi et al. (2015) made performance and method comparisons correspondingly of the previous dependency parser studies. As shown in Table 3.1, the representative dependency parser studies are summarized.

Table 3.1: Previous dependency parsers and methods.

Parser	Approach	Publication
Mate v3.6.1	Maximum spanning tree, 3rd-order features	(Bohnet, 2010)
LTDP v2.0.3	Transition-based, beam-search + dynamic prog	(L. Huang et al., 2012)
ClearNLP v2,3	Transition-based, selectional branching	(Choi & McCallum, 2013)
GN13	Easy-first, dynamic oracle	(Goldberg & Nivre, 2013)
Redshift12	Transition-based, non-monotonic	(Honnibal et al., 2013)
Turbo v2.2	Dual decomposition, 3rd-order features	(Martins et al., 2013)
RBG	Tensor decomposition, randomized hill-climb	(Lei et al., 2014)
SNN	Transition-based, word embeddings	(D. Chen & Manning, 2014)
Yara	Transition-based, beam-search, dynamic oracle	(Rasooli & Tetreault, 2015)
BIST - Transition/Graph	Based on bidirectional-LSTMs (BiLSTMs)	(Kiperwasser & Goldberg, 2016)
Biaffine	Based on a simple graph-based dependency parser with neural attention	(Dozat & Manning, 2016)
Pointer-TD	stack-pointer networks (STACKPTR), based on pointer networks with an internal stack using depth-first search	(Ma et al., 2018)
Pointer-LR	transition-based, parses sentences from left to right by building attachments	(Fernández-González & Gómez-Rodrıguez, 2019)
HPSG	Head-Driven Phrase Structure Grammar, based on self-attention architecture	(J. Zhou & Zhao, 2019)
SeqLab	Incorporate sequence labeling	(Strzyz et al., 2019)
DDParser	Based on neural network and the graph-based Biaffine parser	(S. Zhang et al., 2020)

The dependency parser used in this dissertation for dependency tree retrieval is the SpaCy dependency parser, which is a variant of the parser developed by Honnibal and Johnson (2015). As the parser publication introduced, this SpaCy parser is "a new set of non-monotonic transitions that permits a partial parse state to derive a larger set of completed parse trees than previous work, which allows our parser to escape from a larger set of garden paths", and it achieved a 91.85% directed attachment accuracy.

3.3.2 Constituency Parser

Constituency parsing aims to extract a constituency-based parse tree from a sentence that represents its syntactic structure according to a phrase structure grammar. An assessment[1] is made to evaluate the performance of the representative constituency parsers. The evaluation was carried out based on The Wall Street Journal section of the Penn Treebank, Section 22 is used for development, and Section 23 is used for evaluation. As shown in Table 3.2, the constituency parsing models are evaluated based on F1.

[1]https://nlpprogress.com/english/constituency_parsing.html

Table 3.2: Previous constituency parsers, methods, and evaluation results.

Model	**F1**	**Publication**
Span Attention + XLNet	96.40	(Tian et al., 2020)
Label Attention Layer + HPSG + XLNet	96.38	(Mrini et al., 2020)
Attach-Juxtapose Parser + XLNet	96.34	(K. Yang & Deng, 2020)
Head-Driven Phrase Structure Grammar Parsing (Joint) + XLNet	96.33	(J. Zhou & Zhao, 2019)
Head-Driven Phrase Structure Grammar Parsing (Joint) + BERT	95.84	(J. Zhou & Zhao, 2019)
CRF Parser + BERT	95.69	(Y. Zhang et al., 2021)
Self-attentive encoder + ELMo	95.13	(Kitaev & Klein, 2018)
Model combination	94.66	(Fried et al., 2017)
LSTM Encoder-Decoder + LSTM-LM	94.47	(Takase et al., 2018)
LSTM Encoder-Decoder + LSTM-LM	94.32	(Suzuki et al., 2018)
In-order	94.2	(J. Liu & Zhang, 2017)
CRF Parser	94.12	(Y. Zhang et al., 2021)
Semi-supervised LSTM-LM	93.8	(Charniak et al., 2016)
Stack-only RNNG	93.6	(Kuncoro et al., 2017)
RNN Grammar	93.3	(Dyer et al., 2016)
Transformer	92.7	(Vaswani et al., 2017)
Combining Constituent Parsers	92.4	(Fossum & Knight, 2009)
Semi-supervised LSTM	92.1	(Vinyals et al., 2015)
Self-trained parser	92.1	(McClosky et al., 2006)

3.4 Graph Neural Networks

In recent years, Graph Neural Networks (GNNs) have emerged as a potent tool for the analysis and processing of graph-structured data. The roots of Graphical Neural Networks can be traced back to the evolution of neural networks and their subsequent adaptation to effectively manage and interpret data organized in graph structures. GNNs are developed from Neural Networks, and they are models that capture the dependence of graphs via message passing between the nodes of graphs. GNNs include variants such as Graph Convolutional network (GCN), Graph attention network (GAT), and Graph Recurrent Network (GRN). In this dissertation, the Graph Attention Network (GAT) used in Experiment 3 (Chapter 6) is a type of Graph Neural Network with Attention Layers. In this section, the method development on GNNs is reviewed.

At the early stage during the development stage, the concept of neural networks and connectionism build the foundation of the following development in the area. The origins of neural networks can be traced back to the work of McCulloch and Pitts in the 1940s (McCulloch & Pitts, 1943), which laid

the foundation for artificial neurons and simple models of computation. However, these early neural networks were primarily focused on binary and linear computations. In the 1980s, researchers explored connectionist models that emphasized the use of interconnected nodes. However, these models were often limited to feedforward architectures and lacked the ability to handle complex graph-structured data.

In 2005, The concept of Graph Neural Networks was introduced by Gori et al. (2005). This early work focused on extending neural networks to handle graph-structured data by introducing a Recurrent Neural Network (RNN) architecture that operated on graphs (Frasconi et al., 1998; Sperduti & Starita, 1997).

A significant breakthrough came with the introduction of Graph Convolutional Networks (GCNs) by Kipf and Welling (2016). GCNs extended the convolutional operation to graph-structured data, enabling the aggregation of information from neighboring nodes. Following the introduction of GCNs, there was a surge of interest in GNNs. Researchers explored various architectures, including GraphSAGE (Hamilton et al., 2017), GAT (Graph Attention Networks) (Velickovic et al., 2017), and Graph Isomorphism Networks (GIN) (K. Xu et al., 2018). GNNs were applied to diverse domains, including social network analysis, recommendation systems, and bioinformatics, and in tasks including node classification, link prediction, community detection, and knowledge graph completion.

J. Zhou et al. (2020) and Wu et al. (2020) provided in-depth reviews of the GNNs models and applications. J. Zhou et al. (2020) summarized that the GNNs' application area covered the fields including, but not limited to, Graph Mining, Physics, Chemistry, Biology, Knowledge Graph, Graph Generation, Combinatorial Optimization, Traffic Networks, Recommendation Systems, Text and Image Processing.

As shown in Figure 3.1, J. Zhou et al. (2020) summarized the four steps for GNNs model design pipeline: (1) Find graph structure (*i.e.* structural scenarios and non-structural scenarios); (2) Specify graph type and scale (*i.e.* Directed/Undirected Graphs, Homogeneous/Heterogeneous Graphs); (3) Design loss function; From the learning task perspective, the loss function relies on the level of the task, and can be node-level, edge-level, or graph-level tasks. From the perspective of supervision, the task can be divided into supervised, semi-supervised, or unsupervised tasks. (4) Build a model using computational modules, which include Propagation Module, Sampling Module, and Pooling Module.

As shown in Figure 3.2, J. Zhou et al. (2020) provided a visualization of transferring different types of data into graph data.

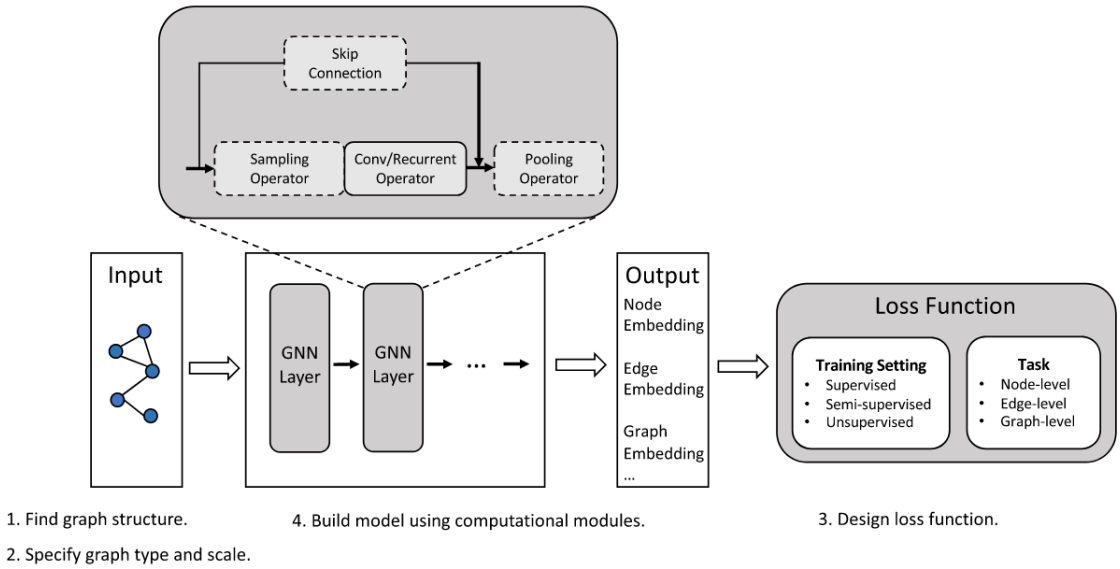


Figure 3.1: The general design pipeline for a GNN model. Figure 2 from J. Zhou et al. (2020)

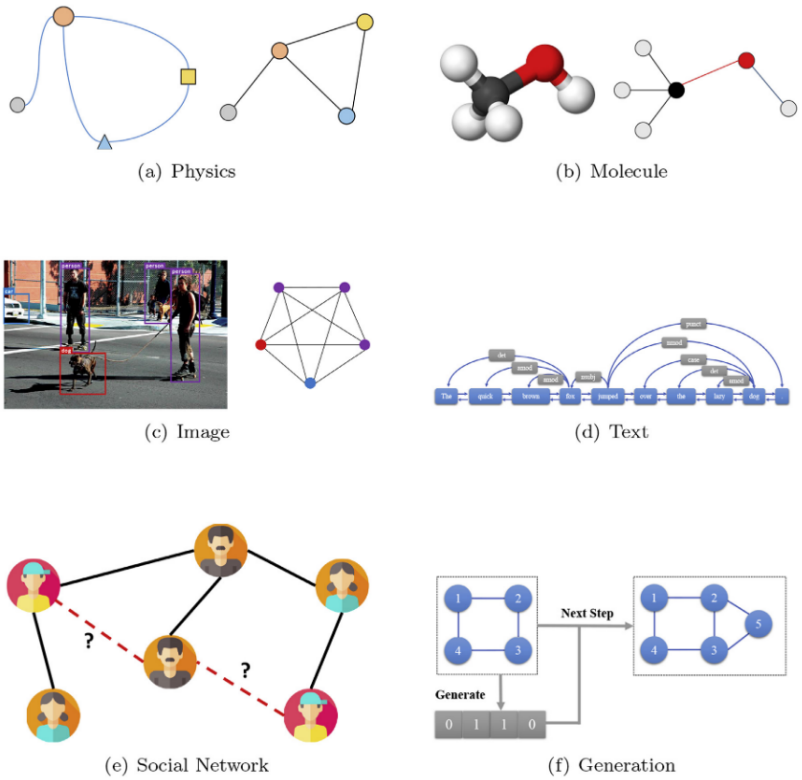


Figure 3.2: Data to graph at different application scenarios. Figure 6 from J. Zhou et al. (2020)

As shown in Figure 3.3, Wu et al. (2020) presented the GNN model structures that are built with Graph Convolutional Layers (GConv), and the subfigures are models (a) ConvGNN with multiple graph convolutional layers; (b) ConvGNN with pooling and readout layers for graph classification; (c) GAE for network embedding; (d) STGNN for spatial–temporal graph forecasting.

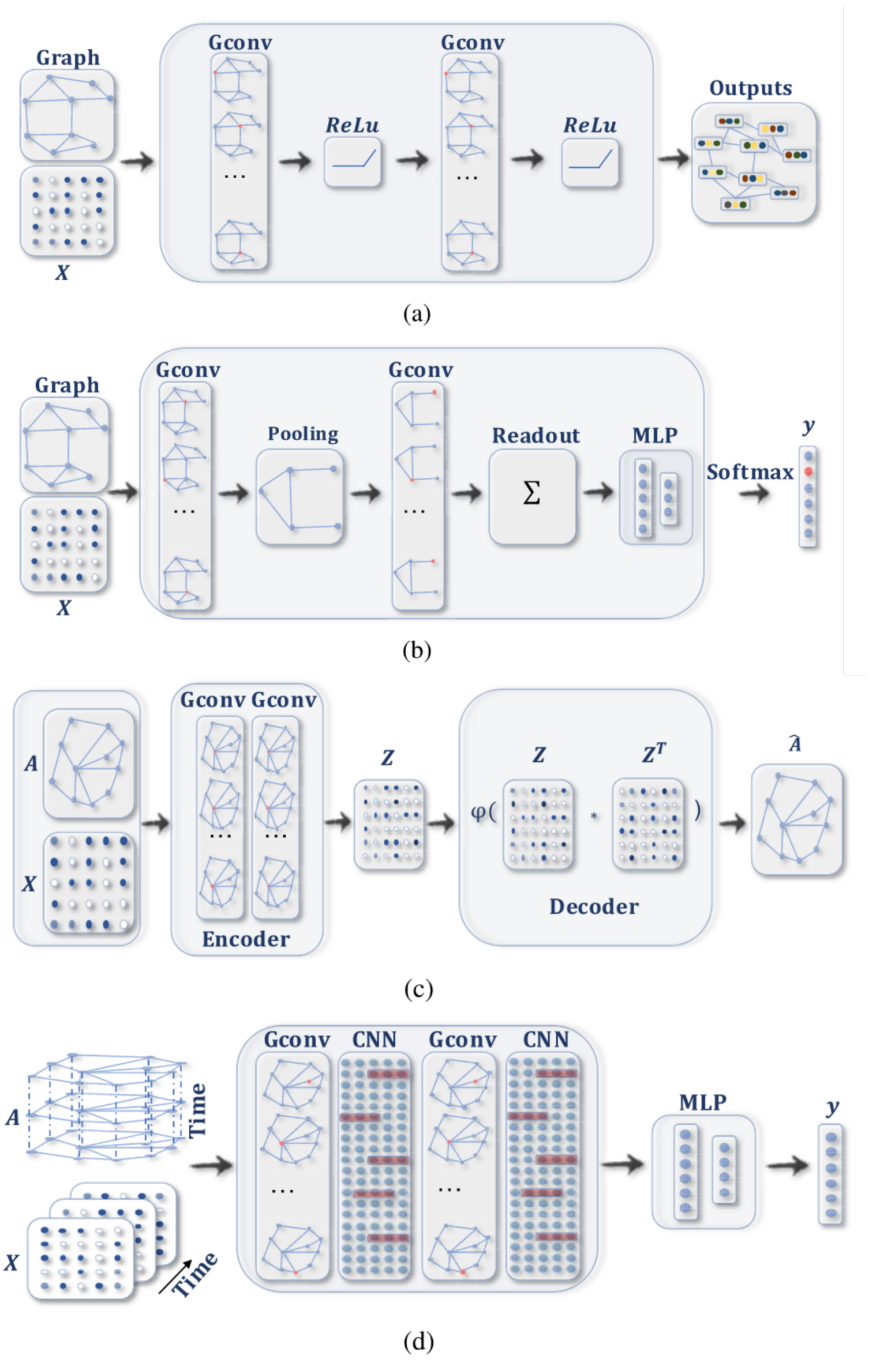


Figure 3.3: Different GNN models built with graph convolutional layers. Figure 2 from Wu et al. (2020)

Chapter 4

Experiment 1: Quantifying Verb Usage Continuity As Discourse Support for Omitted Pronouns

4.1 Introduction

Pro-drop is a phenomenon commonly seen in languages such as Chinese, Spanish, Portuguese, and Italian. In *pro*-drop languages, pronouns, or named entities can be omitted; Native speakers understand these omitted elements based on contextual information or world knowledge. When the omitted element is a pronoun, the phenomena is referred to as *pro*-drop and the dropped word is called a zero pronoun. The main property of zero pronouns is that they are not overtly pronounced, and we are intrigued to understand "what makes them understandable", and that motivates this study to explore the mechanism of *pro*-drop.

Intuitively speaking, native speakers of a *pro*-drop language would drop a pronoun when they think it is "obvious" to be inferred from the context. In other words, the zero pronoun should be salient among all the candidates. There are previous theories proposing the rationale for *pro*-drop, such as Ariel (2001)'s Accessibility Theory, and Almor (1999)'s Informational Load Hypothesis (ILH). This salience effect was observed in our previous study that focused on the character-verb usage continuity (S. Zhang, Li, & Hale, 2022), which revealed that *pro*-drop cases are more salient in verb usage continuity.

(1) 这 是 我 给 他 后来 画 出来 最好 的 一 幅 画像。
zhe shi wo gei ta houlai hua chulai zuihao de yi fu huaxiang
This is I for he later draw out best DE one drawing
"This is the best portrait I drew for him later on."

(2) [我] 六 岁 时, 大人们 使 我 对 我的 画家 生涯 失去 了 勇气。
wo liu sui shi darenmen shi wo dui wode huajia shengya shiqu le yongqi
[I] Six year old grown-ups make I towards my painter career lose LE courage
"When I was six, grown-ups made me lose courage in my painter career."

(3) [我] 除了 画 过 开着 肚皮 和 闭着 肚皮 的 蟒蛇,
wo chule hua guo kaizhe dupi he bizhe dupi de mangshe
[I] except draw PASS opening belly and closing belly DE boa
"Except that I had drawn boas with opening and closing belly,"

(4) [我] 后来 再 没有 学 过 画。
wo houlai zai meiyou xue guo hua
[I] afterwards again not learn PASS draw
"I had never learned drawing afterwards."

Figure 4.1: Example of Chinese omitted pronouns in a topic chain. Omitted pronouns, shown here in green with square brackets are not actually spoken. However, their intended reference is unambiguous for native speakers. Predicates are shown in red, and the overtly expressed entities are shown in blue. Unlike in Romance languages, there is no morphological change (verb inflection change) in verbs to mark the gender or number of omitted elements in Chinese.

Table 4.1: Dependency structure and semantic role annotation table. An annotation example for the sentence "These boas swallow their prey without chewing." The verbs "chew" and "swallow" are located as verbs in the column *V*. Token indices for each verb's Agent and/or Patient are annotated in the columns *V-agent* and *V-patient* respectively, and the character roles they are referring to are annotated in the column *character*.

ID	word	S	V	O	V-agent	V-patient	character
56	这些 (these)						
57	蟒蛇 (boa)	True					ch2_boa
58	把 (BA)						
59	它们 (them)						
60	的 (DE)						
61	猎获物 (prey)			True			
62	不 (not)						
63	加 (with)						
64	咀嚼 (chew)		True		57 (boa)	61 (prey)	
65	地 (DI)						
66	囫囵 (roughly)						
67	吞 (swallow)		True		57 (boa)	61 (prey)	
68	下 (down)						

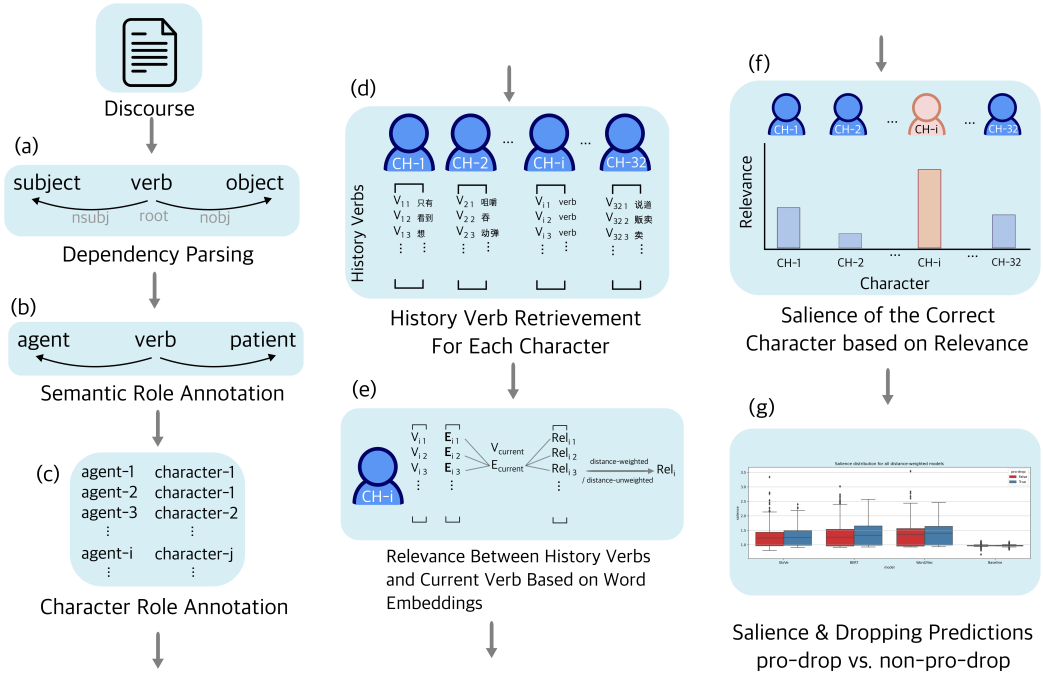


Figure 4.2: Analysis steps adopted in this study: (a) Grammatical subjects and objects of each main verb are identified via dependency parsing on the whole story discourse of The Little Prince (see a sentence example from Table 4.1, columns "S", "V", "O"); (b) Semantic role annotation: for all the subjects and objects, annotate their semantic roles as AGENT or PATIENT (see Table 4.1 column "V-agent" and "V-patient"); (c) Character role annotation: assign story character roles to the entities, see character occurrences in Table 4.2, and Table 4.1 column "character"; (d) History verb retrieval for each story character: for each story character, tabulate the verbs that are its main verbs being used in the discourse (see example Table 4.3); (e) Relevance between history verbs and a current verb: for each current verb, calculate its relevance to the history verbs, and sum with or without their distance weight (see Table 4.6 and 4.7); (f) Salience of the correct character: for each verb, calculate how "salient" the correct character is compared to all other characters (see example Table 4.12); (g) Group test between *pro*-drop verbs vs. non-*pro*-drop verbs, and apply logistic regression to test predictability of character salience on dropping behavior (see group results in Table 4.13, 4.14 and Figure 4.4).

4.1.1 Discourse Coherence

Discourse coherence has been studied in various fields, including linguistics, psychology, sociology, and computer science. These studies of discourse coherence can date back to the publication of *Cohesion in English* by Halliday and Hasan (1976), and theories of discourse coherence have been developed based on assorted theoretical systems.

Halliday and Hasan focused on linguistic features that realize semantic relations and emphasized the role of formal markers ("cohesive devices") to express coherence (Halliday & Hasan, 1976). They categorized "cohesive devices" into: Reference, Conjunction, and Lexical. In their categorization, Reference includes pronominal, demonstrative, comparative, substitution, and ellipsis; Conjunction includes additive, adversative, causal, and temporal; Lexical includes reiteration, synonymy, and hyponymy. Within Halliday and Hason's framework, the zero pronoun itself is also a "reference cohesive device", and theoretically all the other Reference, Conjunction, and Lexical devices are working collaboratively to provide information to make zero pronouns resolvable.

Later theories, such as the ones brought up by Van Dijk, Fries, Brown, and Yule, provide a more dynamic and cognitive view of discourse coherence. Van Dijk considered coherence as a semantic property of discourse based on the interpretation of each sentence relative to the rest of the sentences (Y. Wang & Guo, 2014). He distinguished discourse coherence at the level of "linear coherence" and "global coherence", which provided coherence descriptions at different scopes of the discourse (Van Dijk, 1977). Fries (1983) associated discourse coherence with themantic progression, and claimed that the degree of discourse coherence is influenced by the connectivity of themes. Brown et al. (1983) emphasized the importance of mentally stored knowledge or backward scenes while interpreting the discourse coherence.

Considering the theories mentioned above together, we shall see that the linguistic factors contributing to zero pronoun resolution are working dynamically and collaboratively throughout the discourse, and it would be valuable to explore how the themes are linked together by the rhemes, especially the main verbs since they can be compared correspondingly without the rest of the rhemes being heterogeneous. Therefore, it is necessary to build an information representation of the discourse to locate the themes and their verb chains. In the next section (see Section 4.1.2), the tradition of Information Structure and Prague Dependency Annotation Styles adopted in this study will be introduced.

4.1.2 Discourse Information Structure

Information Structure (IS) originated in the tradition of the Prague School, it is concerned with the structuring of sentences and is grounded in theories of how communication works (Féry & Ishihara, 2016). Krifka (2008) description of IS approaches shows its dynamicity trait: "the speaker accommodates his speech to temporary states of the addressee's mind, rather than to the long-term knowledge of the addressee". In IS theory, the speaker and the addressee share knowledge during communication, and the shared knowledge, referred to as Common Ground (CG) in IS, is continuously modified and updated during the process of communication.

Krifka recognized the differences between "the content of the shared knowledge" and "how these contents should be developed in terms of the **relevance** to the current discourse" (Krifka, 2008). In a Chinese discourse, when *pro*-drop happens, zero pronoun itself cannot add new information to CG or make connections to old CG. Instead, the CG is developing based on the predicate of the zero pronoun and its **relevance** to the previous discourse (*i.e.* shared knowledge). Therefore, it is reasonable for us to assume that the zero pronoun's predicate can make enough amount of relevant connections with previously given information to enable the addressee to develop the shared knowledge.

In this study, the discourse information representation follows the annotation traditions in the Prague Dependency Treebank (Bejček et al., 2013; Hajič et al., 2020; Postolache et al., 2005): (1) Morphological layer, including word segmentation of the discourse; (2) Analytical layer, each word annotated with its main syntactic function, such as SUBJ, PRED/ROOT, OBJ, ADV; (3) Tectogrammatical layer, containing main semantic role information, *i.e.* AGENT and PATIENT; (4) Referential layer, containing the bridging reference between zero pronouns and their antecedents, and the character referencing for NPs.

4.2 Method

4.2.1 Discourse Material

The discourse materials used in this experiment are Chinese, Brazilian Portuguese, and Spanish translations (xiaowangzi.org, 2021) of Saint-Exupéry's *The Little Prince*. The Chinese discourse contains 2802 clauses

and 16010 words, and the word tokenization was manually checked by native Chinese speakers. The Brazilian Portuguese discourse contains 2853 clauses and 13440 words. The Spanish discourse contains 2315 clauses and 12932 words.

4.2.2 *Pro*-drop Annotation

Omitted subjects and objects are manually resolved using numerical indices from 1 to 32 (see Table 4.2). In this study, we focus on just story characters in the Agent semantic role. The subject omission judgment is made based on every verb: As the lexical verbs are recognized by Part-Of-Speech tagging (*i.e.* auxiliary verbs are not included in the analyses), based on whether they have pronounced subject, each verb will be marked as "with omitted subject" or "not with omitted subject". Therefore, the main verb 'want' and the verb in the non-finite-clause 'go' in the case of "[John] wants to go to London" are both marked as verbs with omitted subjects. Similarly, the verbs in the subject-extracted relative clauses (*e.g.* the verb 'came' in "The girl who came to London") are treated as subject-omitted verbs as well. This annotation rule applies to all the experiments presented in this dissertation.

Table 4.2: The number of occurrence of each character in the annotated discourse

Character Label	Story Character	Number of Character Occurrence		
		CN	BP	ES
ch1	the storyteller	475	376	341
ch2	the boa	13	10	6
ch3	the grownups	32	25	29
ch4	the little prince	795	615	583
ch5	the sheep	32	20	24
ch6	the astronomer	9	7	5
ch8	readers	39	17	24
ch7	the ruler	1	1	1
ch9	the baobab	5	2	9
ch10	the seed	10	5	5
ch11	the planet	37	15	4
ch12	the rose	122	104	107
ch13	the children	4	6	8
ch14	the red-faced man	9	7	6
ch15	the tiger	3	1	1
ch16	the drafts	2	1	2
ch17	the volcano	14	11	5
ch18	the king	83	65	65
ch19	the general	5	7	6
ch20	the conceited man	16	14	13
ch21	the drunk man	18	14	14
ch22	the businessman	49	45	42
ch23	the lamplighter	55	45	53
ch24	the geologist	51	42	43
ch25	the explorer	8	7	9
ch26	the snake	42	33	33
ch27	the echo	5	3	3
ch28	the fox	57	53	48
ch29	the switchman	10	9	7
ch30	the train	6	10	6
ch31	the travelers	18	3	10
ch32	the merchant	4	4	3

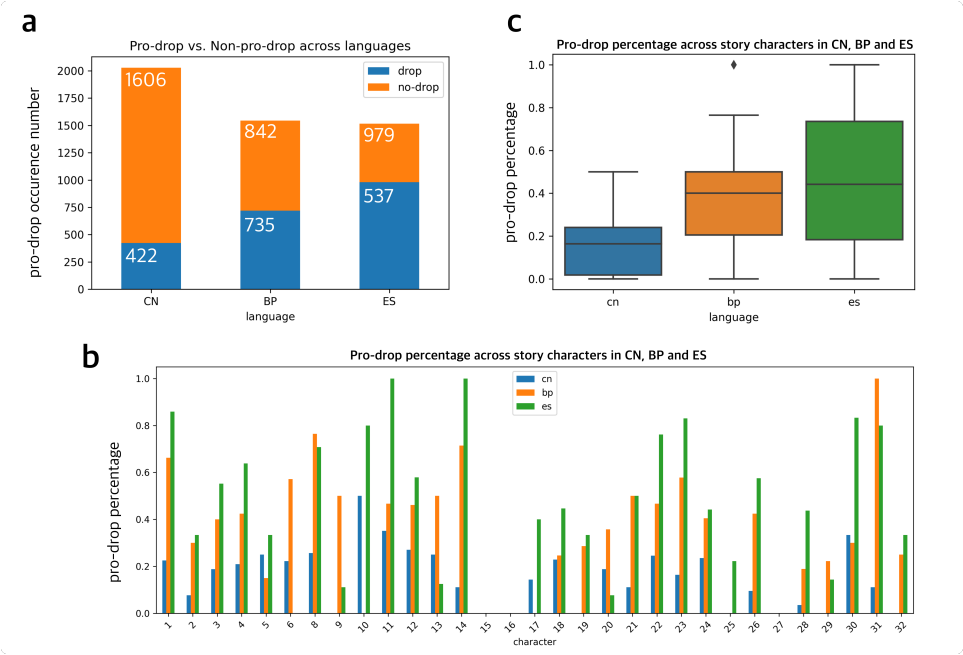


Figure 4.3: Pro-drop rates in Chinese (CN), Brazilian Portuguese (BP), and Spanish (ES). (a) Pro-drop and non-pro-drop occurrences number in CN, BP, and ES; (b) Pro-drop percentage among all story characters; (c) Pro-drop percentage distribution among all story characters.

4.2.3 Dynamic Character-Verb Usage Table

Based on the dependency annotation table, the verbs used for each character are extracted and entered in a second table, the Character-Verb Usage Table (see example in Table 4.3). This table includes the following features: (1) *verb*, the original text of the verb; (2) *verb_id*, the index of the verb in the whole discourse; (3) *agent/patient_character*, the verb's agent or patient story character; (4) *pro_drop*, whether the verb has *pro*-drop; (5) *ch[1-32]_prev_verbs*, for characters 1 through 32, their corresponding previous verbs and indexes are stored as lists.

The dynamic character-verb usage table includes the previous verbs for each story character until a "current verb", and this indicates the verb usage history of each character. By transforming these verb usage histories into numerical vectors, it is possible to use a simple notion of similarity to formalize discourse coherence. For example, as shown in Table 4.3 (translated in Table 4.4), the verb is "回来"(hui2lai2, means "come back"), and it has the following features: (1) *verb_id* shows the verb is the 16008th verb in

Table 4.3: Example of Verb-Character table in Chinese (see a translation of this table in Table 4.4).

verb	回来
verb_id	16008
agent_character	ch4
pro_drop	False
ch1_prev_verbs	[只有, 看到, 想, 用, 画, 画, 让, 画,...]
ch2_prev_verbs	[咀嚼, 吞, 动弹, 消化, 消化, 开, 闭, 闭,...]
ch3_prev_verbs	[理解, 看, 懂, 需要, 解释, 劝, 靠, 弄,...]
ch4_prev_verbs	[朝, 望, 出现, 给, 像, 没有, 像, 干,...]
ch5_prev_verbs	[病, 需要, 像, 睡, 去, 用, 跑, 跑,...]
...	...
ch30_prev_verbs	[运载, 发, 往, 朝着, 开, 过]
ch31_prev_verbs	[寻找, 回来, 满意, 住, 追随, 追随, 睡觉, 打哈欠,...]
ch32_prev_verbs	[说道, 贩卖, 卖, 说]

the discourse; (2) *agent_character* indicates its agent story character is ch4 the little prince (as listed in Table 4.2); (3) *pro_drop* shows this verb does not have *pro*-drop in this sentence; (4) *ch[1-32]_prev_verbs* show the verbs that are correspondingly used with each story character before this verb's presentation.

4.2.4 History-verb and Current-verb Relevance

The idea behind comparing the history verbs and the current verb for each story character is to calculate a numerical similarity level between the current verb and preceding verbs that are part of one or another Topic Chain. Inspired by Sperber and Wilson (1986), we define a quantity called Relevance, a time-weighted function of vector similarity with preceding predicates. The Relevance evaluation process adopts two types of word embeddings (see Section 4.2.4 for details), and steps for the evaluation are introduced in Section 4.2.4.

Word Embeddings Methods

Word embeddings allow each word to be mapped to a single point in a vector space. Under the Distributional Hypothesis (see *e.g.* (Lenci, 2018)), words with similar meanings should be closer in vector space (for a textbook introduction, see Pilehvar and Camacho-Collados (2020)). We use this idea to calculate

Table 4.4: Translation of Table 4.3: Example of Verb-Character table.

verb	come back
verb_id	16008
agent_character	ch4
pro_drop	False
ch1_prev_verbs	[have, see, want, use, draw, draw, let, draw,...]
ch2_prev_verbs	[chew, swallow, move, digest, digest, open, close, close,...]
ch3_prev_verbs	[understand, see, understand, need, explain, advise, lean, play,...]
ch4_prev_verbs	[turn, watch, show up, give, alike, (not) have, alike, do,...]
ch5_prev_verbs	[sick, need, alike, sleep, go, use, run, run,...]
...	...
ch30_prev_verbs	[carry, send, go, turn, drive, pass]
ch31_prev_verbs	[look up, come back, satisfy, live, follow, follow, sleep, yawn,...]
ch32_prev_verbs	[speak, sell, sell, say]

the similarity between the main verb of an omitted pronoun and the verb chains of story characters that might serve as that omitted pronoun's referent.

We use two types of word embeddings: GloVe and BERT. The GloVe model (Pennington et al., 2014) learns word embedding from the term co-occurrence matrix by minimizing the reconstruction error. GloVe has a large context window, which allows it to capture longer-term dependency features. The BERT model (Kenton & Toutanova, 2019) consists of multi-layer bidirectional transformer encoders. BERT is trained on two unsupervised tasks: predict masked tokens, and predict the next sentence, and the BERT embeddings reflect contextual corpus features. All word embeddings we used were trained on large language corpora based on each language correspondingly, and contain contextual word knowledge that carries semantic, syntactic, and pragmatic features.

In this study, BERT and GloVe models are applied with spaCy[1] and Huggingface[2] models for CN, PT, and ES (see Table 4.5 for detail.). A baseline model with 768-dimension random value vectors is adopted to calculate the baseline relevance as compared to the other word embedding models.

[1]https://spacy.io/models
[2]https://huggingface.co/models

Table 4.5: Word embedding models sources and training information.

Model	Language	Pretrained model source	Training material	Dictionary size	Word embedding vector size
GloVe	CN	zh_core_web_lg in spaCy	OntoNotes 5, CoreNLP Universal Dependencies Converter, and Explosion fastText Vectors.	500,000	300
	BP	pt_core_news_lg in spaCy	UD Portuguese Bosque v2.8, WikiNER, and Explosion fastText Vectors.	500,000	300
	ES	es_core_news_lg in spaCy	UD Spanish AnCora v2.8, WikiNER, spaCy looksup data, and Explosion fastText Vectors.	500,000	300
BERT	CN	zh_core_web_trf in spaCy	OntoNotes 5, CoreNLP Universal Dependencies Converter, and bert-base-chinese	-	768
	BP	neuralmind/bert-base-portuguese-cased in Huggingface	brWaC corpora	-	768
	ES	dccuchile/bert-base-spanish-wwm-cased in Huggingface	josecannete/spanish-corpora	-	768

The GloVe word embeddings are obtained from the following models in spaCy for each language: the *zh_core_web_lg* model for Chinese, the *pt_core_news_lg* model for BP, the *es_core_news_lg* model for Spanish. The GloVe model (Pennington et al., 2014) relies on word co-occurrence in the training corpus, and considers the ratios of word-word co-occurrence probabilities to encode semantic information. All three models have 500,000 unique vectors with a dimension size of 300. We obtained the word vectors by searching up the target CN/BP/ES word in the word dictionary.

The BERT word embeddings are retrieved from the following models: (1) Chinese: *zh_core_web_trf* model in spaCy; (2) BP: *neuralmind/bert-base-portuguese-cased* in Huggingface; (3) Spanish: *dccuchile/bert-base-spanish-wwm-cased* in Huggingface. The word embedding vectors were obtained by grouping every 50 words in the discourses, and the model inputs were the 50 words combined as a string (with space between the words). The dimension of the BERT word embedding is 768 for all three BERT models. If there were more than 1 character in a word, their vectors' mean value was used as the word embedding for the whole word. For example, in Chinese, the word "只有" 's embedding was calculated by averaging its subwords' embedding vectors of "只" and "有".

Baseline Word Vectors were 768-dimension vectors generated randomly in the range -1 to 1. The same analysis steps are applied to this model as a baseline.

Relevance Evaluation

The relevance between history verbs and current verbs is calculated based on their word embedding similarities (see Section 4.2.4 for details). At the same time, a weight decay function is applied to the influence of each history verb based on its distance to the current verb, and the function used here is a vanilla value decreasing function (see Equation 4.1), in which ω refers to the weight applying on the similarity, d refers to the clause distance between the verbs being compared, and j, k are the clause numbers the verbs are in:

$$\begin{aligned} \omega(j,k) &= 1/(d+1) \\ d &= |j-k| \end{aligned} \tag{4.1}$$

In this study, the "word embedding similarity" method is realized by calculating the Cosine Similarity between two word embedding vectors. As shown in Equation 4.2, v_{prev} refers to a word embedding vector of a previous verb, and v_{curr} refers to the one for the current verb:

$$R(v_{prev}, v_{curr}) = \frac{v_{prev} \cdot v_{curr}}{||v_{prev}||||v_{curr}||} \tag{4.2}$$

Therefore, the clause-distance-weighted similarity between history verbs and the current verb is shown as Equation 4.3, in which n refers to the number of verbs in the history verb list for a character, and *cl_prev_i* and *cl_curr* refer to the clause numbers that the previous verb and the current verb are in correspondingly.

$$\begin{aligned} &R_{weighted}([v_{prev\_1}, ..., v_{prev\_n}], v_{curr}) = \\ &\sum_{i=1}^{n} \omega(cl\_prev\_i, cl\_curr) * R(v_{prev\_i}, v_{curr}) \end{aligned} \tag{4.3}$$

Via Equation 4.3, for a current verb, each story character has a corresponding relevance value: if the value is higher, the distance-weighted word embedding similarity between history verbs and the current verb is higher; and vice versa.

