Referring expression generation in context

Combining linguistic and computational approaches

Fahime Same

Topics at the Grammar-Discourse Interface 9

language science press

Topics at the Grammar-Discourse Interface

Editors: Philippa Cook (University of Göttingen), Anke Holler (University of Göttingen), Cathrine Fabricius-Hansen (University of Oslo)

In this series:

1. Song, Sanghoun. Modeling information structure in a cross-linguistic perspective.
2. Müller, Sonja. Distribution und Interpretation von Modalpartikel-Kombinationen.
3. Bueno Holle, Juan José. Information structure in Isthmus Zapotec narrative and conversation.
4. Parikh, Prashant. Communication and content.
5. Balogh, Kata, Anja Latrouite & Robert D. Van Valin, Jr. (eds.) Nominal anchoring: Specificity, definiteness and article systems across languages.
6. Næss, Åshild, Anna Margetts & Yvonne Treis (eds.). Demonstratives in discourse.
7. Gergel, Remus & Jonathan Watkins (eds.). Quantification and scales in change.
8. Nakagawa, Natsuko. Information structure in spoken Japanese: Particles, word order, and intonation.
9. Same, Fahime. Referring expression generation in context: Combining linguistic and computational approaches.

ISSN: 2567-3335

Referring expression generation in context

Combining linguistic and computational approaches

Fahime Same

language
science
press

Fahime Same. 2024. *Referring expression generation in context: Combining linguistic and computational approaches* (Topics at the Grammar-Discourse Interface 9). Berlin: Language Science Press.

This title can be downloaded at:
http://langsci-press.org/catalog/book/451

ISBN: 978-3-96110-471-0 (Digital)
978-3-98554-100-3 (Hardcover)

ISSN: 2567-3335
DOI: 10.5281/zenodo.11058114
Source code available from www.github.com/langsci/451
Errata: paperhive.org/documents/remote?type=langsci&id=451

Cover and concept of design: Ulrike Harbort
Typesetting: Fahime Same, Sebastian Nordhoff
Proofreading: Amy Amoakuh
Fonts: Libertinus, Arimo, DejaVu Sans Mono
Typesetting software: XeLaTeX

Language Science Press
xHain
Grünberger Str. 16
10243 Berlin, Germany
http://langsci-press.org

Storage and cataloguing done by FU Berlin

Freie Universität Berlin

Contents

Acknowledgments

This book is a revised and improved version of the doctoral dissertation I submitted in July 2022. The original acknowledgments are reprinted below with minor updates.

I could not have completed this work without the people around me. First and foremost, I would like to thank my supervisors, Nikolaus P. Himmelmann and Kees van Deemter, for their great support and valuable guidance. Thank you, Nikolaus, for always being there for me and guiding me through my PhD project with your deep linguistic understanding. Thank you, Kees, for the great collaboration, the many joyful meetings, and the thoughtful ideas you always have. I really enjoy working with you and learning from you. I can not thank you both enough.

I would like to thank SFB1252 "Prominence in Language" for providing me with a great working environment and giving me the opportunity to work on my PhD. I would like to thank all the wonderful colleagues I have had the pleasure to work with over the years at the University of Cologne, especially in our great House of Prominence (HoP). To name just a few: Thank you to Alessia Cassarà, Barbara Zeyer, Eric Engel, Max Hörl, Constantijn Kaland, and Sara Meuser. Many thanks to everyone from the second floor of the HoP (for coffee, cookies and nice, happy conversations) and to the SFB lunch crew for the funniest (and sometimes absurd) conversations over the years. My very special thanks to Maria Bardaji i Farré (thank you, Maria, for your cheerful and positive presence in the very last phase of my PhD), Christoph Bracks (thank you, Chris, for being a great friend and helping me a lot in the last weeks of writing my PhD), and Guanyi Chen (thank you, Guanyi, for being such a great colleague and coauthor. I really enjoy working with you and have learned so much from you). My very special thanks to T. Mark Ellison (thank you, Mark, for always being there for me, teaching me so much, reading and commenting on my dissertation, and for all the fun conversations.)

I would like to thank my dearest friends, Arefeh (aka Fafafe), Golnoush, Mahsa, and Adineh for keeping me sane during the craziest days of my life. The phase of finishing up a PhD can get pretty lonely. It gets even lonelier when you have to do it during Covid. The best thing is to have great friends to support you, cheer

you up, and dream with you about fun life after the PhD. Thank you all for being there for me. You're the best! I also want to give a special thanks to Rob, for being such a great (and funny) friend and also proofreading parts of my dissertation.

My deepest gratitude goes to my family. Thank you, Zohreh and Mamad, for being the most amazing siblings anyone could ask for, for always having my back and for always believing in me. And my deepest thanks goes to my mom (Maman). Thank you, Maman, for your unconditional love and support. I love you so much!

And a very big "thank you" to my partner, Christian. Thanks, Chris, for granting me full "Narrenfreiheit" over the past few months, for cooking delicious food, and for reading my dissertation more than once. Thanks for being such a wonderful partner.

Abbreviations and conventions

ANTE	Antecedent
BERT	Bidirectional Encoder Representations from Transformers
BF	Bayes Factor
BiGRU	Bi-directional Gated Recurrent Unit neural network
CAtt	Concatenative Attention mechanism
Cb	Backward-looking Center
Cf	Forward-looking Center
CL	Computational Linguistics
Cp	Preferred Center
CT	Centering Theory
CRF	Conditional Random Field
E2E	End-to-end
GloVe	Global Vectors
HierAtt	Hierarchical Attention mechanism
IA	The Incremental Algorithm
LSTM	Long Short-Term Memory
ML	Machine Learning
MLP	Multilayer Perceptron
NeuralRFS	Neural Referential Form Selection.
NLG	Natural Language Generation
NLP	Natural Language Processing
NeuralRFS	Neural Referential Form Selection.
NLG	Natural Language Generation
NLP	Natural Language Processing
NP	Noun Phrase
ONF	OntoNotes Normal File
OOB	Out-of-bag
POS	Part-of-speech
PT	Prominence Theory
RCS	Referential Content Selection
RDF	Resource Description Framework
RE	Referring Expression

REF	Referent
REG	Referring Expression Generation
ReLU	Rectified Linear Unit
RF	Referential Form (Referring Expression Form)
RFI	Random Forest Importance
RFS	Referential Form Selection
RNN	Recurrent Neural Network
SHAP	SHapley Additive exPlanations
Seq2Sep	Sequence-to-sequence decoder
SFS	Sequential Forward Search
SOTA	State-of-the-art
WSJ	Wall Street Journal

This book is written in American English. An exception is the prefix *non-* which is joined to a word by means of a hyphen, e.g., *non-pronominal*. Chapters 5 and 7 are written in first-person plural (*We*). Chapter 8 is a mix of first-person singular and plural (*I* and *We*). Other chapters are mostly written in first-person singular. In what follows, I give an overview of the writing conventions I have followed throughout the book.

In-text examples: In-text examples appear in *italics*, such as *here is an example.*

Translations and glosses: Translations and glosses are surrounded by single quotation marks.

Terms: New terms are introduced in *italics* for the first time. Further use of these terms is in normal type. New terms include linguistic terms (e.g., *coreferential chain*), computational terms (e.g., *Natural Language Generation*), and computational methods used (e.g., *sequential forward search*).

Name of models and corpora: The name of models and corpora, either from the literature or defined in this book, are in SMALL CAPS (e.g.,WEBNLG).

Values: The different values that a linguistic category can take are written in the `typescript` mode (e.g., a 3-way classification consists of three values: `pronoun`, `proper name`, and `description`). The category itself is written mostly in normal font (e.g., grammatical role, animacy). If a value is mentioned in the text flow, it is also written in normal font. In the following example, the values `proper name` and `human` are written in normal font: we use a proper name to refer to a human referent for the first time.

Symbols: The symbols used to represent a category or a feature are in the `typescript` mode. For instance, in tables and graphs, grammatical role is presented as `gm`.

1 Introduction

1.1 Goal of this book

An essential part of successful communication is referring to the things you want to talk about in a manner that is understandable and sounds natural to your reader or listener.

(1) On Monday, **Shinzo Abe** set a record for being Japan's longest-serving prime minister since 1885. Just four days later, **Shinzo Abe** announced **Shinzo Abe** was retiring. **Shinzo Abe**'s term was scheduled to end in September 2021, however, poor health forced an early departure. **Shinzo Abe** has suffered ...

(1) is understandable to the reader, but the repeated use of the term *Shinzo Abe* makes it sound unnatural in English, and the reader may find it difficult to follow the text. In fact, English offers various alternative expressions, the use of which can make the text more readable, as demonstrated in the original version of the above text in (2).[1]

(2) On Monday, **Shinzo Abe** set a record for being Japan's longest-serving prime minister since 1885. Just four days later, **the 65-year-old politician** announced **he** was retiring. **His** term was scheduled to end in September 2021, however, poor health forced an early departure. **Abe** has suffered ...

As (2) illustrates, people constantly make decisions about how to refer to different things. Research on the production of *Referring Expressions* (REs) delves into the choices individuals make and the factors influencing those choices. The topic of reference production also garners significant interest in Computational Linguistics (CL) and Natural Language Generation (NLG), where it is known as *Referring Expression Generation* (REG).

This book aims to present linguistically informed solutions for the task of generating REs within a discourse context, henceforth referred to as *REG-in-context*.

[1] `https://www.dw.com/en/shinzo-abes-departure-signals-the-end-of-an-era-in-tokyo/a-54738816`

Belz & Varges (2007) described REG-in-context as: "Given an intended referent and a discourse context, how do we generate appropriate referential [referring] expressions (REs) to refer to the referent at different points in the discourse?" (p. 9). The term *discourse context*, often abbreviated to *context*, can be interpreted in several ways. At its most basic, context implies that the RE is not generated in isolation. Alternatively, a text extending beyond a single sentence might be viewed as the baseline for defining context (Belz et al. 2010). For instance, in example (2), an algorithm tasked with generating appropriate REs referring to Shinzo Abe can be considered as undertaking a REG-in-context task. Throughout this book, I conduct a systematic analysis of REG-in-context models, examining the approaches adopted, the corpora employed, and the features applied. Additionally, I provide explanations and enhancements to these models, drawing from linguistic insights.

This chapter unfolds as follows: in §1.2, I introduce terminology frequently referenced throughout the book. §1.3 moves beyond the practical (and occasionally oversimplified) definitions of reference-related concepts to delve into the more intricate facets of the theory. Subsequent sections, §1.4 and §1.5, offer succinct overviews of studies that explore the choice of REs from linguistic and computational perspectives. I conclude this chapter in §1.6, outlining the overarching structure of the work and pinpointing the specific research questions this book pursues.