Table 4.3 shows an example of a verb and the history verbs for characters 1 through 32. The GloVe, BERT, Word2Vec, and Baseline embeddings are used to calculate the average relevance of the history verbs to each current verb for each story character.

Regressors obtained from relevance evaluation introduced in this section are shown in Table 4.6. The average similarity is calculated following Equation 4.2 and 4.3. Both distance-weighted and distance-unweighted relevance are explored to see whether clause distance would play a role.

Table 4.6: Regressors obtained after the relevance calculation

Regressor Number	Regressor Name	Regressor Meaning
1	verb	the verb in the discourse acting as a main verb of a clause
2	verb-id	the word order id of this verb in the original discourse
3	agent-character	the story character referred by the agent of the verb
4	pro-drop	whether this agent is dropped in the discourse
5 - 36	ch{1-32}-prev-verbs	the previous verbs used by each story character till the current verb
37 - 68	rel-glove-ch{1-32}	relevance obtained by GloVe word embeddings
69 - 100	rel-bert-ch{1-32}	relevance obtained by BERT word embeddings
102 - 132	rel-baseline-ch{1-32}	relevance obtained by Baseline word vectors

As shown in Table 4.7, the relevance calculation results of the last verb are presented as an example.

4.2.5 Character Salience

With the relevance between history-current verbs computed as described in the previous section, we have a similarity value for each story character to the current verb. This character salience value refers to whether a

Table 4.7: Example of relevance results for the last verb in Chinese

Relevance Regressor	**(Non-weighted relevance, Weighted relevance)**
rel_glove_ch1	(81.89066125531684, 0.32419914580071807)
rel_glove_ch2	(1.8756812506219913, 0.0015036837567098 64)
...	...
rel_glove_ch32	(0.8230171383397842, 0.0012626916691938 39)
rel_bert_ch1	(176.59183087820725, 0.6119750732174682)
rel_bert_ch2	(4.919826668243348, 0.0027848581443943223)
...	...
rel_bert_ch32	(0.867459723760406, 0.0013292740337137 14)
rel_baseline_ch1	(-0.771830408650495, 0.0080051416478193 33)
rel_baseline_ch2	(-0.008373434318707955, 5.9110606393949324e-05)
...	...
rel_baseline_ch32	(0.0882713253972534 4, 0.0001352612744723 8275)

story character stands out compared to other candidate characters. The salience level function is described in Equation 4.4. In Equation 4.4, k refers to character_k, and the relevance values were calculated based on its history-current verbs by Equation 4.3.

$$S(k) = \frac{\sum_{i=1}^{n} \left(\frac{R^{*}_{weighted}(k)}{R^{*}_{weighted}(i) + R^{*}_{weighted}(k)} \right)}{n} \tag{4.4}$$

$$\mathbb{R}^{*}_{weighted} = \mathbb{R}_{weighted} - min(\mathbb{R}_{weighted}) \tag{4.5}$$

Ranged Character Salience

Instead of taking all 32 story characters as candidates for the salience value calculation, the ranged candidates' salience compares the correct character's accumulated relevance value to the ones within a certain number of clauses. We consider candidates within 10, 20, and 30 clauses for this ranged salience.

4.2.6 Clause-Verb Alignment Across Languages

Due to translation and language pragmatics features, the clause structures and verb usages are not perfectly aligned across the three language discourses. Therefore, an alignment annotation step was taken to locate the clauses that use meaning equivalent verbs or verb phrases.

As shown in Table 4.8, it includes the verbs that are aligned in the first two sentences. In each row, if not empty, the verbs are consistent across languages. There are cases that one language use various morphological forms to represent the same meaning: "打扮"(da3ban4, means "dress up") in CN, "embellecerse"(means "beautify") in ES, and "preparar beleza"(means "prepare beauty") in BP. Since the salience calculation is based on *previous* verbs, cases like this are considered verb-matching cases as well as long as they have the same agent.

Table 4.8: Clause-Verb alignment across Chinese, BP, and Spanish discourses example.

id	CN verb	CN id	ES verb	ES id	BP verb	BP id
0	只有	3			tinha	3
1	描写	11	ilustraba	11		
2	叫	16	titulaba	20		
3	看到	23	vi	6	vi	6
4	画	30	representaba	24	representava	21
5	吞食	37	tragaba	31	engolia	26

4.2.7 Verb Suffix Syncretism Level

First of all, it should be noted that there is no conjugation on the verbs in Chinese (*i.e.* no verb inflection change as the verbs' subjects vary), whereas Spanish and Portuguese both have conjugation. In previous studies, Spanish and Portuguese are referred to as "partial *pro*-drop" because of their rich inflection that provides subject-verb agreement features (Holmberg, 2005). The verb conjugations in Spanish and Brazilian Portuguese can reflect Pronoun type (*i.e.* 1st, 2nd, 3rd person pronoun; singularity, plurality), Mood (*i.e.* Indicative, Subjunctive, Imperative), Tense (*e.g.* Present, Imperfect, Conditional, Future), and Verb Form (*e.g.* -IR, -ER, -AR). These morphemes combine with verb stems to provide partial information about *pro*-drop.

Although both Spanish and Brazilian Portuguese have verb conjugations, their verb suffix's type frequencies show different distributions. This difference is supporting their various levels of agreement (see Table 4.9): Spanish has richer agreement information provided by the verb conjugation forms; Compared to Spanish and European Portuguese, Brazilian Portuguese verb conjugation morphemes are more commonly shared across different pronoun types, verb forms, and tenses. For example, conjugations for 2nd.sg and 3rd.sg are commonly the same when their tense and mood are the same. In this sense, the more a morphological syncretism is shared, the less information it can provide to *pro*-drop resolution. It is reasonable to assume that the level of the verb suffix's type frequency is correlated with their *pro*-drop cases' verb-continuity level since more context information is required to understand "who or what" has been dropped.

Table 4.9: Agreement features across languages (J. Zhang, 2016)

Group	***pro-drop languages***	***Agr***
A	Italian, Spanish etc.	++ Agr
B	German, Scandinavian, Modern Hebrew, Turkish, Esperanto, Occitan, Catalan, Portuguese, Romanian (except French), Croation, Brazilian Portuguese, Finnish, Marathi etc.	+ Agr
C	Chinese, Korean, Japanese etc.	- Agr

Verb Suffix and Syncretism Level Retrievment

Verb suffix was obtained by comparing the "word" and its "lemma", and the lemma is retrieved with spaCy's Lemmatizer [3]. The suffix is equal to the word string without its "stem string" in its lemma string: "word - (lemma - ir/er/or/ar)". For example, the verb "engolia" and its lemma "engolir" will result in a stem string as "engol" and a suffix as "ia".

The verb suffix syncretism level is represented by a numerical value of how many times a verb suffix is shared across all the conjugation forms (see Appendix Table A.1, A.2 for conjugation rules in BP and ES). In this study, the suffix syncretism level is calculated based on the number of pronoun types a suffix can be used (see Table 4.10, 4.11). For example, in BP, the suffix "-ia" can be used when its subjects are 1SG/2SG/3SG so that its syncretism level is considered as 3; Similarly, in Spanish, the suffix "-aba" can be used when its subjects are 1SG/3SG so that its syncretism level was considered as 2.

[3] https://spacy.io/api/lemmatizer

suffix	pronoun	syncretism
era	['2SG', '3SG', '1SG']	3
nha	['2SG', '3SG', '1SG']	3
asse	['2SG', '3SG', '1SG']	3
s	['2SG', '3SG', '1SG']	3
ava	['2SG', '3SG', '1SG']	3
ara	['2SG', '3SG', '1SG']	3
e	['2SG', '3SG', '1SG']	3
ia	['2SG', '3SG', '1SG']	3
a	['2SG', '3SG', '1SG']	3
ira	['2SG', '3SG', '1SG']	3
isse	['2SG', '3SG', '1SG']	3
-	['2SG', '3SG', '1SG']	3
esse	['2SG', '3SG', '1SG']	3
puser	['3SG', '1SG']	2
iu	['2SG', '3SG']	2
issem	['3PL', '2PL']	2
essem	['3PL', '2PL']	2
á	['2SG', '3SG']	2
eu	['2SG', '3SG']	2
nham	['3PL', '2PL']	2
em	['3PL', '2PL']	2
aram	['3PL', '2PL']	2
iram	['3PL', '2PL']	2
am	['3PL', '2PL']	2
ou	['2SG', '3SG']	2
assem	['3PL', '2PL']	2
ão	['3PL', '2PL']	2

eram	['3PL', '2PL']	2
avam	['3PL', '2PL']	2
iam	['3PL', '2PL']	2
íramos	['1PL']	1
íssemos	['1PL']	1
imos	['1PL']	1
i	['1SG']	1
ássemos	['1PL']	1
emos	['1PL']	1
mos	['1PL']	1
áramos	['1PL']	1
nho	['1SG']	1
ei	['1SG']	1
eramos	['1PL']	1
puseres	['2SG']	1
éssemos	['1PL']	1
o	['1SG']	1
ávamos	['1PL']	1
êssemos	['1PL']	1
nhamos	['1PL']	1
amos	['1PL']	1
íamos	['1PL']	1

Table 4.10: Verb suffix and syncretism level across types of pronouns for Brazilian Portuguese.

suffix	pronoun	syncretism
a	['2SG', '3SG', '1SG']	3
e	['2SG', '3SG', '1SG']	3

aba	['3SG', '1SG']	2
re	['3SG', '1SG']	2
se	['3SG', '1SG']	2
ra	['3SG', '1SG']	2
ía	['3SG', '1SG']	2
isteis	['2PL']	1
é	['1SG']	1
´ramos	['1PL']	1
en	['3PL']	1
seis	['2PL']	1
sen	['3PL']	1
o	['1SG']	1
an	['3PL']	1
ses	['2SG']	1
ían	['3PL']	1
aste	['2SG']	1
áis	['2PL']	1
es	['2SG']	1
rd	['2PL']	1
´semos	['1PL']	1
asteis	['2PL']	1
imos	['1PL']	1
ras	['2SG']	1
aban	['3PL']	1
á	['3SG']	1
ran	['3PL']	1
í	['1SG']	1
án	['3PL']	1

ás	['2SG']	1
íamos	['1PL']	1
amos	['1PL']	1
ís	['2PL']	1
-	['1SG']	1
éis	['2PL']	1
ábamos	['1PL']	1
emos	['1PL']	1
res	['2SG']	1
rais	['2PL']	1
iste	['2SG']	1
ó	['3SG']	1
abais	['2PL']	1
ren	['3PL']	1
ieron	['3PL']	1
abas	['2SG']	1
íais	['2PL']	1
aron	['3PL']	1
´remos	['1PL']	1
reis	['2PL']	1
ías	['2SG']	1
ió	['3SG']	1
as	['2SG']	1

Table 4.11: Verb suffix and syncretism level across types of pronouns for Spanish.

Table 4.12: Example of salience results for the last verb from three language models and one baseline model with distance-weighted/-unweighted

	Language				
Regressor	**CN-text/lemma**	**BP-text**	**BP-lemma**	**ES-text**	**ES-lemma**
correct-character	回来	voltou	voltar	vuelto	volver
id	16008	15046	15046	15778	15778
pro-drop	False	False	False	False	False
salience_glove (weighted, unweighted)	(0.967689, 0.963752)	(0.969265, 0.968783)	(0.955526, 0.961589)	(0.962786, 0.963063)	(0.957839, 0.959486)
salience_bert (weighted, unweighted)	(0.970634, 0.956799)	(0.96507, 0.96549)	(0.959618, 0.962476)	(0.960672, 0.961095)	(0.95855, 0.96048)
salience_baseline (weighted, unweighted)	(0.126289, 0.081307)	(0.539935, 0.907646)	(0.539935, 0.907646)	(0.032258, 0.662265)	(0.032258, 0.662265)
salience_glove_10 (weighted, unweighted)	(0.59824, 0.859003)	(0.77872, 0.898348)	(0.651214, 0.863248)	(0.729808, 0.869757)	(0.681501, 0.861158)
salience_bert_10 (weighted, unweighted)	(0.626225, 0.838399)	(0.729128, 0.884995)	(0.679851,0.87012)	(0.70354, 0.864203)	(0.686033, 0.862122)
salience_baseline_10 (weighted, unweighted)	(0.807086, 0.349199)	(0.83752, 0.93955)	(0.83752, 0.93955)	(0.333333, 0.872069)	(0.333333, 0.872069)
salience_glove_20 (weighted, unweighted)	(0.59824, 0.859003)	(0.77872, 0.898348)	(0.651214, 0.863248)	(0.729808, 0.869757)	(0.681501, 0.861158)
salience_bert_20 (weighted, unweighted)	(0.626225, 0.838399)	(0.729128, 0.884995)	(0.679851, 0.87012)	(0.70354, 0.864203)	(0.686033, 0.862122)
salience_baseline_20 (weighted, unweighted)	(0.807086, 0.349199)	(0.83752, 0.93955)	(0.83752, 0.93955)	(0.333333, 0.872069)	(0.333333, 0.872069)
salience_glove_30 (weighted, unweighted)	(0.770924, 0.902077)	(0.853577, 0.931085)	(0.76964, 0.903687)	(0.813349, 0.906665)	(0.787952, 0.900067)
salience_bert_30 (weighted, unweighted)	(0.78569, 0.879003)	(0.820576, 0.921579)	(0.789982, 0.909178)	(0.804469, 0.903863)	(0.792983, 0.901728)
salience_baseline_30 (weighted, unweighted)	(0.647478, 0.18402)	(0.705703, 0.923602)	(0.705703, 0.923602)	(0.2, 0.775966)	(0.2, 0.775966)

4.3 Results

4.3.1 Character Salience: *Pro*-drop vs. Non-*pro*-drop

The correct story character's salience compared to all other characters was calculated following Equation 4.4. For each verb, we calculated a salience value for the correct story character. See Table 4.12 for an example of the salience values of the last verb in the Chinese discourse.

The distributions for the salience value obtained from two word embedding models and one baseline model are shown in Figure 4.4. The pro-drop and non-pro-drop salience distribution are compared within three languages: Chinese (CN), Brazilian Portuguese (BP), and Spanish (ES).

Single-sided non-parametric two-sample Wilcoxon Tests are carried out between *pro*-drop and non-*pro*-drop character salience obtained from the two word embedding models and the baseline model. The test results with FDR correction are shown in Table 4.13, 4.14. For distance-weighted models, all BERT

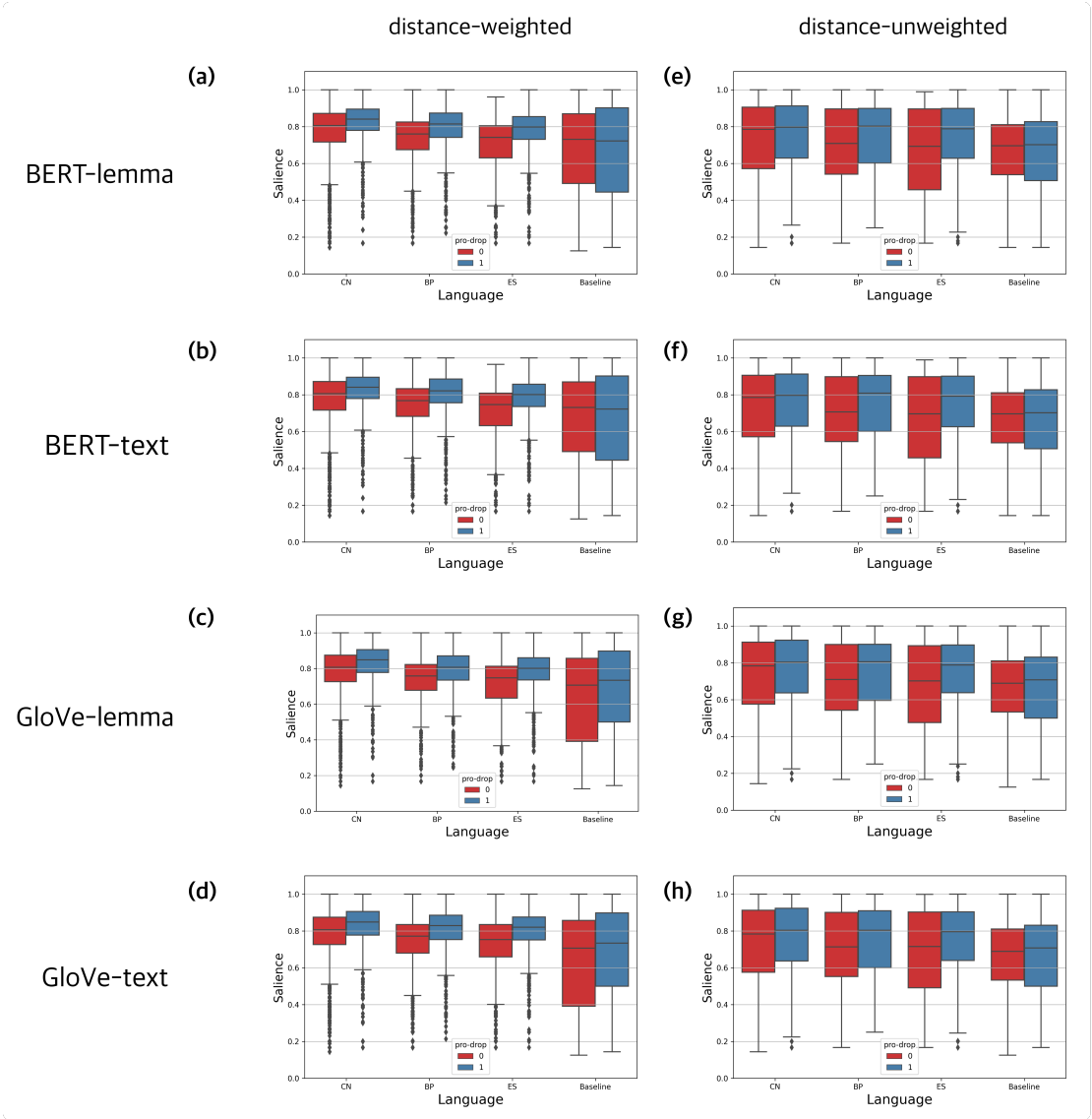


Figure 4.4: Salience distributions from the word embedding models (GloVe, BERT, and Baseline) across three languages (CN, ES, and BP). Salience values shown in this figure were based on "verb distance within 30 clauses". (See the complete statistical test results for all conditions in Table 4.13) Table 4.13 shows salience distribution based on distance-weighted models; Table 4.14 shows salience distribution based on distance-unweighted models. The blue boxes are *pro*-drop salience cases, and the red ones are non-*pro*-drop. The BERT and GloVe models show significant *pro*-drop > non-*pro*-drop effect for all three languages.

and GloVe models show significant results ($p < 0.05$, marked as red numbers). For distance-unweighted models, ES shows significant *pro*-drop > non-*pro*-drop salience effects for all models, and BP shows significant *pro*-drop > non-*pro*-drop salience effects for "distance < 20/30 models" The Baseline model shows null effects on both distance-weighted and distance-unweighted models for all the ranged cases. As shown in Figure 4.4, the boxplots are consistent with the Wilcoxon tests.

Table 4.13: Salience significance result based on distance-weighted models. Single-sided unpaired Wilcoxon test results: *pro*-drop > non-*pro*-drop based on all word embedding models (GloVe, BERT, Baseline; See detailed model information in Section 4.2.4).

		Salience: pro-drop >non-pro-drop (FWER <0.05) Distance Unweighted							
		Distance = all previous clauses		Distance <30 clauses		Distance <20 clauses		Distance <10 clauses	
Language	**Model**	*W-value*	*p-value*	*W-value*	*p-value*	*W-value*	*p-value*	*W-value*	*p-value*
CN	BERT	338206	0.713	352730	0.167	353633	0.149	354721	0.126
	GloVe	325967	0.997	349654	0.254	351075	0.213	355549	0.114
	Baseline	329045	0.946	332315	0.898	334864	0.823	327115	0.987
BP	BERT-text	310121	0.658	329506	0.027	329761	0.026	317693	0.287
	BERT-lemma	305712	0.834	328561	0.034	329664	0.027	317634	0.287
	GloVe-text	307784	0.750	330060	0.025	329544	0.027	317435	0.293
	GloVe-lemma	311840	0.574	329056	0.030	329462	0.027	316840	0.319
	Baseline	307867	0.750	307100	0.781	306423	0.811	296861	1.000
ES	BERT-text	299688	0.000	293735	0.000	288876	0.002	283213	0.015
	BERT-lemma	300097	0.000	293577	0.000	288615	0.002	283382	0.014
	GloVe-text	297744	0.000	290169	0.001	286338	0.005	282005	0.021
	GloVe-lemma	297560	0.000	291889	0.001	287735	0.003	282619	0.018
	Baseline	258346	0.882	238771	1.000	237980	1.000	230918	1.000

Table 4.14: Salience significance result based on distance-unweighted models. Single-sided unpaired Wilcoxon test results: *pro*-drop > non-*pro*-drop based on all word embedding models (GloVe, BERT, Baseline; See detailed model information in Section 4.2.4).

		Salience: pro-drop >non-pro-drop (FWER <0.05) Distance Weighted							
		Distance = all previous clauses		Distance <30 clauses		Distance <20 clauses		Distance <10 clauses	
Language	**Model**	*W-value*	*p-value*	*W-value*	*p-value*	*W-value*	*p-value*	*W-value*	*p-value*
CN	BERT	393042	0.000	402155	0.000	406585	0.000	393962	0.000
	GloVe	370141	0.004	397289	0.000	400478	0.000	391939	0.000
	Baseline	309258	1.000	345048	0.427	352774	0.167	329799	0.945
BP	BERT-text	379238	0.000	378859	0.000	374418	0.000	353763	0.000
	BERT-lemma	365590	0.000	374922	0.000	370314	0.000	349771	0.000
	GloVe-text	358790	0.000	371476	0.000	366461	0.000	346908	0.000
	GloVe-lemma	368860	0.000	375307	0.000	369541	0.000	345825	0.000
	Baseline	301612	0.945	309092	0.713	301152	0.946	290846	1.000
ES	BERT-text	347580	0.000	313245	0.000	318006	0.000	309532	0.000
	BERT-lemma	343897	0.000	311712	0.000	316923	0.000	309103	0.000
	GloVe-text	322940	0.000	302197	0.000	310689	0.000	303319	0.000
	GloVe-lemma	326993	0.000	306717	0.000	313201	0.000	305856	0.000
	Baseline	267221	0.445	231234	1.000	230839	1.000	220278	1.000

4.3.2 *Pro*-drop Character Salience Among Languages

To compare *pro*-drop characters' salience levels across languages, single-sided non-parametric two-sample Wilcoxon Tests were applied. As shown in Figure 4.5 and the statistical results with FDR correction in Table 4.15, the pro-drop characters' salience values are tend to have the group distribution relationship as: CN > ES > BP (in all distance within 10/20/30 and weighted models).

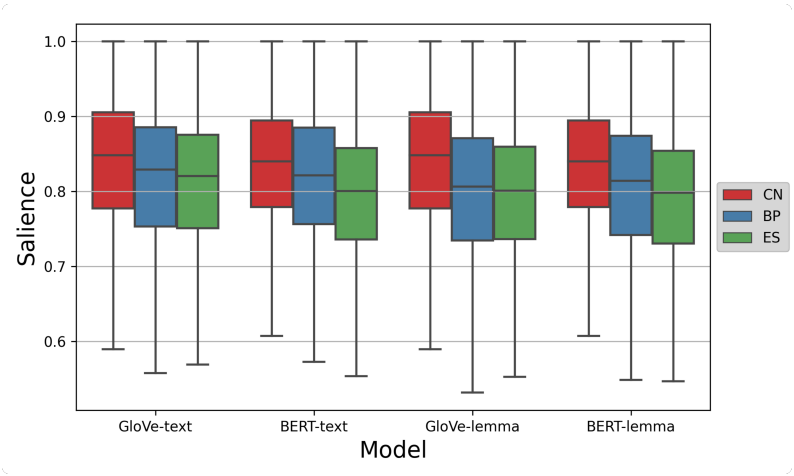


Figure 4.5: *Pro*-drop characters' salience distributions from the word embedding models (GloVe and BERT) across three languages (CN, ES, and BP). Salience calculation was based on verb-distance weighted within 30 clauses.

Table 4.15: Single-sided unpaired Wilcoxon test results: Pro-drop characters' salience between languages based on all word embedding models (GloVe, BERT, Baseline; See detailed model information in Section 4.2.4).

		Pro-drop Characters' Salience Between Languages (FWER <0.05) Distance Weighted							
		Distance = all previous clauses		Distance <30 clauses		Distance <20 clauses		Distance <10 clauses	
Language	**Model**	*W-value*	*p-value*	*W-value*	*p-value*	*W-value*	*p-value*	*W-value*	*p-value*
CN >BP	BERT-text	158597	0.424	207515	0.000	207993	0.000	202315	0.000
	BERT-lemma	151833	1.000	208276	0.000	209633	0.000	205195	0.000
	GloVe-text	148844	1.000	206851	0.000	208447	0.000	204886	0.000
	GloVe-lemma	88185	1.000	188029	0.000	194324	0.000	199040	0.000
CN >ES	BERT-text	82225	1.000	225364	0.007	238077	0.000	242261	0.000
	BERT-lemma	70380	1.000	222444	0.021	236416	0.000	242174	0.000
	GloVe-text	176775	1.000	256095	0.000	260152	0.000	255498	0.000
	GloVe-lemma	94667	1.000	233068	0.000	243066	0.000	247807	0.000
BP >ES	BERT-text	126706	1.000	300849	1.000	314285	1.000	321433	1.000
	BERT-lemma	115896	1.000	296117	1.000	309925	1.000	317146	1.000
	GloVe-text	319094	1.000	330146	1.000	329526	1.000	325580	1.000
	GloVe-lemma	310028	1.000	335847	1.000	335292	1.000	328021	1.000
ES >BP	BERT-text	592860	0.000	418716	0.000	405280	0.000	398132	0.000
	BERT-lemma	603670	0.000	423448	0.000	409640	0.000	402419	0.000
	GloVe-text	400472	0.000	389419	0.004	390039	0.003	393985	0.001
	GloVe-lemma	409538	0.000	383718	0.018	384273	0.016	391544	0.002

4.3.3 *Pro*-drop Character Salience Comparison on Verb-aligned Cases

As introduced in Section 4.2.6, 1487 verb-aligned cases are located across CN, ES, and BP, and 191 cases have character salience values (*i.e.* the rest do not have the aimed story characters in Table 4.2 as agents), and 26 cases are all pro-drop cases among all three languages.

Pro-drop character salience value distributions on 191 aligned verb cases are shown in Figure 4.6.

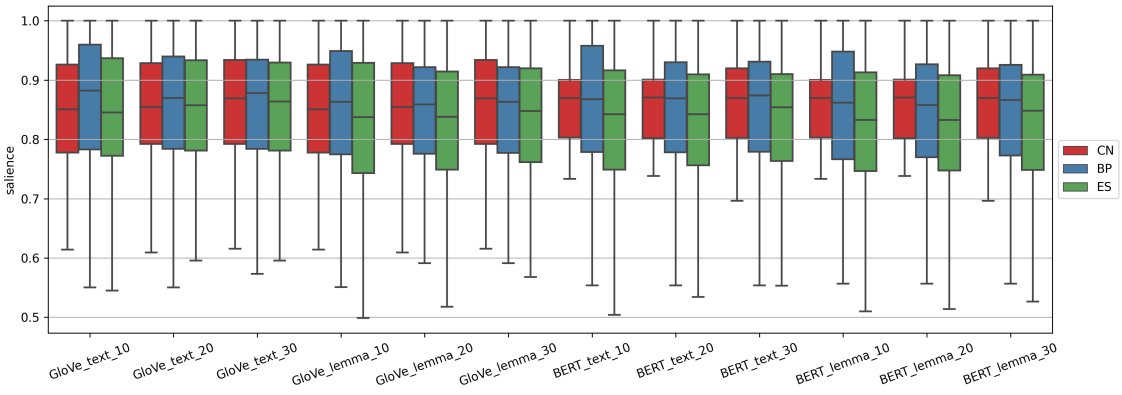


Figure 4.6: *Pro*-drop character salience value distributions on 191 aligned verb cases.

Within the 26 cases of pro-drop and verb-meaning aligned cases, salience difference value distributions were considered based on the difference between CN vs. BP("cn-bp"), CN vs. ES("cn-es"), ES vs. BP("es-bp"). The results are shown in Figure 4.7

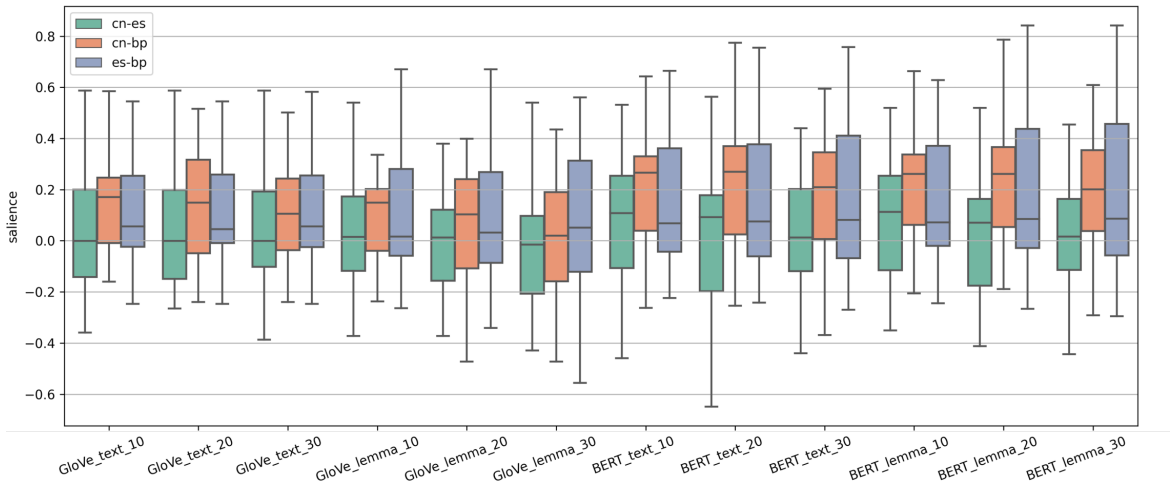


Figure 4.7: *Pro*-drop character salience difference value distributions on 26 all aligned verb cases.

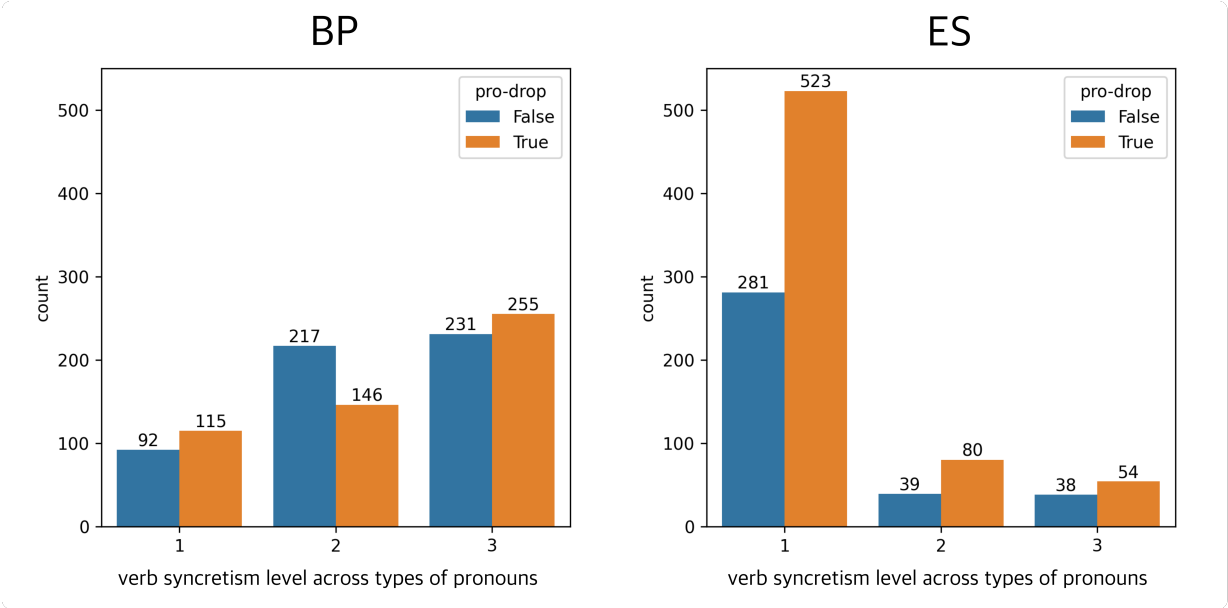


Figure 4.8: *Pro*-drop and non-*pro*-drop occurrences in BP and ES, measured at different verb suffix syncretism levels.

4.3.4 Verb Suffix Syncretism Effect on *Pro*-drop

Verb Suffix Syncretism *vs.* *Pro*-drop Rate

As shown in Figure 4.8, and Table 4.16, although the pro-drop rates in BP and ES do not have a negative linear correlation with the verb syncretism level (obtained from pronoun usage broadness), we observed that the pro-drop rate for verb syncretism levels is ranked as (a) BP, $1 > 3 > 2$, and (b) ES, $2 > 1 > 3$.

Table 4.16: *Pro*-drop and non-*pro*-drop occurrences in BP and ES, measured at different verb suffix syncretism levels.

Language	Verb suffix syncretism level	Pro-drop	Non-pro-drop	Sum	Pro-drop percentage
BP	1	115	92	207	0.556
	2	146	217	363	0.402
	3	255	231	486	0.525
ES	1	523	281	804	0.650
	2	80	39	119	0.672
	3	54	38	92	0.587

Verb Suffix Syncretism *vs.* *Pro*-drop Salience

As shown in Figure 4.9, to examine the suffix syncretism level's effect on pro-drop characters' salience level, all models (verb clause distance < 10) salience values were grouped by their verbs' suffix syncretism level.

Double-sided unpaired Wilcoxon tests with FDR correction were applied on pro-drop and non-pro-drop characters' salience values between verb suffix syncretism levels (*i.e.* level 1 *vs.* level 2, level 1 *vs.* level 3, level 2 *vs.* level 3). The results for BP non-*pro*-drop indicate that the salience values of level 3 are significantly higher than level 1 based on BERT-lemma ($W = 12992$, $p = 0.021$) and BERT-text ($W = 12826$, $p = 0.022$) models, and show marginally significant result that level 3 is higher than level 2 based on BERT-text model ($W = 28413$, $p = 0.057$). No significant result was found among syncretism levels for BP *pro*-drop, and Spanish *pro*-drop and non-*pro*-drop salience values.

4.4 Discussion

In this section, the rationale of quantifying verb continuity as a factor supporting *pro*-drop is reflected in Section 4.4.1, and the interpretation discussion on the results is presented in Section 4.4.2.

4.4.1 Quantifying A Linguistic Factor Across Languages

One of the contributions of this study is the attempt to quantify a linguistic concept based on theories. In previous studies on *pro*-drop, the topic chain theory had pointed out the role of discourse coherence for *pro*-drop resolution, but it left untouched for what discourse elements or features are contributing to this "chain", let alone comparing this effect across languages.

Based on the idea that *pro*-drop cases demand higher discourse consistency in order to be resolved, this study measures on factor, the verb usage continuity, and examines whether *pro*-drop cases tend to have higher verb usage continuity than non-*pro*-drop cases. In this study, the representation of the "verb usage continuity" concept follows the steps including: (1) retrieving the history verb for each story character; (2) calculating the relevance between the history verb and current verb based on word embeddings; (3) calculating the salience of the correct character based on the relevance values in (2). The result of step (2) is straightforwardly measuring what we call "verb usage continuity". The result of step (3) is used to

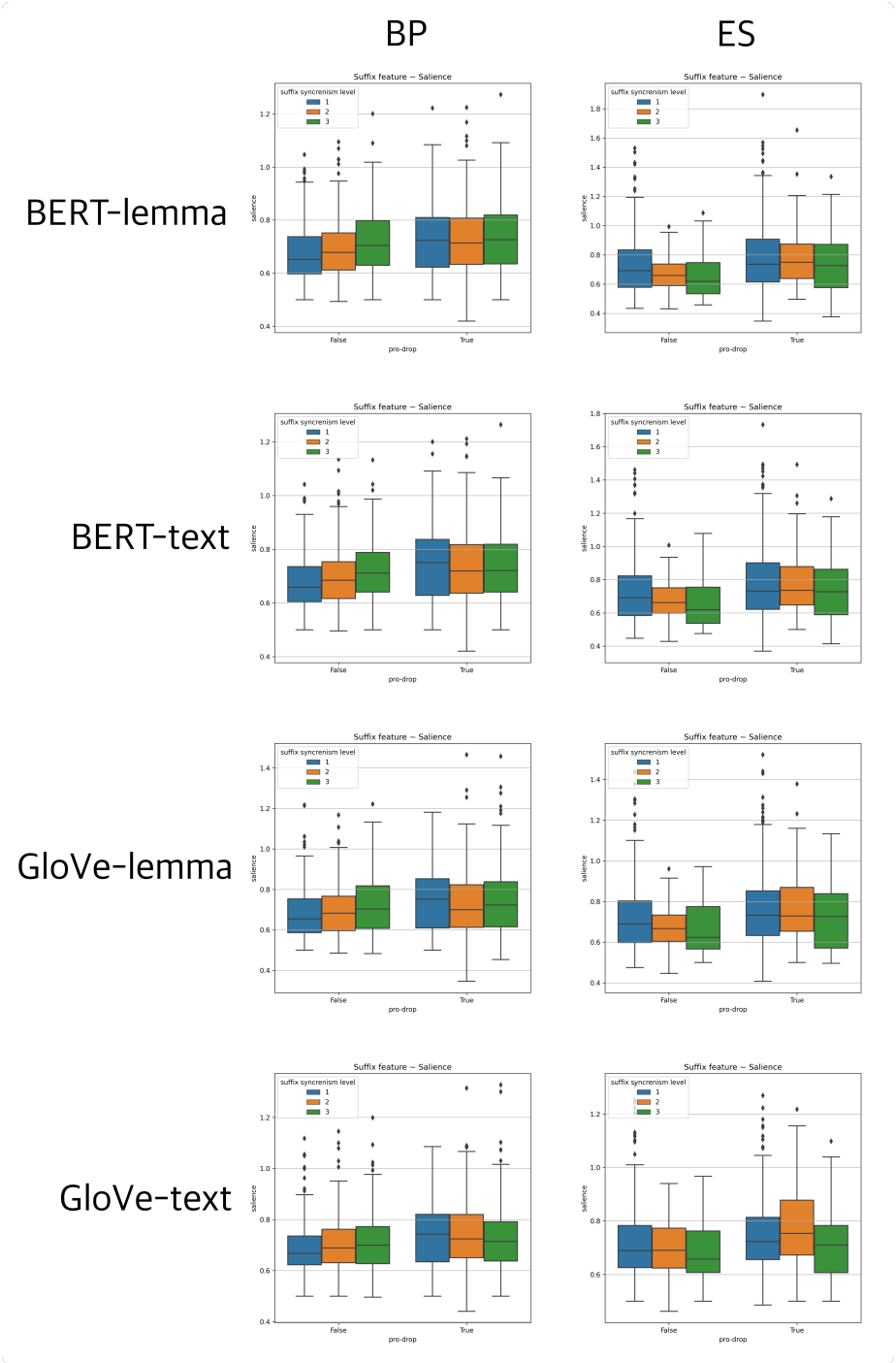


Figure 4.9: *Pro*-drop and non-*pro*-drop Characters' salience distributions from the word embedding models (GloVe and BERT): distribution analysis based on verb syncretism level (see verb syncretism level assessment in Section 4.2.7). Salience calculation was based on verb distance weighted within 10 clauses.

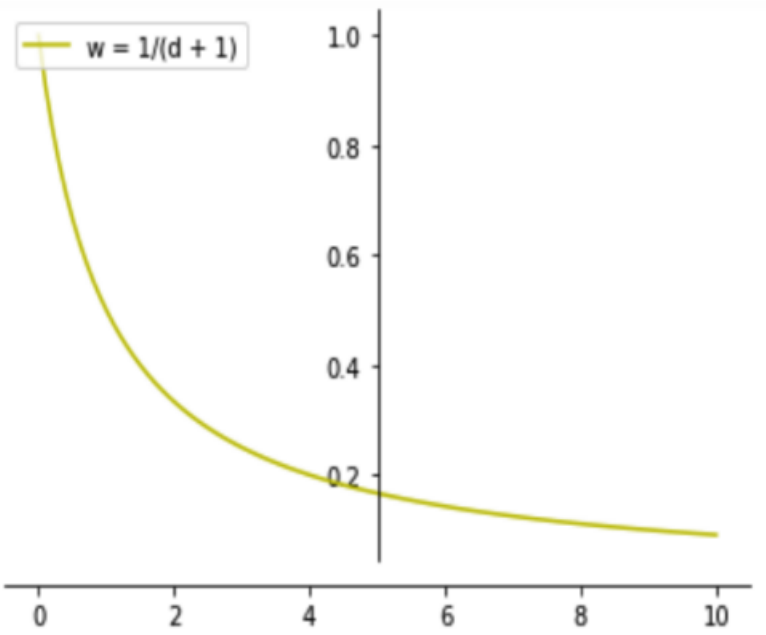


Figure 4.10: Plot of the function $w = 1/(d+1)$. The x-axis is d, and the y-axis is w.

measure "how necessary it is to provide a high consistency context for *pro*-drop cases than non-*pro*-drop cases". To further interpret (3), we can imagine that we are reading a detective story without knowing who is the real agent: All we have is the previous actions (*i.e.* verbs) for each suspect (*i.e. story character candidates*), and as the current action is presented (current verb), we compare this current action with the previous actions to make an inference – the suspect with the highest consistency is the one we are looking for. Hence force, is the one hidden (dropped) required higher consistency to make this inference? That is what this study cares about.

We will explain some formula details used in the mathematical processes, and discuss other possible ways to represent the relationship. Equation 4.1 measures the distance between two verbs and generates a *weight* value to apply to Equation 4.3. As shown in Figure 4.10 for a visualization between distance and weight: as the distance between two verbs (d) increases, the weight that will be applied to the relevance (w) decreases. Of course, this is not the only negative relativity function that can be used (such as $w = -e^{(x)}$), but one of them that can describe the relationship.

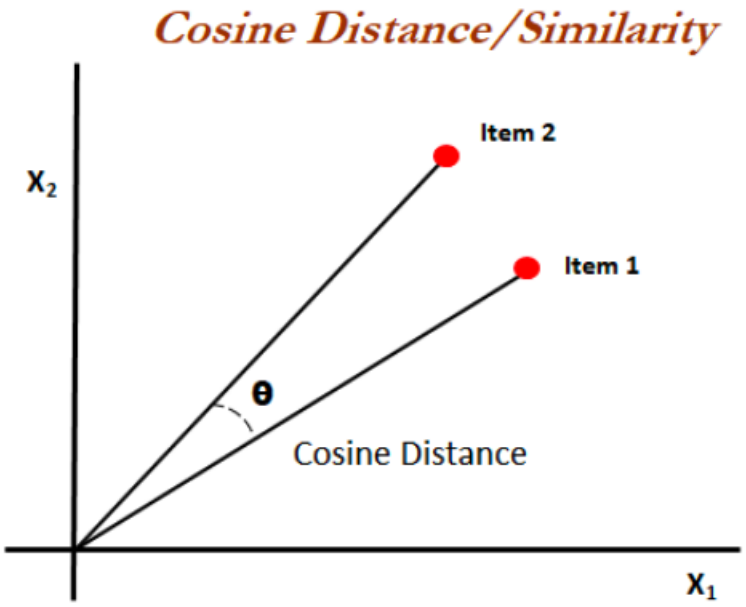


Figure 4.11: Plot of the Cosine Similarity. Item 1 and 2 are two vectors and their distance is θ.

Equation 4.2 calculates the Cosine Similarity (see Figure 4.11) between two vectors (*i.e.* word embeddings). There are also functions such as Euclidean Distance to calculate vector distance, but the Cosine Similarity was chosen here as it is the training standard during the word embedding model training processes.

Equation 4.3 combines the results from Equation 4.1 and Equation 4.2, and it applies to the *history verb list* for a character. This can be understood as adding a penalty to the verbs that are too far away from the *current* verb. As we can see from the results (see Table 4.13,4.14), the weighted results show higher difference between *pro*-drop cases and non-*pro*-drop cases.

Equation 4.4, which seems to be a less straightforward formula, is the process that examines the "suspect" among all the candidate characters. It should be noticed that there are 32 story characters in total in the whole discourse, but as ranged relevance and salience are calculated, only the candidates within a certain range of clauses are taken into consideration. When we talk about the "salience" of the verb usage continuity value, the math representation is expected to be able to show "the highest value is supposed to stand out among the rest". Therefore, as shown in Equation 4.4, the **correct** character's relevance value is

compared in the form of division with the rest of the candidates. However, the number of candidates varies as the current verb and the range of clause changes, and the number n in the equation shows the number of candidates, and the denominator n for $S(k)$ keeps it consistent as the number of ranged candidates varies. The relevance values used in Equation 4.4 are adjusted in Equation 4.5 by deducting the minimum relevance value so that the adjusted weight value is non-negative. As a result, the $S(k)$ value is expected in the range [0,1].

As explained above, the mathematical representation of a relationship is not bound to a single solution, and there are various ways to reach the goal. The features of the quantifying process are: (1) the quantification process is context sensitive, and it uses local discourse information (adding weight to the list of relevance values, and flexible list length as defined); (2) the elements can be improved as future word embeddings are developed; (3) the salience evaluation is normalized so that it can be used as an efficient factor across languages (as CN, BP, and ES are compared in this study).

4.4.2 Results Interpretation

***Pro*-drop rates among languages.** As shown in Figure 4.3, the *pro*-drop rates' order in the discourse for the three languages used in this experiment is CN $<$ BP $<$ ES. This is consistent with the language subject-verb agreement level (N $<$ BP $<$ ES), which indicates that for languages with more local syncretism information provided, the *pro*-drop can happen more frequently.

The main within-language effect: Character salience *pro*-drop $>$ non-*pro*-drop. As shown in Section 4.3.1, for all four word-embedding cases (GloVe-text, GloVe-lemma, BERT-text, BERT-lemma), and for all three languages (CN, BP, and ES), the main effect of verb continuity are all significant (see Table 4.13): *pro*-drop cases tend to have higher verb usage continuity than non-*pro*-drop cases. This main effect is stronger when the distance is weighted during verb usage continuity calculation (See Equation 4.3). First of all, this main effect indicates that the "verb usage continuity" factor that we quantify can distinguish *pro*-drop and non-*pro*-drop cases in the discourse, and showed that the *pro*-drop cases demand higher verb usage continuity be resolved. Second, adding the distance weighted strengthens the main effect, and this supports the accessibility theory that the recency in the discourse plays an effect in *pro*-drop. When the distance is unweighted, the *pro*-drop $>$ non-*pro*-drop tendency is found in ES (all distance cases) and BP

(distance less than 20 and 30 cases). Third, both the text and lemma show significant results, this might reflect the training of word embedding results may have a word's text and lemma being close to each other in the vector space and both of them can represent the word to an extent.

The main cross-language effect: Character salience *pro*-drop across CN, BP, and ES. As we focus on the *pro*-drop cases among the three languages, the order of verb continuity level has the result as CN $>$ BP $>$ ES. One explanation for this trend can be, that this result corresponds to the subject-verb agreement level (CN $<$ BP $<$ ES), and this suggests that the verb usage continuity compensates for the lack of agreement information in radical *pro*-drop (*e.g.* Chinese) and partial *pro*-drop (*e.g.* Brazilian Portuguese) languages than consistent *pro*-drop languages (*e.g.* Spanish).

Verb alignment results. In this follow-up analysis shown in Section 4.3.3, the scope of verbs is limited to those that show up in all three languages, and it ends up in only 26 cases. This small number of cases is not robust enough to make statistical inferences, but it is a direction for future studies with a bigger discourse size to explore.

Verb suffix syncretism level effect on *pro*-drop. As the main cross-language effect indicates the role of subject-verb agreement level playing a role, this follow-up analysis on BP and ES further explores how the information carried by the verb conjugation affects the results. As shown in Figure 4.9, the results for BP non-*pro*-drop indicate that the salience values of level 3 are significantly higher than level 1 based on BERT-lemma and BERT-text models, and show marginally significant results that level 3 is higher than level 2 based on BERT-text model. No significant result was found among syncretism levels for BP *pro*-drop, and Spanish *pro*-drop and non-*pro*-drop salience values. Therefore, this analysis did not confirm the role of verb suffix syncretism affecting the verb usage continuity among *pro*-drop cases. However, the information a verb carries is rooted in many aspects, such as the word frequency in a language, and a verb suffix being common in a conjugation table does not necessarily correspond to its frequency in the language. The verb syncretism level used in this experiment is retrieved from a suffix's occurrence in the conjugation table, and this can lead to a limited representation of the syncretism level. The information carried by a suffix in a language can be affected by other factors such as the word frequency, and this can be explored in future studies to examine what other verb features can affect usage continuity.

4.5 Conclusion

This experiment examined the role of verb continuity as a support factor for *pro*-drop cases in three languages with different agreement levels. The word embedding provides an efficient representation to measure the verb usage continuity in the discourse.

Within each language, it is shown that the verb continuity can distinguish *pro*-drop and non-*pro*-drop cases in all three languages, and the *pro*-drop cases demand higher verb continuity than the non-*pro*-drop cases.

Cross linguistically, the *pro*-drop cases' verb continuity levels among the three languages are CN $>$ BP $>$ ES, which indicates the agreement (Agr) level plays a role, and the higher the agreement level of a language the less it relies on other semantic factors such as verb continuity to resolve a *pro*-drop case.

The quantification process introduced in Experiment 1 signifies a significant step towards bridging the gap between linguistic theories and mathematical formalization. The use of high-dimensional vector values for verbs, provided by pre-trained language models, showcases the potential of leveraging advanced computational tools to quantify linguistic phenomena. While acknowledging that this quantification method may not be the sole or optimal solution, it serves as a proof of concept, illustrating the feasibility of quantifying linguistic theories. This method, rooted in mathematical language, opens avenues for cross-linguistic research, demonstrating that the intricate aspects of language can be systematically analyzed and compared across diverse linguistic contexts. Moreover, Experiment 1 highlights the importance of identifying and quantifying discourse coherence elements in understanding *pro*-drop phenomena. The salience level introduces a quantitative measure, providing a means to assess the competence of story characters in being resolved as the omitted pronoun. This not only enhances our theoretical understanding of *pro*-drop but also sets the stage for future refinements and developments in computational linguistic analyses.

Chapter 5

Experiment 2: Modeling Linguistic Factors for *pro*-drop Using Binomial Logistic Regression and Random Forest Models

5.1 Introduction

In the previous chapter, the effect of "verb usage continuity" on *pro*-drop was explored from a quantitative angle. However, the "verb usage continuity" is not the only factor that contributes to *pro*-drop. In this experiment, more linguistic factors will be considered to explore the question "What factors are supporting *pro*-drop?" and "What are the levels of importance among the factors?". Since *pro*-drop vs. non-*pro*-drop is a binary variable, the binomial logistic regression will be used to model the relationship between the features and the binary output. Besides a commonly used machine learning algorithm, Random Forest is also adopted to measure the importance of the factors from a different angle. In this following section, the linguistic features that are potentially affecting *pro*-drop will be discussed.

Camacho (2013) summarized three typological patterns of Null Subject Languages (NSLs) (M. d. P. P. Barbosa, 1995) (see Chapter 2.1.1 for review):

1. Languages with rich subject agreement morphology, and show a fairly systematic use of null subjects (henceforth consistent NSLs), such as Italian, Spanish, Portuguese, Hungarian, and Greek, among many others. In this type of language, subjects are freely dropped under the appropriate discourse conditions.

2. Languages with agreement and referential null subjects whose distribution is restricted (henceforth partial NSLs) such as Hebrew, Finnish, Marathi, Russian, and colloquial Brazilian Portuguese. Null Subjects in languages are constrained along several dimensions, including the expression of person, tense and referentiality, and antecedent control.

3. Languages that lack agreement, such as Chinese, Japanese, and Korean. These have been described as topic-oriented languages and allow for any argument to be dropped not just subjects. These languages are discourse-related NSLs.

The differences between consistent NSLs (such as European Portuguese and Spanish) and partial NSLs (such as Brazilian Portuguese) are that as P. Barbosa (2011a) mentioned: (i) the Null Subject (NS) is optional in some contexts in which it is mandatory in a consistent NSL; (ii) the NS is excluded in many contexts in which it is possible in a consistent NSL. This indicates the "verbs' morphological" features, such as verb case or tense continuity among clauses, could be a factor that contributes to *pro*-drop.

Apart from the patterns introduced above, Camacho (2013) also summarized a second typology related to NSLs: whether the NSs are restricted to a main or embedded clause. Brazilian Portuguese, Finnish, and Marathi restrict null subjects to controlled instances in embedded clauses.

Pešková (2013) reviewed that the usage of a null or overt subject is mainly motivated by "internal factors" within languages, such as grammatical person, morphological and contextual ambiguity, verb semantics, clause type or switch-reference, in contrast with "external" or social factors. Pešková's Spanish corpus study indicated that the strength of effects of these factors is, listed from highest to lowest: (1) Grammatical person; (2) Type of verb (epistemic vs. perceptive); (3) Type of clause (matrix clause with or without subordinate clause, subordinate clause); (4) Type of sentence (declarative, absolute interrogative, wh-interrogative).

Some *pro*-drop theories-based features can also be considered as potential model features. (1) Centering theory (Walker, Walker, et al., 1998); The "verb usage continuity" idea behind Experiment 1 was inspired

by the Centering Theory, which concerns local coherence and semantic entity salience (Xiao, 2021). The "localness" was incorporated when calculating salience by applying a distance penalty, and the entity salience was directly computed with story characters' verb usage similarity. Also inspired by the Centering Theory, a more straightforward feature, the agent characters' consistency between clauses, can be used as a feature in the model; (2) Discourse representation theory (DRT) (Kamp & Reyle, 2013); In Experiment 2, the between-clause relation will be included as a factor (*e.g.* the discourse relation types such as causal, coordinate, and concession). (3) Accessibility theory (Ariel, 2001); (4) Information Load Hypothesis (Almor, 1999).

Inspired by these previous studies, the following factors will be included in our binomial logistic regression model to examine their contribution effect on *pro*-drop:

Syntactic factors. Clause type (*i.e.* main clause or embedded clause; adverbial clause, relative clause, or noun clause); Depth of the entity in the constituency tree.

Semantic factors. Abstract Meaning Representation (Schneider et al., 2015) suggested sentential discourse relation types; Clause's agent animacy (Jahan et al., 2018; Y. Zhu et al., 2019); Agent character local continuity (Augustyniak et al., 2018; Neumann, 2021; Peng et al., 2022; L. Zhang, Xing, et al., 2020).

Morphological factors. Verb morphological feature continuity between clauses (including verbs' case, mood, number, person, tense, and verb form). This is applicable in languages that have subject-verb agreement such as Spanish and Portuguese, but not in Chinese.

The syntactic, semantic, and morphological features listed above cannot cover all relevant aspects to *pro*-drop. The feature selection and representation in this experiment are expected to be quantifiable (*i.e.* the feature can be represented or measured with numbers) so that they can qualify as independent variables in the binomial logistic regression model.

In the following sections, we will introduce the details for discourse preprocessing, feature selection, and retrieval, the Binomial Logistic Regression model, and the Random Forest model (see Section 5.2), and compare the features' performance in three languages: Chinese, Brazilian Portuguese, and Spanish.

5.2 Method

In this section, the details of discourse preprocessing, feature retrieval, and model information are introduced. Figure 5.1 provides a brief visualization of the preprocessing steps and the corresponding features that are generated after each preprocessing step.

5.2.1 Material

The discourse materials used in this study are Chinese (CN), Brazilian Portuguese (BP), and Spanish (ES) translation (xiaowangzi.org, 2021) of Saint-Exupéry's *The Little Prince*.

5.2.2 Discourse Preprocessing

Clause retrieval and word segmentation. The clauses in all three languages are obtained based on naturalistic punctuation. The word segmentation in BP and ES is their natural word segmented with white spaces, and native speakers manually check the ones in CN.

Clause-level *pro*-drop and story character annotation. Consistent with Experiment 1 (see Chapter 4), *pro*-drop and native speakers of the three languages did story character annotation. The *pro*-drop annotation result is used as the independent variable of the logistic regression model (see details in Section 5.2.4) and the random forest model 5.2.5. The story characters act as "animated agents" in a clause, and this information contributes to obtaining the "*agent_animacy*" feature. If the agent of the verb in a clause belongs to the 32 story characters, it is labeled as "True" as an animated agent; Otherwise, it is labeled as "False" as a non-animated agent.

Word-level Part-Of-Speech (POS) tagging. Each word in the discourse is labeled with a POS tag with SpaCy via the "token.pos_, token.tag_" parameters. In this way, the verbs (tagged as 'VERB') can be extracted to calculate their "depth in the tree" syntactic features.

Word-level morphological features. The morphological features were obtained by SpaCy via "token.morph" for BP and ES. This process is not applied in CN, since there is no morphological change (*i.e.* verb inflection change) in individual words or character change in the verbs. There are 6 morphological

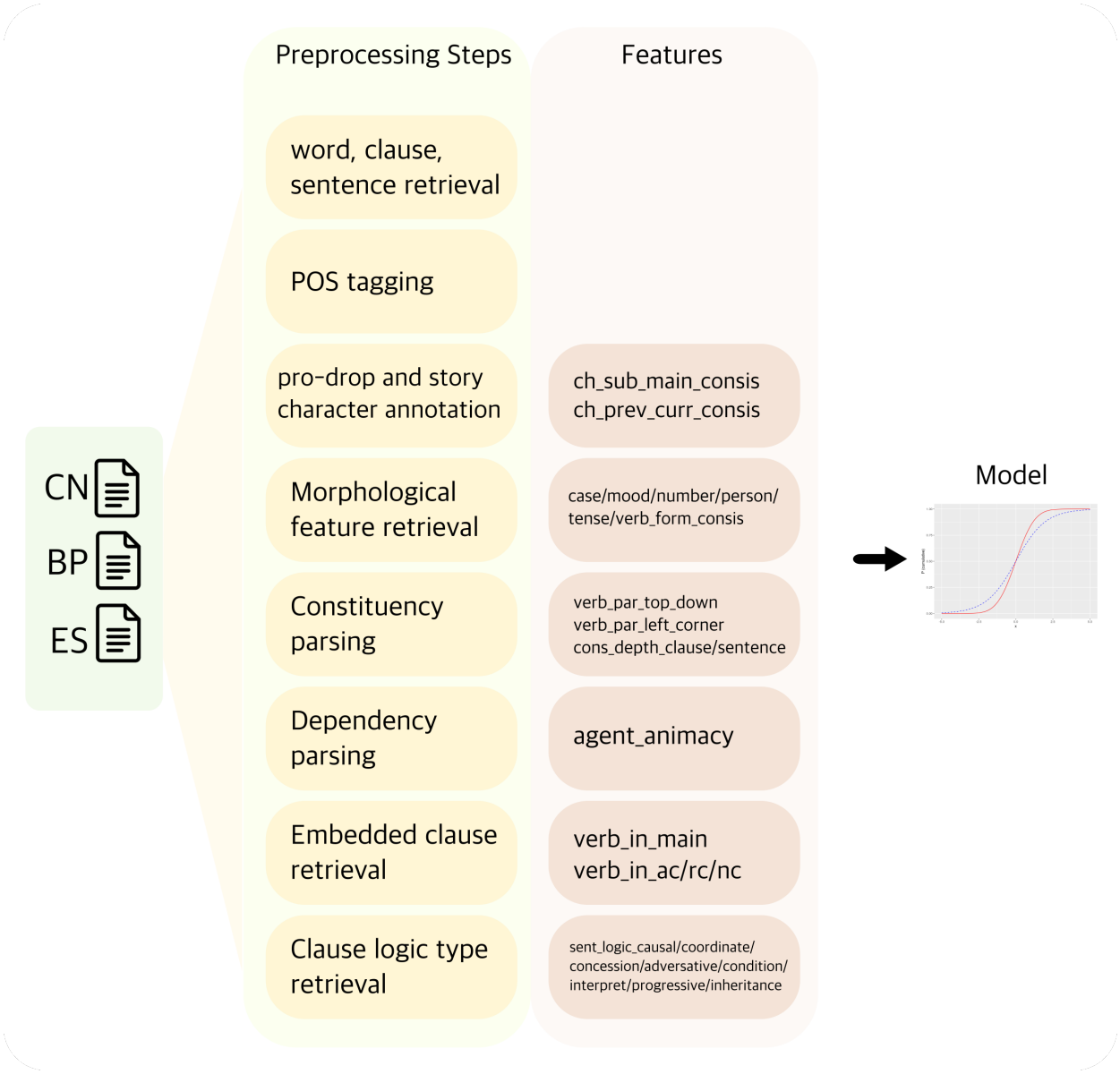


Figure 5.1: The preprocessing steps applied to CN, BP, and ES discourse (see Section 5.2.2), and the features (see Section 5.2.3 and Table 5.1) generated based on each preprocessing step. The model is introduced in Section 5.2.4 and 5.2.5

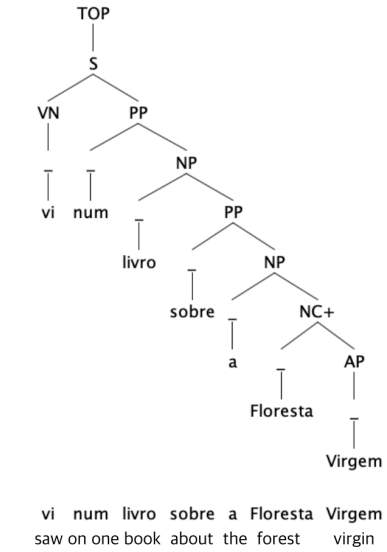
features: 'Case', 'Mood', 'Number', 'Person', 'PronType' (Pronoun Type), 'Tense', and 'VerbForm' (Verb Form). The values for each feature are (1) 'Case': accusative ('Acc'), dative ('Dat'); (2) 'Mood': indicative ('Ind'), subjunctive or conjunctive ('Sub'), conditional ('Cnd'); (3) 'Number': singular ('Sing'), plural ('Plur'), and both singular and plural ('Plur, Sing'); (4) 'Person': first person ('1'), second person ('2'), third person ('3'), first or third person ('1,3'); (5) 'PronType': person ('Prs'); (6) 'Tense': present ('Pres'); past ('Past'); imperfect ('Imp'); future ('Fut'); past perfect ('Pqp'); (7) 'VerbForm': finite ('Fin'); infinite ('Inf'); participle ('Par'); gerund ('Ger').

Constituency Parsing. The constituency parsing results are used in getting the syntactic features including "verb_par_top_down", "verb_par_left_corner", "cons_depth_clause", "cons_depth_sent" (see Section 5.2.3 for details of the features). For both ES and BP, the constituency parsing was realized with SuPar[1](Y. Zhang, Li, et al., 2020; Y. Zhang, Zhou, et al., 2020), which is a Python library designed for structured prediction and has reproductions of many state-of-the-art syntactic/semantic parser with pre-trained models for more than 19 languages. The model used for parsing was "crf-con-xlmr", a multilingual model trained with the SPMRL (Statistical Parsing of Morphologically Rich Languages) dataset by fine-tuning "xlm-roberta-large". The SPMRL dataset (Seddah et al., 2013; Tsarfaty et al., 2010) included Penn Treebank style phrases-structured trees, and have the same shared set of POS tags across languages. As shown in Figure 5.2, the BP and ES clauses are parsed into binary tree structures, and stored as parenthesis structures for further feature extraction purposes. For CN, the constituency parsing was done with ZPar(Y. Zhang & Clark, 2011b) (see Figure 5.3 showing a CN clause parsing result). With the constituency tree information stored in the nested parenthesis form, the depth of a word in the tree can be calculated with the NLTK function "tree.leaf_treeposition". It should be noted that the direct tree parsed with SuPar does not have the POS information for each word (shown as '_' above the word nodes in Figure 5.2), and these nodes were filled with the POS results using Python.

Dependency Parsing. Using dependency parsing, the position of the subject can be located within a clause. The dependency parsing for three languages was realized with the dependency parsers in SpaCy.

[1]https://pypi.org/project/supar/

(1) BP



(2) ES

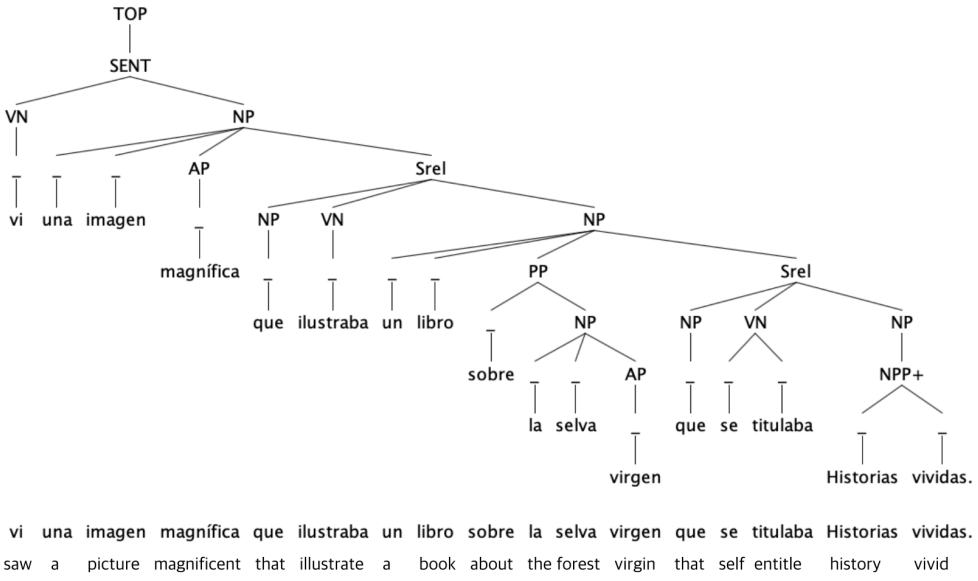


Figure 5.2: Constituency parsing for a clause from BP and ES using SuPar's "crf-con-xlmr" multi-language model. The parsing was applied to the whole discourse on both the sentence and the clause level.

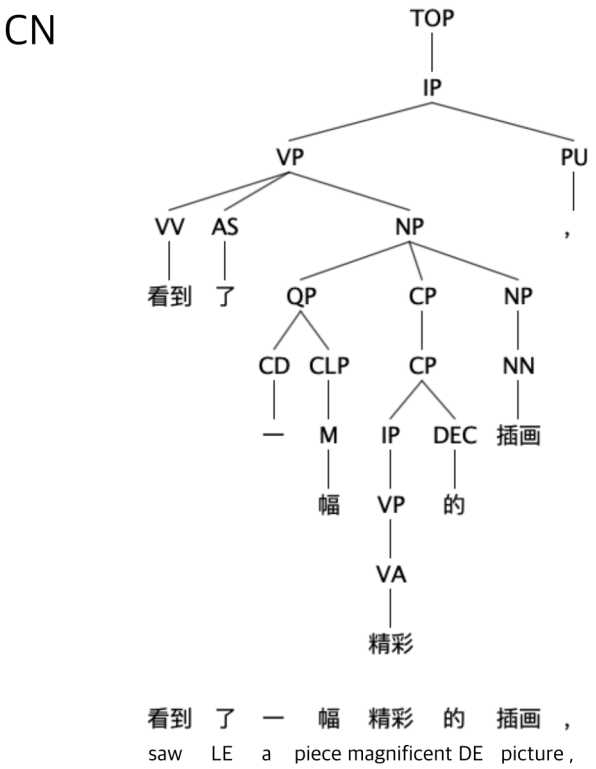


Figure 5.3: Constituency parsing for a clause from CN using ZPar. The parsing was applied to the whole discourse on both the sentence and the clause level.

Embedded clause annotation. There are three embedded clause types annotated: adverbial clause (*e.g.* "When I was six years old, I read a book."), noun clause (*e.g.* "He said that he has a rose."), and relative clause (*e.g.* "I like the picture that you gave me."). For BP and ES, the embedded clause annotation was done based on the constituency tree results. In the constituency parsing results, sub-clauses are led with nodes "Sint", "Ssub", and "Srel". First, we locate these nodes in the tree; Second, the sibling nodes of these labels are located. The relative clauses were located with the sub-tree head "Srel" in the constituency tree parsing results. The adverbial clause and noun clause are identified by their subclause leading words. For ES, the sibling nodes used for locating noun clauses and relative clauses are (1) Noun clause: 'que'; (2) Adverbial clause: 'cuando', 'después', 'hasta que', 'mientras', 'tan pronto como'. For ES, the sibling nodes used for locating noun clauses and relative clauses are (1) Noun clause: 'que'; (2) Adverbial clause: 'cuando', 'después', 'hasta que', 'mientras', 'tan pronto como'. For BP, the sibling nodes used for locating noun clauses and relative clauses are (1) Noun clause: 'que', 'sem que', 'com que'; (2) Adverbial clause: 'quando', 'até que', 'depois que', 'pois', 'de vez em quando', 'enquanto', 'logo que'. As for CN, the AMR (Abstract Meaning Representation)[2], which follows the annotation tradition in Penn Discourse Treebank Version 3.0 (Prasad et al., 2019), for the discourse was adopted to obtain the embedded clause type. The annotation labels used in AMR to identify the embedded clause are (1) Noun clause: 'person' and 'thing'; (2) Adverbial clause: 'before', 'pass', 'temporal', 'now', 'be-located-at', 'relative-position'. Relative clauses in CN were identified by constituency structure using the Tregex rule "$(CP << (IP! <, NP < (VP << (NP\$--VV)))[< DEC| < (CP < DEC)])\$++(NP(! << /时候/! << /情况/)) >!PP$".

Discourse relation type annotation. For ES and BP, The discourse relation type was assigned with a method similar to the one for the embedded clause. The discourse relation determined by the subclauses' previous sibling nodes, and the ones to label in ES are (1) Causal: 'así que', 'como que', 'como si', 'como', 'desde que', 'para que', 'porque', 'pues'; (2) Coordinate: 'y'; (3) Concession: 'aunque', 'bien'; (4) Adversative: 'pero'. The sibling nodes to label in BP are (1) Causal: 'para que', 'a fim de que', 'como', 'porque', 'talvez por que', 'como', 'uma vez que'; (2) Coordinate: 'e', 'ou'; (3) Concession: 'mesmo que', 'mesmo se'; (4) Adversative: 'mas'. As for CN, the AMR(Abstract Meaning Representation)[3] for the discourse

[2]http://www.cs.brandeis.edu/ clp/camr/camr.html
[3]http://www.cs.brandeis.edu/ clp/camr/camr.html

was adopted to obtain the embedded clause type. The annotation labels used in AMR to identify the discourse relation are (1) Causal: 'causation', 'cause'; (2) Coordinate: 'and'; (3) Concession: 'concession'; (4) Adversative: 'contrast'; (5) Condition: 'condition', 'except'; (6) Interpret: 'mean'; (7) Progressive: 'progression'; (8) Inheritance: 'explanation'.

5.2.3 Feature Retrieval

As shown in Table 5.1, the features are retrieved from the discourse for each verb. The "Feature Label" columns are the names for each feature that are shown in the results summaries. The "Feature Meaning" provides a brief explanation for the feature. The "Applied in Languages" column shows the languages that adopt the feature. The "Value" column indicates the form of the feature, and it can be either binary (*i.e.* 0 or 1) or an integer (*e.g.* 1, 2, 3,...). In this section, the details of the feature generation are described in the order of (1) syntactic factors; (2) Morphological factors; and (3) Semantic Features.

Table 5.1: Features applied in the binomial logistic regression models.

Feature Type	Feature Label	Feature Meaning	Applied in Language	Value
syntactic	*verb_in_main*	current verb is in a main clause	CN, BP, ES	0, 1
	verb_in_ac	current verb in an adverbial clause	CN, BP, ES	0, 1
	verb_in_rc	current verb in a relative clause	CN, BP, ES	0, 1
	verb_in_nc	current verb in a noun clause	CN, BP, ES	0, 1
	verb_par_top_down	verb depth in top-town parsing	CN	integer
	verb_par_left_corner	verb depth in left-corner parsing	CN	integer
	cons_depth_clause	verb depth in its clause's constituency tree	BP, ES	integer
	cons_depth_sent	verb depth in its sentence's constituency tree	BP, ES	integer
Semantic	*agent_animacy*	agent is animate or not	CN, BP, ES	0, 1
	ch_sub_main_consis	the characters in the main clause and the subordinate clause are consistent	CN, BP, ES	0, 1
	ch_curr_prev_consis	the characters in the current clause and the previous clause are consistent	CN, BP, ES	0, 1
	sent_logic_causal	the discourse relationship in the current sentence is a causal relationship	CN, BP, ES	0, 1
	sent_logic_coordinate	the discourse relationship in the current sentence is a concession relationship	CN, BP, ES	0, 1
	sent_logic_concession	the discourse relationship in the current sentence is a concession relationship	CN, BP, ES	0, 1
	sent_logic_adversative	the discourse relationship in the current sentence is an adversative relationship	CN, BP, ES	0, 1
	sent_logic_condition	the discourse relationship in the current sentence is a conditional relationship	CN	0, 1
	sent_logic_interpret	the discourse relationship in the current sentence is a interpretation relationship	CN	0, 1
	sent_logic_progressive	the discourse relationship in the current sentence is a progressive relationship	CN	0, 1
	sent_logic_inheritance	the discourse relationship in the current sentence is a inheritance relationship	CN	0, 1
Morphological	*case_consis*	current verb's case is consistent or not with the previous verb	BP, ES	0, 1
	mood_consis	current verb's mood is consistent or not with the previous verb	BP, ES	0, 1
	number_consis	current verb's number is consistent or not with the previous verb	BP, ES	0, 1
	person_consis	current verb's person is consistent or not with the previous verb	BP, ES	0, 1
	tense_consis	current verb's tense is consistent or not with the previous verb	BP, ES	0, 1
	verb_form_consis	current verb's form is consistent or not with the previous verb	BP, ES	0, 1

Syntactic Features

Clause type that the verb is in. There are four features relevant to the clause type that a verb is in (1) *"verb_in_main"*: the verb is in the main clause of the sentence; (2) *"verb_in_ac"*: the verb is in an adverbial

clause of the sentence; (3) *"verb_in_rc"*: the verb is in a relative clause of the sentence; (4) *"verb_in_nc"*: the verb is in a noun clause of the sentence, and the noun clause refers to the clause that functions as a noun.

The verb's depth in its constituency tree. For CN, two features are used to describe the verb's depth: (1) *"verb_par_top_down"*: the verbs' depth in the constituency tree while using top-down parsing strategy; (2) *"verb_par_left_corner"*: the verbs' depth in the constituency tree while using left-corner parsing strategy. For BP and ES, two features are applied to describe the verb's depth: (1) *"cons_depth_clause"*: the verbs' depth in its clause's constituency tree by counting branch depth; (2) *"cons_depth_sent"*: the verbs' depth in its sentence's constituency tree by counting branch depth.

Semantic Features

Agent animacy. This factor shows whether the verb's agent is animate or not, represented with the label *"agent_animacy"*.

Story character consistency. This factor shows the relationship between the current verb's agent character and its previous clause or its main clause's agent character, and it has two features applied in the model: (1) *"ch_sub_main_consis"*: agent character consistency between the subordinate clause and its main clause; (2) *"ch_curr_prev_consis"*: agent character consistency between the current clause and its previous clause.

Sentence logic. According to the discourse relation type the sentence contains, 8 features are used in the model (4 for BP and ES, 8 for CN): (1) *"sent_logic_causal"*: the discourse relation of the sentence that contains the clause is a causal relationship; (2) *"sent_logic_coordinate"*: the discourse relation of the sentence that contains the clause is a coordinate relationship; (3) *"sent_logic_concession"*: the discourse relation of the sentence that contains the clause is a concession relationship; (4) *"sent_logic_adversative"*: the discourse relation of the sentence that contains the clause is an adversative relationship; (5) *"sent_logic_condition"*: the discourse relation of the sentence that contains the clause is a conditional relationship; (6) *"sent_logic_interpret"*: the discourse relation of the sentence that contains the clause is an interpretive relationship; (7) *"sent_logic_progressive"*: the discourse relation of the sentence that contains

the clause is a progressive relationship; (8) *"sent_logic_inheritance"*: the discourse relation of the sentence that contains the clause is an inheritance relationship.

Morphological Features

Verb morphological feature consistency. There are six features generated to reflect the morphological factors of the verb and its consistency with the previous verb: (1) *"case_consis"*: the case consistency between the current verb and its previous verb; (2) *"mood_consis"*: the mood consistency between the current verb and its previous verb; (3) *"number_consis"*: the number consistency between the current verb and its previous verb; (4) *"person_consis"*: the person consistency between the current verb and its previous verb; (5) *"tense_consis"*: the tense consistency between the current verb and its previous verb; (6) *"verb_form_consis"*: the verb form consistency between the current verb and its previous verb.

5.2.4 Binomial Logistic Regression Model

Binomial logistic regression is a binary classification algorithm and a standard corpus linguistics method, as introduced in Brezina's textbook (Brezina, 2018) (see Chapter 4.4). In this experiment, the binomial logistic regression model was realized with the "*glm*" function in the R studio.

In our case, the prediction values (*i.e.* dependent variable or "*y*" value) are "True or False" based on a case including *pro*-drop or not. The independent variables (*i.e.* the "*x*" values. See Section 5.2.3 for the features applied in the model) are the factors we took into consideration that are potentially relevant to *pro*-drop. It should be noted that these independent variables do not necessarily predict the pro-drop cases perfectly, and we do not focus on the prediction accuracy, instead, we care more about the comparative importance level of these features.