1.2 Setting the terminology

(2) displays various REs, such as *Shinzo Abe* and *the 65-year-old politician*, which are employed to *refer* to the *referent* or *(discourse) entity* Shinzo Abe (hereafter SHINZO ABE). Alternatively, we can say that these REs are different *mentions* of SHINZO ABE.

REs referring to the same referent are described as being *coreferential*, and collectively form a *coreferential chain*. For instance, in the context of (2), the coreferential chain consists of the REs: {Shinzo Abe, the 65-year-old politician, he, His, Abe}.

SHINZO ABE is, for the first time, introduced in the initial sentence of (2). This introductory RE is termed a *first mention*. Any mentions that follow are labeled *subsequent mentions*, *anaphors*, or *anaphoric expressions*. If an entity is mentioned only once in a text, that RE is termed a *singleton* (Jurafsky & Martin 2021). With the exception of first-mention referents and singletons, all REs have at least one coreferential antecedent within the text. In (2), if we consider the target expres-

sion to be *he* from the second sentence, its immediate antecedent is *the 65-year-old politician*, while its secondary antecedent is *Shinzo Abe* from the first sentence. Henceforth, the immediate antecedent will be termed *antecedent*.

As illustrated in (2), a variety of REs are used to denote the referent. The term *Referential Form* (RF) pertains to the different forms REs can take. For instance, the initial RE in (2) combines the first (*Shinzo*) and last name (*Abe*) of Japan's former prime minister. This expression is categorized as a *proper name*, or simply *name*. The second RE is the definite noun phrase *the 65-year-old politician*, which is called a *description*. Both definite and indefinite *Noun Phrases (NPs)* fall into this category.[2] The third RE in (2) is the *pronoun he*. These three RF classes (`name`, `description`, `pronoun`) are central to this book, though more fine-grained classifications are also feasible.[3]

1.3 The study of reference: Definition and challenges

The preceding section laid the groundwork by introducing terminology that will be consistently used throughout this book. The terms provided offer a simplified and practical understanding of concepts related to reference. The primary focus of this work will be on the core types of reference that are frequently observed in corpora. This means that I will bypass a number of lively debates and intricate questions about reference, especially those posed by logicians and philosophers of language (Frege 1960, Russell 1905, Strawson 1950, Donnellan 1966, Searle 1969, Récanati 1993).[4]

To illustrate, Frege, one of the pioneering figures in the contemporary discourse on reference, differentiated between *sense* (in German: Sinn) and *reference* (Bedeutung) of an RE. The former can be understood as the meaning or mode of presentation of an RE, while the latter pertains to the actual referent of the expression. With this distinction, he elucidated why the statement *The Morning Star is The Evening Star* is informative, whereas *The Morning Star is The Morning Star*

[2]Some scholars view proper names as instances of descriptions. For example, as highlighted in Donnellan (1972), Russell argues that proper names are concealed definite descriptions, asserting that "the name 'Romulus' is not really a name [that is, in the 'narrow logical sense'] but a sort of truncated description. It stands for a person who did such-and-such things, who killed Remus, and founded Rome, and so on." However, in the majority of experimental and computational studies, proper names are distinguished from descriptions. I also adhere to this tradition throughout this book.

[3]Another RE type to note is the zero or null REs (Scott 2019), where REs are conveyed by inflection or implied pragmatically. This RE type is not covered in this book, given that one of the main corpora examined is not annotated for null instances.

[4]It is worth noting that Frege's original article dates back to 1892.

is not. These two REs share the same reference but possess distinct senses. Similarly, this distinction was used to account for the difference in the use of REs in *extensional* and *intensional* contexts (van Deemter 2016). In extensional contexts, the truth-value of a sentence depends only on the references of its REs, not their senses. Hence, the statement *The Morning Star is a planet* also implies *The Evening Star is a planet.* Conversely, in intensional contexts, the truth-value of a proposition depends on both the references and senses of its REs. A case of intensional context is when we use verbs such as "believe", "know" or "think". For instance, if it is true that *John believes that the Morning Star is a planet*, the sentence *John believes that the Evening Star is a planet* may or may not hold true. The truth-value of the former does not necessarily guarantee the truth-value of the latter. While the distinction between intensional and extensional contexts is pivotal in theories of reference, a discernible knowledge gap exists between theory-oriented accounts of reference and its computational modeling (van Deemter 2016).

Searle (1969) offered a definition of reference that aligns with our needs in most scenarios. Based on this definition, the referent of an expression is ultimately determined by what the speaker has in mind on a given occasion. He characterized REs as follows:

> Any expression which serves to identify any thing, process, event, action, or any other kind of individual or particular I shall call a referring expression. Referring expressions point to particular things; they answer the questions Who?, What?, Which? (Searle 1969: 27).

While many prior works have primarily presented complicated referential cases, the definition provided by Searle offers practical appeal. Nonetheless, it is not without its shortcomings. To illustrate, let us consider the last part of the above definition, viz. *they [REs] answer the questions Who?, What?, Which?* [hereafter *3W*] as a test of referentiality. The bold expressions in the following examples indeed respond to 3W, yet none picks up an individual referent.

(3) a. Who is going to the office tomorrow? **No one**. It's a bank holiday.
b. What should we bring to the party? **Nothing**.
c. (pointing at two artworks) Which one should I buy? **None of them**.

Another challenging distinction arises between the *attributive* and *referential* readings of REs (Donnellan 1966, van Deemter 2016). Consider the following sentence: *Smith's murderer must be insane.* Under an attributive reading, someone might utter this upon witnessing Smith's monstrous and disturbing crime scene,

without having a specific individual in mind as the murderer. While the attributive cases do not refer to an individual per se, a question like *Who is insane?* can still elicit *Smith's murderer* as a response. Conversely, under a referential reading, someone might utter the phrase during Smith's murder trial, referring to the alleged killer's unusual behavior on the stand.

As outlined above, various scenarios present challenges in the study of reference. To address the generation of REs within a discourse context, I use several *real-life* corpora, notably the Wall Street Journal (WSJ) portion of the OntoNotes corpus (hereafter referred to as ONTONOTES, Weischedel, Ralph et al. 2013).[5] This corpus encompasses news articles as well as the insights of the respective authors. Given this composition, one can anticipate encountering more complex cases of reference. For instance, the pronoun *it* in (4) illustrates an attributive instance which is annotated as a coreferential RE.

(4) [WSJ-1424] Nobody is sure what will come next in Somalia or whom the successor might be. But as one expert tells me : "Whoever **it** is will have to work pretty damn hard to be worse than Barre."

The attributive example mentioned above and many other theoretical challenges have not yet found their way into computational modeling, as these models require concrete plans to implement a concept. To date, computational models have not fully considered non-literal, attributive, and intensional cases. As van Deemter (2016) points out regarding intensional contexts, "theories have interesting things to say about these contexts, but they do not yet offer the detail and precision required by computational REG models". He further continues that "computational models tend to lag behind pure theory, with theories exploring issues long before they are addressed by means of algorithms and computer programs" (p. 36). Given the intricate nature of defining reference and identifying REs, this book will lean on the existing annotations of REs in the corpora under discussion, without delving deeply into complex edge cases. Nonetheless, future studies should take into account the prevalence and characteristics of these cases, recognizing the potential impact of these phenomena on research outcomes.

1.4 The linguistic tradition and the choice of RE

The study of reference involves two distinct processes: *production* and *comprehension* (Hendriks 2016). This book primarily focuses on the production of referring

[5] For the studies discussed in this book, OntoNotes 5.0 is utilized. OntoNotes 5.0 is licensed by the Linguistic Data Consortium (LDC): `https://catalog.ldc.upenn.edu/LDC2013T19`.

expressions. The comprehension of REs will be addressed only when deemed essential.

Theoretical studies examining the production of REs offer diverse explanations for the referential choices individuals make when talking about a referent. According to the *Accessibility Theory* (Ariel 2001), a leading theory in reference production, the more accessible a referent becomes, the more attenuated its corresponding RE is. This theory also delineates a detailed hierarchy of RFs, categorizing them from the least to the most accessible. Other theoretical approaches tread a similar path, associating this choice with the referent's salience, givenness, centrality, and prominence (Gundel et al. 1993, Grosz et al. 1995, Chiarcos 2011, von Heusinger & Schumacher 2019).

Empirical studies put these theories to the test to discern which factors influence the status of referents (e.g., increasing their prominence) and, in turn, influence the choice of REs. Linguistic studies have pinpointed factors such as grammatical role, recency, competition, animacy, thematic role, coherence, and order of mention as influential determinants (Stevenson et al. 1994, Brennan 1995, Arnold 2001, Arnold & Griffin 2007, Kehler et al. 2008, Kaiser & Trueswell 2011, Fukumura & van Gompel 2011). Take the recency factor as an example: it posits that the closer a referent is to its antecedent, the more likely it is to be realized as a pronoun (Givón 1992). Similarly, the animacy factor suggests that animate referents have a higher likelihood of being realized as pronouns (Fukumura & van Gompel 2011). To validate these factors, researchers can employ a variety of methods. These include analyzing corpora of written and spoken language, conducting offline experiments like surveys, and utilizing real-time measurement techniques such as eye-tracking.

As illustrated in the preceding paragraphs, studies within the *linguistic tradition* seek to elucidate why speakers choose different RFs when invoking a referent. They also try to identify and explain the factors that cause these RF alternations. While there are distinctions between theoretical, corpus-based, and experimental approaches to studying reference production, I will not delve deeply into these differences. This is because the primary focus of this book lies on algorithmic solutions. Throughout the book, I will use the term "linguistic tradition" to encompass these approaches and will draw from their findings to enrich the REG-in-context algorithms, aiming to develop more accurate and informed models.

1.5 The computational tradition and the choice of RE

Reference production, often termed REG in computational terms, has also attracted much attention in the field of NLG. NLG is concerned with the generation of natural language text from non-linguistic input (Krahmer & van Deemter 2012, Gatt & Krahmer 2018). Its practical applications span a broad spectrum (Mei et al. 2016, Reiter 2017), including the generation of financial and medical reports (Gatt et al. 2009), weather forecasts (Reiter et al. 2005), and sports predictions (van der Lee et al. 2017).

REG encompasses two distinct yet related tasks (Krahmer & van Deemter 2012, Gatt & Krahmer 2018): (1) *one-shot REG*, and (2) REG-in-context. One-shot REG focuses on conceptualization or the selection of properties of a referent to produce a unique description of it. An example of this task is to single out a referent from a set of competing referents in a visual scene. REG-in-context, which is the main focus of this book, is concerned with the choice of (anaphoric) referring expressions within a discourse context.

REG-in-context is the task of determining the form and semantic content of REs within a given context (Reiter & Dale 2000). *Referential Form Selection* (RFS) is the task of determining the form, and *Referential Content Selection* (RCS) is the task of determining the semantic content of each RE. Suppose we want to generate REs for the referent SHINZO ABE of (2) repeated below:

(2) On Monday, **SHINZO ABE** set a record for being Japan's longest-serving prime minister since 1885. Just four days later, **SHINZO ABE** announced **SHINZO ABE** was retiring. **SHINZO ABE**'s term was scheduled to end in September 2021, however, poor health forced an early departure. **SHINZO ABE** has suffered ...

In RFS, the algorithm's goal is to predict the class of RF from a set of forms for a specific reference slot in the text. For instance, it determines whether a referent should be realized as a `proper name`, a `description`, or a `pronoun`.[6] In RCS, the task is to generate the actual content of an RE. For example, deciding whether to

[6] Both RFS and RCS are not strictly *deterministic*. Multiple forms or expressions might be suitable in various parts of a text. Research that examines the non-deterministic generation of referring expressions includes Castro Ferreira et al. (2016b) and van Gompel et al. (2019). While the non-deterministic generation of referring expressions is important, this book primarily focuses on the deterministic generation of REs and evaluates the models' performance against gold-standard corpora. Given that this book examines three different facets of the task – choice of corpora, feature sets, and REG approaches – adding another dimension could reduce the transparency of the comparisons.

refer to Shinzo Abe by his full name (*Shinzo Abe*), his modified full name (*Shinzo Abe, the former president of Japan*), or his last name (*Abe*).

Classic REG-in-context models primarily employ *rule-based* and *feature-based machine learning (ML)* methods. They typically approach REG in two steps: (1) deciding on the form, and (2) populating it with content. In contrast, the more recent *neural end-to-end (E2E)* models tackle both steps at once.

As implied by their name, rule-based models generate content based on a predefined set of rules (McCoy & Strube 1999, Henschel et al. 2000, Poesio 2004). As such, crafting precise rules is crucial. These models draw heavily from insights gained in linguistic studies. In feature-based models (Belz et al. 2010, Greenbacker & McCoy 2009a, Kibrik et al. 2016), each data point is represented as a set of feature–value pairs taken from a dataset. A machine learning algorithm then uses these pairs to determine the prediction rules. Consequently, these data-driven models need feature engineering and the choice of corpus, features, and machine learning algorithm plays a pivotal role in determining the efficacy of these models. E2E models (Castro Ferreira et al. 2018a, Cao & Cheung 2019), another subset of data-driven models, stand out as they bypass the need for feature engineering. They map directly from input to output. Therefore, the architecture of the model and the quality of training data become paramount for these models.

In the preceding sections, I provided a concise summary of both linguistic and computational studies related to the choice of RE. Moving forward, I will outline the framework of this book and offer an overview of its chapters, detailing the research questions and hypotheses presented in each.

1.6 Outline of the book

This book comprises eight chapters, each largely self-contained with its own introduction and discussion. As a result, some overlap between chapters is inevitable. All the analyses presented are based on English language corpora. In this work, I present seven distinct studies in Chapters 4 through 7, labeled alphabetically from study A to study G. The research questions and hypotheses for each study are numbered according to its respective label.

Chapter 2 delves deeply into reference from a linguistic viewpoint. It introduces several theories of reference production, connecting the choice of RF to factors like a referent's cognitive accessibility, text coherence, the dynamicity of a referent, and the relational properties of referents. This chapter further elaborates on the diverse factors influencing this choice. These factors are notably diverse; while some factors emphasize the inherent traits of referents, others establish a relation to the preceding context.

Chapter 3 looks into computational theories of REG. Following a brief introduction to one-shot REG, the chapter addresses issues associated with REG-in-context. It distinguishes the methodological variances among REG-in-context models, offering a chronological review of rule-based, feature-based, and E2E neural models. The chapter also underscores the significance of choosing the right corpora for the task and establishing robust baselines for fair comparisons. By the chapter's conclusion, readers should gain a comprehensive understanding of several aspects of the REG-in-context task warranting further reconsideration. These aspects encompass the choice of (1) corpora, (2) features, and (3) REG approaches. In conjunction with Chapter 2, this chapter lays the groundwork for the discussions in subsequent chapters.

Chapter 4 introduces the initial study of the work, denoted as study A. It tackles the primary REG-in-context consideration: the selection of a corpus. Given the diverse corpora employed in REG-in-context models to date, the overarching question this study poses is:

> **QA.** Does the choice of corpus matter for REG-in-context studies?

I posit that the choice of corpus matters for REG-in-context studies, leading to the following hypotheses:

> **HA1.** The corpora used in the previous REG-in-context studies are not adequate for the task.
>
> **HA2.** The lessons learned in previous REG-in-context studies are not generally valid.

The term *previous REG-in-context studies* specifically refers to the *GREC (Generating Referring Expressions in Context)* shared tasks (Belz et al. 2010), a series of shared tasks dedicated to generating REs in context. To test HA1, I analyze the two corpora used in GREC, assessing their adequacy for the RFS task. To put the findings into a broader perspective and address HA2, I curate a new dataset derived from the WSJ portion of ONTONOTES. Study A reconstructs the systems submitted to the GREC shared tasks across the three corpora. The performance of the models is evaluated using *Bayes Factor* (BF) analysis and *per-class evaluation*. Furthermore, study A emphasizes the significance of employing metrics beyond mere accuracy to assess the performance of computational models.

Chapter 5 presents the second and third studies of this work, denoted as B and C. These studies investigate the choice of features, an essential aspect of

feature-based REG-in-context models, providing a macro and micro overview of this choice, respectively. Given that these studies draw from two co-authored published articles (Same & van Deemter 2020a,b), they adopt a first-person plural narrative.

Study B conducts a systematic evaluation of the features employed in prior feature-based REG-in-context models. Given the labor-intensive nature of feature engineering, a systematic assessment of features becomes imperative to craft simple yet effective feature-based models. Therefore, the study addresses the following question:

> **QB.** In previous feature-based REG-in-context studies, do all features used in their feature sets contribute equally to the success of the models?

We posit that certain features, previously employed in feature-based REG-in-context studies, do not make a substantial contribution to the task.

> **HB1.** A reduced subset of features from each set can perform comparably to the full feature set.

To validate this hypothesis, we undertake a structured evaluation of earlier feature-based REG-in-context models, aiming to determine which features contribute most to the task. We create models with varying subsets from each feature set, employing techniques such as *variable importance*, *sequential forward search* (SFS), and an array of subsetting rules. The subsequent hypothesis we aim to examine in this study is:

> **HB2.** A small set of features drawn from previously published datasets can form a model that is substantially as accurate as the best-performing existing model.

To assess this hypothesis, we combine and examine various subsets of the most important features from different feature sets. A BF analysis is subsequently carried out to contrast the performance of the newly developed model, which uses the optimal subset of features, with the top-performing model from prior REG-in-context studies.

In study C of this chapter, we delve deeper into the concept of recency, which is characterized as the distance between a target referent and its preceding antecedent. Among the motivations for a more in-depth exploration of recency are its diverse measurement methods and the significant attention it has garnered in both theoretical and computational research. Our study seeks to answer the following question:

QC. What is the best notion of recency for the RFS task?

To address this question, we first introduce a taxonomy of recency metrics employed in previous ML studies, emphasizing the varied implementations of this concept. We then put forward the following hypothesis:

HC1. Recency metrics that encode *higher-level* distances contribute more to RFS than those based on *lower-level* distances.

To assess HC1, we carry out a series of experiments employing the *Multilayer Perceptron* (MLP) algorithm, as well as an SFS experiment. This is followed by a BF analysis of the results. These analyses aim to ascertain which recency metrics have the most significant contribution to the RFS task.

HC2. The effectiveness of recency metrics can vary depending on corpus-specific characteristics, such as the genre and structure of texts.

To evaluate HC2, we apply the previously mentioned analyses to two corpora that exhibit distinctly different characteristics.

Chapter 6 introduces studies D and E, which explore the importance of paragraph structure – a broader contextual factor– for REG-in-context. The central question posed by the studies in this chapter is:

QDE. Does paragraph structure have an impact on the the choice of RF?

Study D presents an exhaustive corpus analysis of paragraph structure, thoroughly examining both *intra-paragraph* and *inter-paragraph* factors. The term "Intra-paragraph" pertains to factors that affect the internal structure of paragraphs, whereas "inter-paragraph" denotes factors signaling transitions between paragraphs. This study will test the subsequent hypotheses:

HD1. *Paragraph-prominent* entities are substantially more likely to become pronominalized.

HD2. *Paragraph-new* and *paragraph-initial* referring expressions are substantially more likely to be non-pronominal.

HD3. Paragraph-new REs are more likely to be pronominal if the referent is prominent in the current (P_i) and the previous (P_{i-1}) paragraph.

Subsequently, study E addresses question QDE from a computational standpoint. The hypothesis tested in this study is:

> **HE1.** The incorporation of paragraph-related information substantially improves the performance of feature-based REG-in-context models.

To assess this hypothesis, I conduct a feature-based RFS study, incorporating various paragraph-related features. Beyond evaluating the performance of models with and without paragraph-related information, study E also seeks to increase the explainability of the referential form predictions by offering an in-depth *error analysis*.

Chapter 7 introduces two studies, F and G. The former offers a systematic evaluation of different REG-in-context approaches, while the latter delves into an interpretability experiment concerning neural RFS models. Both studies draw from two coauthored published articles (Same et al. 2022, Chen et al. 2021) and are thus written in first-person plural.

Neural models have often replaced classic rule-based and feature-based approaches in recent years. Study F poses the following question:

> **QF.** Do neural REG models live up to the hype?

This study argues that well-designed classic models should not be overlooked, hypothesizing:

> **HF1.** Neural REG models are not always better than rule-based and feature-based models.

To systematically contrast different REG-in-context approaches, we examine two very different English-language datasets, assessing each algorithm through both *automatic* and *human evaluations*. Consequently, this study not only discusses the REG approaches employed, but also emphasizes the choice of corpus.

A widely recognized challenge with neural models is their "black box" nature, which inherently lacks *explainability*. Study G in this chapter is one of the first attempts to bring explainability to Deep Learning (DL) REG-in-context models. A well-established method to determine if a neural model's latent representations encode certain information is *probing*. Consequently, we introduce a suite of probing tasks to inspect neural REG-in-context models. The central question we pose is:

> **QG.** Which linguistic features are encoded by neural models?

The question we pose is inherently exploratory. In our investigation, we focus on various (1) probing tasks, (2) RF classifications, and (3) neural model architectures. We construct eight probing classifiers to discern which linguistic features, influential in RF determination, are learned and captured by neural RFS models. Our neural RFS models are designed to handle three different RFS classification tasks: 2-way (`pronoun`, `non-pronoun`), 3-way (`pronoun`, `description`, `proper name`), and 4-way (`pronoun`, `description`, `demonstrative`, `proper name`). Given the varying complexity of these tasks (for instance, a 2-way classification might be more straightforward than a 4-way classification), we aim to determine if these models capture different contextual features. We are also interested to know how RFS benefits from different neural architectures.