Math Behind Binomial Logistic Regression

Binomial logistic regression is in a class of general linear models (GLMs), and it is applied when the outcome variable is binary (*e.g.* "0 and 1", "True and False" or "positive and negative"). The difference between logistic regression and linear regression is in their fitting functions, and the logistic regression fits on a non-linear logistic function (see Equation 5.1 as an example) to represent the non-linearity relationship between the variables.

$$P(y=1) = \frac{1}{1+e^{-k(x-x_0))}} \tag{5.1}$$

Equation 5.1 can be transformed into Equation 5.2. The *P(y=1)* refers to the probability of the outcome variable is 1. β_0 and β_1 are constant values that the model will calculate.

$$P(y=1) = \frac{1}{1+e^{-(\beta_0+\beta_1 x))}} \tag{5.2}$$

As there are two possible outcomes for the binary model, the probability distribution of *y=0* can be represented as in Equation 5.3.

$$P(y=0) = 1 - \frac{1}{1+e^{-(\beta_0+\beta_1 x))}} \tag{5.3}$$

Therefore, the possibility of *y=1* compared to *y=0* can be represented as in Equation 5.4. Therefore, the log-odds value e_1^{β} represents: with x being positive (or increasing by one unit), y has e_1^{β} more possibility to be positive.

$$\frac{P(y=1)}{P(y=0)} = e^{\beta_0+\beta_1 x} = e^{\beta_0} e^{\beta_1 x} \tag{5.4}$$

As we include more independent variables (*i.e.* the series of *s*: $x_1, x_1, ..., x_p$), the binomial logistic regression formula becomes Equation 5.5. e^{β_0} represents the odds of the outcome when all inputs are zero, and e^{β_i} represents the odds ratio associated with a unit increase in x_i assuming no change in the other inputs (that is, a unit increase in x_i multiplies the odds of our outcome by e^{β_i}.

$$\frac{P(y=1)}{P(y=0)} = e^{\beta_0+\beta_1 x_1+\beta_2 x_2+...+\beta_p x_p} = e^{\beta_0} e^{\beta_1 x_1} e^{\beta_2 x_2} ... e^{\beta_p x_p} \tag{5.5}$$

5.2.5 Random Forest

Random Forest(Breiman, 2001) is a commonly used machine learning algorithm for tasks such as classification and regression (see Chapter 11 of *Hands on Machine Learning with R* by Boehmke and Greenwell (2019) for an introduction for applying the method). Random Forest is made up of multiple Decision Trees (see Figure 5.4 for visualization of decision trees to random forest), which use algorithms to find the

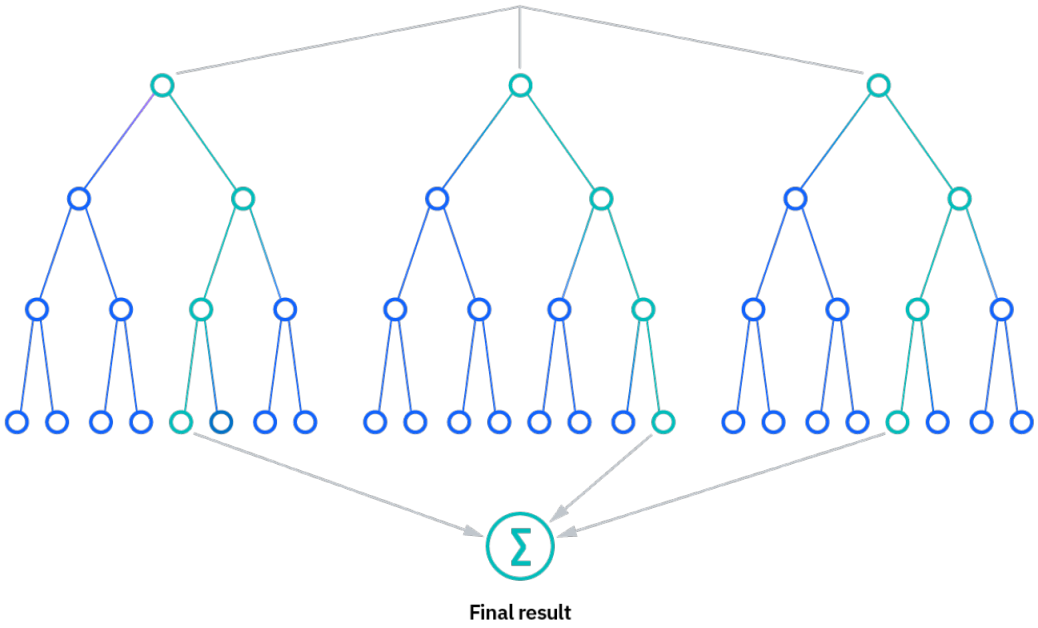


Figure 5.4: From Decision Trees to Random Forest.
Figure source: https://www.ibm.com/topics/random-forest

best split of datasets. Within the Random Forest, each decision tree can "vote" for the classification of a case, and for example, in our case, it would vote on *pro*-drop *vs.* non-*pro*-drop by the trees. All the "votes" will be considered together to make a comparably best decision by the Random Forest to have a more robust prediction result. The metrics used in the algorithm include Gini Impurity, Information Gain, or Mean Square Error (MSE).

In this experiment, Random Forest was realized by the "RandomForestClassifier" package in the Scikit-Learn (Pedregosa et al., 2011) Python library. The features (see Table 5.1)are consistent with the ones used in Binomial Logistic Regression. 80% data were used as the training dataset, and the rest of the 20% were the testing dataset. The number of estimators was 100, the random seed size was set as 42, and the max depth of the prediction tree was set as 3.

Math Behind Random Forest Model

As the random forest is made up of multiple decision trees, it is essential to understand how a decision tree makes the decision, which means how it divides the data samples into the labels (in our case, *pro*-drop and non-*pro*-drop represented in binary form as 1 and 0) and what features they use to make the divisions.

The goal of branch split is to obtain higher purity children nodes. There are two common algorithms used in the decision tree to measure the branch split impurity level: Gini Index and Entropy. The algorithm used in random forest in this experiment is the Gini index, which is computationally faster than Entropy. As shown in Equation 5.6, *p(x)* is a fraction of examples in a given class.

$$Gini = 1 - \sum_{i=1}^{C}(p_i^2) \tag{5.6}$$

The evaluation of the decision tree will be how well the division was done, and this is realized with the Information Gain, represented in Equation 5.7. f is the feature to split on, D_p is the dataset of the parent node, D_{left} is the dataset of the left child node, D_{right} is the dataset of the right child node, I is the impurity criterion (Gini Index as shown in Equation 5.6), N is the total number of samples, N_{left} is the number of samples of the left child node, and N_{right} is the number of samples of the right child node. The attribute with the largest information gain will be chosen as the decision node, and the dataset will be divided into its child nodes. This process will repeat until a brain reaches a leaf node (full purity of one label) or the depth limitation of the tree.

$$IG(D_p, f) = I(D_p) - \frac{N_{left}}{N} I(D_{left}) - \frac{N_{right}}{N} I(D_{right}) \tag{5.7}$$

The feature importance in the random forest is a measurement of how much including a variable increases accuracy. The feature importance used in this experiment (see Figure 5.6 and 5.5 in results) is based on the mean decrease in impurity, and is computed as the mean and standard deviation of accumulation of the impurity decrease within each tree.

5.3 Results

In this section, the results from Binomial Logistic Regression are presented in Section 5.3.1 and 5.3.2.

5.3.1 Binomial Logistic Regression Results

In this section, the statistical results from applying binomial logistic regression on the features from CN, BP, and ES are presented. Binomial logistic regression (see detailed introduction in Section 5.2.4) analyzed the relationship between the predictor variables shown in Table 5.1 and the response variable "*pro*-drop".

CN model results

As shown in Table 5.2, the coefficients summary from the model shows that the following variables meet the significance level of less than 0.05 (in column $\Pr(>|z|)$): "verb_in_ac", "ch_sub_main_consis", "ch_curr_prev_consis", "sent_logic_coordinate", "sent_logic_concession".

We can interpret the CN model results as follows: (1) All else being equal, the factor that a verb showing up in an adverbial clause ("verb_in_ac") has a significant positive effect on the likelihood of *pro*-drop, with the variable being positive increasing the odds of *pro*-drop by 102%; (2) All else being equal, the factor that the character being consistent between the subclause and its main clause ("ch_sub_main_consis") has a significant positive effect on the likelihood of *pro*-drop, with the variable being positive increasing the odds of *pro*-drop by 170%; (3) All else being equal, the factor that the verb being consistent between the current clause and its previous clause ("ch_curr_prev_consis") has a significant positive effect on the likelihood of *pro*-drop, with the variable being positive increasing the odds of *pro*-drop by 63%; (4) All else being equal, the factor that the verb occurring in a clause with coordinate discourse relation ("sent_logic_coordinate") has a significant positive effect on the likelihood of *pro*-drop, with the variable being positive increasing the odds of *pro*-drop by 56%; (5) All else being equal, the factor that the verb occurring in a clause with concession discourse relation ("sent_logic_concession") has a significant positive effect on the likelihood of *pro*-drop, with the variable being positive increasing the odds of *pro*-drop by 38%.

Table 5.2: Binomial logistic regression results for Chinese.

	Estimate	**Std.Error**	**z value**	**odds_ratio**	**Pr(>|z|)**
(Intercept)	-1.78033	0.34305	-5.190	0.1686	2.11e-07 ***
verb_in_ac	0.70122	0.24723	2.836	2.0162	0.00456 **
verb_in_rc	-0.25202	0.30855	-0.817	0.7772	0.41405
verb_in_nc	0.05393	0.26681	0.202	1.0554	0.83982
agent_animacy	-0.32625	0.16849	-1.936	0.7216	0.05284 .
verb_par_top_down	0.46957	0.60701	0.774	1.5993	0.43918
verb_par_left_corner	-0.73836	0.61890	-1.193	0.4779	0.23286
ch_sub_main_consis	0.99209	0.12683	7.823	2.6969	5.18e-15 ***
ch_curr_prev_consis	0.48896	0.11931	4.098	1.6306	4.16e-05 ***
sent_logic_causal	-0.20310	0.18142	-1.119	0.8162	0.26294
sent_logic_coordinate	0.44520	0.12051	3.694	1.5608	0.00022 ***
sent_logic_concession	0.32431	0.15971	2.031	1.3831	0.04229 *
sent_logic_adversative	-0.04192	0.21030	-0.199	0.9589	0.84200
sent_logic_condition	0.53785	0.42239	1.273	1.7123	0.20289
sent_logic_interpret	0.04971	0.15886	0.313	1.0510	0.75436
sent_logic_progressive	0.51254	0.44196	1.160	1.6695	0.24617
sent_logic_inheritance	309.11981	-0.041	0.96747	0.0000	0.96747
	Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 Null deviance: 2183.6 on 2242 degrees of freedom Residual deviance: 2030.6 on 2226 degrees of freedom AIC: 2064.6 Number of Fisher Scoring iterations: 12				

BP model results

As shown in Table 5.3, the coefficients summary from the model shows that the following variables meet the significance level of less than 0.05 (in column Pr(>|z|)): "cons_depth_sent", "agent_animacy", "ch_sub_main_consis", "ch_curr_prev_consis", "Tense_consis".

We can interpret the BP model results as follows: (1) All else being equal, the depth of a verb in its sentence's constituency tree ("cons_depth_sent") has a significant negative effect on the likelihood of *pro*-drop, with the variable depth increase by 1, the odds of *pro*-drop will likely to decrease by 77%;

(2) All else being equal, the verb's agent animacy ("agent_animacy") has a significant positive effect on the likelihood of *pro*-drop, with the variable being positive increasing the odds of *pro*-drop by 200%; (3) All else being equal, the factor that the character being consistent between the subclause and its main clause ("ch_sub_main_consis") has a significant positive effect on the likelihood of *pro*-drop, with the variable being positive increasing the odds of *pro*-drop by 74%; (4) All else being equal, the factor that the character being consistent between the current clause and its previous clause ("ch_curr_prev_consis") has a significant positive effect on the likelihood of *pro*-drop, with the variable being positive increasing the odds of *pro*-drop by 75%; (5) All else being equal, the factor that the current verb and its previous verb have consistent tense("Tense_consis") has a significant positive effect on the likelihood of *pro*-drop, with the variable being positive increasing the odds of *pro*-drop by 33%.

Table 5.3: Binomial logistic regression results for Brazilian Portuguese.

	Estimate	**Std.Error**	**z value**	odds_ratio	**Pr(>|z|)**
(Intercept)	-1.982161	0.310999	-6.374	0.1378	1.85e-10 ***
cons_depth_clause	0.689839	0.694992	0.993	1.9934	0.3209
cons_depth_sent	-1.478165	0.677089	-2.183	0.2281	0.0290 *
agent_animacy	1.099610	0.158482	6.938	3.0030	3.97e-12 ***
verb_in_rc	0.296622	0.251778	1.178	1.3453	0.2388
verb_in_nc	0.012827	0.277535	0.046	1.0129	0.9631
verb_in_ac	0.322068	0.395785	0.814	1.3800	0.4158
sent_logic_causal	-0.460371	0.456318	-1.009	0.6310	0.3130
sent_logic_concession	-11.917643	367.935462	-0.032	0.0000	0.9742
sent_logic_coordinate	0.054864	0.166957	0.329	1.0564	0.7425
sent_logic_adversative	-0.151125	0.228229	-0.662	0.8597	0.5079
ch_sub_main_consis	0.556413	0.114220	4.871	1.7444	1.11e-06 ***
ch_curr_prev_consis	0.561698	0.114592	4.902	1.7536	9.50e-07 ***
Case_consis	0.718838	0.505667	1.422	2.0520	0.1552
Mood_consis	0.066100	0.214138	0.309	1.0683	0.7576
Number_consis	0.199957	0.123871	1.614	1.2214	0.1065
Person_consis	-0.007485	0.134393	-0.056	0.9925	0.9556
Tense_consis	0.287998	0.141555	2.035	1.3338	0.0419 *
VerbForm_consis	-0.305184	0.211294	-1.444	0.7370	0.1486
	Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1 Null deviance: 2606.6 on 2210 degrees of freedom Residual deviance: 2365.9 on 2192 degrees of freedom AIC: 2403.9 Number of Fisher Scoring iterations: 12				

ES model results

As shown in Table 5.4, the coefficients summary from the model shows that the following variables meet the significance level of less than 0.05 (in column Pr(>|z|)): “cons_depth_clause”, “cons_depth_sent”, “sent_logic_coordinate”, “ch_sub_main_consis”, “ch_curr_prev_consis”, “Case_consis”, “Number_consis”, “Person_consis”.

We can interpret the BP model results as follows: (1) All else being equal, the depth of a verb in its clause's constituency tree ("cons_depth_clause") has a significant negative effect on the likelihood of *pro*-drop, with the variable depth increase by 1, the odds of *pro*-drop will likely to decrease by 80%; (2) All else being equal, the depth of a verb in its sentence's constituency tree ("cons_depth_sent") has a significant negative effect on the likelihood of *pro*-drop, with the variable depth increase by 1, the odds of *pro*-drop will likely to decrease by 79%; (3) All else being equal, the factor that the verb occurring in a clause with coordinate discourse relation ("sent_logic_coordinate") has a significant positive effect on the likelihood of *pro*-drop, with the variable being positive increasing the odds of *pro*-drop by 80%; (4) All else being equal, the factor that the verb being consistent between the subclause and its main clause ("ch_sub_main_consis") has a significant positive effect on the likelihood of *pro*-drop, with the variable being positive increasing the odds of *pro*-drop by 243%; (5) All else being equal, the factor that the character being consistent between the current clause and its previous clause ("ch_curr_prev_consis") has a significant positive effect on the likelihood of *pro*-drop, with the variable being positive increasing the odds of *pro*-drop by 321%; (6) All else being equal, the factor that the current verb and its previous verb have a consistent case ("Case_consis") has a significant positive effect on the likelihood of *pro*-drop, with the variable being positive increasing the odds of *pro*-drop by 1200%; (7) All else being equal, the factor that the current verb and its previous verb have a consistent number ("Number_consis") has a significant positive effect on the likelihood of *pro*-drop, with the variable being positive increasing the odds of *pro*-drop by 31%; (8) All else being equal, the factor that the current verb and its previous verb have a consistent person ("Person_consis") has a significant negative effect on the likelihood of *pro*-drop, with the variable being positive decreasing the odds of *pro*-drop by 54%.

Table 5.4: Binomial logistic regression results for Spanish.

	Estimate	**Std.Error**	**z value**	**Odds_ratio**	**Pr(>|z|)**
(Intercept)	0.54730	0.28691	1.908	1.7286	0.05645 .
cons_depth_clause	-1.61881	0.79372	-2.040	0.1981	0.04140 *
cons_depth_sent	-1.55681	0.72140	-2.158	0.2108	0.03092 *
agent_animacy	0.05202	0.16081	0.323	1.0534	0.74635
verb_in_rc	-0.33605	0.26662	-1.260	0.7146	0.20752
verb_in_nc	0.17580	0.29231	0.601	1.1922	0.54756
verb_in_ac	-0.09635	0.44548	-0.216	0.9081	0.82877
sent_logic_causal	0.43278	0.39589	1.093	1.5415	0.27431
sent_logic_concession	-13.36820	284.85486	-0.047	0.0000	0.96257
sent_logic_coordinate	0.59008	0.16504	3.575	1.8041	0.00035 ***
sent_logic_adversative	0.17501	0.22274	0.786	1.1913	0.43203
ch_sub_main_consis	1.23525	0.10861	11.373	3.4392	<2e-16 ***
ch_curr_prev_consis	1.43924	0.11536	12.476	4.2175	<2e-16 ***
Case_consis	2.63842	1.10650	2.384	13.9911	0.01710 *
Mood_consis	0.31904	0.18332	1.740	1.3758	0.08179 .
Number_consis	0.27203	0.12065	2.255	1.3126	0.02415 *
Person_consis	-0.77239	0.12219	-6.321	0.4619	2.6e-10 ***
Tense_consis	-0.19106	0.12598	-1.517	0.8261	0.12938
VerbForm_consis	-0.09724	0.18226	-0.534	0.9073	0.59366
	Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 Null deviance: 2795.2 on 2026 degrees of freedom Residual deviance: 2317.8 on 2008 degrees of freedom AIC: 2355.8 Number of Fisher Scoring iterations: 12				

Binomial Logistic Regression Model Goodness-of-fit Levels

Pseudo-R^2 measurement analyses were applied to the three languages' binomial logistic regression models to estimate the fit level. As shown in Table 5.5, four Pseudo-R^2 measurements were used: McFadden's R^2, Cox and Snell's R^2, Nagelkerke's R^2, and Tjur's R^2.

The CN model fit level sits in the range of 6% to 10%. The BP model fit level sits between 9% to 15%. The ES model fit level is in the field of 17% to 28%. The order of fitness level among the three languages is: ES > BP > CN.

Table 5.5: Binomial logistic regression R-square values.

Pseudo-R^2	**CN**	**BP**	**ES**
McFadden	0.07008709	0.09234818	0.1708013
CoxSnell	0.06595537	0.10315542	0.2098521
Nagelkerke	0.10599526	0.14898359	0.2804872
Tjur	0.07023654	0.10319113	0.2206713

5.3.2 Random Forest Results

As shown in Table 5.6 and Figure 5.5, the feature importance levels within each language's Random Forest model results are ordered as (1) CN: "ch_sub_main_consis" > "sent_logic_coordinate" > "verb_in_ac" >"sent_logic_condition"; (2) BP: "agent_animacy" > "ch_sub_main_consis" >"ch_curr_prev_consis" > "cons_depth_sent"; (3) ES: "ch_curr_prev_consis" > "ch_sub_main_consis" > "Person_consis" > "con_depth_clause".

Table 5.6: Feature importance for CN, BP, and ES based on Random Forest results. The features shown in this table are the ones in which at least one language has non-zero results, and the all-zero features across languages are omitted.

	CN	**BP**	**ES**
agent_animacy	0.0031	0.4745	0.0004
sent_logic_causal	0.0043	0.0117	0.0060
sent_logic_condition	0.0395	0.0	0.0
sent_logic_coordinate	0.1791	0.0101	0.0006
sent_logic_concession	0.0107	0.0	0.0
sent_logic_adversative	0.0	0.0176	0.0
sent_logic_inheritance	0.0005	0.0	0.0
sent_logic_interpretation	0.0145	0.0	0.0
sent_logic_progressive	0.0291	0.0	0.0
ch_sub_main_consis	0.5251	0.2143	0.2580
verb_in_nc	0.0	0.0009	0.0009
verb_in_ac	0.0860	0.0068	0.0011
verb_in_rc	0.0197	0.0100	0.0358
verb_par_left_corner	0.0316	0.0	0.0
ch_curr_prev_consis	0.0216	0.1058	0.4408
verb_par_top_down	0.0351	0.0	0.0
cons_depth_clause_clean	0.0	0.0227	0.0822
cons_depth_sent	0.0	0.0653	0.0494
Mood_consis	0.0	0.0064	0.0147
Case_consis	0.0	0.0144	0.0018
Person_consis	0.0	0.0085	0.0843
VerbForm_consis	0.0	0.0140	0.0054
Tense_consis	0.0	0.0043	0.0007
Number_consis	0.0	0.0128	0.0181

As shown in Figure 5.6, the levels of feature importance are compared among all the non-all-zero features among the three languages. For the cases that have an importance level larger than 0.1, the results show that (1) BP stands out on the "agent_animacy" feature compared to the other languages; (2) CN stands out on the "sent_logic_coordinate" compared to the other languages; (3) Both BP and ES stand out on the feature "ch_curr_prev_consis" compared to CN, and all three languages stand out on the "ch_sub_main_consis" feature.

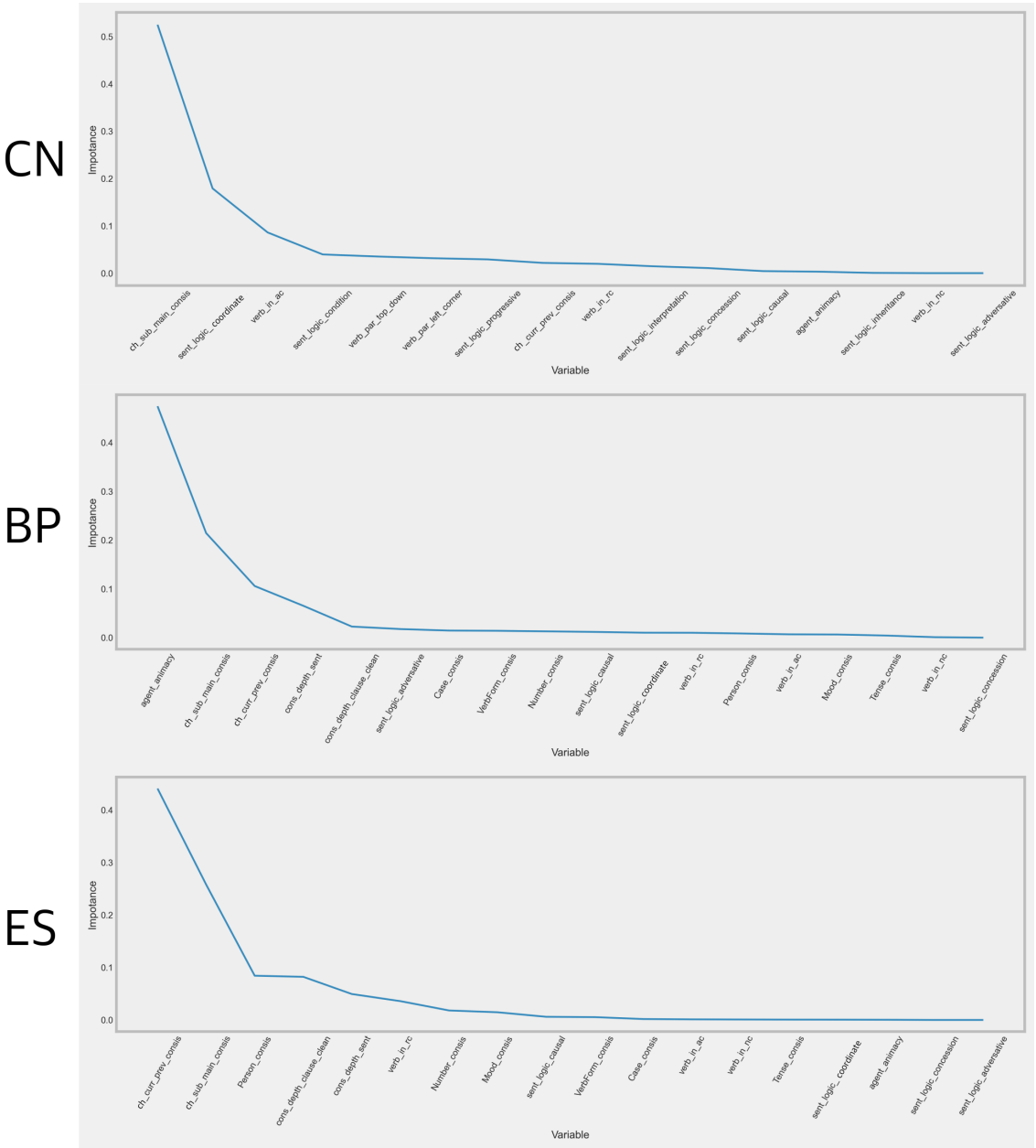


Figure 5.5: The feature importance distribution across CN, BP, and ES based on Random Forest results. The subfigures are (1) CN: The feature importance distribution for all features modeling Chinese *pro*-drop *vs.* non-*pro*-drop; (2) BP: The feature importance distribution for all features modeling Brazilian Portuguese *pro*-drop *vs.* non-*pro*-drop; (3) ES: The feature importance distribution for all features modeling Spanish *pro*-drop *vs.* non-*pro*-drop. See a comparison across features among all languages in Figure 5.6, and the detailed value in Table 5.6

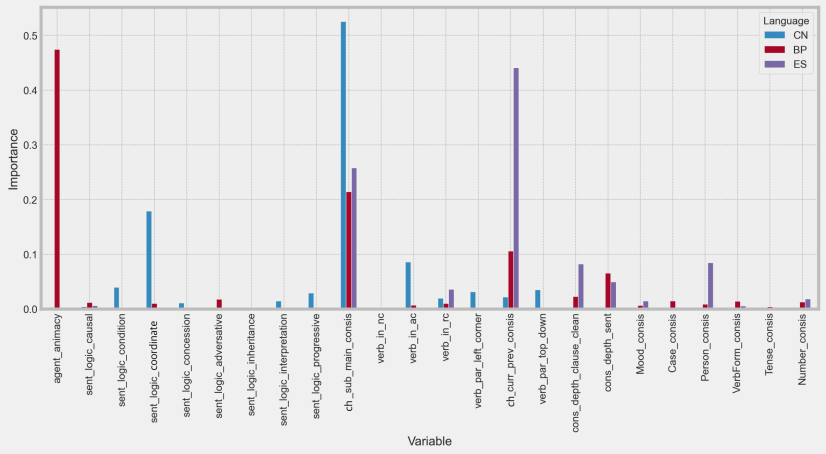


Figure 5.6: The feature importance distribution across CN, BP, and ES based on Random Forest results. The features shown in this figure are the ones with at least one language having non-zero results, and the all-zero features across languages are omitted.

As shown in Table 5.7, the prediction accuracy values of the Random Forest models are CN 83%, BP 71%, and ES 72%. The model performance for BP and ES are balanced between *pro*-drop and non-*pro*-drop cases. However, the model performance for CN only has acceptable accuracy on non-*pro*-drop cases, but is bad for *pro*-drop cases. Although model performance, similar to R^2 in Binomial Logistic Regression, is not what we are mainly curious about in this experiment, these accuracy results are consistent with the ones in Binomial Logistic Regression: the models have better explanation and prediction capability for ES and better than BP and CN.

Table 5.7: Random Forest model performance for the three languages.

Language	Label	Precision	Recall	f1-score	Accuracy
CN	non-pro-drop	0.83	1.00	0.90	0.83
	pro-drop	0.00	0.00	0.00	
BP	non-pro-drop	0.72	0.99	0.84	0.71
	pro-drop	0.71	0.04	0.07	
ES	non-pro-drop	0.71	0.70	0.70	0.72
	pro-drop	0.74	0.75	0.74	

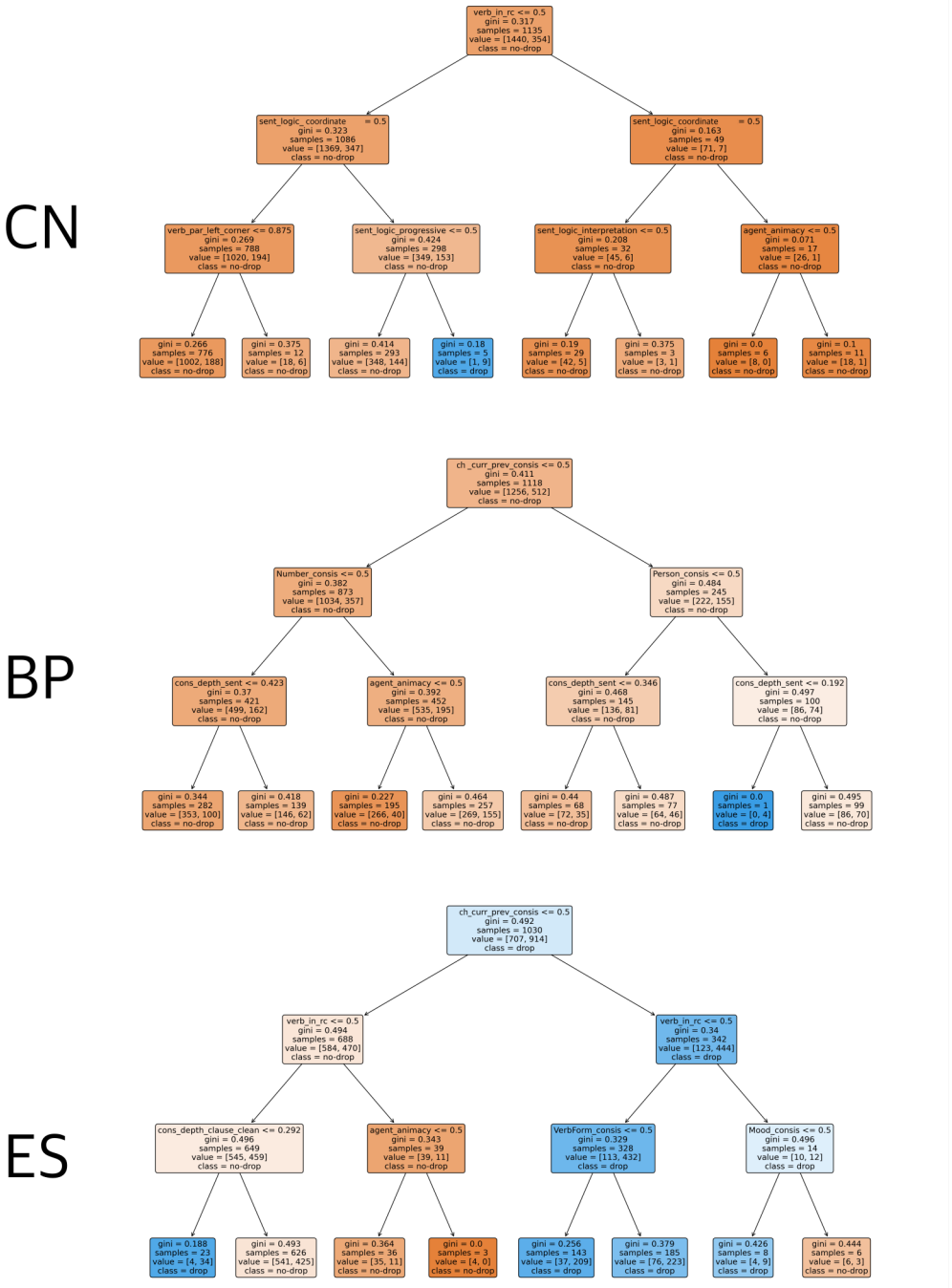


Figure 5.7: Random Forest first estimator's decision trees for CN, BP, and ES. There are 10 estimators for each Random Forest model, and only the first one is shown here as a representation.

5.4 Discussion

As detailed in Section 5.3, the features found more important by the two models, Binomial Logistic Regression and Random Forest, have majority consistency and minority differences. As shown in Table 5.8, the features that are found significant by the Binomial Logistic Regression model and the important features recognized by the Random Forest model are listed and compared across languages. In the following subsections, attempts are made to explain the features and their roles in *pro*-drop.

Table 5.8: Results comparison between the Binomial Logistic Regression Model and Random Forest Model across three languages.

	CN	**BP**	**ES**
Binomial Logistic Regression Significant Features Ranked by p-value	ch_sub_main_consis ch_curr_prev_consis sent_logic_coordinate verb_in_ac sent_logic_concession	agent_animacy ch_curr_prev_consis ch_sub_main_consis cons_depth_sent Tense_consis	ch_curr_prev_consis ch_sub_main_consis Person_consis sent_logic_coordinate Case_consis Number_consis cons_depth_sent cons_depth_clause
Random Forest Features Ranked by Importance level	ch_sub_main_consis sent_logic_coordinate verb_in_ac sent_logic_condition	agent_animacy ch_sub_main_consis ch_curr_prev_consis cons_depth_sent	ch_curr_prev_consis ch_sub_main_consis Person_consis cons_depth_clause

In the following subsections, we will discuss the features that were found significant by the models, and that the results indicate and related to linguistic features within the three languages.

5.4.1 Interpreting The Binomial Logistic Regression Model Results

Significant Features for *Pro*-drop

As summarized in Section 5.3, while we fit the features of interest (see Table 5.1) with *pro*-drop, the three languages CN, BP, and ES showed different significant preferences towards the features. The significant features for the three languages are: (1) CN: "verb_in_ac", "ch_sub_main_consis", "ch_curr_prev_consis", "sent_logic_coordinate", "sent_logic_concession"; (2) BP: "cons_depth_sent", "agent_animacy", "ch_sub_main_consis", "ch_curr_prev_consis", "Tense_consis"; (3) ES: "cons_depth_clause",

"cons_depth_sent", "sent_logic_coordinate", "ch_sub_main_consis", "ch_curr_prev_consis", "Case_consis", "Number_consis", "Person_consis".

For all three languages, the features of subclause and main clause character consistency ("ch_sub_main_consis"), and previous and current clause character consistency ("ch_curr_prev_consis") are significant. This indicates the local repetitive usage of characters can lead to *pro*-drop. This is consistent with the results in the 2_{nd} experiment, which showed that the *pro*-drop cases have higher local continuity in the discourse than non-*pro*-drop cases.

Both BP and ES models have the verb's depth in their sentence's constituency tree as a significant factor, and both have negative coefficients. The negative coefficients indicate that as all the other factors are kept the same, the deeper the verb is in the constituency structure, the less likely *pro*-drop would take place.

Both CN and ES have the coordinate discourse relation as a significant factor. This relation can be represented in structures such as "Clause A, and Clause B". In ES, the word *'y'* was used to locate the coordinate structures, and the CAMR label *'and'* was used to find the cases in the Chinese discourse. The coordinate relation in the sentence, especially when the agents are the same across clauses, can encourage *pro*-drop to happen.

As for the morphological feature consistency, it is found to take a role in BP and ES: BP's Tense; ES's Case, Number, and Person. The verbs' morphological features in BP and ES can carry information for its agents, whereas CN does not have this property. These morphological features can indicate the plurality, gender, and tense information, and the consistency of these features across clauses can narrow down the agent candidates. The reflection complexity order of these three languages is $ES > BP > CN$, and it is reasonable that ES relied more on morphological features to reveal the dropped pronoun, whereas CN and BP are less likely to have these resources.

Models' Fit Levels

As shown in Table 5.5, the Pseudo R^2 results of the three languages' models are not very high (all less than 30%). This indicates that our features are insufficient to accurately predict the *pro*-drop in the discourse. However, as mentioned in previous sections, this experiment is not aimed at predicting *pro*-drop, but to explore the comparative importance of the features within and among the languages we study.

The range of the Pseudo R^2 values of the three languages have the order of $ES > BP > CN$. There are several indications we can draw from this trend. First, the features used in this experiment are more effective in capturing *pro*-drop tendency in ES; Compared to CN, ES and BP both have extra verb morphological features included in the binomial logistic regression model, and these features are found significantly effective to predict *pro*-drop. Second, the low fit levels in all three models hint that there are more linguistic features that were not covered in this experiment that are relevant to *pro*-drop.

It should be noted that the motivation of *pro*-drop varies from case to case, clause to clause. That is to say, in some cases, the reason for dropping an agent pronoun is discourse relation (such as coordinate), whereas in some other cases, it is due to an embedded clause (such as an adverbial clause in Chinese). Therefore, it is hard for the model to capture a feature as "promising" for *pro*-drop prediction. This can be considered as a drawback of the regression method itself, or can be understood as the nature of discourse as being "dynamic".

5.4.2 Interpreting The Random Forest Model Results

As shown in the summary Table 5.8, the majority of the important features found by Random Forest models are consistent with the ones in Binomial Logistic Regression models, and this applies to all three languages. These consistent results from two various algorithms indicate the features found by them are promising.

5.5 Conclusion

This experiment adopted Binomial Logistic Regression and Random Forest to explore the linguistic factors and their roles in contributing to *pro*-drop. The factors are set up according to syntactic, semantic, and morphological aspects, and focused on local consistency in the discourse. The results indicate that (1) the character consistency between sub-main, current-previous clauses is significant in all three languages; (2) Spanish and Brazilian Portuguese have richer verb agreement systems to help recover the dropped pronoun, and potentially can lead to their models' higher fit level than Chinese; (3) Sentential discourse relation features such as coordinate structures can encourage *pro*-drop, and it was found significant in both CN and ES.

Experiment 2 extends the analytical horizon, uncovering a rich tapestry of intricate linguistic features influencing *pro*-drop patterns. The incorporation of machine learning models, coupled with a diverse array of features, provides a multifaceted perspective on *pro*-drop phenomena. This integration not only reveals universal factors but also accentuates the nuanced language-specific aspects shaping *pro*-drop occurrences. The comprehensive nature of this exploration serves as a foundational stepping stone, paving the way for in-depth inquiries into the intricate interplay of linguistic factors. Moreover, the utilization of machine learning models introduces a dynamic approach that complements traditional linguistic analyses, offering a data-driven lens through which to unravel the complexities of language phenomena.

Expanding on this, the nuanced perspective afforded by Experiment 2 enables a deeper understanding of the intricate dynamics at play in *pro*-drop languages. By incorporating a diverse set of features spanning syntactic, semantic, morphological, and logical dimensions, the experiment navigates the complexity of language structures. The significance of these features in distinguishing *pro*-drop from non-*pro*-drop cases contributes not only to theoretical linguistics but also to practical applications, such as language processing and machine learning. This exploration sets the stage for future research endeavors to delve into specific language pairs and regional variations, offering a comprehensive view of the factors influencing *pro*-drop phenomena across diverse linguistic contexts.

Furthermore, the integration of machine learning models in Experiment 2 sparks conversations about the evolving role of computational approaches in linguistic research. The combination of traditional linguistic analyses with advanced computational tools not only enhances our ability to unravel language complexities but also opens avenues for predictive modeling and automated pattern recognition. The results of Experiment 2, showcasing the capability of machine learning models to discern linguistic features, emphasize the potential of these models as valuable tools for investigating intricate language phenomena.