In Chapter 8, I summarize the principal findings of this book and delve into various facets of the REG-in-context task. This includes the significance of selecting appropriate corpora, linguistic features, computational methodologies, and evaluation methods. I also offer a detailed comparison between the one-shot REG and REG-in-context tasks, shedding lights on their commonalities and distinctions. Furthermore, this chapter underscores insights from the linguistic perspective that can be helpful for the computational generation of REs, and vice versa. The chapter concludes by discussing the primary contributions of this work to the study of reference.

2 Choosing referring expressions in context: Linguistic studies

2.1 Introduction

Referring is a fundamental aspect of communication and plays a crucial role in how we discuss and interact with the world around us. In conversations and written texts, we often employ a variety of forms to mention people, objects, and concepts. Intriguingly, the same referent can be addressed in multiple ways, depending on the context, the speaker's intention, and the level of formality or familiarity involved. Consider the following examples to illustrate this diversity in referential expression:

(1) a. **A woman in a black dress** just walked past the building.
b. **The woman you were talking to the other day** just walked past the building.
c. **Emily Smith**, **the famous novelist**, just walked past the building.
d. **Emily Smith** just walked past the building.
e. **Emily** just walked past the building.
f. **The novelist** just walked past the building.
g. **That woman over there** just walked past the building.
h. **She** just walked past the building.
i. **She** entered the alley and ∅ just walked past the building.

The bold referring expressions can all be used in various contexts to refer to the fictional novelist, *Emily Smith*. Below, I outline scenarios where each sentence might be appropriate:

Scenario 1: Sentence 1a describes a situation where the speaker is not familiar with the novelist and simply wants to report an event to the listener.

Scenario 2: Sentence 1c could be used when the speaker knows Emily Smith but is not sure if the listener would recognize her without additional details.

Scenario 3: If both the speaker and the listener know, or are friends with, the novelist, then sentence 1e is suitable.

Scenario 4: When the speaker and the listener are talking about the novelist and she is the topic of their conversation, sentence 1h is appropriate.

This chapter presents various theories surrounding the choice of referential forms and the linguistic factors at play. Section §2.2 introduces linguistic theories that provide insights into the reasons behind choosing specific RFs. These insights connect the choice of RF with the status of a referent in a given context, such as whether a referent is accessible, prominent, or salient. The empirical studies discussed in §2.3 highlight several factors influencing a referent's status. This section does not differentiate between corpus-based analyses and psycholinguistic experiments, focusing instead on the factors themselves rather than the methods employed to study them. This chapter lays the groundwork for Chapters 5 and 6, which further explore the importance of different linguistic features for the REG-in-context task.

2.2 Theories about RF choice

From the provided examples, it becomes clear that multiple ways exist to refer to a specific entity, and the felicity of these REs varies depending on the context. For example, in SCENARIO 4 (1h), using a pronoun is entirely appropriate due to the established topic of conversation. In contrast, employing the RE from (1a) would be inappropriate, and using the RE from (1c) would come across as overinformative and peculiar.

What aspects of these scenarios make the use of different REs appear more or less felicitous? This question might be addressed from a cognitive standpoint. Another approach would be to answer it based on the characteristics of the text or the existing relationships between REs within the text.

Givenness Theory (Gundel et al. 1993) and *Accessibility Theory* (Ariel 1990, 2001) offer cognitive explanations for the choice of referring expressions. Both theories also present a hierarchy that associates various RFs with the cognitive status or accessibility each form mediates. *Centering Theory* (CT) (Grosz et al. 1995) and *Prominence Theory* (PT) (von Heusinger & Schumacher 2019) explain this choice by examining the properties of a text and existing relationships between the REs. I will first provide a brief overview of the two cognitive theories and then delve into centering and prominence perspectives on reference.

Givenness Theory addresses the cognitive status of a referent in an addressee's mind. According to Gundel et al. (1993), there are six distinct cognitive statuses that can explain the use of various RFs. These cognitive statuses and the forms that express them are organized in a hierarchy known as the *Givenness Hierarchy* (see Figure 2.1). The statuses are arranged in such a way that each status also entails all statuses beneath it. Therefore, if a referent is *uniquely identifiable*, it is also necessarily *referential* and *type identifiable*.

in focus	>	activated	>	familiar	>	uniquely identifiable	>	referential	>	type identifiable
{it}		{that / this / this N}		{that N}		{the N}		{indefinite this N}		{a N}

Figure 2.1: The Givenness Hierarchy of Gundel et al. (1993).

The least restrictive cognitive status is termed *type identifiable*. This means that upon hearing the expression, the addressee should be able to recognize the correct type of a referent. For instance, a sentence such as *I bought a car* should provide the addressee with a representation of a car rather than any other object, such as *a bag*. The linguistic form associated with this cognitive status in English is `indefinite determiner + N`.

Conversely, at the opposite end of the spectrum, the most restrictive cognitive status is termed *in focus*. Being "in focus" signifies that the referent is both in short-term memory and is the current center of attention. This status allows the use of pronominal forms.

The following examples from Gundel (2003: 129) are all continuations of the sentence *I could not sleep last night*. These continuations show that different REs are associated with different cognitive statuses according to the Givenness Hierarchy.

(2) a. **A train** kept me awake.
Type Identifiable – identify what kind of thing this is.

b. **This train** kept me awake.
Referential – associate a unique representation by the time the sentence is processed.

c. **The train** kept me awake.
Uniquely identifiable – associate a unique representation by the time the nominal is processed.

d. **That train** kept me awake.
   Familiar – associate a representation already in memory.
e. **This train/this/that** kept me awake.
   Activated – associate a representation from working memory.
f. **It** kept me awake.
   In focus – associate a representation that your attention is currently focused on.

A significant limitation of this hierarchy is its failure to account for REs in the form of a proper name. To address this limitation, Mulkern (1996) expanded the hierarchy to include proper names. She noted that a proper name might be employed if a referent meets the unique identifiability criterion. Typically, the most extended form of the proper name (for instance, the complete name with modification) is used when the referent is first introduced. Mulkern also observed that a single name (either first or last) is commonly used to reference a referent that satisfies, at a minimum, the familiarity criterion.

In combination with Mulkern's extension, the Givenness Theory offers cognitive explanations for the use of various RFs. However, it does not account for the use of modifications, such as in the case of modified NPs.

Accessibility Theory, a cognitive theory rooted in the concept of a referent's accessibility, introduces a more comprehensive hierarchy of referring expressions (Ariel 1990, 2001). Ariel (1990, 2001) posited that each RF encodes a distinct degree of mental accessibility. Furthermore, REs cue the addressee on "how to retrieve the appropriate mental representation in terms of degree of mental accessibility" (Ariel 2001: 31). Figure 2.2 shows the Accessibility Hierarchy, as described in Ariel (2001).

As illustrated in Figure 2.2, the hierarchy provides information regarding the accessibility degrees of proper names, descriptions, demonstrative NPs, and pronouns. Each primary category is further divided into more fine-grained subcategories. Table 2.1 displays the degree of accessibility of proper names, arranged from the least to the most accessible.

Table 2.1: The degree of accessibility of proper names, arranged from the least to the most accessible.

Full name+modifier	<	Full name	<	Last name	<	First name
Joe Biden, the US president	<	*Joe Biden*	<	*Biden*	<	*Joe*

According to Ariel (1990), three criteria – *informativity*, *rigidity*, and *attenuation* – control the linguistic coding of accessibility degrees. These criteria, which

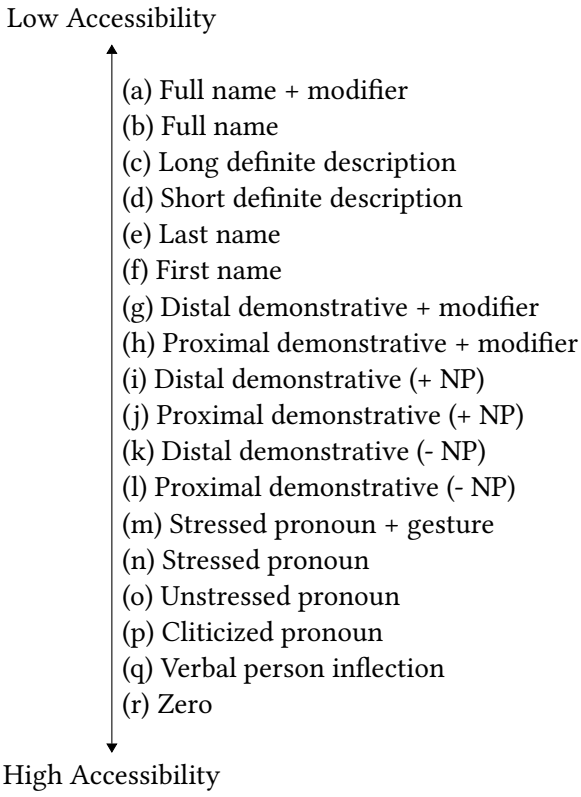

Figure 2.2: The Accessibility Hierarchy of Ariel (2001).

partially overlap, translate the cognitive concept of accessibility into RFs with varying degrees of accessibility.

Based on the informativity criterion, expressions containing more lexical information are employed to retrieve referents with a lower degree of accessibility. Informativity serves as the primary criterion for determining the use of full names (e.g., *Joe Biden*) versus partial names (e.g., *Biden*). Nevertheless, it does not explain the distinction between first names (e.g., *Joe*) and last names (e.g., *Biden*), as both possess comparable amounts of lexical information.

The second criterion is rigidity, which pertains to the uniqueness of an expression. In Western countries, at least, last names (e.g., *Biden*) are more distinctive than first names (e.g., *Joe*). Thus, their difference in accessibility can be attributed to rigidity.

The final criterion is attenuation, which refers to the phonological size of an expression. As per Ariel (1991), attenuation and informativity significantly over-

lap. Nonetheless, this criterion distinguishes REs that are equally informative and rigid. For instance, consider the expressions *the United States of America* and *US*. Both expressions are equally unique and informative, but *the United States of America* is less attenuated, suggesting a lower degree of accessibility. The Accessibility Theory posits that forms that are more informative, more rigid, and less attenuated are utilized to refer to discourse referents less accessible to the addressee.

The Givenness and Accessibility hierarchies provide cognitive explanations for the utilization of various REs, linking this use to the referent's cognitive status in the addressee's mind. A notable strength of these two theories is their comprehensive inventory of RFs, which they associate with cognitive explanations. However, a significant limitation of these theories is their evaluation of each referent's cognitive status or accessibility in isolation, overlooking potential competition within a given context.

Centering Theory, which I present next in this chapter, seeks to address the aforementioned two shortcomings by (1) considering the contextual competition and relational properties of referents and (2) offering concrete implementation rules for predicting RFs. CT models the relationship between the choice of RE, the focus of attention, and coherence within a discourse segment (Grosz et al. 1995). This theory posits that the choice of RE is constrained by the centrality of the discourse referents to an utterance.

Each utterance (U_i) in a discourse segment (D) possesses a *ranked* set of *forward-looking centers* (Cfs). These are the entities that are mentioned in an utterance and are potential candidates to become the center of the next utterance. They are ranked based on their grammatical roles, with subjects usually ranked higher than objects. The highest-ranked Cf in U_i is termed the *preferred center* (Cp).

Furthermore, non-initial utterances have a *backward-looking center* (Cb). This is the entity that is the current center of attention in an utterance and the current utterance is "about." The Cb of U_i is the highest-ranked Cf of its antecedent U_{i-1}, which is realized in U_i. A list of various utterances is provided in Table 2.2, along with their Cf, Cp, and Cb.

In addition, CT develops a typology for transitions from U_{i-1} to U_i based on the interaction between the centers. These transitions can be distinguished by two factors: (1) whether Cb and Cp of U_i are the same, and (2) whether Cb of U_{i-1} and U_i are the same.

Table 2.3 shows that in the *continue* transition, the speaker has talked about an entity in the previous utterance and now continues to talk about it. The Cbs in the previous and current utterances are identical, and the Cb appears to be in the

Table 2.2: Centering Theory and the instantiation of centers. Cf, Cp, and Cb stand for forward-looking, preferred, and backward-looking centers, respectively.

Utterance	Cf	Cp	Cb
[$U1$] Lena is a passionate biker.	LENA	LENA	-
[$U2$] She bikes everyday to the university.	LENA, UNIVERSITY	LENA	LENA
[$U3$] Now, she wants to go on a cycling trip.	LENA, CYCLING TRIP	LENA	LENA
[$U4$] Maria likes to join her.	MARIA, LENA	MARIA	LENA
[$U5a$] She calls her to check the dates.	MARIA, LENA, DATES	MARIA	MARIA
[$U5b$] Nina told her not to.	NINA, MARIA	NINA	MARIA

Table 2.3: Centering transitions (Walker & Prince 1996).

Transition types	Cb(U_i)=Cb(U_{i-1})	Cb(U_i)=Cp(U_i)
Continue	+	+
Retain	+	−
Smooth shift	−	+
Rough shift	−	−

subject position, the highest-ranked grammatical role. In Table 2.2, the transition from $U2$ to $U3$ is a continue transition.

In a *retain* transition, the speaker continues to discuss the same referent as in the previous utterance but intends to shift to a new referent in the subsequent sentence. The change is motivated by positioning the referent in a less preferred grammatical position in the utterance. The transition from $U3$ to $U4$ exemplifies a retain transition.

A *shift* transition causes the Cb of the current utterance to change. If the entity is realized as Cp, the transition is *smooth* and signals that the speaker is interested in continuing to discuss the current Cb. It is a *rough* transition if the entity is realized in a less preferred syntactic position. $U5a$ and $U5b$ represent two different continuations of $U4$. While the transition between $U4$ and $U5a$ is smooth, the transition between $U4$ and $U5b$ is rough.

According to CT, "sequences of continuation are preferred over sequences of retaining and sequences of retaining are to be preferred over sequences of shifting" (Grosz et al. 1995: 17). Therefore, one basis for achieving local coherence would be to shift toward new centers using retention.

The RF choice also plays a crucial role in maintaining coherent discourse through transitions. A key rule in CT is the *pronoun rule.* It posits that if any elements of the Cf of the previous utterance are realized as pronouns in the current sentence, the Cb of the current sentence must also be realized as a pronoun. This rule explains why a continuation like (4a), in which the Cb is pronominalized, reads better than (4b). Moreover, this rule suggests that pronouns are the preferred form of referring when the transition between the two utterances is a continue transition.

(3) Lena invited Maria for dinner.

(4) a. She asked her to be there at 7 p.m.
    b. Lena asked her to be there at 7 p.m.

Further supporting this, experimental studies have demonstrated that when Cb is realized as a non-pronominal form in a continuation scenario, a processing penalty (known as the *Repeated Name Penalty*) occurs (Almor 1999). Moreover, the application of the pronoun rule extends beyond theoretical analysis. It has been implemented in various computational studies (Kibble & Power 1999, Poesio et al. 2004) to enhance the naturalness and coherence of generated texts.

As the previous paragraphs have shown, CT provides concrete rules for predicting the RF based on the coherence of the text and the available transitions between utterances. However, for the implementation to work, CT has reduced the dimension of the referential inventory to only two forms, unlike the previous two cognitive theories. Therefore, the predictive power of CT is limited to the pronominalization problem. In addition to its limited scope, CT can only explain the choice of RF in *local contexts*, that is, in two adjacent utterances. As von Heusinger & Schumacher (2019) point out, however, the ranking of referents in discourse is not limited to the local context; rather, they exhibit a global effect. In what follows, I present the Prominence Theory of von Heusinger & Schumacher (2019). The advantage of this theory over the Accessibility and Givenness theories is that it considers the relational properties of referents. Compared to CT, PT does not limit its scope to the local context but extends its boundaries to a broader context. Moreover, it concretely provides three defining characteristics of a theory of reference.

Based on the concept of prominence discussed in Himmelmann & Primus (2015), PT (von Heusinger & Schumacher 2019) addresses both the *dynamic* and *relational* aspects of referents. Utilizing the following criteria, this theory views prominence as a "structure-building principle" to explain the representation of different referents in discourse (p. 119):

Definition 1 (Singling-out). Prominence is a relational property that singles out one element from a set of elements of equal type and structure.

Definition 2 (Dynamicity). Prominence status shifts in time (as discourse unfolds).

Definition 3 (Structural attraction). Prominent elements are structural attractors; i.e., they serve as anchors for the larger structures they are constituents of, and they may license more operations than their competitors.

The first rule (Def. 1) accounts for the relational nature of reference. It states that the prominence status of a referent is determined by comparison with other elements of similar type and structure, that is, other discourse referents. Thus, in contrast to Givenness and Accessibility, referents are not considered in isolation, but their prominence status is determined in relation to other competing referents.

The second definition (Def. 2) addresses the dynamic nature of referents. As discourse unfolds, the prominence status of referents changes dynamically. This means that the most prominent referent may lose prominence and regain it later in the discourse. Both PT and CT account for dynamicity; however, unlike CT, PT relates dynamicity to a broader context.

The third definition (Def. 3) states that more variation is observed when a referent is prominent. For less prominent entities, we use enriched forms with more semantic content; for more prominent entities, we can use various referential strategies. Thus, a broader inventory of forms is available for reference to prominent entities. Therefore, in line with Givenness Theory, PT can also explain the varied use of referring expressions in context.

Based on the aforementioned three principles, PT successfully combines different aspects of the previous theories of reference, namely, Givenness, Accessibility, and CT. The fact that it considers the relational properties of reference renders it more powerful than the two cognitive theories. With its second rule, the dynamicity principle, PT can explain the shifts that occur in discourse. Unlike CT, it also has the advantage of considering a broader context, which is more suitable for studying natural language.

Another important point is that PT does not introduce a rigid inventory of RFs, nor does it limit its applicability to a specific class of RFs. Instead, it proposes three principles as the structure-building elements of discourse. Consequently, the theory is more adaptable and applicable to a variety of cases. A notable advantage of CT, however, is its definitive rules and straightforward implementation. However, this aspect also introduces limitations. The initial pronominalization

rule of CT heavily relies on a single factor, specifically the grammatical role. Nevertheless, as we will explore, numerous other factors and their interactions play a vital role in predicting RFs.

For the reasons stated above, I employ the terminology of PT in the remainder of this book. Having presented various theoretical explanations, I now turn to the factors influencing the choice of RE. In their work, von Heusinger & Schumacher (2019) describe these features as "prominence-lending cues" because "they boost the prominence value of their respective referent to a certain extent" (p. 119).

2.3 Prominence-lending cues

In the previous section, I mentioned that the RF choice reflects the prominence status of referents in context. The factors influencing prominence have frequently been discussed in both theoretical and empirical studies. A reduced form, such as a pronoun, is often employed to refer to prominent entities, whereas a semantically richer expression is used for less prominent entities. Factors influencing prominence include grammatical function, animacy, recency, thematic role, the presence of competing referents, and coherence relations, among others (Brennan 1995, Fukumura & van Gompel 2011, Arnold & Griffin 2007, Ariel 1990). The impact of these factors on the prominence of referents has been studied either in isolation (Arnold & Griffin 2007) or in combination (Ariel 1990).

In the remainder of this section, I will discuss prominence-lending cues and their influence on the prominence status of referents in context. The cues under discussion in this section include syntax (§2.3.1), thematic role (§2.3.2), givenness (§2.3.3), competition (§2.3.4), animacy (§2.3.5), and recency (§2.3.6).

2.3.1 Syntax

This section highlights the syntax-related factors that influence the prominence status of referents. Below, I discuss the effects of *grammatical role*, *first-mention bias*, and *syntactic parallelism*.

2.3.1.1 Grammatical role

Psycholinguistic studies have shown that subjects are more prominent than objects or adjuncts (Stevenson et al. 1994, Arnold et al. 2000, Fukumura & van Gompel 2010). There are several reasons for this, including (1) subjects often acting as agents in a sentence, (2) subjects being the topic of a sentence, and (3) subjects being mentioned first in a canonical structure in a language like English.

These factors contribute to the prominence of a subject in a sentence. Therefore, the subsequent mention of the subject is more likely to be pronominal (Brennan 1995, Arnold 2008, 2010). For example, as a continuation of sentence (5), sentence (6a) is more natural than sentence (6b). Since *John* is the subject of sentence (5), it is more prominent than *David*, which means that it is more likely to be pronominalized in the following sentence.

(5) $\text{John}_{\text{SUBJ}}$ invited $\text{David}_{\text{OBJ}}$ for dinner.

(6) a. **He** asked David to be there at 7 p.m.
b. **John** asked David to be there at 7 p.m.
c. David asked **him/John** to cook pasta.

2.3.1.2 First-mention bias

In a language like English, the subject is typically the first entity mentioned in a sentence. The entity's prominence may be attributed to being the first-mentioned entity in the sentence, rather than because of its role as subject. Gernsbacher (1989) argues that the first-mention position confers an advantaged cognitive status on entities. The addressees construct a mental representation of the information they receive. The referent presented in the first position serves as the foundation for this mental representation. Kaiser & Trueswell (2011) conducted a series of sentence-completion and eye-tracking experiments in Finnish, demonstrating that grammatical role and order of mention are two independent factors influencing the choice of RF.