In summary, Experiment 2's expansion of the analytical scope unveils the intricate features influencing *pro*-drop, blending traditional linguistic analyses with advanced computational methodologies. This approach provides a nuanced understanding of both universal and language-specific aspects, laying the foundation for future investigations into the complex interplay of linguistic factors. The integration of machine learning models not only enriches theoretical linguistics but also points towards the potential applications of these models in deciphering the intricacies of language phenomena.

Chapter 6

Experiment 3: Modeling *pro*-drop Features With Attention Values In Graph Neural Networks

6.1 Introduction

Retrieving subject information while reading a story is naturalistic for us readers. As shown in Figure 6.1, these five sentences are clauses that all appear in the Chinese version of The Little Prince story, and they share the same subject story character "the little prince". As a reader, when we have the discourse context information, even though the subjects are presented in different surface forms (*i.e.* pronouns, proper names, pro-drops), we can still understand who is the story character in these scenarios. However, what linguistic factors have made it clear for us to understand it? If we present these clauses to an AI agent, would it "understand" them? If so, what would be its strategy to "understand" it? And would the strategy taken by the AI algorithm share any consistency with linguistic theory for semantic parsing and pronoun resolution?

To understand the method motivation of this experiment, we can first think about the application of image classification in the computer vision field. This study by Zhu et al. L. Zhu et al. (2018) applied a deep neural network called AlexNet to classify vegetable pictures. As shown in Figure 6.2, the original broccoli picture is used as an input, and more features are learned to "understand" to classify this picture correctly

Figure 6.1: Clauses (1) to (5) all have "the little prince" as subject story character. The text highlighted in red are the subjects in their clauses. The square bracket '[ ]' in example (3) refers to a pro-drop.

after multiple layers of matrix convolution and transformation. The model can grab these features better and better when a lot of pictures are used as training materials. In this image classification task, the model is able to learn abstract image contour and shade features as the convolutional layers go deeper.

Inspired by this classification concept in machine learning, if we could train a language neural network to classify the subject story characters, and take the clauses as the input, we could explore what features the model relies on to retrieve this semantic information. This would benefit our understanding of the factors for semantic information retrieval in languages and can enable us to compare the features between different languages with statistical methods.

There are many ways in linguistics to factorize a language to show its inner structure, such as part-of-speech (POS) tag, constituency tree, and dependency tree. As shown in Figure 6.3, the clauses are parsed as dependency trees. These parsing processes provide a possibility to generate universal features for different languages, and focusing on features such as dependency types can help us understand story character resolution from a new perspective.

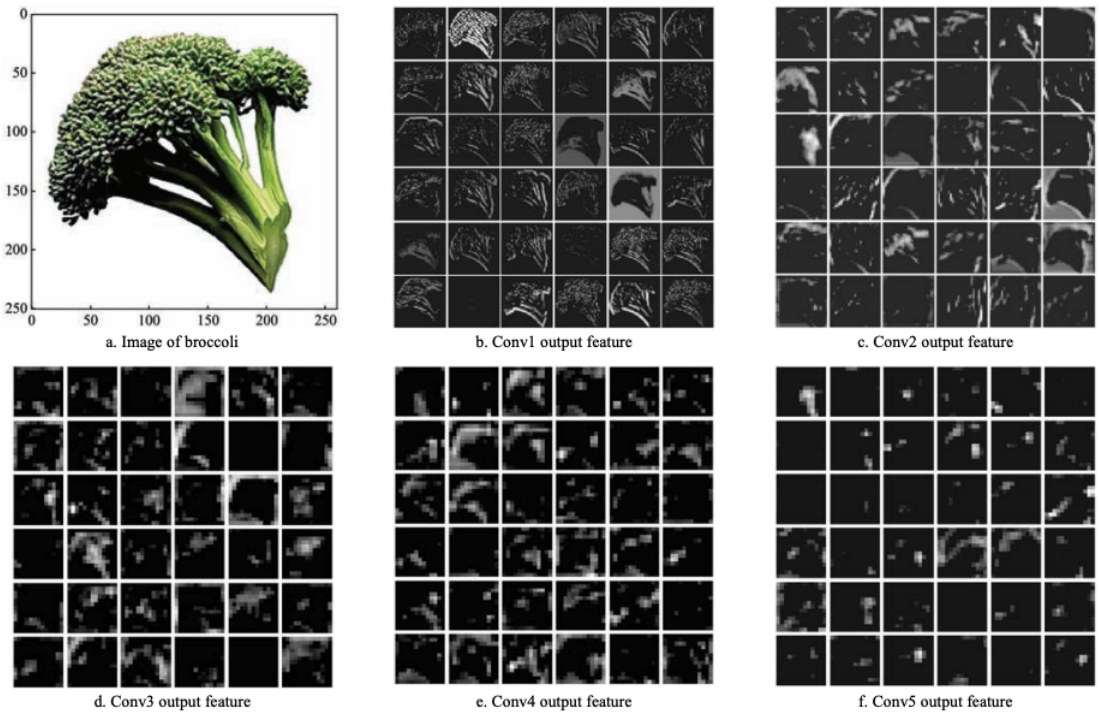


Figure 6.2: Figure 3 from L. Zhu et al. (2018)'s study on using AlexNet to classify vegetable pictures.

Since 2010, neural networks have been playing a significant role in natural language processing (NLP). However, these methods often consider text as a bag of words without taking semantic or syntactic structure into consideration. Recently, the rapidly developing field of Graph Neural Networks (GNN) (J. Zhou et al., 2020) has made it possible to take structural information as its components, and show promising outcomes in NLP applications (B. Liu & Wu, 2022).

In this experiment, carefully designed Graph Attention Networks (GATs) were adopted to realize our goal of exploring what semantic dependency factors are more important for the model to classify clauses' subject story characters correctly in Chinese, Brazilian Portuguese, and Spanish. The results (see Section 6.3 for details) indicated that GATs valued highly on the main verb, the subject, and the object for all three languages. The marker and auxiliary are ranked higher in Chinese since they carry tense information. The determiner and case, on the other hand, are more important in Brazilian Portuguese and Spanish since they provide gender information. As for pro-drop cases trained models, the "thicker tail" effect indicated

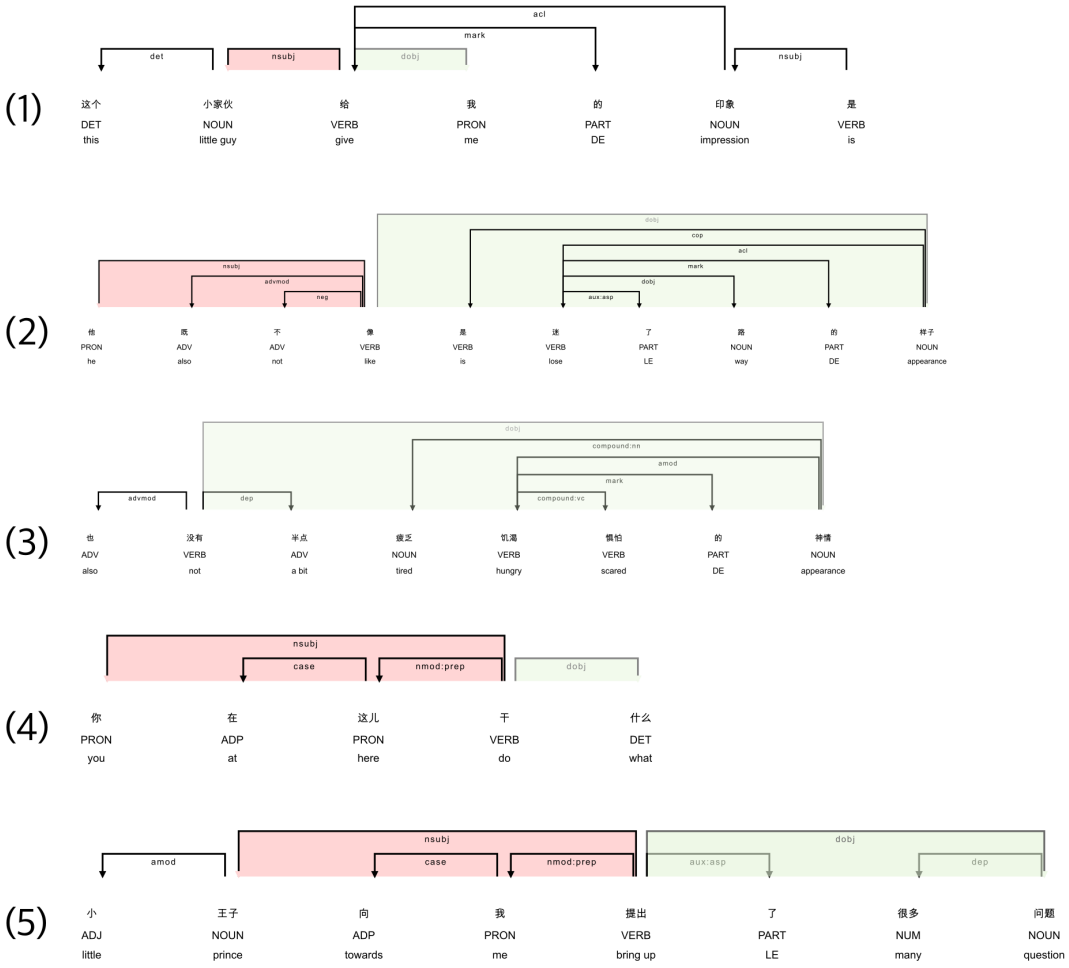


Figure 6.3: Dependency parsing results for the clauses in Figure 6.1. The parsing was realized with SpaCy's Chinese model (see Section 6.2.2 for details).

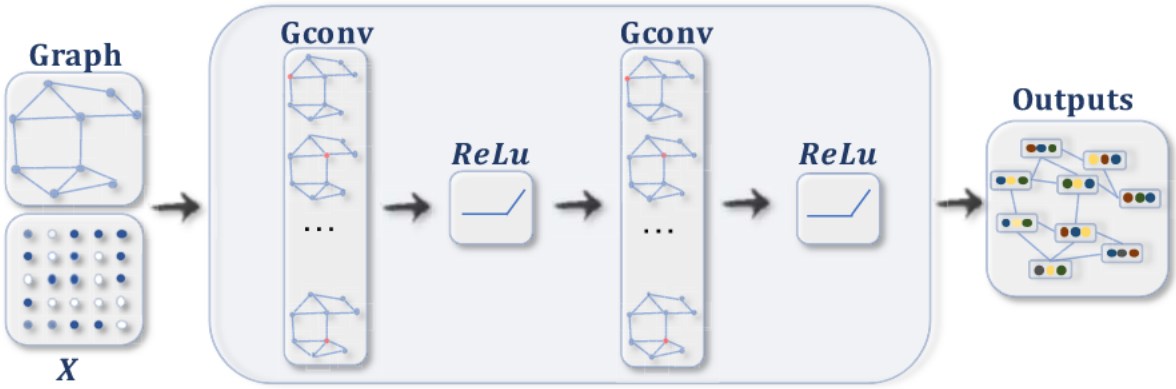


Figure 6.4: Figure visualizing the structure of a GNN (Figure 2 from Wu et al. (2020)).

that the model seeks information among the lower ranked dependency types harder to obtain the missing story character information.

In the following subsections, we will review the application of GNN, and discuss why and how GNN can be an excellent tool to be applied to structural language data.

6.1.1 Graph Neural Networks

Graph Neural Networks is a type of neural network that takes graphs as the data structure for input, perceptrons, and output. Compared with other data structures (*e.g.* number vectors or matrix, sequence of words), the graph structures are able to show non-Euclidean relationships such as nodes and edges between them (Asif et al., 2021; Miller et al., 2010; Singh, 2022). As shown in Figure 6.4, the inputs, outputs, and neural networks' nodes are all graphs in this GNN. These graphs contain nodes and edges/arcs structure, and it is feasible for us to apply linguistic trees in the network as graphs.

6.1.2 Graph Attention Networks

Graph Attention Networks (GATs), brought up by Velickovic et al. (2017), add attention layers in GNNs so that nodes in the graph are able to attend features from their neighborhoods' features. The GATs are

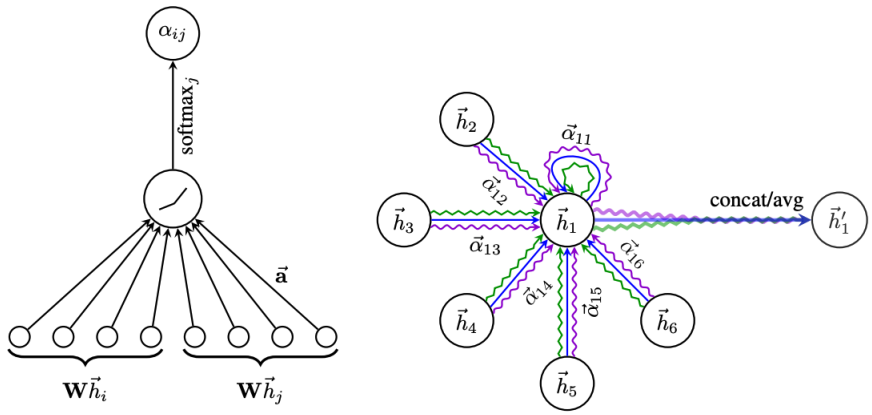


Figure 6.5: Figure visualizing one node from attention layer in GATs (Figure 1 from Velickovic et al. (2017)).

trained to specify different weights to different nodes in a neighborhood. At each attention layer, multiple attention heads can be used in the network to increase the system complexity in different tasks.

In this experiment, GATs are used to explore how different types of dependency structures support the task of subject story character resolution in multiple languages. The attention activation in the GATs is used as an indicator to examine the rationale and possibilities behind this method. One special design used in this study was the graph construction process (see Section 6.2.2), which allows all clauses represented in graphs that share the same dependency types order. With this design, all input graphs' consistent dependency types make it possible to retrieve the attention layer results and be analyzed together for group statistical analyses.

6.1.3 Application of GNN in Linguistics

The sub-fields in linguistics have already wisely used the graph structure to represent relationships in languages, as human language has innate compositional, hierarchical, and flexible structures. For example, the syntax tree and semantic representation such as AMR can both be considered graphs.

Closely relevant to our study, GNN had been adopted to solve problems such as pronoun resolution and coreference resolution (Guo et al., 2022; Jiang & Cohn, 2022; L. Liu et al., 2020; Phung et al., 2021;

Y. Xu & Yang, 2019), relation extraction (P. Chen et al., 2021; Park et al., 2022), and even zero pronoun resolution (Pandit et al., 2020; J. Yang et al., 2022; J. Yang et al., 2020; J. Yang et al., 2021).

In Chapter 5, we provide a model that includes linguistic factors "consciously" to predict *pro*-drop in the discourse. However, these discrete factors can hardly capture all the inner features of the language, especially the link between linguistic elements. Therefore, in this chapter, a GNN model will be used as the core algorithm to learn the pattern of *pro*-drop, while semantic and syntactic structures act as graph inputs. These graph structures will replace the selected features and keep more naturalistic information in the discourse.

6.2 Method

6.2.1 Discourse Material

The discourse materials used in this study are Chinese (CN), Brazilian Portuguese (BP), and Spanish (ES) translation (xiaowangzi.org, 2021) of Saint-Exupéry's *The Little Prince*.

6.2.2 Discourse Preprocessing and Graph Construction

As shown in Figure 6.6's "Preprocessing" and "Graph Construction" sub-figure, the following preprocessing steps were taken to prepare a raw discourse into GNN input graphs: (1) Preprocessing: Select clauses with story character subjects; Generate word embeddings with multiple linguistic features; Obtain dependency feature for each clause. (2) Graph construction: Generate node feature, edge feature, and graph feature files. The details for these preprocessing steps are described in the following subsections.

Discourse Preprocessing

Select clauses with story character subjects. There are 32 story characters (see Table 4.2) used in this experiment. The clauses containing these story characters as subjects are selected for further analysis. The numbers of clauses for the three languages are as follows: CN 1636 clauses, BP 1229 clauses, and ES 1220 clauses. Each clause's corresponding subject character is represented as a 36-digit length vector, and it will be used as the output when training the GATs. As for making the label, the story character i will have the

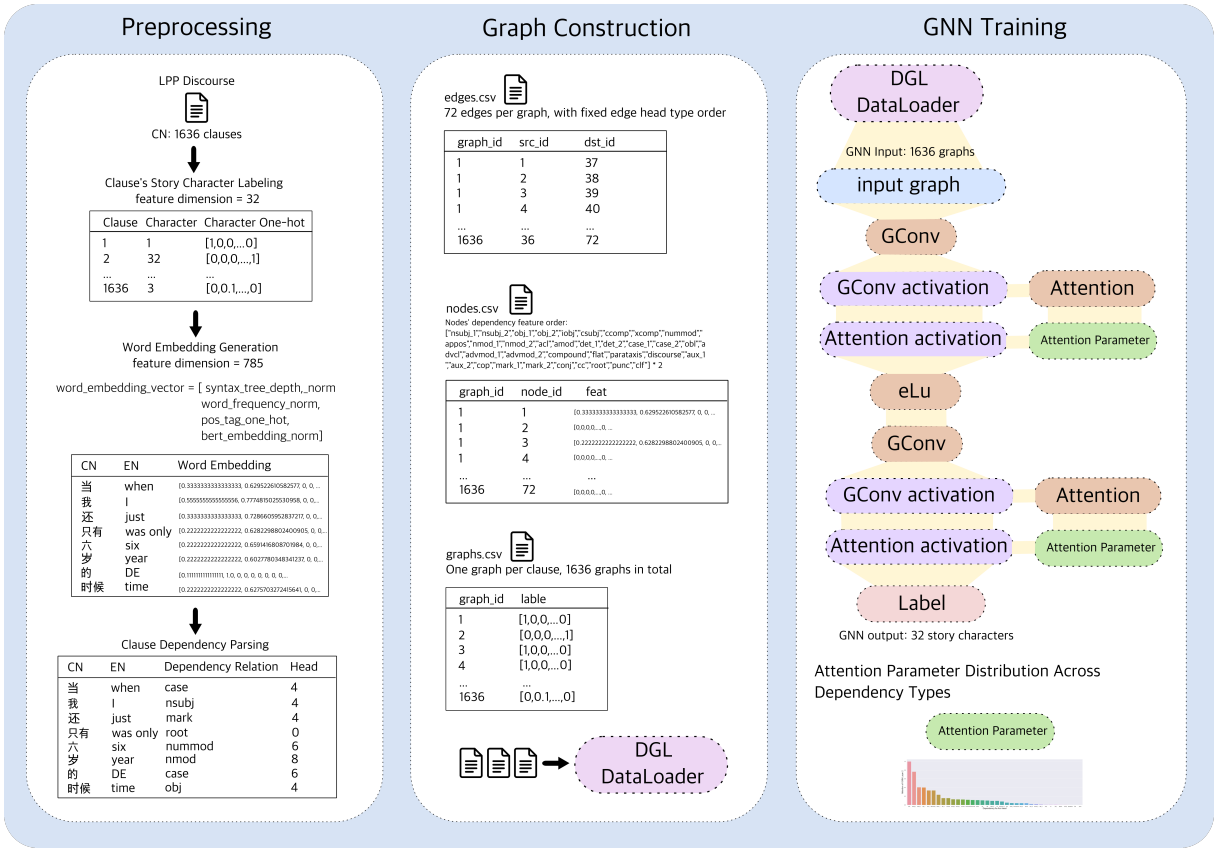


Figure 6.6: Data processing procedures: (1) Preprocessing; (2) Graph construction; (3) GNN training and Attention parameter statistical analysis (see detailed vector transformation in GNN in Figure 6.7).

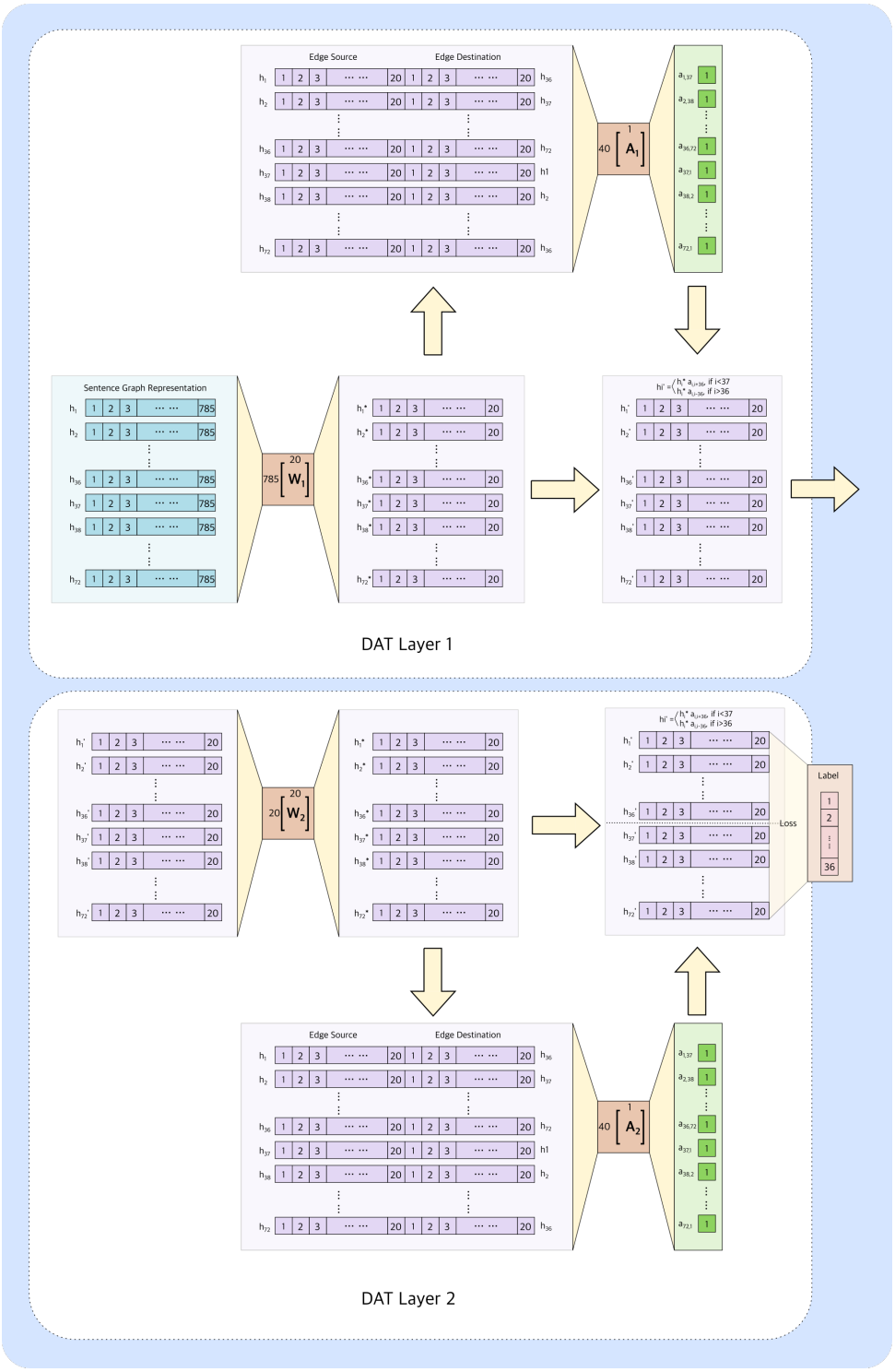


Figure 6.7: GNN layer visualization with detailed vector shape information. The input vector size shown in this figure is the mixed Chinese word embedding (as described in Section 6.2.2), which is 785, and it would be 768 for the ones for all original BERT cases in all three languages

(a)

word	word_en	word_id	src_id	dst_id	dep_rel
当	when	0	3	0	case
我	I	1	3	1	nsubj
还	just	2	3	2	mark
只有	only have	3	3	3	root
六	six	4	5	4	nummod
岁	year's old	5	7	5	nmod
的	DE	6	5	6	case
时候	time	7	3	7	obj

(b)

node_id_reorder	dep_rel	src_id	dst_id
0	nsubj	3	1
1	nsubj		
2	obj	3	7
3	obj		
4	iobj		
5	csubj		
6	ccomp		
7	xcomp		
8	nummod	5	4
9	appos		
10	nmod	7	5
11	nmod		
12	acl		
13	amod		
14	det		
15	det		
16	case	3	0
17	case	5	6
18	obl		
19	advcl		
20	advmod		
21	advmod		
22	compound		
23	flat		
24	parataxis		
25	discourse		
26	aux		
27	aux		
28	cop		
29	mark	3	2
30	mark		
31	conj		
32	cc		
33	root	3	3
34	punc		
35	clf		

(c)

node_id_final	dep_rel	word_id	word
0	nsubj	3	只有
1	nsubj		
2	obj	3	只有
3	obj		
4	iobj		
5	csubj		
6	ccomp		
7	xcomp		
8	nummod	5	岁
9	appos		
10	nmod	7	
11	nmod		
12	acl		
13	amod		
14	det		
15	det		
16	case	3	只有
17	case	5	
18	obl		
19	advcl		
20	advmod		
21	advmod		
22	compound		
23	flat		
24	parataxis		
25	discourse		
26	aux		
27	aux		
28	cop		
29	mark	3	只有
30	mark		
31	conj		
32	cc		
33	root	3	只有
34	punc		
35	clf		
36	nsubj	1	
37	nsubj		
38	obj	7	
39	obj		
40	iobj		
41	csubj		
42	ccomp		
43	xcomp		
44	nummod	4	
45	appos		
46	nmod	5	
47	nmod		
48	acl		
49	amod		
50	det		
51	det		
52	case	0	
53	case	6	
54	obl		
55	advcl		
56	advmod		
57	advmod		
58	compound		
59	flat		
60	parataxis		
61	discourse		
62	aux		
63	aux		
64	cop		
65	mark	2	
66	mark		
67	conj		
68	cc		
69	root	3	只有
70	punc		
71	clf		

(d)

graph_id	src_id	dst_id
1	0	36
1	1	37
1	2	38
1	3	39
1	4	40
1	5	41
1	6	42
1	7	43
1	8	44
1	9	45
1	10	46
1	11	47
1	12	48
1	13	49
1	14	50
1	15	51
1	16	52
1	17	53
1	18	54
1	19	55
1	20	56
1	21	57
1	22	58
1	23	59
1	24	60
1	25	61
1	26	62
1	27	63
1	28	64
1	29	65
1	30	66
1	31	67
1	32	68
1	33	69
1	34	70
1	35	71
1	36	0
1	37	1
1	38	2
1	39	3
1	40	4
1	41	5
1	42	6
1	43	7
1	44	8
1	45	9
1	46	10
1	47	11
1	48	12
1	49	13
1	50	14
1	51	15
1	52	16
1	53	17
1	54	18
1	55	19
1	56	20
1	57	21
1	58	22
1	59	23
1	60	24
1	61	25
1	62	26
1	63	27
1	64	28
1	65	29
1	66	30
1	67	31
1	68	32
1	69	33
1	70	34
1	71	35

Figure 6.8: An example of the process to convert an original clause into a graph as GNN input. The following steps are taken to generate a dependency-order-consistent edge graph file: (a) Original clause with naturalistic word order, dependency relationship type (*i.e.* "dep_rel"), dependency head (*i.e.* govener of the dependency relation, or source, "src_id"), and tail (*i.e.* dependent of the dependency relation, or destination, "dst_id"); (b) A dataframe generated with a fixed dependency relation type order (36 dependency types in total), as shown in the dep_rel column, repetitive dependency types are allowed since one relationship can show up multiple times in a single clause; The word_ids in (a) are filled in (b) based on each word's "dep_rel" type; (c) A fill list of nodes containing all head nodes and tail nodes are constructed, the blank cells are the ones that do not show up in this clause; This is final node order that is used to generate the "nodes.csv" file so that the words in such an order that the edge for each clause will maintain the same throughout the discourse; (4) This is the consistent form for each clause in the "edges.csv" file, which means that the node_id order will be the same for all 1608 clauses.

i_{th} element in the vector as 1, and all the rest elements would be zero. For example, story character 4 is labeled as $[0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]$.

Generate word embeddings for each word. BERT word embeddings generated by SpaCy language models with a length of 768 were used for each language. For Chinese, word segmentation was done by native speakers (as mentioned in Chapter 4). Besides the original BERT embedding, an extra "mixed-factor" word embedding was made for CN to explore the effect of a mixture of linguistic factors and a 785-dimension vector was generated for each word. The following elements were included in the 785-dimension word embedding: (1) The word's depth in its clause's syntax tree, which was parsed with Zpar(Y. Zhang & Clark, 2011a) constituency parser; (2) Word frequency, estimated using Google ngrams (Version 20120701) and the SUBTLEX corpora for Chinese (Cai & Brysbaert, 2010); (3) Part-Of-Speech (POS) type one-hot vector, obtained via SpaCy[1]; (4) BERT embedding, obtained via SpaCy using model *zh_core_web_trf*. These four features were normalized to range from 0 to 1 individually and were concatenated into a 785-dimension vector as the final word embedding.

Clause dependency parsing. The dependency parsing for the clauses provides a graph structure, and it shows how word elements are connected. The dependency parsing was completed with SpaCy, and it provided a dependency relationship (*dep_rel*) and relationship head for each word. In other words, the *dep_rel* indicates the destination nodes' dependency feature, and *head* indicates the head node's location in the clause. The dependency types in CN are slightly different than the ones in BP and ES, and they have a joined subset of 31 dependencies. As shown in the results in Figure 6.9, the 31 dependencies on the x-axis are the subset of dependencies that were commonly used in all three languages.

Graph Construction

After the preprocessing steps, we have had access to the "graph structure" information for each clause. However, more steps need to be done so that we have the format that can be used directly in Deep Graph Library (DGL)[2] for GATs training purposes. As shown in Figure 6.6, the structure of a graph is stored in three separate CSV files, and they individually contain the nodes' features, the edges' features, and the

[1] https://spacy.io/models
[2] https://www.dgl.ai/

graph's features. The hierarchical structures between these three files are: One discourse contains multiple clauses, and one clause is represented as one graph; Each graph contains multiple edges, and each edge links two nodes.

Remember that the final goal of training the GATs is to obtain the activation in the attention layers, and the order of the input graph's dependency features needs to remain consistent across all the graphs. Therefore, we need to convert the original word order to fit them into a fixed dependency list. If a type of dependency is not showing up in a clause, that certain dependency element will be filled with NULL to keep the spot occupied, and its word embedding feature is filled with all a zero vector. As a result, the edge feature for each graph will be the same, and the *node_id* in the node feature will be converted to satisfy this (see Figure 6.8 for a visualization of this process).

The detailed steps for making the three CSV subfiles are:

(1) Edge Feature. The *edges.csv* includes three columns: *graph_id*, *src_id*, *dst_id*. (1) *graph_id*: This is the same as clause id, and each clause will get an id as its naturalistic order in the discourse; (2) *src_id* is the head node's id; (3) *dst_id* is the tail node's id.

(2) Node Feature. The *nodes.csv* includes three columns: *graph_id*, *node_id*, *feat*. (1) *graph_id*: This is the same as clause id; (2) *node_id* is an id assigned to each word; Note that this id is different from the naturalistic word order since the nodes are reordered to make edges consistent across graphs. (3) *feat* is the 768 or 785-dimension word embedding feature for each word as described in the previous subsection.

(3) Graph Feature. The *graphs.csv* includes two columns: *graph_id*, *label*. (1) *graph_id*: This is the same as clause id; (2) *label* is the 36-dimension vector story character label as described in Section 6.2.2.

6.2.3 GATs Training

The Graph Attention Networks (GATs) used in this experiment are a two-layer GNN with one attention head on each layer (see a brief structure visualization in Figure 6.6, and a detailed vector transformation in Figure 6.7).

The input and output for the GATs are: (1) Input: all graphs generated from all clauses; As each clause is represented as a single graph, this graph will act as the input of the GAT in the form of a [72, 768] tensor

([72, 785] for the CN mix-factor word embedding model), and it contains 72 nodes and each node has 768-dimensional (or 785 dimensional) word embedding features. (2) Output: After going through two fully connected graph convolutional layers and two attention layers, the output of the GATs is a [1, 72] size vector. The loss of the network is calculated via cross-entropy between the output and the [1,36] story character label.

Equation 6.1 shows the input is a [768, 72] matrix, and h_i is the i_{th} node's word embedding in the graph:

$$matrix_{L1\_input} = [h_1, h_2, ..., h_{72}] \tag{6.1}$$

Equation 6.2 is the first fully connected graph convolutional layer in the GATs. The matrix W_1 is learned during the training process, and $matrix_{L1\_FC}$ is the activation of this layer:

$$matrix_{L1\_FC} = [h_1, h_2, ..., h_{72}] * W_1 = [h_1^*, h_2^*, ..., h_{72}^*], \tag{6.2}$$

Equation 6.3 is the first attention layer in the GATs. As defined in the graph edge information, the head and tail nodes' activation are concatenated to obtain a vector $[h_i^*, h_{i+36}^*]$ to be convolved with $e_{i,j}$. The $e_{i,j}$ is an unnormalized attention score between two neighbors that form an edge. The $matrix_{L1\_A1}$ is the activation result of the first attention layer. The raw attention score vector $e_{i,j}$ is used for our attention activation analysis:

$$attention\_vector_{L1\_A1} = [[h_1^*, h_{37}^*] * e_{(1,36)}, [h_2^*, h_{38}^*] * e_{(2,38)}, ..., [h_{36}^*, h_{72}^*] * e_{(36,72)}] \tag{6.3}$$

$$\alpha\_1_{(ij)} = exp(e_{(i,j)}) / \sum_{k \in N(i))} exp(e_{(i,k)}) \tag{6.4}$$

$$matrix_{L1\_A1} = attention\_vector_{L1\_A1} * matrix_{L1\_FC} = \sum_{j \in N(i))} \alpha_{ij} hj^* = [h_1^{'}, h_2^{'}, ..., h_{72}^{'}] \tag{6.5}$$

Equation 6.6, and Equation 6.9 are the functions for the second fully connected graph convolutional layer and the second attention layer, which are similar to the ones in the first layers:

$$matrix_{L2\_FC} = [h_1^{'}, h_2^{'}, ..., h_{72}^{'}] * W_2 = [h_1^{**}, h_2^{**}, ..., h_{72}^{**}] \tag{6.6}$$

$$attention\_vector_{L2\_A2} = [[h_1^{**}, h_{37}^{**}] * e_{(1,36)}, [h_2^{**}, h_{38}^{**}] * e_{(2,38)}, ..., [h_{36}^{**}, h_{72}^{**}] * e_{(36,72)}] \tag{6.7}$$

$$\alpha\_2_{(ij)} = exp(e_{(i,j)}) / \sum_{k \in N(i))} exp(e_{(i,k)}) \tag{6.8}$$

$$matrix_{L2\_A2} = attention\_vector_{L2\_A2} * matrix_{L2\_FC} = \sum_{j \in N(i))} \alpha_{ij} hj^{**} = [h_1^{''}, h_2^{''}, ..., h_{72}^{''}] \tag{6.9}$$

The output of the GATs, as shown in Equation 6.10 is the activation after the second attention layer (Equation 6.9). The shape of the output vector is [1,72], and it is "folded" as Equation 6.11. On the one hand, this folding operation will be descriptive for a dependency node while it is acting as a head and a tail in the edge. On the other hand, the folding output provides a [1,36] shaped vector to calculate the loss with the label. The loss function is cross entropy as shown in Equation 6.12. It should be noted that another solution that can be applied to achieve the shape matching between the second graph convolutional layer and the label is to add a linear layer, and it can learn the parameters for head and tail nodes automatically after training.

$$matrix_{output} = matrix_{L2\_A2} = [h_1^{''}, h_2^{''}, ..., h_{72}^{''}] \tag{6.10}$$

$$matrix_{output\_fold}^{'} = [(h_1^{''} + h_{37}^{''}), (h_2^{''} + h_{38}^{''}), ..., (h_{36}^{''} + h_{72}^{''})] \tag{6.11}$$

$$Loss = Cross\_Entropy(matrix'_{output\_fold}, matrix_{story\_character\_label}) \tag{6.12}$$

The training of the GATs goes through 20 epochs on each batch (size of 50 graphs), and the learning rate is 1e-3. The 1_{st} and 2_{nd} attention layers' activation at the 20_{th} epoch was retrieved for further statistical analysis. The distributions of the attention layers' activation show how much attention has been given to an edge during the training process to realize the story character recognition for all clauses. The results of the distribution are introduced in Section 6.3.

6.2.4 Attention Activation Retrieval

The attention activation to be obtained from the trained GATs are the ones in the first and second attention layers, and they are described in Equation 6.3 and Equation 6.7 as $e_{(i,j)}$. These two vectors both have the size of [1, 72], and they are recorded when the 20_{th} epoch is done training.

The attention activation for the correct predicted graphs was first normalized to range 0 to 1 (as shown in Equation 6.13, e_i refers to the i_{th} element in the attention score vector). An average attention activation value was calculated across all the graphs, and T-tests were carried out to compare the activation values across languages and *pro*-drop conditions. The numbers of correct predicted graphs for each language in the 20_{th} batch are: CN, 1433 (out of 1600); BP, 948 (out of 1200); ES, 1038 (out of 1200).

$$e_i = (e_i - min([e_]) / (max([e_1, ..., e_{72}]) - min([e_1, ..., e_{72}])) \tag{6.13}$$

To make the attention activation value more descriptive, the percentage of each attention activation is calculated as in Equation 6.14. To simplify the description, it will still be mentioned as "attention activation" as in the following sections, but it is actually a percentage value. In Equation 6.14, e_i refers to the i_{th} element's normalized activation, and value n is the number of nodes that are taken into consideration (n = 31 since there are 31 common dependencies among all the languages).

$$e_i = e_i / \sum_{i=1}^{n} e_i \tag{6.14}$$

6.3 Result

As described in Section 6.2.2, the input nodes are reordered so that the dependency types for the GATs nodes remain the same across all graphs. Table 6.1 shows the dependency types and their labels used in this experiment (see https://universaldependencies.org/en/dep/index.html for a full list of dependency node types), and the column *"Dependency Label"* is the dependency types analyzed across conditions.

Table 6.1: Dependency types and labels. The attention value comparisons listed in this table are obtained from the T-tests results as presented in Table 6.2, 6.3, and 6.4.