2.3.1.3 Syntactic parallelism

Syntactic parallelism may also increase the prominence of a referent and the likelihood of pronominalization. Pronouns, as reduced forms, are more likely to be used when the target referent occupies the same syntactic position as its coreferential antecedent. Therefore, if we choose to continue sentence 5 by discussing *John* in the subject position (sentence 6a), there is a strong preference for using a pronoun compared to a scenario where *John* appears as the object of the sentence (sentence 6c).

2.3.2 Thematic role

Several experimental studies have examined the effect of *thematic role* alternation on reference production (Stevenson et al. 1994, Arnold 2001, Fukumura &

van Gompel 2010, Rosa 2015, Vogels 2019). These studies investigated (1) whether certain thematic roles increase the likelihood of a referent being mentioned again in a subsequent context, known as *the next mention bias*, and (2) whether certain thematic roles enhance the likelihood of pronominalization.

Stevenson et al. (1994) tested various thematic role pairs in two text–completion experiments and discovered that after a goal–source sentence, people showed a preference for continuing the text with the goal referent rather than the source referent. In other thematic role pairs, the patient was preferred over the agent, and the stimulus was preferred over the experiencer. However, the study identified no significant effect of thematic role on the choice of RF. The critical factor in choosing the RF was the first-mention bias, that is, whether the antecedent referent was mentioned first or second in the sentence.

In the same vein, Arnold (2001) examined the role of thematic roles in reference production, focusing specifically on the thematic roles goal and source. The experiments employed *transfer to possession verbs*, such as *send–receive*. The advantage of using these verbs lies in the fact that for some, like *send*, the subject is the source, while for others, such as *receive*, the subject is the goal. Consequently, both the source and the goal appeared in the subject position. The following examples from Arnold (2001) demonstrate the source–goal and goal–source conditions.

(7) a. [Source–Goal] The drama club was worried that no one would come to the opening performance of their play. Everyone agreed to try to get all their friends to come. **Erin**$_{\text{SOURCE}}$ sent an invitation to **Bill**$_{\text{GOAL}}$.

b. [Goal–Source] Getting a telegram always scares me. It has to be either great news or awful news. **Juan**$_{\text{GOAL}}$ received a telegram from **Claire**$_{\text{SOURCE}}$ when their mother died.

Similar to Stevenson et al. (1994), Arnold (2001) demonstrated that speakers tend to refer to goal entities more frequently than to source entities. Both studies suggest that end-states, or what Stevenson et al. termed *consequences*, are the most prominent elements in a sentence. Consequently, the most predictable next mentions are end-state entities, such as the goal in source–goal sentences and the patient in agent–patient constructions.

Contrasting with Stevenson et al., Arnold observed an effect of thematic roles on referential form choice: speakers used pronouns for goal entities more often than for source entities. However, since the likelihood of continuing with the subject referent is much higher than that of continuing with the goal referent,

the subject bias is stronger. In other words, thematic roles influence referent accessibility only in situations where other factors, such as subject effects, are less dominant (Arnold 2001).

Fukumura & van Gompel (2010) replicated this line of experiments, employing stimulus–experiencer verbs within implicit causality contexts. They observed a next-mention bias for stimulus entities but found no significant effect on pronominalization, aligning with the findings of Stevenson et al. (1994).

(8) a. [stimulus–experiencer (SE)] **Glen**$_{\text{STIMULUS}}$ annoyed **July**$_{\text{EXPERIENCER}}$ when the two-minute silence took place in the yard. This was because...

b. [experiences–stimulus (ES)] **Glen**$_{\text{EXPERIENCER}}$ despised **July**$_{\text{STIMULUS}}$ when the two-minute silence took place in the yard. This was because...

According to the experimental results outlined in this section, thematic role influences the likelihood of a referent appearing as the next mention, but its impact on the RF choice is open to debate.

2.3.3 Givenness

Gundel (2003) defines *referential givenness* as "a relation between a linguistic expression and a corresponding non-linguistic (conceptual) entity in (a model of) the speaker/hearer's mind" (p. 125). According to Gundel et al.'s Givenness Hierarchy, as presented in §2.2, different RFs convey varied information about the presumed cognitive status of entities in the addressee's mind. The speaker assesses whether the addressee has a mental representation of a referent and selects forms accordingly. If the speaker presumes the referent is new to the addressee or its mental representation is inactive, a full form, such as a proper name or description, is used. Conversely, if the entity is the focus of the current sentence or the speaker believes the addressee is familiar with it, a reduced form is employed for reference.

As implied in the previous paragraphs, the *newness–givenness* distinction is not binary and should be considered a gradient notion. Various studies have adopted parameters such as degree of salience, familiarity (Prince 1992), accessibility (Ariel 2001), activation, and identifiability (Chafe 1976) as key characteristics of the newness-givenness distinction. Chafe (1976) initially made a distinction based on the identifiability of referents, that is, whether or not the mental representations of referents are identifiable to the addressee. Subsequently, he assigned three activation states to each class: `given`, `accessible`, `new`. Prince (1992),

on the other hand, distinguished two levels: (1) *hearer-old* versus *hearer-new*, and (2) *discourse-old* versus *discourse-new*. According to this distinction, an entity can be both new and given on two different dimensions: it can be new in discourse but old to the hearer. Suppose (9) is the introductory sentence of a sports article. *Lionel Messi* is the first mention of the Argentinian soccer player in the text, making it discourse-new. However, it is very likely that the reader of a sports magazine is already familiar with *Lionel Messi*, making this referent hearer-old.

(9) **Lionel Messi** joined Paris Saint-Germain. He ...

Although it is highly likely that the addressee is familiar with the target in the previous example, there is still the possibility that the referent is unknown to the addressee. According to Baumann & Riester (2012: 127), "persons, places or other entities are rarely ever objectively known or unknown but only with respect to some intended recipient". Generally, speakers and writers do not have access to the thoughts of the addressee, especially when addressing a large audience. Since a simplified notion of givenness will be explored in the following chapters, I will not delve further into the intricacies of givenness and its various interpretations.

2.3.4 Competition

Generally, the stronger the competition between a referent and other referents, the lower the likelihood of using an attenuated RE for that referent. In this section, I discuss two forms of competition, namely *gender* and *additional character* effects.

2.3.4.1 The gender effect

Studies have shown that the likelihood of employing attenuated REs diminishes when a referent of the same gender is present in the immediate context of the target referent. One possible explanation could be the desire to avoid ambiguity. In order to circumvent ambiguity in these situations, the speaker opts for more specific forms, such as proper names and descriptions (Karmiloff-Smith 1985, Arnold & Griffin 2007, Fukumura et al. 2013, Rosa 2015). Consider the following:

(10) a. Mary had an appointment with John. **She** turned up half an hour late.
b. Mary had an appointment with Emily. **She** turned up half an hour late.

The referents in (10a) possess different genders, and the pronoun *she* unambiguously refers to *Mary*. Conversely, in (10b), both referents share the same

gender. Therefore, the pronoun becomes ambiguous, potentially referring to either *Mary* or *Emily*.

Arnold & Griffin (2007) consider semantic competition as another plausible explanation for the increased use of more specific forms in gender-congruent settings. According to this perspective, a same-gender competing referent is semantically more similar to the target than an opposite-gender referent. Consequently, it is likely that same-gender referents experience higher competition, leading to a reduced prominence status for the target. Fukumura et al. (2013) tested this theory in an experiment conducted in Finnish, a gender-neutral language where the same pronoun (*hän*) is used for both males and females. They observed a lower frequency of pronouns when both referents shared the same gender.

Similar to Arnold & Griffin (2007), Fukumura et al. (2013) suggest that a competing referent, which is semantically very similar to the target referent, impairs the memory retrieval of the non-linguistic representation of the target. Consequently, speakers resort to using more specific forms to resolve this interference. "The fact that gender congruence reduced the use of Finnish pronouns suggests that gender is one of the non-linguistic properties that speakers take into account, even when the language does not express the referent's gender and hence the presence of a same-gender competitor does not make the use of a pronoun ambiguous" (Fukumura et al. 2013: p, 1017). In summary, the gender effect can be attributed to several factors, most notably ambiguity avoidance and semantic competition.

2.3.4.2 The additional character effect

In addition to the gender effect, Arnold & Griffin (2007) mention a second type of competition, hereafter referred to as the additional character effect: the presence of additional characters in the immediate context of the target, regardless of whether the target and competitors share the same gender, reduces the likelihood of pronominalization. In a storytelling experiment, participants observed two-panel cartoons and listened to the first sentence under two conditions: (1) the sentence included *solely* the main referent [Condition A], and (2) alongside the main character, another character of a different gender was present [Condition B].

(11) a. **Mickey** went for a walk in the hills one day. [Condition A]
b. **Mickey** went for a walk **with Daisy** in the hills one day. [Condition B]

Participants were instructed to recite the first sentence and continue the story with a second sentence. The results of the experiment showed that in condition A, the likelihood of using a pronoun for the main referent was greater than in condition B. Unlike the gender-related effect, the outcome observed in this experiment was not due to ambiguity avoidance. According to Arnold & Griffin (2007), participants were 30% more inclined to use *Mickey* instead of *he* in Condition B as opposed to Condition A. In condition B, the two characters engaged in discourse share the available attentional resources, resulting in diminished activation for each within the speaker's internal representation. Thus, the speaker tends to use a more specific form to activate the representation of the target referent.

This explanation aligns with the semantic competition effect. Although the semantic similarity between the two gender-incongruent characters is lower than that of the same-gender referents, they still share identical animacy values. The similarity between the two characters might intensify competition and decrease the target referent's prominence status. Therefore, animacy can also be regarded as a factor associated with semantic competition.

2.3.5 Animacy

Various linguistic theories propose that *animate* entities hold more prominence than *inanimate* entities (Comrie 1989, Aissen 2003). This prominence influences numerous linguistic choices. For instance, animate entities are more likely to be chosen as the subject or topic of a sentence compared to inanimate entities (Givón 1983, Dahl & Fraurud 1996). Dahl & Fraurud (1996) examined the effect of animacy on the choice of RF in a Swedish text corpus, revealing that in 36% of instances, pronouns were used to refer to third-person human referents. By contrast, only 8% of pronominal instances referred to non-human referents. However, distinguishing between the effects of subjecthood and animacy in such a corpus study is challenging. To clarify, it is necessary to determine whether the increased frequency of pronominalization is attributable solely to animacy, or whether animate referents tend to occupy the subject position, which inherently carries a greater likelihood of pronominalization.

To unravel these effects, Fukumura & van Gompel (2011) investigated the impact of animacy on RF choice through a series of controlled story-completion experiments. Regarding (12a), the study discovered that speakers tended to use pronouns for animate referents, specifically *the hikers*, more frequently than for inanimate referents, like *the canoes*. This pattern persisted even when the conditions were reversed, as in (12b), where the positions of the NPs were switched:

speakers more frequently used pronouns for animate objects over inanimate subjects. Consequently, the influence of animacy is distinct from that of the grammatical role. The elevated rate of pronominalization for animate referents suggests their greater conceptual prominence in the speaker's mind. As a result, the speaker needs to retrieve less semantic content to refer to them (Fukumura & van Gompel 2011, Vogels 2014).

(12) a. The hikers carried the canoes a long way downstream. Sometimes, ...
b. The canoes carried the hikers a long way downstream. Sometimes, ...

Fukumura & van Gompel also investigated whether pronominalization rates were affected by the presence of animacy-congruent competitors in the previous sentence. They observed a decline in the likelihood of pronominalization when both the referent and its competitor were animate. However, this effect did not occur among inanimate referents.

The diminished use of pronouns when animate competitors are present may be partially attributed to the semantic competition described in §2.3.4. This explanation is partial because animacy congruence affects only the pronominalization ratio for animate referents. The presence of an animate competitor diminishes the prominence of the animate target referent. Consequently, the speaker employs more explicit forms to activate the referent's representation. This effect is absent in congruent pairs of inanimate referents.

2.3.6 Recency

Recency is defined as the distance between the current mention of a referent and its antecedent. The larger the distance between the two mentions, the more likely it is to use a full noun phrase anaphora (Vonk et al. 1992, Givón 1992, Arnold 2010). Conversely, the smaller the distance between the two mentions, the more likely it is to use pronouns. This section outlines the three most common interpretations of recency employed in linguistic and computational studies.

2.3.6.1 Immediate context

Research focusing on the occurrence of pronominal forms typically defines *short distance* or *immediate context* as scenarios where the antecedent appears in the same sentence or is separated by just one sentence. On the other hand, if the antecedent is positioned more than one sentence away, it is categorized as *long distance* (Hobbs 1978, Ariel 1990, Hitzeman & Poesio 1998, Poesio et al. 2004).

The corpus analysis conducted by Hobbs (1978) revealed that in 98% of instances, the antecedent of a pronoun anaphora is found in either the preceding or the same sentence. Similarly, Ariel (1990) investigated the distribution of pronouns, proper names, descriptions, and demonstratives in a corpus analysis and found that in over 80% of the cases, pronouns exhibit a preference for short distances, meaning, antecedents located within the same sentence or just one sentence away.

2.3.6.2 Non-local context

Different lines of study examine recency in a broader context. These studies incorporate the concept of *non-local context*, that is, a larger span of text, in their definition of recency. In a comprehensive study of topic continuity in discourse by Givón (1983), the measurement of distance to the previous mention extended as far back as 20 clauses. This research represents one of the initial efforts to quantify the role of distance in discourse. McCoy & Strube (1999) proposed in a computational pronominalization study that "when the last mention of an item is several sentences back in the text, a definite description is preferred" (p 64). Using a corpus of New York Times articles, the study revealed that definite descriptions were nearly always used in long distance situations.

2.3.6.3 Unit boundary

While the distance patterns described in the previous paragraphs can explain many instances of pronominalization, Fox (1987b) argues that these patterns do not encompass all varieties of anaphoric references. The study by Fox demonstrates that pronouns can be used to refer to a referent over long stretches of distance until the goal of a narrative changes (cited in Smith 2003). Building on this idea, Ariel (1990) introduces the concept of *unity*, defined as an antecedent existing within the same frame, segment, or paragraph. Additionally, Vonk et al. (1992) and Tomlin (1987a) highlight the significance of *episode* and *unit boundaries*, typically regarded as *paragraph boundaries* in written texts, as contributing factors to the principle of recency. In summary, this section has elucidated three distinct interpretations of recency. The first two involve measuring distance in sentences or clauses, whereas the third interpretation extends beyond the sentence level, emphasizing paragraphs.

2.4 Summary and discussion

This chapter introduced several theories concerning the choice of REs. The Givenness and Accessibility hierarchies offer cognitive explanations for how the cognitive status and accessibility of discourse referents' mental representations affect this choice. However, a limitation of these cognitive theories is their failure to account for the relational properties of discourse referents.

In addition to these two dominant theories, Centering Theory elucidates various pronominalization decisions by associating the choice of REs with the coherence of discourse. However, CT in its initial form adopts a more localized perspective and overlooks the global context in the choice of referring expression forms.

A more recent approach, Prominence Theory, seeks to delineate the choice of REs, incorporating the dynamicity and relational properties of referents. Furthermore, this theory extends the interpretation of REs beyond merely local contexts. Prominence Theory posits that the choice of referential form is shaped by multiple factors, also termed prominence-lending cues.

In §2.3, I explored a variety of prominence-lending cues. These cues exhibit differences in multiple aspects. Notably, Prat-Sala & Branigan (2000) identified two distinct types of accessibility: *inherent* and *derived*. Inherent accessibility originates from the inherent characteristics of a referent and remains unchanged throughout the discourse. This category includes cues like animacy or gender, as a referent's animacy status or gender remains unchanged in a text. Conversely, derived accessibility varies within the discourse and responds to contextual factors. This form of accessibility emerges from the prominence of a referent within the discourse context. An example of this is subjecthood, since a referent does not possess innate subjecthood but becomes a subject in a sentence. A valid question to ask is which interpretation holds greater significance in determining the form of referring expressions.

In this chapter, six important prominence-lending cues have been discussed, along with their various interpretations and implications. I have not discussed other prominence-lending cues such as coherence relations (Hobbs 1979, Kehler 2002) and information status (Lambrecht 1994) because they are not discussed in the subsequent chapters. Although I presented and discussed these cues individually, it is likely that a combination of them plays a role in determining the prominence status of the referents. As noted, "they might weigh in differently in different contexts" (De la Fuente 2015: 24). Chapter 5 examines different implementations of these factors in a feature-based REG-in-context experiment to assess their importance in predicting the form of REs in context.

3 Generating referring expressions in context: Computational studies

3.1 Introduction

Chapter 2 explained the linguistic theories pertaining to the choice of referring expressions and the factors that influence that choice. The current chapter concentrates on computational theories of reference. Chapters 2 and 3 collectively establish the context for the studies detailed later in this work.

This chapter begins with an introduction to Natural Language Generation in §3.2 and delineates the primary subtasks of an NLG pipeline. Subsequently, I explore the task of Referring Expression Generation in §3.3. Two distinct subtasks of REG, specifically one-shot REG and REG-in-context, receive comprehensive explanations in §3.3.1 and §3.3.2. The aim is to understand the methodologies employed in addressing and resolving REG challenges. In §3.4, I introduce the various approaches for evaluating REG algorithms, highlighting the most commonly employed evaluation techniques. The chapter concludes in §3.5 with a discussion of the most important aspects of the REG-in-context task that need further examination.

3.2 Natural language generation

NLG is concerned with the generation of natural language text from non-linguistic input (Gatt & Krahmer 2018). These systems are used in practical applications, such as the generation of weather forecasts (Mei et al. 2016), clinical reports (Portet et al. 2009, Gatt et al. 2009), and soccer reports (van der Lee et al. 2017).

The primary subtasks of NLG systems encompass *content determination*, *text structuring*, *sentence aggregation*, *lexicalization*, *referring expression generation*, and *linguistic realization* (Reiter & Dale 2000). Consider the objective of automatically generating a report on the Brazil–Germany match in the 2014 FIFA World Cup semifinals. The following describes the pipeline structure for generating such a report:

Content determination: Initially, one must decide which information to include in the report. Match statistics encompass details about corners, passes, fouls, and more. However, it might not be necessary to detail every pass and foul in the report. After deciding which content to include, raw data is transformed into data objects or messages for inclusion in the final output.

Text structuring: The aim of this task is to establish the sequence for presenting the information. It is logical to begin a soccer report with general information, such as the location and time of the game, followed by significant events (e.g., goals, penalties) in chronological order. In this phase, a discourse, text, or document plan is developed, serving as a structured and ordered representation of the messages (Gatt & Krahmer 2018). Content determination and text structuring constitute the *macro-planning* phases of the pipeline, where decisions regarding *what to say* are made.

Sentence aggregation: Mapping messages to sentences in a one-to-one ratio results in excessively lengthy and challenging-to-read text. Aggregating relevant sentences allows for the construction of more complex sentences. For instance, Toni Kroos scored a goal in the 24th minute of the Brazil vs. Germany match and another in the 26th minute. These two goal-scoring events can be synthesized into a single, more concise sentence.

(1) Toni Kroos scored in the 24th and 26th minutes.

Lexicalization: A single event can often be verbalized in multiple ways. For instance, the scoring event in (1) might be expressed as *to score a goal* or *to kick a goal*. These verbalization choices occur during the lexicalization stage.

Referring expression generation: This task involves choosing expressions to refer to domain entities. The process resembles lexicalization; however, the chosen expression must differentiate the target referent, Toni Kroos, from other referents. Therefore, if Toni Kroos is not a prominent referent in the context, using his name is advisable. Conversely, if he is prominent, such as being mentioned in the preceding sentence, a pronoun may be more appropriate. Sentence aggregation, lexicalization, and referring expression generation constitute the *micro-planning* stages of the pipeline, where decisions are made regarding *how to express* the content.

Linguistic realization: The final step involves combining all the selected words into well-formed sentences. For an in-depth overview of the NLG subtasks outlined above, see Reiter & Dale (2000) and Gatt & Krahmer (2018).

It is important to note that the sequence and number of tasks employed in various NLG modular systems can vary considerably. Additionally, a specific task might be segmented into subtasks and executed at different stages of the generation process (Mellish et al. 2006). However, the modular pipeline structure described above is merely one of the many approaches to NLG. For information on alternative NLG architectures, such as planning-based methods, refer to Gatt & Krahmer (2018).

A particular NLG architecture that has become popular in recent years is the neural end-to-end data-to-text approach. With the rapid advancement of neural methods, direct mapping from input to output has become feasible (Goodfellow et al. 2016, Goldberg 2017). These models learn input-output mappings directly and rely much less on explicit intermediate representations such as the ones outlined previously (Castro Ferreira et al. 2019). Despite the increasing popularity and efficacy of end-to-end NLG approaches, pipeline-based methods remain widespread (Gatt & Krahmer 2018). For one, they align well with linguistic and psycholinguistic research (van Gompel et al. 2019). They also predominate in commercial applications of NLG (for further argumentation, see Reiter 2016b, 2017). Furthermore, a systematic comparison of pipeline and end-to-end REG systems has shown that "having explicit intermediate steps in the generation process results in better texts than the ones generated by end-to-end approaches" Castro Ferreira et al. (2019: 552). For these reasons, it is crucial to have a thorough understanding of the subtasks of the modular architecture.