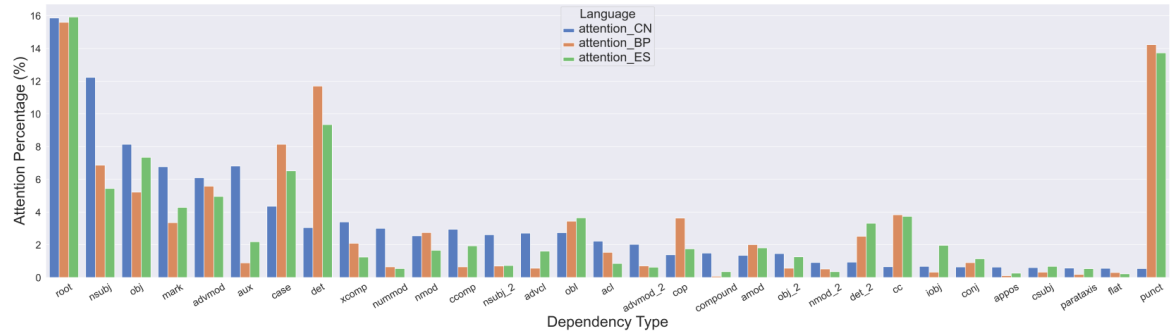
Dependency Category	Dependency	Dependency Label	Attention all graphs	Attention pro-drop graphs	Attention non-pro-drop graphs
Core dependents of clausal predicates	nominal topic	nsubj	CN >BP >ES	CN >BP, ES	CN >BP >ES
	object	obj	CN, BP >ES	ES >BP	CN >BP, ES
	oblique object	iobj	ES >CN, BP	-	ES >BP >CN
	clausal topic	csubj	CN >ES	BP >CN >ES	ES >BP >CN
	clausal complement	ccomp	CN, BP >ES	-	CN, ES >BP
	open clausal complement	xcomp	CN, BP >ES	CN, BP >ES	-
Noun dependents	numeric modifier	nummod	CN >BP, ES	CN, BP >ES	CN, ES >BP
	appositional modifier	appos	-	CN, BP >ES	ES >BP >CN
	nominal modifier	nmod	-	CN, BP >ES	-
	clausal modifier of a noun (adnominal clause)	acl	CN >ES	CN, BP >ES	-
	adjectival modifier	amod	BP, ES >CN	-	BP, ES >CN
	determiner	det	BP, ES >CN	ES >CN	ES >BP >CN
Case-marking, prepositions, possessive	case-marking	case	BP, ES >CN	-	-
Non-core dependents of clausal predicates	indirect nominal	obl	BP, ES >CN	BP, ES >CN	-
	adverbial clause modifier	advcl	CN >ES >BP	-	ES >BP
	adverbial modifier	advmod	CN >ES	-	CN >BP >ES
Compounding and unanalyzed	compound	compound	CN >ES >BP	CN, BP >ES	ES >CN, BP
	flat multiword expression	flat	-	BP >CN >ES	ES >BP >CN
Loose joining relations	parataxis	parataxis	ES >CN, BP	BP >CN, ES	ES >BP >CN
Special clausal dependents	auxiliary	aux	CN >ES >BP	CN >BP, ES	CN >ES >BP
	copula	cop	BP >CN, ES	-	BP, ES >CN
	marker	mark	CN >ES >BP	-	CN >BP, ES
Coordination	conjunct	conj	BP, ES >CN	ES >BP >CN	ES >BP >CN
	coordinating conjunction	cc	BP, ES >CN	BP, ES >CN	BP, ES >CN
Sentence head	root	root	CN, BP >ES	ES >CN >BP	CN >BP >ES
Punctuation	punctuation	punct	BP >ES >CN	ES >BP >CN	ES >BP >CN

6.3.1 Attention Activation Distribution: All Clauses

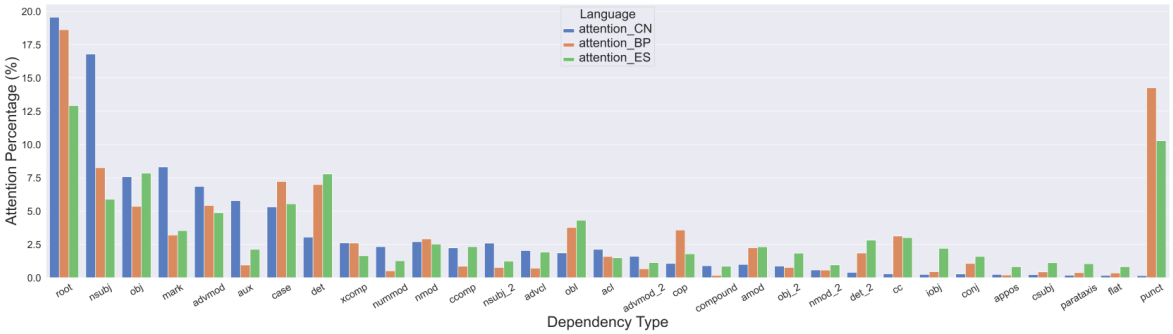
Figure 6.9 and Figure 6.10 show the attention activation distribution across dependency types in all three languages while all the models were trained on all clauses' graphs. As introduced in previous sections, the GATs were trained on 72 nodes, which contain 36 head nodes and 36 tail nodes. As a node can be a head on one edge and a tail on another, it is possible for the same word embedding to show up multiple times in a graph. While represented in Equation 6.11, the output was folded between head nodes and tail nodes, so that each edge formed one element in the final output. Figure 6.9 shows the activation sum between head nodes and tail nodes, whereas Figure 6.9 shows the head node and tail node activation separately.

Table 6.2 shows the T-test results for all three languages using head and tail sum activation values. Bonferroni correction was used for all p-values.

(a) Attention from Layer 1 – nodes as head + tail



(b) Attention from Layer 2 – nodes as head + tail



(c) Attention from Layer 1 + Layer 2 – nodes as head + tail

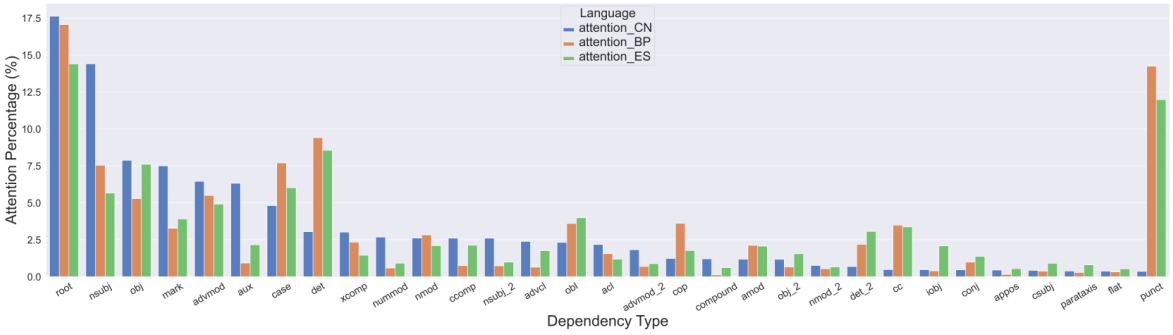
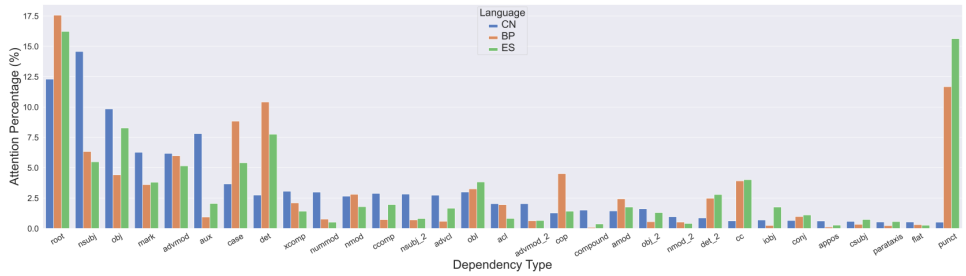
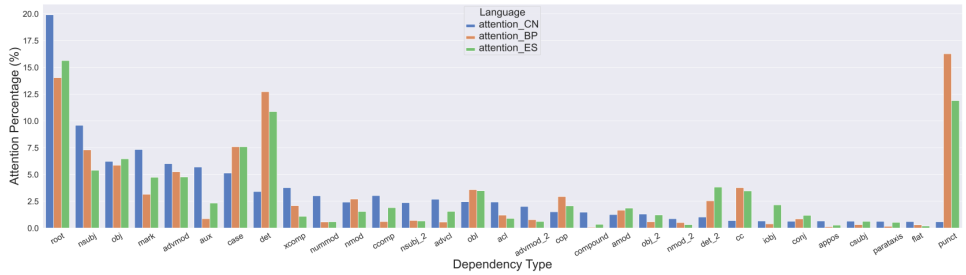


Figure 6.9: Graph Attention Neural Network's attention layer activation distribution across dependency types, and compared between CN, BP, ES. The attention percentages were calculated based on each dependency (31 types) type within each language (*i.e.* CN, BP, and ES). The attention values were added from the nodes as the head ($0 - 35$) and nodes as the tail ($36 - 71$) (corresponding to the red and green sections in Figure 6.8, see Figure 6.10 for head and tail separate results). The subfigures are the attention percentage distribution calculated based on: (a) The 1st attention layer's activation in the GAT; (b) The 2nd attention layer's activation in the GAT; (c) The sum of the 1st and the 2nd layers' activation in the GAT.

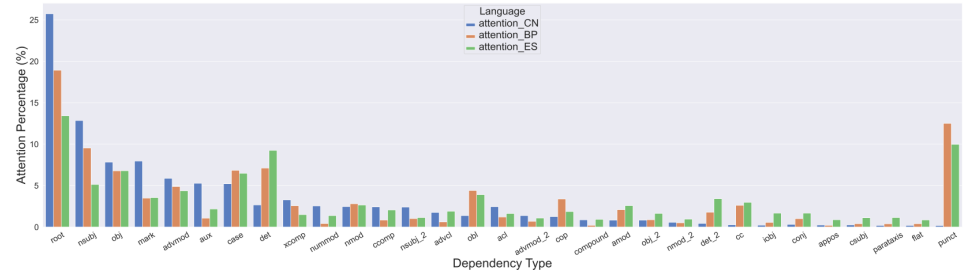
(a) Attention from Layer 1 – nodes as head



(b) Attention from Layer 1 – nodes as tail



(c) Attention from Layer 2 – nodes as head



(d) Attention from Layer 2 – nodes as tail

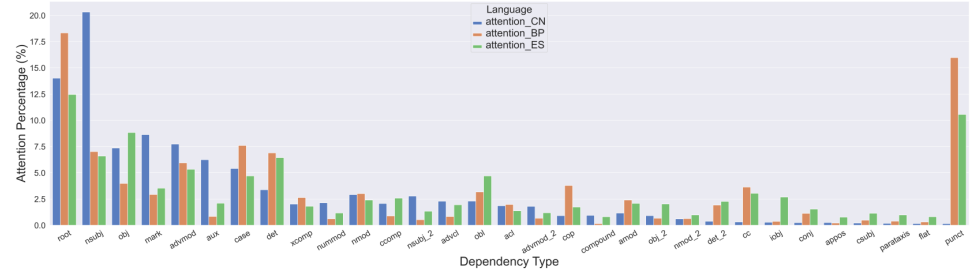


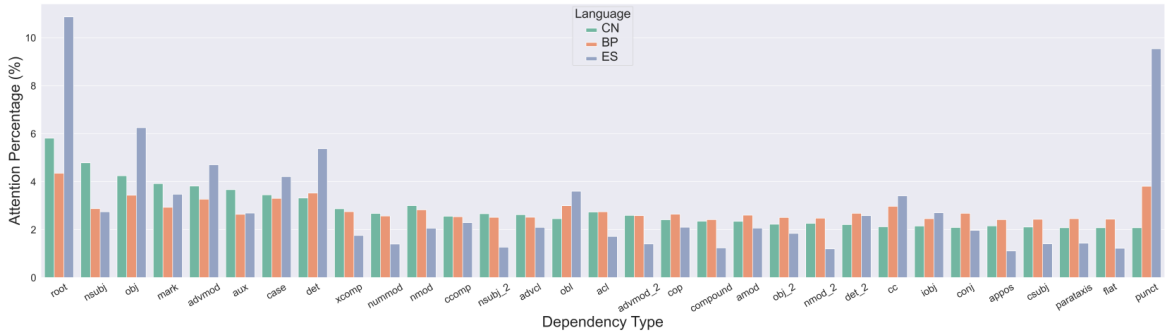
Figure 6.10: Graph Attention Neural Network's attention layer activation distribution across dependency types, and compared between CN, BP, ES. The attention percentages were calculated based on each dependency (31 types) type within each language (*i.e.* CN, BP, and ES). The attention values were the nodes as the head ($0 - 35$) and nodes as the tail ($36 - 71$) (corresponding to the red and green sections in Figure 6.8. The subfigures are attention activation from (a) 1st attention layer's head nodes; (b) 1st attention layer's tail nodes; (c) 2nd attention layer's head nodes; (d) 2nd attention layer's tail nodes.

6.3.2 Attention Activation Distribution: *Pro*-drop vs. Non-*pro*-drop Clauses

Figure 6.11 shows the attention activation for (a) pro-drop cases' trained GATs and (b) non-pro-drop cases' trained GATs for three languages. Figure 6.12, using the same values as Figure 6.11, from the angle of each language, shows how *pro*-drop models' and non-*pro*-drop models' attention activation.

Table 6.3 shows the *pro*-drop trained GATs attention activation T-tests between languages on all dependency types. Table 6.4 shows the non-*pro*-drop trained GATs attention activation T-tests between languages on all dependency types. Table 6.5 shows pro-drop vs. non-pro-drop attention activation across all dependency types within each language.

(a) Pro-drop Attention from Layer 1 + Layer 2 - nodes as head + tail



(b) Non-pro-drop Attention from Layer 1 + Layer 2 - nodes as head + tail

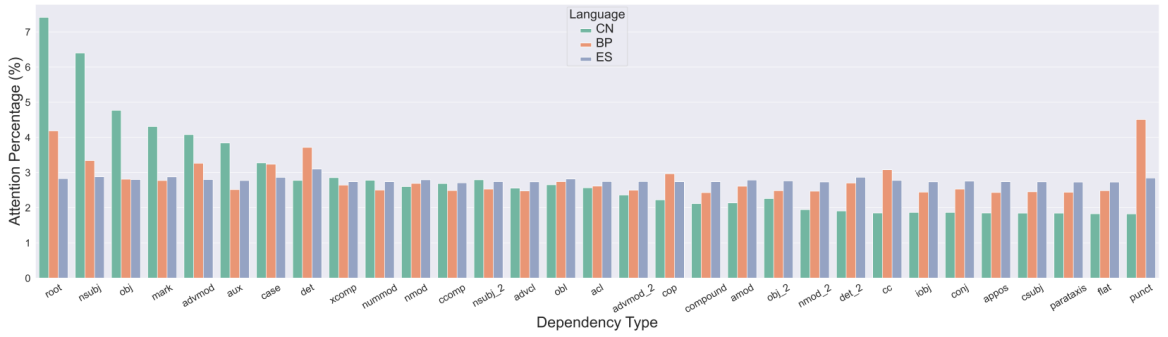
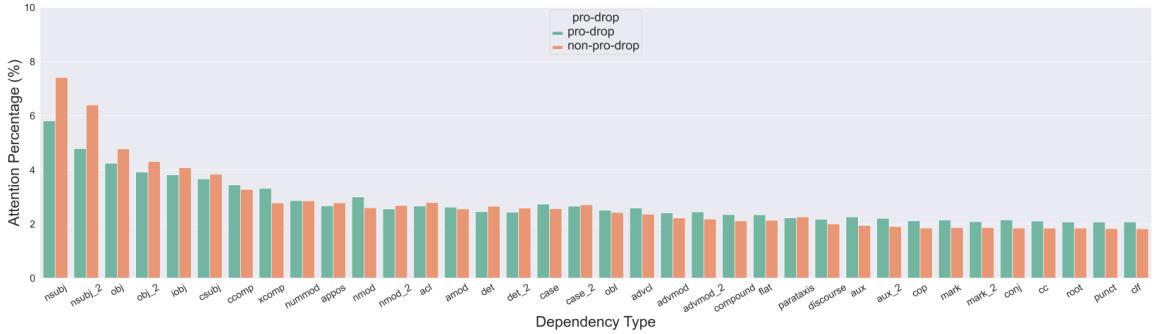
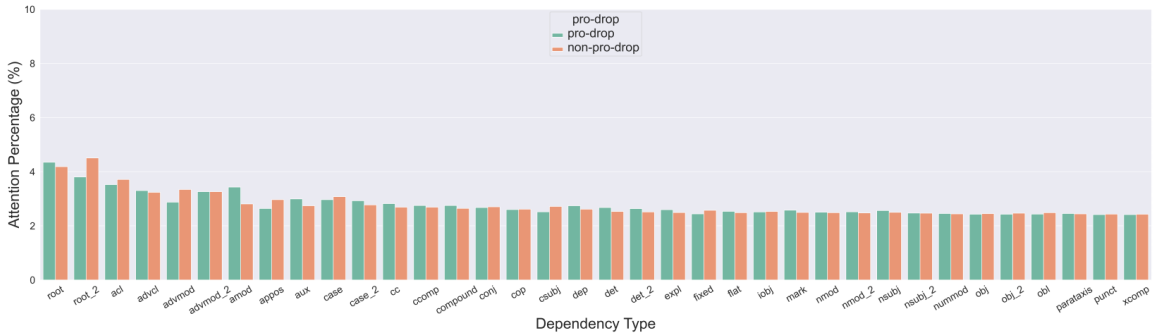


Figure 6.11: Graph Attention Neural Network's attention layer activation distribution across all dependency types. Pro-drop results are GATs trained on *pro*-drop graphs, and non-*pro*-drop results are GATs trained on non-*pro*-drop graphs. The attention values were calculated based on Layer 1 + Layer 2, and head + tail within each model. The subfigures are comparisons among all three languages: (a) GAT attention when trained on *pro*-drop cases; (b) GAT attention when trained on non-*pro*-drop cases. See a comparison result between pro-drop and non-pro-drop within each language in Figure 6.12.

(a) CN: pro-drop vs. non-pro-drop



(b) BP: pro-drop vs. non-pro-drop



(c) ES: pro-drop vs. non-pro-drop

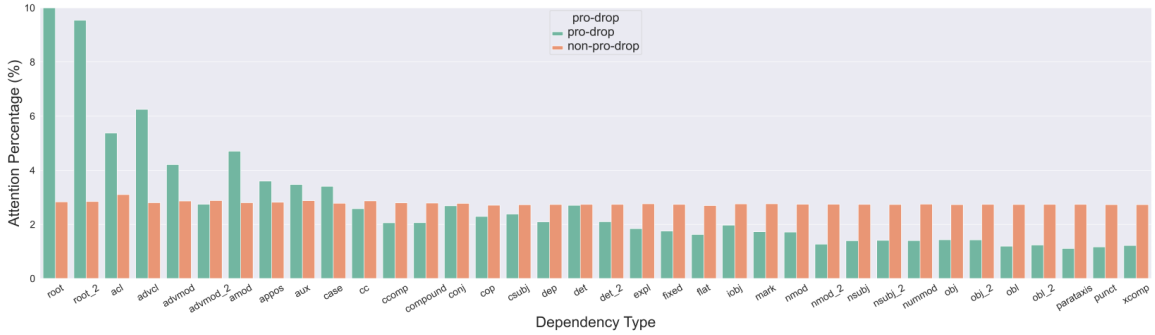


Figure 6.12: Graph Attention Neural Network's attention layer activation distribution across all dependency types. Pro-drop results are GATs trained on *pro*-drop graphs, and non-*pro*-drop results are GATs trained on non-*pro*-drop graphs. The attention values were calculated based on Layer 1 + Layer 2, and head + tail within each model. The subfigures are comparisons within each language between *pro*-drop and non-*pro*-drop: (a) CN; (b) BP; (c) ES.

Table 6.2: Statistical results for attention activation (Layer 1 + Layer 2, head + tail) across dependency types among three languages, and GATs trained on all graphs. The results were independent t-tests with Bonferroni correction. The colors indicate the direction between the languages: If it is green, it is true that the column's direction is correct, otherwise, it is red; If there is no color, the statistical test was not significant (*i.e.* p-value > 0.05).

			All cases					
			CN >BP		**CN >ES**		**BP >ES**	
ID	**Dependency Type**	**Order**	***t-value***	***p-value***	***t-value***	***p-value***	***t-value***	***p-value***
0	root	CN, BP >ES	2.367	1.0	7.767	0.0	4.958	0.0
1	nsubj	CN >BP >ES	18.492	0.0	23.745	0.0	4.661	0.0
2	obj	CN, BP >ES	6.399	0.0	-1.318	1.0	-7.616	0.0
3	mark	CN >ES >BP	12.729	0.0	9.236	0.0	-3.682	0.0
4	advmod	CN >ES	2.797	0.465	3.376	0.093	0.368	1.0
5	aux	CN >ES >BP	16.188	0.0	12.062	0.0	-5.824	0.0
6	case	BP, ES >CN	-8.546	0.0	-5.251	0.0	3.459	0.093
7	det	BP, ES >CN	-21.915	0.0	-23.12	0.0	2.058	1.0
8	xcomp	CN, BP >ES	2.568	0.93	6.253	0.0	3.541	0.0
9	nummod	CN >BP, ES	9.343	0.0	8.506	0.0	-1.654	1.0
10	nmod		0.21	1.0	2.937	0.279	2.722	0.651
11	ccomp	CN, BP >ES	8.272	0.0	1.102	1.0	-7.702	0.0
12	nsubj_2	CN >BP, ES	8.999	0.0	8.341	0.0	-1.684	1.0
13	advcl	CN >ES >BP	8.049	0.0	3.161	0.186	-6.219	0.0
14	obl	BP, ES >CN	-4.189	0.0	-6.987	0.0	-2.345	1.0
15	acl	CN >ES	3.124	0.186	5.231	0.0	1.851	1.0
16	advmod_2	CN >BP, ES	5.715	0.0	5.242	0.0	-0.834	1.0
17	cop	BP >CN, ES	-9.436	0.0	-2.039	1.0	7.375	0.0
18	compound	CN >ES >BP	7.054	0.0	4.272	0.0	-5.303	0.0
19	amod	BP, ES >CN	-6.029	0.0	-6.1	0.0	0.384	1.0
20	obj_2	ES >BP	3.237	0.093	-2.796	0.465	-5.436	0.0
21	nmod_2		1.89	1.0	0.847	1.0	-1.138	1.0
22	det_2	ES >BP >CN	-9.285	0.0	-14.605	0.0	-3.817	0.0
23	cc	BP, ES >CN	-14.568	0.0	-15.319	0.0	0.768	1.0
24	iobj	ES >CN, BP	0.02	1.0	-10.885	0.0	-8.815	0.0
25	conj	BP, ES >CN	-5.78	0.0	-8.101	0.0	-1.363	1.0
26	appos		2.25	1.0	-1.647	1.0	-2.77	0.558
27	csubj	CN >ES	-1.09	1.0	-5.4	0.0	-2.886	0.372
28	parataxis	ES >CN, BP	0.13	1.0	-5.002	0.0	-3.761	0.0
29	flat		-0.951	1.0	-2.722	0.651	-1.373	1.0
30	punct	BP >ES >CN	-45.294	0.0	-50.723	0.0	4.55	0.0

Table 6.3: Statistical results for attention activation (Layer 1 + Layer 2, head + tail) across dependency types among three languages, and GATs trained on pro-drop graphs. The results were independent t-tests with Bonferroni correction. The colors indicate the direction between the languages: If it is green, it is true that the column's direction is correct, otherwise, it is red; If there is no color, the statistical test was not significant (*i.e.* p-value > 0.05).

			pro-drop					
			CN >BP		**CN >ES**		**BP >ES**	
ID	**Dependency Type**	**Order**	***t-value***	***p-value***	***t-value***	***p-value***	***t-value***	***p-value***
0	root	ES >CN >BP	3.655	0.0	-6.387	0.0	-7.899	0.0
1	nsubj	CN >BP, ES	7.251	0.0	12.509	0.0	1.701	1.0
2	obj	ES >BP	2.185	1.0	-2.857	0.372	-4.045	0.0
3	mark		3.33	0.093	2.685	0.651	-0.644	1.0
4	advmod		1.26	1.0	-0.899	1.0	-1.738	1.0
5	aux	CN >BP, ES	4.174	0.0	4.736	0.0	-0.378	1.0
6	case		0.051	1.0	-0.575	1.0	-0.485	1.0
7	det	ES >CN	-2.212	1.0	-4.904	0.0	-1.888	1.0
8	xcomp	CN, BP >ES	0.531	1.0	5.23	0.0	3.801	0.0
9	nummod	CN, BP >ES	0.817	1.0	7.66	0.0	5.661	0.0
10	nmod	CN, BP >ES	0.832	1.0	5.489	0.0	4.001	0.0
11	ccomp		0.278	1.0	1.058	1.0	0.606	1.0
12	nsubj_2	CN, BP >ES	1.533	1.0	10.513	0.0	8.421	0.0
13	advcl		1.358	1.0	2.769	0.558	0.983	1.0
14	obl	BP, ES >CN	-4.102	0.0	-4.521	0.0	-1.018	1.0
15	acl	CN, BP >ES	-0.092	1.0	6.419	0.0	5.622	0.0
16	advmod_2	CN, BP >ES	-0.25	1.0	7.496	0.0	6.357	0.0
17	cop		-1.142	1.0	1.84	1.0	2.433	1.0
18	compound	CN, BP >ES	0.409	1.0	6.544	0.0	5.625	0.0
19	amod		-2.855	0.465	0.103	1.0	1.822	1.0
20	obj_2		-2.828	0.465	1.094	1.0	2.594	0.93
21	nmod_2	CN, BP >ES	-1.167	1.0	8.745	0.0	10.498	0.0
22	det_2	BP >CN	-5.182	0.0	-2.216	1.0	1.372	1.0
23	cc	BP, ES >CN	-7.562	0.0	-5.343	0.0	-0.472	1.0
24	iobj		-2.261	1.0	-3.216	0.093	-1.644	1.0
25	conj	ES >BP >CN	-6.314	0.0	-0.068	1.0	3.639	0.0
26	appos	CN, BP >ES	-2.269	1.0	10.555	0.0	13.23	0.0
27	csubj	BP >CN >ES	-3.526	0.0	4.86	0.0	6.416	0.0
28	parataxis	BP >CN, ES	-4.57	0.0	2.876	0.372	4.596	0.0
29	flat	BP >CN >ES	-4.576	0.0	4.383	0.0	5.815	0.0
30	punct	ES >BP >CN	-13.307	0.0	-19.387	0.0	-7.792	0.0

Table 6.4: Statistical results for attention activation (Layer 1 + Layer 2, head + tail) across dependency types among three languages, and GATs trained on non-pro-drop graphs. The results were independent t-tests with Bonferroni correction. The colors indicate the direction between the languages: If it is green, it is true that the column's direction is correct, otherwise, it is red; If there is no color, the statistical test was not significant (*i.e.* p-value > 0.05).

			non-pro-drop					
			CN >BP		**CN >ES**		**BP >ES**	
ID	**Dependency Type**	**Order**	***t-value***	***p-value***	***t-value***	***p-value***	***t-value***	***p-value***
0	root	CN >BP >ES	8.914	0.0	16.165	0.0	9.192	0.0
1	nsubj	CN >BP >ES	16.087	0.0	15.063	0.0	4.713	0.0
2	obj	CN >BP, ES	11.476	0.0	7.89	0.0	-0.6	1.0
3	mark	CN >BP, ES	11.933	0.0	6.725	0.0	-2.778	0.558
4	advmod	CN >BP >ES	3.782	0.0	6.305	0.0	3.875	0.0
5	aux	CN >ES >BP	11.385	0.0	5.608	0.0	-6.502	0.0
6	case		-1.697	1.0	1.651	1.0	3.203	0.093
7	det	ES >BP >CN	-11.033	0.0	-4.745	0.0	4.601	0.0
8	xcomp		3.33	0.093	1.224	1.0	-2.593	0.93
9	nummod	CN, ES >BP	4.334	0.0	-0.365	1.0	-7.785	0.0
10	nmod		-0.19	1.0	-1.407	1.0	-2.203	1.0
11	ccomp	CN, ES >BP	4.141	0.0	-0.589	1.0	-9.474	0.0
12	nsubj_2	CN, ES >BP	3.928	0.0	0.456	1.0	-5.69	0.0
13	advcl	ES >BP	3.029	0.186	-1.388	1.0	-9.16	0.0
14	obl		-0.537	1.0	-2.017	1.0	-2.573	0.93
15	acl		0.09	1.0	-1.934	1.0	-3.294	0.093
16	advmod_2	ES >CN, BP	0.178	1.0	-3.52	0.0	-6.013	0.0
17	cop	BP, ES >CN	-8.876	0.0	-4.779	0.0	3.283	0.093
18	compound	ES >CN, BP	-1.016	1.0	-5.878	0.0	-13.051	0.0
19	amod	BP, ES >CN	-6.719	0.0	-9.585	0.0	-3.25	0.093
20	obj_2	ES >CN, BP	-1.581	1.0	-6.629	0.0	-9.933	0.0
21	nmod_2	ES >BP >CN	-8.89	0.0	-15.917	0.0	-10.341	0.0
22	det_2	ES >BP >CN	-13.95	0.0	-21.721	0.0	-4.139	0.0
23	cc	BP, ES >CN	-14.442	0.0	-23.137	0.0	2.737	0.558
24	iobj	ES >BP >CN	-11.431	0.0	-20.707	0.0	-11.791	0.0
25	conj	ES >BP >CN	-11.421	0.0	-18.015	0.0	-6.616	0.0
26	appos	ES >BP >CN	-12.027	0.0	-26.131	0.0	-9.805	0.0
27	csubj	ES >BP >CN	-12.069	0.0	-25.6	0.0	-8.789	0.0
28	parataxis	ES >BP >CN	-13.471	0.0	-27.974	0.0	-10.763	0.0
29	flat	ES >BP >CN	-13.998	0.0	-30.112	0.0	-7.187	0.0
30	punct	ES >BP >CN	-25.014	0.0	-27.694	0.0	9.599	0.0

Table 6.5: Statistical results for attention activation (Layer 1 + Layer 2, head + tail) across dependency types among three languages, and GATs trained on pro-drop vs. non-pro-drop graphs. The results were independent t-tests with Bonferroni correction. The colors indicate the direction between the languages: If it is green, it is true that the column's direction is correct, otherwise, it is red; If there is no color, the statistical test was not significant (*i.e.* p-value > 0.05).

			pro-drop >non-pro-drop					
			CN		**BP**		**ES**	
ID	**Dependency Type**	**Order**	***t-value***	***p-value***	***t-value***	***p-value***	***t-value***	***p-value***
0	root	BP >CN, ES	-3.851	0.0	-1.914	1.0	17.767	0.0
1	nsubj	BP >CN	-5.718	0.0	-3.288	0.093	-3.471	0.093
2	obj	CN, BP >ES	-0.523	1.0	5.597	0.0	10.582	0.0
3	mark		-1.35	1.0	2.238	1.0	1.331	1.0
4	advmod	BP >ES	-0.938	1.0	0.063	1.0	6.188	0.0
5	aux		0.435	1.0	2.571	0.93	-0.148	1.0
6	case	BP >ES	1.455	1.0	0.17	1.0	3.95	0.0
7	det	CN >BP >ES	3.567	0.0	-1.014	1.0	5.876	0.0
8	xcomp	ES >BP	0.275	1.0	2.431	1.0	-5.496	0.0
9	nummod	ES >BP	-0.508	1.0	1.861	1.0	-12.862	0.0
10	nmod	ES >BP	2.652	0.744	1.773	1.0	-6.228	0.0
11	ccomp		-0.959	1.0	2.133	1.0	-3.186	0.093
12	nsubj_2	ES >BP	-0.884	1.0	0.195	1.0	-20.378	0.0
13	advcl	ES >BP	0.363	1.0	0.977	1.0	-4.934	0.0
14	obl		-1.433	1.0	2.613	0.837	2.568	0.93
15	acl	ES >BP	1.418	1.0	1.904	1.0	-9.803	0.0
16	advmod_2	ES >BP	1.793	1.0	2.457	1.0	-14.114	0.0
17	cop	ES >BP	1.307	1.0	-2.585	0.93	-5.63	0.0
18	compound	ES >BP	1.546	1.0	0.991	1.0	-15.805	0.0
19	amod	ES >BP	1.933	1.0	0.458	1.0	-4.791	0.0
20	obj_2	ES >BP	-0.517	1.0	1.477	1.0	-8.29	0.0
21	nmod_2	CN, ES >BP	4.266	0.0	0.968	1.0	-26.07	0.0
22	det_2	CN, ES >BP	4.165	0.0	0.432	1.0	-4.675	0.0
23	cc	CN >BP	4.355	0.0	-0.815	1.0	2.029	1.0
24	iobj	CN >BP	4.38	0.0	1.153	1.0	0.524	1.0
25	conj	ES >BP	2.78	0.558	2.424	1.0	-7.673	0.0
26	appos	CN, ES >BP	5.667	0.0	0.462	1.0	-33.476	0.0
27	csubj	CN, ES >BP	4.883	0.0	0.365	1.0	-17.918	0.0
28	parataxis	CN, ES >BP	4.64	0.0	0.981	1.0	-13.075	0.0
29	flat	CN, ES >BP	5.234	0.0	0.02	1.0	-15.269	0.0
30	punct	CN >BP >ES	5.392	0.0	-2.51	1.0	18.354	0.0

6.4 Discussion

In this experiment, the Graph Attention Networks (GATs) were used to explore what dependency types are recognized as contributive to story character resolution. Although the GATs used in this experiment were not very complex, as compared to other large language models (such as GPT-3), it showed its potential to contribute to linguistic theory development from a brand new angle. This section will discuss how the results can be interpreted from the following perspectives: (1) How to understand the "attention activation" values in the model; (2) Understanding the attention activation distribution across dependency types within each language, and among three languages; (3) Interpreting the differences between pro-drop and non-pro-drop cases trained models; (4) The choice of the language tree structure (*i.e.* dependency tree) in this experiment, and potential structural descriptions that could be used; (5) Potential usage of GNN and GAT on other linguistic topics.

6.4.1 Understanding Attention in The Graph Attention Networks

In machine learning models, attention is a popular method applied in the algorithms to encourage the model to obtain preference for certain features. Originally as a concept in cognitive psychology, attention in these models is a learnable parameter during model training, and it has been used in tasks such as natural language processing (W. Yin et al., 2016), and computer vision (Zheng et al., 2017). Some features are more important to realize a regression or classification goal, and the attention layers can strengthen the importance of these features for a better result.

GNNs become GATs when adding attention layers. In this experiment, two single-head attention layers were added after each fully connected graph convolution layer. As shown in Figure 6.8, in the attention layer, a pair of head and tail nodes from an edge is concatenated and convolved with the attention matrix, and the output of this step is a number which indicates the importance of this edge. Therefore, the attention values retrieved indicate the importance of the edges in the graph classification task. In our case, attention activation reveals the roles of the dependency connections.

The graph construction (see Section 6.2.2) unfolded an edge into [head, tail], and its symmetrical [tail, head]. The loss calculation (see Equation 6.12) refold these two nodes as a single averaged one. As a result,

the Layer1 + Layer2, head + tail results are closer to describing the dependency types' importance in the results, and we will focus on discussing these results (as shown in Figure 6.9).

6.4.2 Attention Activation: CN, BP, ES

As shown in Table 6.6, the dependency types are ranked based on the attention activation retrieved from Layer1 + Layer2, and head + tail. The cells with red numbers are the dependency types that take more than 5% among all types. We can see that the top six dependencies for each language take at least 50% of the whole group.

For all three languages, the following dependencies show commonly in the first six places: the main verb ("root"), the first subject ("nsubj"), and the first object ("obj"). These three dependency types leading the role are consistent with the core of pronoun resolution, since knowing "who is doing what to whom" is indeed essential to achieving subject story character resolution, and these elements form the S-V-O main stem for its clause. As all three languages are pro-drop languages, to understand the subject character, the main verbs have taken the leading role compared to the other dependency structures. This is consistent with the results in the first experiment (see Chapter 4), which discussed the verb usage continuity and its role in zero pronoun resolution.

For CN, marker ("mark") and auxiliary ("aux") take the 3rd and 5th places in the result. In the statistical results in Table 6.2, both marker and auxiliary show that they are taking more percentage in CN compared to BP and ES (*i.e.* CN $>$ BP, ES). In CN discourse, the marker nodes are words including '就'(jiu4, means "just"), '也'(ye3, means "just"), '又'(you4, means "again"), '都'(dou1, means "all"), '还'(hai2, means "still"). These marker words, together with auxiliary words (*e.g.* '了'(leo, means "already"), '着'(zheo, means "going on")) are able to hint time information in the clause. Since the verb normally does not provide tense change in CN, unlike the verb conjugation in other languages like ES and BP, the marker and auxiliary words in CN are essential to distinguish the subjects.

Determiner ("det") and case ("case"), on the other hand, take less significant roles in CN, but are ranked high in both BP and ES (*i.e.* BP, ES $>$ CN). Determiners in BP and ES (such as 'el/los') can provide gender information, and the ones in CN (except the third person pronouns, "ta") are not taking this role.

Punctuation ("punct") in BP and ES is taking a significantly higher role than in CN. This is partly due to parsing process differences. The clauses in CN did not include the original punctuation, but they included punctuation in BP and ES. The punctuation such as conversation-starting indicators (*e.g.* '-' in ES) can provide hints of the preference for who is more likely to get involved in a conversation in the discourse.

Table 6.6: Dependency type ranking for CN, BP ES based on attention activation average value on Layer1 + Layer2, head + tail. The cells with red text are the cases with $> 5\%$.

		Attention Activation Percentage (%)		
	Dependency Type	CN	BP	ES
0	root	17.630	17.069	14.403
1	nsubj	14.411	7.544	5.668
2	obj	7.877	5.288	7.607
3	mark	7.507	3.280	3.904
4	advmod	6.461	5.502	4.914
5	aux	6.323	0.924	2.162
6	case	4.815	7.697	6.028
7	det	3.048	9.419	8.555
8	xcomp	3.020	2.341	1.456
9	nummod	2.683	0.587	0.913
10	nmod	2.619	2.830	2.098
11	ccomp	2.609	0.755	2.136
12	nsubj_2	2.604	0.736	0.992
13	advcl	2.390	0.643	1.768
14	obl	2.322	3.603	3.990
15	acl	2.180	1.566	1.184
16	advmod_2	1.824	0.693	0.885
17	cop	1.236	3.611	1.774
18	compound	1.209	0.122	0.615
19	amod	1.184	2.125	2.069
20	obj_2	1.181	0.666	1.553
21	nmod_2	0.758	0.541	0.665
22	det_2	0.685	2.195	3.065
23	cc	0.481	3.493	3.368
24	iobj	0.474	0.389	2.083
25	conj	0.470	0.986	1.377
26	appos	0.449	0.155	0.549
27	csubj	0.427	0.380	0.907
28	parataxis	0.387	0.283	0.801
29	flat	0.377	0.328	0.529
30	punct	0.358	14.249	11.981
	Sum of cells >5 (%)	60.210	66.768	54.243

Table 6.7: Dependency type ranking for CN, BP ES based on attention activation average value on Layer1 + Layer2, head + tail. The dependency types are color-coded.

Dependency ranking	CN	BP	ES
1	root	root	root
2	nsubj	punct	punct
3	obj	det	det
4	mark	case	obj
5	advmod	nsubj	case
6	aux	advmod	nsubj
7	case	obj	advmod
8	det	cop	obl
9	xcomp	obl	mark
10	nummod	cc	cc
11	nmod	mark	det_2
12	ccomp	nmod	nmod
13	nsubj_2	xcomp	amod
14	advcl	det_2	aux
15	obl	amod	ccomp
16	acl	acl	advcl
17	advmod_2	conj	iobj
18	cop	aux	cop
19	compound	ccomp	obj_2
20	amod	nsubj_2	xcomp
21	obj_2	advmod_2	conj
22	nmod_2	obj_2	acl
23	det_2	advcl	nsubj_2
24	cc	nummod	nummod
25	iobj	nmod_2	csubj
26	conj	iobj	advmod_2
27	appos	csubj	parataxis
28	csubj	flat	nmod_2
29	parataxis	parataxis	compound
30	flat	appos	appos
31	punct	compound	flat

6.4.3 Attention Activation: *Pro*-drop vs. Non-*pro*-drop

As shown in Table 6.8, the first three columns show the attention activation percentage for all three languages when their GATs models are trained on *pro*-drop cases, and the fourth to sixth columns are the ones trained on non-*pro*-drop cases and the last three columns are the percentage difference (*i.e. pro*-drop - non-*pro*-drop) for each language.

Compared to the "all cases" results in Table 6.6, the pro-drop and non-pro-drop results have "thicker tails", which means that the higher ranked dependencies take less proportions and the percentages are more evenly distributed towards the lower ranked dependency types. This indicates that pro-drop and non-pro-drop cases trained models rely more on other features to classify subject characters.

Table 6.8: Dependency type ranking for CN, BP ES based on attention activation average value on Layer1 + Layer2, head + tail, when models are trained on pro-drop and non-pro-drop cases. The percentage difference columns are color highlighted when the results are significant ($p < 0.05$) as consistent with Table 6.5

		pro-drop Attention Activation Percentage (%)			non-pro-dop Attention Activation Percentage (%)			pro-drop - non-pro-drop Attention Activation Difference (%)		
ID	**Dependency Type**	**CN**	**BP**	**ES**	**CN**	**BP**	**ES**	**CN**	**BP**	**ES**
0	root	6.622	4.981	11.857	8.414	4.806	3.276	-1.792	0.175	8.580
1	nsubj	5.450	3.293	2.990	7.261	3.838	3.336	-1.812	-0.545	-0.346
2	obj	4.835	3.933	6.816	5.416	3.228	3.244	-0.582	0.706	3.572
3	mark	4.462	3.356	3.785	4.892	3.186	3.329	-0.430	0.170	0.456
4	advmod	4.344	3.741	5.128	4.631	3.749	3.241	-0.286	-0.008	1.887
5	aux	4.174	3.018	2.931	4.359	2.888	3.211	-0.186	0.131	-0.280
6	case	3.926	3.784	4.589	3.719	3.718	3.316	0.207	0.066	1.273
7	det	3.781	4.040	5.862	3.154	4.267	3.590	0.627	-0.228	2.272
8	xcomp	3.266	3.150	1.914	3.243	3.037	3.171	0.023	0.113	-1.256
9	nummod	3.042	2.939	1.521	3.156	2.873	3.170	-0.114	0.067	-1.650
10	nmod	3.416	3.233	2.244	2.953	3.090	3.236	0.463	0.143	-0.992
11	ccomp	2.912	2.907	2.499	3.049	2.857	3.134	-0.136	0.050	-0.635
12	nsubj_2	3.033	2.877	1.383	3.173	2.906	3.177	-0.140	-0.029	-1.795
13	advcl	2.987	2.883	2.284	2.903	2.845	3.166	0.084	0.038	-0.881
14	obl	2.796	3.429	3.927	3.013	3.149	3.262	-0.217	0.280	0.665
15	acl	3.115	3.144	1.869	2.914	3.005	3.178	0.201	0.139	-1.309
16	advmod_2	2.949	2.957	1.527	2.681	2.868	3.180	0.268	0.089	-1.652
17	cop	2.746	3.024	2.288	2.523	3.408	3.172	0.223	-0.383	-0.884
18	compound	2.675	2.765	1.350	2.404	2.791	3.171	0.271	-0.026	-1.822
19	amod	2.668	2.981	2.249	2.426	3.000	3.228	0.241	-0.019	-0.979
20	obj_2	2.535	2.868	2.009	2.566	2.855	3.198	-0.032	0.012	-1.189
21	nmod_2	2.574	2.836	1.303	2.209	2.836	3.165	0.364	-0.000	-1.862
22	det_2	2.515	3.069	2.815	2.166	3.103	3.319	0.349	-0.034	-0.504
23	cc	2.410	3.399	3.715	2.099	3.537	3.215	0.311	-0.138	0.500
24	iobj	2.446	2.812	2.949	2.120	2.807	3.168	0.326	0.005	-0.219
25	conj	2.378	3.066	2.146	2.120	2.906	3.192	0.258	0.159	-1.046
26	appos	2.451	2.765	1.211	2.100	2.794	3.172	0.350	-0.028	-1.961
27	csubj	2.403	2.785	1.536	2.094	2.817	3.168	0.308	-0.033	-1.632
28	parataxis	2.365	2.812	1.562	2.095	2.804	3.160	0.270	0.008	-1.598
29	flat	2.365	2.791	1.331	2.077	2.854	3.159	0.288	-0.063	-1.828
30	punct	2.365	4.361	10.408	2.071	5.178	3.295	0.294	-0.817	7.113

6.4.4 Language's Graph Representation

The graph structure used in this experiment was from dependency parsing results, and this type of structure provides semantic relationships within the clauses. In linguistics, more representation methods can be considered as graphs, such as constituency trees, morphological dependency trees, and Abstract Meaning Representation (AMR). These tree structures can be applied in future studies using GNNs to explore how they are contributing to language phenomenons.

6.4.5 GNN and Language Studies

This experiment takes a pioneer exploration of applying GNNs in linguistic studies. Using the model features, such as the attention layer in this case, we can learn how a linguistic phenomenon is processed in the model, and there are many new angles to be explored in the future.

6.5 Conclusion

In this experiment, Graph Attention Networks (GATs) are trained to classify subject story characters, and the training material was built with dependency structure graphs from CN, BP, and ES discourse material. First, the results in GATs are consistent with the main elements in pronoun resolution: the GATs found the elements including the subject, the main verb, and the object take leading roles among all other dependency types. Second, the elements shown importance in CN are different from the ones in BP and ES: marker and auxiliary are more important in CN; determiner and case are more important in BP and ES. This is consistent with the innate language features.

This experiment provided the rationale and proved the possibility of using machine learning models to study linguistic phenomena. It is possible to find out the consistency and difference between the models and linguistic theories and even neurolinguistic studies.

In conclusion, Experiment 3 stands as a groundbreaking initiative, illustrating the potential of Graph Neural Network (GNN) models to investigate linguistic phenomena from a language structural perspective. This pioneering step provides valuable insights into the alignment and divergence between computational models and established linguistic theories, fostering a more comprehensive understanding of language intricacies. The significance of this approach extends beyond the realm of *pro*-drop, laying a foun-

dation for future investigations that delve into the convergence of computational models and theoretical linguistics studies.

Expanding on this, the utilization of GNN models in Experiment 3 represents a paradigm shift in the study of language structure. By harnessing the power of graph-based representations, the experiment delves into the intricate dependencies and relationships within discourse material across languages. This not only deepens our understanding of *pro*-drop but also prompts broader questions about the applicability of graph-based models in deciphering complex linguistic phenomena. The observed consistency and divergence between GNN models and linguistic theories provide a rich terrain for further exploration, encouraging researchers to scrutinize the interplay between structural representations and linguistic principles.

In summary, Experiment 3 marks a transformative moment, showcasing the potential of GNN models to unravel language structural intricacies. Beyond its immediate implications for *pro*-drop, this approach sets the stage for broader inquiries into the convergence of computational models, theoretical linguistics, and neurolinguistic studies. The interplay between GNN models and linguistic theories opens up a dynamic space for exploration, offering new perspectives on the representation and processing of language structure.

Chapter 7

Conclusion

7.1 Summary

Research on *pro*-drop has a substantial history dating back to the 1980s, exploring various dimensions such as syntax, semantics, morphology, pragmatics, typology, and applications. Recent corpus studies have provided insights into the quantitative distributional features of *pro*-drop within discourse. However, there has been a limited focus on a direct statistical analysis comparing the linguistic factors that contribute to *pro*-drop.

This dissertation presents three computational experiments investigating the *pro*-drop mechanism in Chinese, Brazilian Portuguese, and Spanish, which are correspondingly radical, partial, and consistent *pro*-drop languages. The discourse materials utilized are translations (xiaowangzi.org, 2021) of Saint-Exupéry's *The Little Prince* in Chinese, Brazilian Portuguese, and Spanish. The employed statistical methods facilitate (1) a statistical examination of linguistic factors within each language and (2) a numerical representation of features that allows for cross-language comparisons, encompassing different types of *pro*-drop languages.

The dissertation focuses on three key questions: (1) How does discourse coherence support *pro*-drop languages, and how can we quantify discourse coherence? (2) What is the order of importance among linguistic factors relevant to *pro*-drop? (3) If an AI model is trained to predict *pro*-drop based on language dependency structures, what factors are crucial cross-linguistically? These questions are addressed in three experiments.

7.2 Experiment 1: Quantifying Verb Usage Continuity as Discourse Coherence Factor

Experiment 1 (Chapter 4) reveals a notable correlation among the three languages used as materials. The distribution analyses on *pro*-drop rate among Chinese, Brazilian Portuguese, and Spanish are carried out, and the results indicate that the *pro*-drop rate is linked to the level of agreement richness. Concurrently, a higher agreement level corresponds to a diminished reliance on semantic context, specifically the factor of verb usage continuity, for *pro*-drop in the language.

This experiment draws inspiration from the Topic Chain theory (Pu, 2019b; Pu & Pu, 2014), aiming to mathematically describe the existing "chain" throughout the discourse. Since *pro*-drop involves the omission of the subject, the "S-V-O" structure's "V" (verb) component becomes a promising element to support the chain. Hence, the verb continuity factor is designed to create a mathematical representation describing how this "chain" forms across verbs. Verb continuity is measured using the cosine similarity of verbs' word embeddings (BERT and GloVe). The results indicate that within each language, verb continuity effectively distinguishes *pro*-drop from non-*pro*-drop cases, with *pro*-drop cases demanding higher verb continuity. Cross-linguistically, the order of *pro*-drop cases' verb continuity levels is CN $>$ BP $>$ ES. This suggests that the agreement level plays a significant role, and the higher the agreement level of a language, the less it relies on additional semantic factors, such as verb continuity, to resolve a *pro*-drop case.

These findings underscore the nuanced relationship between agreement richness and the reliance on semantic factors in *pro*-drop languages. The experimental approach, grounded in the Topic Chain theory, sheds light on the intricate dynamics of verb continuity as a crucial element in supporting the coherence of *pro*-drop discourse. The observed cross-linguistic variation in the order of verb continuity levels emphasizes the influence of agreement richness, offering valuable insights into how different languages navigate the balance between agreement features and semantic factors in the resolution of *pro*-drop. Further research could explore the implications of these findings on our understanding of discourse structure and language-specific preferences in *pro*-drop phenomena.

In terms of contribution, Experiment 1 represents a pioneering effort to articulate a language theoretical concept using mathematical language. The initiation of this process stems from the intuition

that discourse coherence is inherent in *pro*-drop languages. By examining the element within a null-subject clause, the "verb chain" emerges as a plausible candidate to encapsulate this coherence. The subsequent mathematical description of the "verb chain" involves calculating verb usage continuity for **all** story character candidates. Their aptitude for being resolved as the dropped pronoun is quantified through a "salience level". Notably, this quantification relies on pre-trained large language models, generating high-dimensional vector values for the verbs.

In summary, Experiment 1's contribution lies in its innovative approach to translating a language theoretical concept into mathematical terms. The process demonstrates the viability of quantifying linguistic theories and introduces a methodological pathway for cross-linguistic research. The utilization of pre-trained language models and the emphasis on discourse coherence elements lay a foundation for further investigations into the intersection of linguistic theory and computational modeling, showcasing the potential for advanced methodologies to deepen our understanding of language phenomena.

7.3 Experiment 2: Statistical Modeling of Linguistic Factors on *Pro*-drop

Experiment 2 (Chapter 5) delves into a more comprehensive exploration of linguistic features influencing *pro*-drop, employing two machine learning models: Binomial Logistic Regression and Random Forest. The factors considered in this experiment span syntactic, semantic, morphological, and logical dimensions. Both models exhibit significant capability in differentiating cases between *pro*-drop and non-*pro*-drop throughout the discourse. The results highlight several key findings: (1) Character consistency between sub-main (or embedded-matrix) and current-previous clauses is a significant factor across all three languages; (2) Spanish and Brazilian Portuguese demonstrate richer verb agreement systems, contributing to potentially higher model fit levels compared to Chinese; (3) Sentential discourse relation features, such as coordinate structures, prove influential in encouraging *pro*-drop, and their significance is observed in both Chinese and Spanish.

The incorporation of a diverse array of linguistic features in Experiment 2 enhances our understanding of the multifaceted nature of *pro*-drop phenomena. The consistent significance of verb consistency across languages underscores its universal role in influencing *pro*-drop occurrences. Furthermore, the varied

verb agreement systems in Spanish and Brazilian Portuguese shed light on language-specific mechanisms aiding in the recovery of dropped pronouns. The implications of these findings extend beyond the theoretical framework, offering practical insights into the predictive capabilities of machine learning models in discerning *pro*-drop patterns.

The identified role of sentential discourse relation features, particularly coordinate structures, in encouraging *pro*-drop adds a layer of complexity to our understanding. This suggests that the discourse-level organization and relationships between sentences play a pivotal role in shaping *pro*-drop tendencies. This can be related to the Parallel Hypothesis, which explains the phenomenon where different forms of parallelism among conjuncts in coordinate structures aid the processor (Frazier et al., 2000; Frazier et al., 1984). Exploring these discourse-level features not only contributes to the refinement of predictive models but also deepens our appreciation for the interconnectedness of syntactic, semantic, and discourse-level factors in *pro*-drop phenomena.

In conclusion, Experiment 2 broadens the scope of analysis, revealing intricate linguistic features that contribute to *pro*-drop patterns. The integration of machine learning models with a diverse set of features offers a nuanced perspective, highlighting both universal and language-specific aspects of *pro*-drop phenomena. This comprehensive exploration lays the groundwork for further investigations into the interplay of linguistic factors and the potential applications of machine learning in understanding complex language phenomena.

7.4 Experiment 3: Involving Dependency Structures in Graph Neural Networks

In Experiment 3 (Chapter 6), Graph Attention Networks (GATs) were utilized to train models for classifying subject story characters. The training material incorporated dependency structure graphs from Chinese, Brazilian Portuguese, and Spanish discourse material. The results yielded valuable insights on three key aspects. Firstly, the outcomes align with the central elements in pronoun resolution, with the GATs emphasizing the importance of elements such as the subject, main verb, and object among all other dependency types. Secondly, notable differences emerged in the linguistic importance of elements between Chinese, Brazilian Portuguese, and Spanish. Specifically, marker and auxiliary elements hold greater impor-

tance in Chinese, while determiner and case elements take precedence in Brazilian Portuguese and Spanish. This alignment with innate language features showcases the language-specific nuances influencing dependency structures. Thirdly, the models trained on *pro*-drop and non-*pro*-drop scenarios exhibit a more evenly distributed dependency importance. This suggests a tendency for the models to seek additional resources to accomplish the task, hinting at the complexity of linguistic phenomena.

Experiment 3 not only provided a rationale but also demonstrated the feasibility of employing Graph Neural Network models to investigate linguistic phenomena. The ability to uncover consistencies and differences between these models and linguistic theories, as well as potential applications in neurolinguistic studies, underscores the versatility and potential of this approach. By leveraging machine learning, researchers can bridge the gap between theoretical linguistics and empirical observations, offering a data-driven perspective on the intricate interplay of linguistic elements. Furthermore, the language-specific variations in the importance of different elements emphasize the need for tailored approaches in studying *pro*-drop across diverse linguistic contexts.

This experiment opens avenues for further exploration, encouraging researchers to delve into the comparative analysis of machine learning models and traditional linguistic theories. Understanding how these models align with or diverge from established linguistic frameworks can enrich our comprehension of language processing mechanisms. Additionally, the observed tendency for models to seek extra resources prompts considerations of the adaptability and resourcefulness of machine learning in capturing the subtleties of linguistic phenomena.

In conclusion, Experiment 3 serves as a pioneering step in showcasing the potential of GNN models to study linguistic phenomena from the language structural perspective, offering insights into the consistency and divergence between models and linguistic theories. This approach not only enhances our understanding of *pro*-drop but also lays the groundwork for future investigations exploring the intersection of computational models, theoretical linguistics, and neurolinguistic studies.

7.5 Final Remarks

This dissertation takes pioneering strides in the exploration of linguistic factors by ingeniously creating original mathematical representations of linguistic theories and applying cutting-edge AI models. This

innovative approach not only contributes to the advancement of linguistic research but also signals a compelling experiment with implications for the burgeoning field of digital humanities. By forging a bridge between traditional linguistic theories and state-of-the-art computational methodologies, this research opens avenues for the reexamination of established theories, showcasing their relevance in a novel light.

In the swiftly evolving landscape of Artificial Intelligence, where advancements are rapidly reshaping the computational field, there arises a crucial need for discourse on how these technologies can be effectively applied to linguistics. The dissertation emphasizes the importance of such discussions, highlighting the potential synergy between AI and linguistics. As computational methods continue to revolutionize our understanding of language, it becomes imperative to explore how these technologies can not only aid in linguistic research but also foster a deeper comprehension of the humanistic aspects of language and communication.

Furthermore, this research aligns with the trajectory of digital humanity, a domain where the intersection of technology and humanities opens up unprecedented possibilities. The experiment of creating mathematical representations and employing AI models in linguistic exploration serves as a testament to the transformative potential of interdisciplinary collaboration. As theories are revisited and rejuvenated through the lens of computational tools, the digital humanities direction becomes a fertile ground for unveiling novel insights into language, cognition, and communication.

In essence, this dissertation sparks a paradigm shift in how linguistic factors are studied, embracing the symbiosis of mathematical representation and AI models. It not only contributes to the evolving landscape of linguistic research but also sets the stage for a more profound engagement with the digital humanities. The ongoing dialogue between Artificial Intelligence and linguistics holds the promise of not only advancing our understanding of language but also enriching the broader field of humanity studies.

Appendix A

mood	tense	pronoun	suffix	ending
indicative	Present	1SG	o	AR
indicative	Present	2SG	a	AR
indicative	Present	3SG	a	AR
indicative	Present	1PL	amos	AR
indicative	Present	2PL	am	AR
indicative	Present	3PL	am	AR
indicative	Present	1SG	o	ER
indicative	Present	2SG	e	ER
indicative	Present	3SG	e	ER
indicative	Present	1PL	emos	ER
indicative	Present	2PL	em	ER
indicative	Present	3PL	em	ER
indicative	Present	1SG	o	IR
indicative	Present	2SG	e	IR
indicative	Present	3SG	e	IR
indicative	Present	1PL	imos	IR
indicative	Present	2PL	em	IR

indicative	Present	3PL	em	IR
indicative	Present	1SG	nho	R
indicative	Present	2SG	e	R
indicative	Present	3SG	e	R
indicative	Present	1PL	mos	R
indicative	Present	2PL	em	R
indicative	Present	3PL	em	R
indicative	Imperfective Past	1SG	ava	AR
indicative	Imperfective Past	2SG	ava	AR
indicative	Imperfective Past	3SG	ava	AR
indicative	Imperfective Past	1PL	ávamos	AR
indicative	Imperfective Past	2PL	avam	AR
indicative	Imperfective Past	3PL	avam	AR
indicative	Imperfective Past	1SG	ia	ER
indicative	Imperfective Past	2SG	ia	ER
indicative	Imperfective Past	3SG	ia	ER
indicative	Imperfective Past	1PL	íamos	ER
indicative	Imperfective Past	2PL	iam	ER
indicative	Imperfective Past	3PL	iam	ER
indicative	Imperfective Past	1SG	ia	IR
indicative	Imperfective Past	2SG	ia	IR
indicative	Imperfective Past	3SG	ia	IR
indicative	Imperfective Past	1PL	íamos	IR
indicative	Imperfective Past	2PL	iam	IR
indicative	Imperfective Past	3PL	iam	IR
indicative	Imperfective Past	1SG	nha	R
indicative	Imperfective Past	2SG	nha	R
indicative	Imperfective Past	3SG	nha	R

indicative	Imperfective Past	1PL	nhamos	R
indicative	Imperfective Past	2PL	nham	R
indicative	Imperfective Past	3PL	nham	R
indicative	Perfective Past	1SG	ei	AR
indicative	Perfective Past	2SG	ou	AR
indicative	Perfective Past	3SG	ou	AR
indicative	Perfective Past	1PL	amos	AR
indicative	Perfective Past	2PL	aram	AR
indicative	Perfective Past	3PL	aram	AR
indicative	Perfective Past	1SG	i	ER
indicative	Perfective Past	2SG	eu	ER
indicative	Perfective Past	3SG	eu	ER
indicative	Perfective Past	1PL	emos	ER
indicative	Perfective Past	2PL	eram	ER
indicative	Perfective Past	3PL	eram	ER
indicative	Perfective Past	1SG	i	IR
indicative	Perfective Past	2SG	iu	IR
indicative	Perfective Past	3SG	iu	IR
indicative	Perfective Past	1PL	imos	IR
indicative	Perfective Past	2PL	iram	IR
indicative	Perfective Past	3PL	iram	IR
indicative	Perfective Past	1SG	s	R
indicative	Perfective Past	2SG	s	R
indicative	Perfective Past	3SG	s	R
indicative	Perfective Past	1PL	emos	R
indicative	Perfective Past	2PL	eram	R
indicative	Perfective Past	3PL	eram	R
indicative	Plus-Perfect Past	1SG	ara	AR

indicative	Plus-Perfect Past	2SG	ara	AR
indicative	Plus-Perfect Past	3SG	ara	AR
indicative	Plus-Perfect Past	1PL	áramos	AR
indicative	Plus-Perfect Past	2PL	aram	AR
indicative	Plus-Perfect Past	3PL	aram	AR
indicative	Plus-Perfect Past	1SG	era	ER
indicative	Plus-Perfect Past	2SG	era	ER
indicative	Plus-Perfect Past	3SG	era	ER
indicative	Plus-Perfect Past	1PL	eramos	ER
indicative	Plus-Perfect Past	2PL	eram	ER
indicative	Plus-Perfect Past	3PL	eram	ER
indicative	Plus-Perfect Past	1SG	ira	IR
indicative	Plus-Perfect Past	2SG	ira	IR
indicative	Plus-Perfect Past	3SG	ira	IR
indicative	Plus-Perfect Past	1PL	íramos	IR
indicative	Plus-Perfect Past	2PL	iram	IR
indicative	Plus-Perfect Past	3PL	iram	IR
indicative	Plus-Perfect Past	1SG	era	R
indicative	Plus-Perfect Past	2SG	era	R
indicative	Plus-Perfect Past	3SG	era	R
indicative	Plus-Perfect Past	1PL	eramos	R
indicative	Plus-Perfect Past	2PL	eram	R
indicative	Plus-Perfect Past	3PL	eram	R
indicative	Future	1SG	ei	AR
indicative	Future	2SG	á	AR
indicative	Future	3SG	á	AR
indicative	Future	1PL	emos	AR
indicative	Future	2PL	ão	AR

indicative	Future	3PL	ão	AR
indicative	Future	1SG	ei	ER
indicative	Future	2SG	á	ER
indicative	Future	3SG	á	ER
indicative	Future	1PL	emos	ER
indicative	Future	2PL	ão	ER
indicative	Future	3PL	ão	ER
indicative	Future	1SG	ei	IR
indicative	Future	2SG	á	IR
indicative	Future	3SG	á	IR
indicative	Future	1PL	emos	IR
indicative	Future	2PL	ão	IR
indicative	Future	3PL	ão	IR
indicative	Future	1SG	ei	R
indicative	Future	2SG	á	R
indicative	Future	3SG	á	R
indicative	Future	1PL	emos	R
indicative	Future	2PL	ão	R
indicative	Future	3PL	ão	R
indicative	Would-Conditional	1SG	ia	AR
indicative	Would-Conditional	2SG	ia	AR
indicative	Would-Conditional	3SG	ia	AR
indicative	Would-Conditional	1PL	íamos	AR
indicative	Would-Conditional	2PL	iam	AR
indicative	Would-Conditional	3PL	iam	AR
indicative	Would-Conditional	1SG	ia	ER
indicative	Would-Conditional	2SG	ia	ER
indicative	Would-Conditional	3SG	ia	ER

indicative	Would-Conditional	1PL	íamos	ER
indicative	Would-Conditional	2PL	iam	ER
indicative	Would-Conditional	3PL	iam	ER
indicative	Would-Conditional	1SG	ia	IR
indicative	Would-Conditional	2SG	ia	IR
indicative	Would-Conditional	3SG	ia	IR
indicative	Would-Conditional	1PL	íamos	IR
indicative	Would-Conditional	2PL	iam	IR
indicative	Would-Conditional	3PL	iam	IR
indicative	Would-Conditional	1SG	ia	R
indicative	Would-Conditional	2SG	ia	R
indicative	Would-Conditional	3SG	ia	R
indicative	Would-Conditional	1PL	íamos	R
indicative	Would-Conditional	2PL	iam	R
indicative	Would-Conditional	3PL	iam	R
subjunctive	Present	1SG	e	AR
subjunctive	Present	2SG	e	AR
subjunctive	Present	3SG	e	AR
subjunctive	Present	1PL	emos	AR
subjunctive	Present	2PL	em	AR
subjunctive	Present	3PL	em	AR
subjunctive	Present	1SG	a	ER
subjunctive	Present	2SG	a	ER
subjunctive	Present	3SG	a	ER
subjunctive	Present	1PL	amos	ER
subjunctive	Present	2PL	am	ER
subjunctive	Present	3PL	am	ER
subjunctive	Present	1SG	a	IR

subjunctive	Present	2SG	a	IR
subjunctive	Present	3SG	a	IR
subjunctive	Present	1PL	amos	IR
subjunctive	Present	2PL	am	IR
subjunctive	Present	3PL	am	IR
subjunctive	Present	1SG	a	R
subjunctive	Present	2SG	a	R
subjunctive	Present	3SG	a	R
subjunctive	Present	1PL	amos	R
subjunctive	Present	2PL	am	R
subjunctive	Present	3PL	am	R
subjunctive	Imperfective Past	1SG	asse	AR
subjunctive	Imperfective Past	2SG	asse	AR
subjunctive	Imperfective Past	3SG	asse	AR
subjunctive	Imperfective Past	1PL	ássemos	AR
subjunctive	Imperfective Past	2PL	assem	AR
subjunctive	Imperfective Past	3PL	assem	AR
subjunctive	Imperfective Past	1SG	esse	ER
subjunctive	Imperfective Past	2SG	esse	ER
subjunctive	Imperfective Past	3SG	esse	ER
subjunctive	Imperfective Past	1PL	êssemos	ER
subjunctive	Imperfective Past	2PL	essem	ER
subjunctive	Imperfective Past	3PL	essem	ER
subjunctive	Imperfective Past	1SG	isse	IR
subjunctive	Imperfective Past	2SG	isse	IR
subjunctive	Imperfective Past	3SG	isse	IR
subjunctive	Imperfective Past	1PL	íssemos	IR
subjunctive	Imperfective Past	2PL	issem	IR

subjunctive	Imperfective Past	3PL	issem	IR
subjunctive	Imperfective Past	1SG	esse	R
subjunctive	Imperfective Past	2SG	esse	R
subjunctive	Imperfective Past	3SG	esse	R
subjunctive	Imperfective Past	1PL	éssemos	R
subjunctive	Imperfective Past	2PL	essem	R
subjunctive	Imperfective Past	3PL	essem	R
subjunctive	Future	1SG	-	AR
subjunctive	Future	2SG	-	AR
subjunctive	Future	3SG	-	AR
subjunctive	Future	1PL	mos	AR
subjunctive	Future	2PL	em	AR
subjunctive	Future	3PL	em	AR
subjunctive	Future	1SG	-	ER
subjunctive	Future	2SG	-	ER
subjunctive	Future	3SG	-	ER
subjunctive	Future	1PL	mos	ER
subjunctive	Future	2PL	em	ER
subjunctive	Future	3PL	em	ER
subjunctive	Future	1SG	-	IR
subjunctive	Future	2SG	-	IR
subjunctive	Future	3SG	-	IR
subjunctive	Future	1PL	mos	IR
subjunctive	Future	2PL	em	IR
subjunctive	Future	3PL	em	IR
subjunctive	Future	1SG	puser	R
subjunctive	Future	2SG	puseres	R
subjunctive	Future	3SG	puser	R

subjunctive	Future	1PL	mos	R
subjunctive	Future	2PL	em	R
subjunctive	Future	3PL	em	R
imperative		1SG	-	AR
imperative		2SG	e	AR
imperative		1PL	emos	AR
imperative		2PL	em	AR
imperative		1SG	-	ER
imperative		2SG	a	ER
imperative		1PL	amos	ER
imperative		2PL	am	ER
imperative		1SG	-	IR
imperative		2SG	a	IR
imperative		1PL	amos	IR
imperative		2PL	am	IR
imperative		1SG	-	R
imperative		2SG	nha	R
imperative		1PL	nhamos	R
imperative		2PL	nham	R

Table A.1: Verb conjuration rules for Brazilian Portuguese.

mood	tense	pronoun	suffix	ending
indicative	present	1SG	o	AR
indicative	present	2SG	as	AR
indicative	present	3SG	a	AR
indicative	present	1PL	amos	AR
indicative	present	2PL	áis	AR

indicative	present	3PL	an	AR
indicative	present	1SG	o	ER
indicative	present	2SG	es	ER
indicative	present	3SG	e	ER
indicative	present	1PL	emos	ER
indicative	present	2PL	éis	ER
indicative	present	3PL	en	ER
indicative	present	1SG	o	IR
indicative	present	2SG	es	IR
indicative	present	3SG	e	IR
indicative	present	1PL	imos	IR
indicative	present	2PL	ís	IR
indicative	present	3PL	en	IR
indicative	Imperfect	1SG	aba	AR
indicative	Imperfect	2SG	abas	AR
indicative	Imperfect	3SG	aba	AR
indicative	Imperfect	1PL	ábamos	AR
indicative	Imperfect	2PL	abais	AR
indicative	Imperfect	3PL	aban	AR
indicative	Imperfect	1SG	ía	ER
indicative	Imperfect	2SG	ías	ER
indicative	Imperfect	3SG	ía	ER
indicative	Imperfect	1PL	íamos	ER
indicative	Imperfect	2PL	íais	ER
indicative	Imperfect	3PL	ían	ER
indicative	Imperfect	1SG	ía	IR
indicative	Imperfect	2SG	ías	IR
indicative	Imperfect	3SG	ía	IR

indicative	Imperfect	1PL	íamos	IR
indicative	Imperfect	2PL	íais	IR
indicative	Imperfect	3PL	ían	IR
indicative	past	1SG	é	AR
indicative	past	2SG	aste	AR
indicative	past	3SG	ó	AR
indicative	past	1PL	amos	AR
indicative	past	2PL	asteis	AR
indicative	past	3PL	aron	AR
indicative	past	1SG	í	ER
indicative	past	2SG	iste	ER
indicative	past	3SG	ió	ER
indicative	past	1PL	imos	ER
indicative	past	2PL	isteis	ER
indicative	past	3PL	ieron	ER
indicative	past	1SG	í	IR
indicative	past	2SG	iste	IR
indicative	past	3SG	ió	IR
indicative	past	1PL	imos	IR
indicative	past	2PL	isteis	IR
indicative	past	3PL	ieron	IR
indicative	future	1SG	é	AR
indicative	future	2SG	ás	AR
indicative	future	3SG	á	AR
indicative	future	1PL	emos	AR
indicative	future	2PL	éis	AR
indicative	future	3PL	án	AR
indicative	future	1SG	é	ER

indicative	future	2SG	ás	ER
indicative	future	3SG	á	ER
indicative	future	1PL	emos	ER
indicative	future	2PL	éis	ER
indicative	future	3PL	án	ER
indicative	future	1SG	é	IR
indicative	future	2SG	ás	IR
indicative	future	3SG	á	IR
indicative	future	1PL	emos	IR
indicative	future	2PL	éis	IR
indicative	future	3PL	án	IR
indicative	conditional	1SG	ía	AR
indicative	conditional	2SG	ías	AR
indicative	conditional	3SG	ía	AR
indicative	conditional	1PL	íamos	AR
indicative	conditional	2PL	íais	AR
indicative	conditional	3PL	ían	AR
indicative	conditional	1SG	ía	ER
indicative	conditional	2SG	ías	ER
indicative	conditional	3SG	ía	ER
indicative	conditional	1PL	íamos	ER
indicative	conditional	2PL	íais	ER
indicative	conditional	3PL	ían	ER
indicative	conditional	1SG	ía	IR
indicative	conditional	2SG	ías	IR
indicative	conditional	3SG	ía	IR
indicative	conditional	1PL	íamos	IR
indicative	conditional	2PL	íais	IR

indicative	conditional	3PL	ían	IR
subjunctive	present	1SG	e	AR
subjunctive	present	2SG	es	AR
subjunctive	present	3SG	e	AR
subjunctive	present	1PL	emos	AR
subjunctive	present	2PL	éis	AR
subjunctive	present	3PL	en	AR
subjunctive	present	1SG	a	ER
subjunctive	present	2SG	as	ER
subjunctive	present	3SG	a	ER
subjunctive	present	1PL	amos	ER
subjunctive	present	2PL	áis	ER
subjunctive	present	3PL	an	ER
subjunctive	present	1SG	a	IR
subjunctive	present	2SG	as	IR
subjunctive	present	3SG	a	IR
subjunctive	present	1PL	amos	IR
subjunctive	present	2PL	áis	IR
subjunctive	present	3PL	an	IR
subjunctive	imperfect	1SG	ra	AR
subjunctive	imperfect	2SG	ras	AR
subjunctive	imperfect	3SG	ra	AR
subjunctive	imperfect	1PL	´ramos	AR
subjunctive	imperfect	2PL	rais	AR
subjunctive	imperfect	3PL	ran	AR
subjunctive	imperfect	1SG	se	ER
subjunctive	imperfect	2SG	ses	ER
subjunctive	imperfect	3SG	se	ER

subjunctive	imperfect	1PL	´semos	ER
subjunctive	imperfect	2PL	seis	ER
subjunctive	imperfect	3PL	sen	ER
subjunctive	imperfect	1SG	ra	IR
subjunctive	imperfect	2SG	ras	IR
subjunctive	imperfect	3SG	ra	IR
subjunctive	imperfect	1PL	´ramos	IR
subjunctive	imperfect	2PL	rais	IR
subjunctive	imperfect	3PL	ran	IR
subjunctive	imperfect	1SG	se	AR
subjunctive	imperfect	2SG	ses	AR
subjunctive	imperfect	3SG	se	AR
subjunctive	imperfect	1PL	´semos	AR
subjunctive	imperfect	2PL	seis	AR
subjunctive	imperfect	3PL	sen	AR
subjunctive	imperfect	1SG	ra	ER
subjunctive	imperfect	2SG	ras	ER
subjunctive	imperfect	3SG	ra	ER
subjunctive	imperfect	1PL	´ramos	ER
subjunctive	imperfect	2PL	rais	ER
subjunctive	imperfect	3PL	ran	ER
subjunctive	imperfect	1SG	se	IR
subjunctive	imperfect	2SG	ses	IR
subjunctive	imperfect	3SG	se	IR
subjunctive	imperfect	1PL	´semos	IR
subjunctive	imperfect	2PL	seis	IR
subjunctive	imperfect	3PL	sen	IR
subjunctive	future	1SG	re	AR

subjunctive	future	2SG	res	AR
subjunctive	future	3SG	re	AR
subjunctive	future	1PL	´remos	AR
subjunctive	future	2PL	reis	AR
subjunctive	future	3PL	ren	AR
subjunctive	future	1SG	re	ER
subjunctive	future	2SG	res	ER
subjunctive	future	3SG	re	ER
subjunctive	future	1PL	´remos	ER
subjunctive	future	2PL	reis	ER
subjunctive	future	3PL	ren	ER
subjunctive	future	1SG	re	IR
subjunctive	future	2SG	res	IR
subjunctive	future	3SG	re	IR
subjunctive	future	1PL	´remos	IR
subjunctive	future	2PL	reis	IR
subjunctive	future	3PL	ren	IR
imperative	affirmative	1SG	-	AR
imperative	affirmative	2SG	a	AR
imperative	affirmative	3SG	e	AR
imperative	affirmative	1PL	emos	AR
imperative	affirmative	2PL	rd	AR
imperative	affirmative	3PL	en	AR
imperative	affirmative	1SG	-	ER
imperative	affirmative	2SG	e	ER
imperative	affirmative	3SG	a	ER
imperative	affirmative	1PL	amos	ER
imperative	affirmative	2PL	rd	ER

imperative	affirmative	3PL	an	ER
imperative	affirmative	1SG	-	IR
imperative	affirmative	2SG	a	IR
imperative	affirmative	3SG	a	IR
imperative	affirmative	1PL	amos	IR
imperative	affirmative	2PL	rd	IR
imperative	affirmative	3PL	an	IR
imperative	negative	1SG	-	AR
imperative	negative	2SG	es	AR
imperative	negative	3SG	e	AR
imperative	negative	1PL	emos	AR
imperative	negative	2PL	éis	AR
imperative	negative	3PL	en	AR
imperative	negative	1SG	-	ER
imperative	negative	2SG	as	ER
imperative	negative	3SG	a	ER
imperative	negative	1PL	amos	ER
imperative	negative	2PL	áis	ER
imperative	negative	3PL	an	ER
imperative	negative	1SG	-	IR
imperative	negative	2SG	as	IR
imperative	negative	3SG	a	IR
imperative	negative	1PL	amos	IR
imperative	negative	2PL	áis	IR
imperative	negative	3PL	an	IR

Table A.2: Verb conjuration rules for Spanish.

Bibliography

Almor, A. (1999). Noun-phrase anaphora and focus: The informational load hypothesis. *Psychological review*, *106*(4), 748.

Almor, A., Nair, V. A., Boiteau, T. W., & Vendemia, J. M. (2017). The n400 in processing repeated name and pronoun anaphors in sentences and discourse. *Brain and language*, *173*, 52–66.

Almor, A., Smith, D. V., Bonilha, L., Fridriksson, J., & Rorden, C. (2007). What is in a name? Spatial brain circuits are used to track discourse references. *Neuroreport*, *18*, 1215–1219.

Anderson, M., & Gómez-Rodrıguez, C. (2020). Distilling neural networks for greener and faster dependency parsing. *Proceedings of the 16th International Conference on Parsing Technologies and the IWPT 2020 Shared Task on Parsing into Enhanced Universal Dependencies*, 2–13.

Ariel, M. (1990). *Accessing noun-phrase antecedents*. Routledge.

Ariel, M. (2001). Accessibility theory: An overview. *Text representation: Linguistic and psycholinguistic aspects*, *8*, 29–87.

Aroonmanakun, W. (2000). Zero pronoun resolution in thai: A centering approach. *Burnham, Denis, et. al. Interdisciplinary Approaches to Language Processing: The International Conference on Human and Machine Processing on Human and Machine Processing of Language and Speech. NECTEC: Bangkok*, 127–147.

Asif, N. A., Sarker, Y., Chakrabortty, R. K., Ryan, M. J., Ahamed, M. H., Saha, D. K., Badal, F. R., Das, S. K., Ali, M. F., Moyeen, S. I., et al. (2021). Graph neural network: A comprehensive review on non-euclidean space. *IEEE Access*, *9*, 60588–60606.

Augustyniak, Ł., Kajdanowicz, T., & Kazienko, P. (2018). Extracting aspects hierarchies using rhetorical structure theory. *Proceedings of the 2018 International Conference on Algorithms, Computing and Artificial Intelligence*, 1–5.

Ayres, M. R., & de Ávila Othero, G. (2021). Contexts for null subjects in contemporary Brazilian Portuguese. *Revista Linguıtica*, *17*(3), 100–124.

Barbosa, M. d. P. P. (1995). *Null subjects* (Doctoral dissertation). Massachusetts Institute of Technology.

Barbosa, P. (2011a). Partial pro-drop as null NP anaphora. *Presented at NELS 41, UPenn, 2010 and Romania Nova, Campos do Jordão*. https://repositorium.sdum.uminho.pt/bitstream/1822/16200/1/Barbosa.pdf

Barbosa, P. (2011b). Pro-drop and theories of pro in the minimalist program part 2: Pronoun deletion analyses of null subjects and partial, discourse and semi pro-drop. *Language and Linguistics Compass*, *5*(8), 571–587.

Barbosa, P. P. (2011). Pro-drop and theories of pro in the minimalist program part 1: Consistent null subject languages and the pronominal-agr hypothesis. *Language and Linguistics Compass*, *5*(8), 551–570.

Barbosa, P. P. (2019). Pro as a minimal nP: Toward a unified approach to pro-drop. *Linguistic Inquiry*, *50*(3), 487–526.

Bejček, E., Hajičová, E., Hajič, J., Jınová, P., Kettnerová, V., Kolářová, V., Mikulová, M., Mırovskỳ, J., Nedoluzhko, A., Panevová, J., et al. (2013). Prague dependency treebank 3.0.

Bengio, Y., Ducharme, R., & Vincent, P. (2000). A neural probabilistic language model. *Advances in neural information processing systems*, *13*.

Bianchi, V., & Frascarelli, M. (2010). Is topic a root phenomenon?

Biberauer, T., Holmberg, A., Roberts, I., & Sheehan, M. (2009). *Parametric variation: Null subjects in minimalist theory*. Cambridge University Press.

Biezma, M. (2014). Multiple focus strategies in pro-drop languages: Evidence from ellipsis in Spanish. *Syntax*, *17*(2), 91–131.

Boehmke, B., & Greenwell, B. M. (2019). *Hands-on machine learning with R*. CRC press.

Bohnet, B. (2010). Top accuracy and fast dependency parsing is not a contradiction. *Proceedings of the 23rd international conference on computational linguistics (coling 2010)*, 89–97.

Boiteau, T. W., Bowers, E., Nair, V. A., & Almor, A. (2014). The neural representation of plural discourse entities. *Brain and Language*, *137*, 130–141.

Breiman, L. (2001). Random forests. *Machine learning*, *45*, 5–32.

Brezina, V. (2018). *Statistics in corpus linguistics: A practical guide*. Cambridge University Press.

Brodbeck, C., Gwilliams, L., & Pylkkänen, L. (2016). Language in context: MEG evidence for modality-general and -specific responses to reference resolution. *eNeuro*, *3*, e0145-16.2016 1–16.

Brodbeck, C., & Pylkkänen, L. (2017). Language in context: Characterizing the comprehension of referential expressions with MEG. *NeuroImage*, *147*, 447–460.

Brown, G., Brown, G. D., Brown, G. R., Yule, G., & Gillian, B. (1983). *Discourse analysis*. Cambridge university press.

Cai, Q., & Brysbaert, M. (2010). Subtlex-ch: Chinese word and character frequencies based on film subtitles. *Plos ONE*, *5*, e10729.

Camacho, J. A. (2013). *Null subjects.* Cambridge University Press. https://search.ebscohost.com/login.aspx?direct=true&AuthType=ip,shib&db=e000xna&AN=545060&site=eds-live&custid=uga1

Carlson, K. (2001). The effects of parallelism and prosody in the processing of gapping structures. *Language and Speech*, *44*(1), 1–26.

Carlson, K. (2013). *Parallelism and prosody in the processing of ellipsis sentences*. Routledge.

Chai, H., & Strube, M. (2022). Incorporating centering theory into neural coreference resolution. *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2996–3002.

Charniak, E., et al. (2016). Parsing as language modeling. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2331–2336.

Chen, D., & Manning, C. D. (2014). A fast and accurate dependency parser using neural networks. *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 740–750.

Chen, K., Wang, R., Utiyama, M., Sumita, E., & Zhao, T. (2018). Syntax-directed attention for neural machine translation. *Proceedings of the AAAI conference on artificial intelligence*, *32*(1).

Chen, P., Ding, H., Araki, J., & Huang, R. (2021). Explicitly capturing relations between entity mentions via graph neural networks for domain-specific named entity recognition. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*.

Chen, P. (1986). *A discourse analysis of third person zero anaphora in Chinese* (Vol. 281). Indiana University Linguistics Club.

Chen, S., Gu, B., Qu, J., Li, Z., Liu, A., Zhao, L., & Chen, Z. (2021). Tackling zero pronoun resolution and non-zero coreference resolution jointly. *Proceedings of the 25th Conference on Computational Natural Language Learning*, 518–527.

Cheng, S., Fang, K., & Guodong, Z. (2017). Towards better Chinese zero pronoun resolution from discourse perspective. *National CCF Conference on Natural Language Processing and Chinese Computing*, 406–418.

Choi, J. D., & McCallum, A. (2013). Transition-based dependency parsing with selectional branching. *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1052–1062.

Choi, J. D., Tetreault, J., & Stent, A. (2015). It depends: Dependency parser comparison using a web-based evaluation tool. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 387–396.

Chomsky, N. (1981). *Lectures on government and binding*. Foris.

Church, K., & Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational linguistics*, *16*(1), 22–29.

Clifton, L. F. C. (1998). Comprehension of sluiced sentences. *Language and Cognitive Processes*, *13*(4), 499–520.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., & Kuksa, P. (2011). Natural language processing (almost) from scratch. *Journal of machine learning research*, *12*(ARTICLE), 2493–2537.

Coopmans, C. W., & Nieuwland, M. S. (2020). Dissociating activation and integration of discourse referents: Evidence from ERPs and oscillations. *cortex*, *126*, 83–106.

Cyrino, S. M. L., Duarte, M. E. L., & Kato, M. A. (2000). Visible subjects and invisible clitics in Brazilian Portuguese. *Visible subjects and invisible clitics in Brazilian Portuguese*, 55–73.

De Vincenzi, M. (1991a). Filler-gap dependencies in a null subject language: Referential and nonreferential whs. *Journal of Psycholinguistic Research*, *20*, 197–213.

De Vincenzi, M. (1991b). *Syntactic parsing strategies in italian: The minimal chain principle* (Vol. 12). Springer Science & Business Media.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American society for information science*, *41*(6), 391–407.

Dozat, T., & Manning, C. D. (2016). Deep biaffine attention for neural dependency parsing. *International Conference on Learning Representations*.

Duarte, I., & Silva, M. C. F. (2016). The null subject parameter and the structure of the sentence in european and Brazilian Portuguese. *The handbook of Portuguese linguistics*, 234–253.

Duarte, M. E. L., & KATO, M. A. (2017). Do pronome nulo ao pronome pleno: A trajetória do sujeito nulo no português do brasil. *Português brasileiro: uma viagem diacrônica*.

Duarte, M. E. L. (1996). A perda do princıpio" evite pronome" no português brasileiro. *Sınteses-ISSN 1981-1314*, *1*.

Duarte, M. E. L., & Marins, J. E. (2021). Brazilian Portuguese: A 'partial' null subject language? *Cadernos de Estudos Linguısticos*, *63*, e021021–e021021.

Duguine, M. (2014). Argument ellipsis: A unitary approach to pro-drop. *The Linguistic Review*, *31*(3-4), 515–549.

Dyer, C., Kuncoro, A., Ballesteros, M., & Smith, N. A. (2016). Recurrent neural network grammars. In K. Knight, A. Nenkova, & O. Rambow (Eds.), *Proceedings of the 2016 conference of the north American chapter of the association for computational linguistics: Human language technologies* (pp. 199–209). Association for Computational Linguistics. https://doi.org/10.18653/v1/N16-1024

El Ouardi, L., Yeou, M., & Faroqi-Shah, Y. (2023). Neural correlates of pronoun processing: An activation likelihood estimation meta-analysis. *Brain and Language*, *246*, 105347.

Ethayarajh, K. (2019). How contextual are contextualized word representations? comparing the geometry of bert, elmo, and gpt-2 embeddings. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 55–65.

Farouk, M. (2019). Measuring sentences similarity: A survey. *Indian Journal of Science and Technology*, *12*, 25.

Fellbaum, C. (1998). *WordNet: An electronic lexical database*. MIT press.

Fernández-González, D., & Gómez-Rodrıguez, C. (2019). Left-to-right dependency parsing with pointer networks. *2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 710–716.

Féry, C., & Ishihara, S. (2016). *The Oxford handbook of information structure*. Oxford University Press.

Fossum, V., & Knight, K. (2009). Combining constituent parsers. *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, 253–256.

Frascarelli, M., & Casentini, M. (2019). The interpretation of null subjects in a radical pro-drop language: Topic chains and discourse-semantic requirements in Chinese. *Studies in Chinese Linguistics*, *40*(1), 1–46.

Frascarelli, M., & Jiménez-Fernández, Á. L. (2019). Understanding partiality in pro-drop languages: An information-structure approach. *Syntax*, *22*(2-3), 162–198.

Frasconi, P., Gori, M., & Sperduti, A. (1998). A general framework for adaptive processing of data structures. *IEEE transactions on Neural Networks*, *9*(5), 768–786.

Frazier, L., Munn, A., & Clifton, C. (2000). Processing coordinate structures. *Journal of Psycholinguistic Research*, *29*, 343–370.

Frazier, L., Taft, L., Roeper, T., Clifton, C., & Ehrlich, K. (1984). Parallel structure: A source of facilitation in sentence comprehension. *Memory & cognition*, *12*, 421–430.

Fried, D., Stern, M., & Klein, D. (2017). Improving neural parsing by disentangling model combination and reranking effects. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 161–166.

Fries, P. H. (1983). On the status of theme in English: Arguments from discourse. *Micro and macro connexity of texts*, *45*.

Geeslin, K. (1999). A typological investigation of the pro-drop parameter. *Proceedings of Electronic Conference: Language Typology*.

Gelormini-Lezama, C., & Almor, A. (2011). Repeated names, overt pronouns, and null pronouns in Spanish. *Language and cognitive processes*, *26*(3), 437–454.

Goldberg, Y., & Nivre, J. (2013). Training deterministic parsers with non-deterministic oracles. *Transactions of the association for Computational Linguistics*, *1*, 403–414.

Gori, M., Monfardini, G., & Scarselli, F. (2005). A new model for learning in graph domains. *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.*, *2*, 729–734.

Grefenstette, G. (2012). *Explorations in automatic thesaurus discovery* (Vol. 278). Springer Science & Business Media.

Grice, P. (n.d.). Logic and conversation. In *syntax and semantics: Speech acts*.

Grosz, B. J., Joshi, A. K., & Weinstein, S. (1995). Centering: A framework for modeling the local coherence of discourse.

Guo, D., Gupta, A., Agarwal, S., Kao, J.-Y., Gao, S., Biswas, A., Lin, C.-W., Chung, T., & Bansal, M. (2022). GRAVL-BERT: Graphical visual-linguistic representations for multimodal coreference resolution. *Proceedings of the 29th International Conference on Computational Linguistics*, 285–297.

Hajič, J., Bejček, E., Hlaváčová, J., Mikulová, M., Straka, M., Štěpánek, J., & Štěpánková, B. (2020). Prague dependency treebank–consolidated 1.0. *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 5208–5218.

Hall, K., & Yoshida, M. (2021). Coreference and parallelism. *Language, Cognition and Neuroscience*, *36*(3), 296–319.

Halliday, M. A. K., & Hasan, R. (1976). Cohesion in English. *English Language Series, Longman, London*.

Hamilton, W., Ying, Z., & Leskovec, J. (2017). Inductive representation learning on large graphs. *Advances in neural information processing systems*, *30*.

Hammer, A., Goebel, R., Schwarzbach, J., Münte, T. F., & Jansma, B. M. (2007). When sex meets syntactic gender on a neural basis during pronoun processing. *Brain Research*, *1146*, 185–198.

Hammer, A., Jansma, B., Tempelmann, C., & Münte, T. F. (2011). Neural mechanisms of anaphoric reference revealed by fMRI. *Frontiers in Psychology*, *2*, 32.

Hammer, A., Jansma, B. M., Lamers, M., & Münte, T. F. (2008). Interplay of meaning, syntax and working memory during pronoun resolution investigated by ERPs. *Brain research*, *1230*, 177–191.

Hammer, A., Jansma, B. M., Tempelmann, C., & Münte, T. F. (2011). Neural mechanisms of anaphoric reference revealed by fMRI. *Frontiers in Psychology*, *2*, 1–9.

Heinat, F., & Klingvall, E. (2020). Set focus and anaphoric reference: An ERP study. *Brain and Language*, *206*, 104808.

Herbeck, P. (2021). Perspectival factors and pro-drop: A corpus study of speaker/addressee pronouns with creer 'think/believe' and saber 'know' in spoken Spanish. *Glossa: a journal of general linguistics*, *6*(1).

Holmberg, A. (2005). Is there a little pro? evidence from Finnish. *Linguistic inquiry*, *36*(4), 533–564.

Honnibal, M., Goldberg, Y., & Johnson, M. (2013). A non-monotonic arc-eager transition system for dependency parsing. *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, 163–172.

Honnibal, M., & Johnson, M. (2015). An improved non-monotonic transition system for dependency parsing. In L. Màrquez, C. Callison-Burch, & J. Su (Eds.), *Proceedings of the 2015 conference on empirical methods in natural language processing* (pp. 1373–1378). Association for Computational Linguistics. https://doi.org/10.18653/v1/D15-1162

Huang, C.-T. J. (1984). On the distribution and reference of empty pronouns. *Linguistic inquiry*, 531–574.

Huang, C.-T. J. (1989). Pro-drop in Chinese: A generalized control theory. In O. Jaeggli & K. Safir (Eds.), *The null subject parameter* (pp. 185–214). Springer.

Huang, L., Fayong, S., & Guo, Y. (2012). Structured perceptron with inexact search. *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 142–151.

Huang, Y. (1994). *The syntax and pragmatics of anaphora: A study with special reference to Chinese*. Cambridge University Press.

Huang, Z. (1997). Hnc 理论概要 (hierachical network of concepts). 中文信息学报 *(Journal of Chinese Information Processing)*, *11*(4), 12–21.

Iacobacci, I., Pilehvar, M. T., & Navigli, R. (2015). Sensembed: Learning sense embeddings for word and relational similarity. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 95–105.

Iida, R., Inui, K., & Matsumoto, Y. (2006). Exploiting syntactic patterns as clues in zero-anaphora resolution. *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, 625–632.

Iida, R., Inui, K., & Matsumoto, Y. (2007). Zero-anaphora resolution by learning rich syntactic pattern features. *ACM Transactions on Asian Language Information Processing (TALIP)*, *6*(4), 1–22.

Iida, R., Torisawa, K., Hashimoto, C., Oh, J.-H., & Kloetzer, J. (2015). Intra-sentential zero anaphora resolution using subject sharing recognition. *EMNLP*, 2179–2189.

Jahan, L., Chauhan, G., & Finlayson, M. A. (2018). A new approach to animacy detection. *Proceedings of the 27th International Conference on Computational Linguistics*.

Jeon, S., & Strube, M. (2020). Centering-based neural coherence modeling with hierarchical discourse segments. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 7458–7472.

Jiang, F., & Cohn, T. (2022). Incorporating constituent syntax for coreference resolution. *Proceedings of the AAAI Conference on Artificial Intelligence*, *36*(10), 10831–10839.

Joshi, A. K., & Weinstein, S. (1981). Control of inference: Role of some aspects of discourse structure-centering. *IJCAI*, 385–387.

Joulin, A., Grave, E., & Mikolov, P. B. T. (2017). Bag of tricks for efficient text classification. *EACL 2017*, 427.

Kameyama, M., Passonneau, R. J., & Poesio, M. (1993). Temporal centering. *31st Annual Meeting of the Association for Computational Linguistics*, 70–77.

Kamp, H., & Reyle, U. (2013). *From discourse to logic: Introduction to modeltheoretic semantics of natural language, formal logic and discourse representation theory* (Vol. 42). Springer Science & Business Media.

Kato, M. A. (2000). The partial pro-drop nature and the restricted vs order in Brazilian Portuguese. *The partial pro-drop nature and the restricted VS order in Brazilian Portuguese*, 223–258.

Kato, M. A., & Duarte, E. (2018). Pre-verbal position in bp: A reinterpretation of the" avoid pronoun principle". *Revista Diadorim*, *20*, 610–626.

Kawahara, D., & Kurohashi, S. (2004). Zero pronoun resolution based on automatically constructed case frames and structural preference of antecedents. *International Conference on Natural Language Processing*, 12–21.

Kenton, J. D. M.-W. C., & Toutanova, L. K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of NAACL-HLT*, 4171–4186.

Kim, Y., Ra, D., & Lim, S. (2021). Zero-anaphora resolution in Korean based on deep language representation model: BERT. *ETRI Journal*, *43*(2), 299–312.

Kiperwasser, E., & Goldberg, Y. (2016). Simple and accurate dependency parsing using bidirectional LSTM feature representations. *Transactions of the Association for Computational Linguistics*, *4*, 313–327.

Kipf, T. N., & Welling, M. (2016). Semi-supervised classification with graph convolutional networks. *International Conference on Learning Representations*.

Kitaev, N., & Klein, D. (2018). Constituency parsing with a self-attentive encoder. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2676–2686.

Koeneman, O., & Zeijlstra, H. (2019). Morphology and pro drop. In *Oxford research encyclopedia of linguistics*.

Kong, F., Zhang, M., & Zhou, G. (2019). Chinese zero pronoun resolution: A chain-to-chain approach. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, *19*(1), 1–21.

Kong, F., Zhou, G., & Zhu, Q. (2009). Employing the centering theory in pronoun resolution from the semantic perspective. *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, 987–996.

Konno, R., Matsubayashi, Y., Kiyono, S., Ouchi, H., Takahashi, R., & Inui, K. (2020). An empirical study of contextual data augmentation for Japanese zero anaphora resolution. *Proceedings of the 28th International Conference on Computational Linguistics*, 4956–4968.

Krifka, M. (2008). Basic notions of information structure. *Acta Linguistica Hungarica*, *55*(3-4), 243–276.

Kuncoro, A., Ballesteros, M., Kong, L., Dyer, C., Neubig, G., & Smith, N. A. (2017). What do recurrent neural network grammars learn about syntax? *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, 1249–1258.

Lei, T., Xin, Y., Zhang, Y., Barzilay, R., & Jaakkola, T. (2014). Low-rank tensors for scoring dependency structures. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1381–1391.

Lenci, A. (2018). Distributional models of word meaning. *Annual review of Linguistics*, *4*, 151–171.

Levinson, S. C. (1987). Minimization and conversational inference. *International Conference on Pragmatics*, 61–129.

Li, C. N. (1976). Subject and topic: A new typology of language. *Subject and topic*.

Li, C. N., & Thompson, S. A. (1979). Third-person pronouns and zero-anaphora in Chinese discourse. In *Discourse and syntax* (pp. 311–335). Brill.

Li, C. N., & Thompson, S. A. (1989). *Mandarin Chinese: A functional reference grammar* (Vol. 3). Univ of California Press.

Li, J., Wang, S., Luh, W.-M., Pylkkänen, L., Yang, Y., & Hale, J. (2021). Cortical processing of reference in language revealed by computational models. *bioRxiv 2020.11.24.396598*.

Li, W. (2004). Topic chains in Chinese discourse. *Discourse Processes*, *37*(1), 25–45.

Liceras, J. M. (1989). On some properties of the pro-drop parameter: Looking for missing subjects in non-native Spanish. *Linguistic perspectives on second language acquisition*, 109–133.

Liejiong, X. (1986). Free empty category. *Linguistic Inquiry*, *17*(1), 75–93.

Liu, B., & Wu, L. (2022). Graph neural networks in natural language processing. In *Graph neural networks: Foundations, frontiers, and applications* (pp. 463–481). Springer.

Liu, D. (2008). Adequate language description in L2 research/teaching: The case of pro-drop language speakers learning L2 English. *International Journal of Applied Linguistics*, *18*(3), 274–292.

Liu, J., & Zhang, Y. (2017). In-order transition-based constituent parsing. *Transactions of the Association for Computational Linguistics*, *5*, 413–424.

Liu, L., Song, Z., & Zheng, X. (2020). Improving coreference resolution by leveraging entity-centric features with graph neural networks and second-order inference. *arXiv preprint arXiv:2009.04639*.

Ma, X., Hu, Z., Liu, J., Peng, N., Neubig, G., & Hovy, E. (2018). Stack-pointer networks for dependency parsing. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1403–1414.

Manjuan, D., & Ping, J. (2010). An empirical study of centering in Chinese anaphoric resolution. *2010 International Conference on Artificial Intelligence and Computational Intelligence*, *1*, 373–377.

Martins, A. F., Almeida, M. B., & Smith, N. A. (2013). Turning on the turbo: Fast third-order non-projective turbo parsers. *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 617–622.

Matchin, W., Sprouse, J., & Hickok, G. (2014). A structural distance effect for backward anaphora in Broca's area: An fMRI study. *Brain and Language*, *138*, 1–11.

McClosky, D., Charniak, E., & Johnson, M. (2006). Effective self-training for parsing. *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, 152–159.

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, *5*, 115–133.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Miller, B. A., Bliss, N. T., & Wolfe, P. J. (2010). Toward signal processing theory for graphs and non-Euclidean data. *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, 5414–5417.

Mrini, K., Dernoncourt, F., Tran, Q. H., Bui, T., Chang, W., & Nakashole, N. (2020). Rethinking self-attention: Towards interpretability in neural parsing. *Findings of the Association for Computational Linguistics: EMNLP 2020*, 731–742.

Müller, G. (2007). *Some consequences of an impoverishment-based approach to morphological richness and pro-drop*. na.

Navigli, R., & Martelli, F. (2019). An overview of word and sense similarity. *Natural Language Engineering*, *25*(6), 693–714.

Navigli, R., & Ponzetto, S. P. (2012). BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial intelligence*, *193*, 217–250.

Neeleman, A., & Szendröi, K. (2008). *Case morphology and radical pro-drop*. na.

Neeleman, A., & Szendrői, K. (2007). Radical pro-drop and the morphology of pronouns. *Linguistic Inquiry*, *38*(4), 671–714.

Neumann, A. (2021). Using and comparing rhetorical structure theory parsers with rst-workbench. *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, 1–6.

Nieuwland, M. S., Coopmans, C. W., & Sommers, R. (2019). Distinguishing old from new referents during discourse comprehension: Evidence from ERPsand oscillations. *Frontiers in human neuroscience*, *13*, 398.

Nieuwland, M. S., & Martin, A. E. (2017). Neural oscillations and a nascent corticohippocampal theory of reference. *Journal of cognitive neuroscience*, *29*(5), 896–910.

Nieuwland, M. S., Petersson, K. M., & Van Berkum, J. J. (2007). On sense and reference: Examining the functional neuroanatomy of referential processing. *NeuroImage*, *37*(3), 993–1004.

Otheguy, R., Zentella, A. C., & Livert, D. (2007). Language and dialect contact in Spanish in New York: Toward the formation of a speech community. *Language*, 770–802.

Pandit, O., Denis, P., & Ralaivola, L. (2020). Integrating knowledge graph embeddings to improve mention representation for bridging anaphora resolution. *CRAC 2020-Third Workshop on Computational Models of Reference, Anaphora and Coreference*.

Park, S., Yoon, D., & Kim, H. (2022). Improving graph-based document-level relation extraction model with novel graph structure. *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, 4379–4383.

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., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Peng, S., Liu, Y. J., & Zeldes, A. (2022). Gcdt: A Chinese RST treebank for multigenre and multilingual discourse parsing. *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, 382–391.

Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for word representation. *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 1532–1543.

Pešková, A. (2013). Experimenting with pro-drop in Spanish. *SKY Journal of Linguistics*, *26*, 117–149.

Peters, M. E., Neumann, M., Zettlemoyer, L., & Yih, W.-t. (2018). Dissecting contextual word embeddings: Architecture and representation. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 1499–1509.

Phung, D., Nguyen, T. N., & Nguyen, T. H. (2021). Hierarchical graph convolutional networks for jointly resolving cross-document coreference of entity and event mentions. *Proceedings of the*

Fifteenth Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-15), 32–41. https://doi.org/10.18653/v1/2021.textgraphs-1.4

Pilehvar, M. T., & Camacho-Collados, J. (2020). Embeddings in natural language processing: Theory and advances in vector representations of meaning. *Synthesis Lectures on Human Language Technologies*, *13*(4), 1–175.

Piñango, M. M., Finn, E., Lacadie, C., & Constable, R. T. (2016). The localization of long-distance dependency components: Integrating the focal-lesion and neuroimaging record. *Frontiers in psychology*, *7*, 1434.

Poesio, M., Stevenson, R., Eugenio, B. D., & Hitzeman, J. (2004). Centering: A parametric theory and its instantiations. *Computational linguistics*, *30*(3), 309–363.

Poria, S., Cambria, E., Winterstein, G., & Huang, G.-B. (2014). Sentic patterns: Dependency-based rules for concept-level sentiment analysis. *Knowledge-Based Systems*, *69*, 45–63.

Postolache, O., Kruijff-Korbayová, I., & Kruijff, G.-J. M. (2005). Data-driven approaches for information structure identification. *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, 9–16.

Prasad, R., Webber, B., Lee, A., & Joshi, A. (2019). Penn Discourse Treebank version 3.0, ldc2019t05. *Web Download. Philadelphia: Linguistic Data Consortium. URL: https://catalog. ldc. upenn. edu/LDC2019T05*.

Pu, M.-M. (2019a). *Zero anaphora and topic chain in Chinese discourse* (C. Shei, Ed.). Routledge.

Pu, M.-M. (2019b). Zero anaphora and topic chain in Chinese discourse. In *The routledge handbook of chinese discourse analysis* (pp. 188–200). Routledge.

Pu, M.-M., & Pu, Q. (2014). Zero anaphora and topic chain: A cross-linguistic study. *International Journal of Linguistics and Communication*, *2*(1), 27–44.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, *1*(8), 9.

Rao, S., Ettinger, A., Daumé III, H., & Resnik, P. (2015). Dialogue focus tracking for zero pronoun resolution. *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 494–503.

Rasooli, M. S., & Tetreault, J. (2015). Yara parser: A fast and accurate dependency parser. *arXiv preprint arXiv:1503.06733*.

Rello, L., & Ilisei, I. (2009). A comparative study of Spanish zero pronoun distribution. *Proceedings of the International Symposium on Data and Sense Mining, Machine Translation and Controlled Languages (ISMTCL)*, 1–5.

Rizzi, L., Jaeggli, O., & Silva-Corvalan, C. (1986). On the status of subject clitics in Romance. *Studies in Romance linguistics*, 391–419.

Roeper, T., & Rohrbacher, B. (2000). Null subjects in early child English and the theory of economy of projection. In *The acquisition of scrambling and cliticization* (pp. 345–396). Springer.

Rothe, S., & Schütze, H. (2015). Autoextend: Extending word embeddings to embeddings for synsets and lexemes. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 1793–1803.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *nature*, *323*(6088), 533–536.

Rus, V., & Niraula, N. (2012). Automated detection of local coherence in short argumentative essays based on centering theory. *Computational Linguistics and Intelligent Text Processing: 13th International Conference, CICLing 2012, New Delhi, India, March 11-17, 2012, Proceedings, Part I 13*, 450–461.

Santi, A., & Grodzinsky, Y. (2012). Broca's area and sentence comprehension: A relationship parasitic on dependency, displacement or predictability? *Neuropsychologia*, *50*, 821–832.

Schneider, N., Flanigan, J., & O'Gorman, T. (2015). The logic of AMR: Practical, unified, graph-based sentence semantics for NLP. *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorial Abstracts*, 4–5.

Seddah, D., Tsarfaty, R., Kübler, S., Candito, M., Choi, J. D., Farkas, R., Foster, J., Goenaga, I., Galletebeitia, K. G., Goldberg, Y., et al. (2013). Overview of the SPMRL 2013 shared task: A cross-framework evaluation of parsing morphologically rich languages. *Proceedings of the Fourth Workshop on Statistical Parsing of Morphologically-Rich Languages*, 146–182.

Shi, D. (1989). Topic chain as a syntactic category in Chinese. *Journal of Chinese Linguistics*, *17*, 223–261.

Singh, P. K. (2022). Data with non-Euclidean geometry and its characterization. *Journal of Artificial Intelligence and Technology*, *2*(1), 3–8.

Soares, E. C., Miller, P., & Hemforth, B. (2020). The effect of semantic and discourse features on the use of null and overt subjects-a quantitative study of third person subjects in Brazilian Portuguese. *DELTA: Documentação de Estudos em Lingüistica Teórica e Aplicada*, *36*, 2020360107.

Speas, M. (1995). Economy, agreement and the representation of null arguments. *Ms. University of Massachusetts, Amherst*.

Sperber, D., & Wilson, D. (1986). *Relevance: Communication and cognition* (Vol. 142). Citeseer.

Sperduti, A., & Starita, A. (1997). Supervised neural networks for the classification of structures. *IEEE Transactions on Neural Networks*, *8*(3), 714–735.

Strzyz, M., Vilares, D., & Gómez-Rodrıguez, C. (2019). Viable dependency parsing as sequence labeling. *NAACL HLT 2019-2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, *1*, 717–723.

Sun, K. (2019). The integration functions of topic chains in Chinese discourse. *Acta Linguistica Asiatica*, *9*(1), 29–57.

Sun, S.-j. (2022). Self-attention enhanced CNNs with average margin loss for Chinese zero pronoun resolution. *Applied Intelligence*, *52*(5), 5739–5750.

Suzuki, J., Takase, S., Kamigaito, H., Morishita, M., & Nagata, M. (2018). An empirical study of building a strong baseline for constituency parsing. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 612–618.

Takase, S., Suzuki, J., & Nagata, M. (2018). Direct output connection for a high-rank language model. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

Tian, Y., Song, Y., Xia, F., & Zhang, T. (2020). Improving constituency parsing with span attention. *Findings of the Association for Computational Linguistics, ACL 2020: EMNLP 2020*, 1691–1703.

Tomioka, S. (2003). The semantics of Japanese null pronouns and its cross-linguistic implications. *The interfaces: Deriving and interpreting omitted structures*, *61*, 321.

Tong, J., Yang, J., Li, S., & Gao, S. (2019). Dropped pronoun recovery in Chinese conversations with knowledge-enriched neural network. *China National Conference on Chinese Computational Linguistics*, 545–557.

Tsao, F. (1977). *A functional study of topic in Chinese: The first step towards discourse analysis* (Doctoral dissertation). USC. Los Angeles, California.

Tsarfaty, R., Seddah, D., Goldberg, Y., Kübler, S., Versley, Y., Candito, M., Foster, J., Rehbein, I., & Tounsi, L. (2010). Statistical parsing of morphologically rich languages (SPMRL) what, how and whither. *Proceedings of the NAACL HLT 2010 First Workshop on Statistical Parsing of Morphologically-Rich Languages*, 1–12.

Van Dijk, T. A. (1977). Semantic macro-structures and knowledge frames in discourse comprehension. *Cognitive processes in comprehension*, *332*, 3–31.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Velickovic, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., Bengio, Y., et al. (2017). Graph attention networks. *stat*, *1050*(20), 10–48550.

Vilares, D., Gómez-Rodríguez, C., & Alonso, M. A. (2017). Universal, unsupervised (rule-based), uncovered sentiment analysis. *Knowledge-Based Systems*, *118*, 45–55.

Vinyals, O., Kaiser, Ł., Koo, T., Petrov, S., Sutskever, I., & Hinton, G. (2015). Grammar as a foreign language. *Advances in neural information processing systems*, *28*.

Walker, J. P., Walker, M. I., et al. (1998). *Centering theory in discourse*. Oxford University Press.

Wang, L., Tu, Z., Zhang, X., Liu, S., Li, H., Way, A., & Liu, Q. (2017). A novel and robust approach for pro-drop language translation. *Machine Translation*, *31*(1), 65–87.

Wang, Y., & Guo, M. (2014). A short analysis of discourse coherence. *Journal of Language Teaching and Research*, *5*(2), 460.

White, L. (1985). The "pro-drop" parameter in adult second language acquisition. *Language learning*, *35*(1), 47–61.

Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., & Philip, S. Y. (2020). A comprehensive survey on graph neural networks. *IEEE transactions on neural networks and learning systems*, *32*(1), 4–24.

Xian, H., & Keliang, Z. (2009). Zero anaphora in Chinese: The state of art. (汉语零形回指研究综述). *Journal of Chinese Information Processing (*中文信息学报*)*, *23*(4), 10–16.

Xiao, C. (2021). A literature review on Centering Theory. *Studies in Literature and Language*, *22*(3), 5–12.

xiaowangzi.org. (2021). 小王子网站 [Accessed: 2021-04-03].

Xu, K., Hu, W., Leskovec, J., & Jegelka, S. (2018). How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826*.

Xu, Y., & Yang, J. (2019). Look again at the syntax: Relational graph convolutional network for gendered ambiguous pronoun resolution. *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, 96–101.

Yamada, K., & Miyamoto, Y. (2017). On the interpretation of null arguments in L2 japanese by european non-pro-drop and pro-drop language speakers. *Journal of the European Second Language Association*, *1*(1).

Yamashiro, S., Nishikawa, H., & Tokunaga, T. (2018). Neural Japanese zero anaphora resolution using smoothed large-scale case frames with word embedding. *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation*.

Yang, J., Li, S., Gao, S., & Guo, J. (2022). CorefDPR: A joint model for coreference resolution and dropped pronoun recovery in Chinese conversations. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *30*, 571–581.

Yang, J., Xu, K., Xu, J., Li, S., Gao, S., Guo, J., Wen, J.-R., & Xue, N. (2020). Transformer-GCRF: Recovering Chinese dropped pronouns with general conditional random fields. In T. Cohn, Y. He, & Y. Liu (Eds.), *Findings of the association for computational linguistics: Emnlp 2020* (pp. 137–147). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.findings-emnlp.13

Yang, J., Xu, K., Xu, J., Li, S., Gao, S., Guo, J., Xue, N., & Wen, J.-R. (2021). A joint model for dropped pronoun recovery and conversational discourse parsing in Chinese conversational speech. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 1752–1763.

Yang, K., & Deng, J. (2020). Strongly incremental constituency parsing with graph neural networks. *Advances in Neural Information Processing Systems*, *33*, 21687–21698.

Yeh, C.-L., & Chen, Y.-C. (2003). Zero anaphora resolution in Chinese with partial parsing based on centering theory. *International Conference on Natural Language Processing and Knowledge Engineering, 2003. Proceedings. 2003*, 683–688.

Yeh, C.-L., & Chen, Y. (2004). Topic identification in Chinese based on centering model. *Proceedings of the Conference on Reference Resolution and Its Applications*, 103–109.

Yeh, C.-L., & Chen, Y.-C. (2007). Zero anaphora resolution in Chinese with shallow parsing. *J. Chin. Lang. Comput.*, *17*(1), 41–56.

Yin, Q., Zhang, W., Zhang, Y., & Liu, T. (2017). A deep neural network for Chinese zero pronoun resolution. *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, 3322–3328.

Yin, W., Schütze, H., Xiang, B., & Zhou, B. (2016). Abcnn: Attention-based convolutional neural network for modeling sentence pairs. *Transactions of the Association for computational linguistics*, *4*, 259–272.

Ynoa, M. R. (2020). Analysis of PRO-drop errors in L2 English by L1 Spanish speakers.

Yoshida, S., & Nagata, M. (2009). Utilizing features of verbs in statistical zero pronoun resolution for Japanese speech. *Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation, Volume 2*, 867–874.

Yun, Z., & JIN, Y.-h. (2012). A Chinese-English patent machine translation system based on the theory of hierarchical network of concepts. *The Journal of China Universities of Posts and Telecommunications*, *19*, 140–146.

Zhang, J. (2016). Cross-linguistic variations of pro licensing conditions. *Journal of Language Teaching and Research*, *7*(3), 499.

Zhang, L., Shen, Z., & Shao, Y. (2020). Semantic-aware Chinese zero pronoun resolution with pre-trained semantic dependency parser. *China National Conference on Chinese Computational Linguistics*, 17–29.

Zhang, L., Xing, Y., Kong, F., Li, P., & Zhou, G. (2020). A top-down neural architecture towards text-level parsing of discourse rhetorical structure. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 6386–6395.

Zhang, S., Wang, L., Sun, K., & Xiao, X. (2020). A practical Chinese dependency parser based on a large-scale dataset. *arXiv preprint arXiv:2009.00901*.

Zhang, S., Li, J., & Hale, J. (2022). Quantifying discourse support for omitted pronouns. *Proceedings of the Fourth Workshop on Computational Models of Reference, Anaphora and Coreference*, 1–12.

Zhang, S., Li, J., Yang, Y., & Hale, J. (2022). Decoding the silence: Neural bases of zero pronoun resolution in Chinese. *Brain and Language*, *224*, 105050.

Zhang, Y., Li, Z., & Min, Z. (2020). Efficient second-order TreeCRF for neural dependency parsing. *Proceedings of ACL*, 3295–3305. https://www.aclweb.org/anthology/2020.acl-main.302

Zhang, Y., Zhou, H., & Li, Z. (2020). Fast and accurate neural CRF constituency parsing. *Proceedings of IJCAI*, 4046–4053. https://doi.org/10.24963/ijcai.2020/560

Zhang, Y., Zhou, H., & Li, Z. (2021). Fast and accurate neural CRF constituency parsing. *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, 4046–4053.

Zhang, Y., & Clark, S. (2008). A tale of two parsers: Investigating and combining graph-based and transition-based dependency parsing. *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, 562–571.

Zhang, Y., & Clark, S. (2011a). Syntactic processing using the generalized perceptron and beam search. *Computational Linguistics*, *37*(1). https://doi.org/10.1162/coli_a_00037

Zhang, Y., & Clark, S. (2011b). Syntactic processing using the generalized perceptron and beam search. *Computational linguistics*, *37*(1), 105–151.

Zhang, Y., Qi, P., & Manning, C. D. (2018). Graph convolution over pruned dependency trees improves relation extraction. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2205–2215.

Zheng, H., Fu, J., Mei, T., & Luo, J. (2017). Learning multi-attention convolutional neural network for fine-grained image recognition. *Proceedings of the IEEE international conference on computer vision*, 5209–5217.

Zhou, J., Cui, G., Hu, S., Zhang, Z., Yang, C., Liu, Z., Wang, L., Li, C., & Sun, M. (2020). Graph neural networks: A review of methods and applications. *AI Open*, *1*, 57–81.

Zhou, J., & Zhao, H. (2019). Head-driven phrase structure grammar parsing on Penn Treebank. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2396–2408.

Zhu, L., Li, Z., Li, C., Wu, J., & Yue, J. (2018). High performance vegetable classification from images based on AlexNet deep learning model. *International Journal of Agricultural and Biological Engineering*, *11*(4), 217–223.

Zhu, Y., Song, W., Liu, X., Liu, L., & Zhao, X. (2019). Improving anaphora resolution by animacy identification. *2019 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA)*, 48–51.