This book focuses on the REG subtask, which is one of the essential steps in the micro-planning stage. Additionally, REG can be relatively easily separated from a specific application domain and studied on its own (Gatt & Krahmer 2018). Therefore, several stand-alone solutions to the REG problem exist, making the evaluation of such models feasible and necessary. In the rest of this chapter, the focus will be exclusively on different aspects of the REG step.

3.3 Referring expression generation

REG studies address one of these two tasks (Gatt et al. 2014):

- *Conceptualization* – the choice of properties to represent in an RE that is a full definite noun phrase.

- *Choice of (anaphoric) REs* in discourse (e.g., full definite NP, reduced NP, pronoun).

Suppose you wish to identify the figures enclosed by the black rectangle in both visual scenes of Figure 3.1: In the context of the first visual scene, the property profession, as indicated in the referring expression *the judge*, is sufficient for identifying the figure. In the second visual scene, an additional distinctive feature such as wig color is necessary to provide a distinguishing description: *the judge in a dark wig*. The conceptualization task focuses on selecting the attributes required to create unique descriptions. The descriptions generated by this task are also known as *one-shot*, *non-anaphoric* expressions. As mentioned in Chapter 1, this task is referred to as one-shot REG.

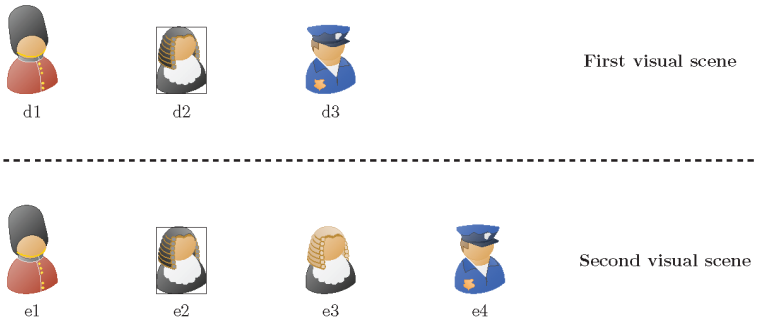

Figure 3.1: Referents in two different visual scenes. To describe `d2`, the RE *the judge* is distinctive, while a more distinctive RE such as *the judge in a dark wig* is needed to describe `e2`.

Now, imagine the enclosed figure of the second visual scene is already prominent in the discourse, for example, through its mention in (2a). A reduced NP, such as *the judge*, or the pronoun, *him*, can then be used to refer to the judge `e2` in (2b).

(2) a. **The judge in a dark wig** asked the police about the crime scene.
b. The police's response made **the judge/him** very concerned.

Here, the decision concerning the form and content of a referring expression is influenced by a range of factors that affect the prominence status of a referent. Owing to the contextual nature of these decisions, as previously discussed in Chapter 1, this task is known as REG-in-context. §3.3.1 introduces several well-known approaches to one-shot REG, and §3.3.2 provides an in-depth overview of methods for handling REG-in-context.

3.3.1 One-shot REG

In this section, I concentrate on one-shot REG. The discussion begins with an explanation of the task and then transitions to the primary methods employed in the task, along with its various extensions.

One-shot REG is defined as follows: Given a target object `r` in a finite domain `D` (`r` ∈ `D`), the goal is to find a set of attribute–value pairs, `L`, whose conjunction is true of the target, but not of any of the distractors. `L` is referred to as *a distinguishing description* of the target (Dale & Reiter 1995, Krahmer & van Deemter 2019). For instance, consider the context of Figure 3.2, where the objective is to identify a distinguishing description for duck `d1` (`r=d1`). The finite domain and attributes are outlined below:

$$\text{D}=\{\texttt{d1} : \texttt{d7}\}$$
$$\text{Attributes}=\{\texttt{body color}, \texttt{hair color}, \texttt{bill color}\}$$

Figure 3.2: Set of ducks.

If the sole objective is identification, a fault-proof strategy involves combining all the properties of duck `d1` to create a distinguishing description, assuming the referent is distinguishable. In such a scenario, *the duck with a blue body, black hair, and a green bill* serves as an apt description for `d1`. However, in addition to creating a distinctive description, it is also important to consider *humanlikeness*. The vast majority of REG solutions attempt to generate referring expressions that closely resemble those of humans (van Deemter et al. 2012, van Deemter 2016). Below, some of the most seminal approaches are explored.

The Full Brevity Algorithm *The Full Brevity Algorithm* produces the shortest distinguishing description for an intended referent (Dale 1989). This algorithm adheres to Grice's maxim of quantity (Grice 1975) and consistently uses the minimum number of properties necessary to identify a given referent. As a result, the algorithm generates *the duck with a green bill* as the distinguishing description of duck `d1`. However, this algorithm is not widely implemented due to two primary reasons: First, conducting an exhaustive search for the shortest distinguishing description is not computationally efficient; second, humans often produce non-minimal descriptions (Krahmer & van Deemter 2019).

The Greedy Heuristic Algorithm *The Greedy Heuristic Algorithm* is an extension of the Full Brevity Algorithm, which attempts to generate humanlike expressions with more redundant properties (Dale 1989). This algorithm first determines the property of the referent that excludes the most distractors. It then adds this property to the description. In subsequent iterations, the algorithm evaluates which of the remaining properties excludes the most distractors. This process continues until no distractors are left. The algorithm lacks *backtracking*; that is, once a property is added to the description, it remains, even if it becomes redundant. To describe `d3`, the first attribute eliminating the most distractors (four ducks) is skin color. Bill color, possessing the next highest discriminatory power, excludes the remaining two distractors. Consequently, the distinguishing description generated by this algorithm is *the blue duck with a red bill*.

Although the Greedy Heuristic Algorithm may generate more natural results, its performance is not yet comparable to that of humans. According to Dale & Reiter (1995), humans might begin to utter an RE before they have fully scanned the set of distractors. Moreover, several studies have shown that people tend to use color attributes more frequently than any other attribute (Pechmann 1989, Viethen et al. 2017), even in situations where the color does not lead to a discriminative description. If one shows individuals a picture that includes a white bird, a black cup, and a white cup, and asks them to come up with a distinguishing description for the white bird, they tend to refer to it as *the white bird*, even though *the bird* is sufficient (Pechmann 1989). *The white bird* in this scenario exemplifies an *overspecified* description, where an excess of attributes is used.

The Incremental Algorithm and its extensions As just mentioned, people exhibit preference for certain properties over others when referring to entities. For instance, to refer to one of the ducks described earlier, people may prefer body color over bill color. This preference order constitutes one of the parameters of the *Incremental Algorithm* (IA) (Dale & Reiter 1995). A simplified representation of IA is illustrated in Algorithm 1 (Krahmer & van Deemter 2019).

Suppose we want to generate a description of `d3` in the context of Figure 3.2. The input of the algorithm is an object `r`, a domain `D` consisting of all objects `d1`:`d7`, and a list of preferred attributes shown in the first line of the algorithm 1. Assume that `Pref` = body color > hair color > bill color. Line (2) shows that the description is initialized with an empty set. As shown in (3), the context set `C` of distractors (everything except `d3`) is initialized as `D`−{`d3`}. In (4), the algorithm iterates over the list of attributes listed in `Pref`, and for each attribute, it looks up the value of the target referent (5). It then checks whether this attribute–value

Algorithm 1: Sketch of the core Incremental Algorithm (Krahmer & van Deemter 2019).

```
1  Incremental Algorithm ({r}, D, Pref){
2  L ← ∅
3  C ← D-{r}
4  for each A_i in list Pref do
5      V = Value (r, A_i)
6      if C ∩ RulesOut(⟨ A_i, V ⟩) ≠ ∅
7      then L ← L ∪ {⟨ A_i, V ⟩}
8               C ← C − RulesOut(⟨ A_i, V ⟩)
9      endif
10     if C ≠ ∅
11     then return L
12     endif
13
14 return failure }
```

pair excludes any of the distractors (6). The function `RulesOut` ($\langle$ A_i, V $\rangle$) returns a set of objects that have different values for the attribute A_i than the target object. If one or more distractors are excluded, the attribute–value pair $\langle$ A_i, V $\rangle$ is added to the description under construction (7), and a new set of distractors is computed (8). Body color is the first attribute to be considered, for which `d3` has the value `blue`. According to this rule, all ducks except `d1` and `d5` are excluded and the attribute–value pair ⟨bodyColor, blue⟩ is added to the description `L`. The new set of distractors is `C` = {`d1`, `d5`}, and the next attribute ⟨hair color⟩ is tested. Since they all have the same hair color, no distractor can be ruled out, so the attribute–value pair ⟨hairColor, black⟩ is not included in the description. Next, the third attribute is checked. The target's bill is red while the remaining distractors' bills are not, so the attribute–value pair ⟨billColor, red⟩ is also included. At this point, all distractors have been ruled out (10), a set of properties has been discovered that uniquely characterize the target, and the task is complete (11). The algorithm would have failed if it had reached the end of the `Pref` list without eliminating all the distractors (14).

The Incremental Algorithm does not incorporate backtracking, enhancing its computational efficiency. Simultaneously, this approach permits redundant descriptions that are psycholinguistically more plausible and align more closely with human-produced language. Consequently, IA has become the most widely

implemented REG algorithm. However, the original IA, along with all its predecessors, is not capable of generating more complex expressions, such as references to sets of objects. Here is a brief overview of some of the extensions to IA.

Van Deemter (2002) enhanced the IA algorithm, enabling it to generate expressions with negated properties (3), expressions referring to sets of objects (4), and expressions containing a logical disjunction of properties (5). These algorithms operate in stages, attempting to generate a longer disjunction of properties when shorter descriptions fail to create a distinguishing description. The subsequent expressions refer to the various ducks depicted in Figure 3.2.

(3) [d4]: The yellow duck that does not have black hair

(4) [d1, d3, d5]: The blue ducks

(5) [d1, d4]: The duck with a green bill and the duck with gray hair

A more recent extension of IA considers the conceptual naturalness of sets and generates set expressions that are conceptually coherent. The idea behind this approach is that the felicity of a description is influenced by the conceptual relatedness of the elements of the set (Gatt & van Deemter 2007b). Imagine a set consisting of three people: an Italian composer, a Greek chef, and a German engineer. The phrase *the Greek and the German* forms a more coherent expression than *the Greek and the engineer*, since the former is derived from a single coherent perspective, namely the nationality of the set's members (Gatt & van Deemter 2007b). In contrast, the latter phrase is derived from two perspectives. Various extensions of IA have been developed to incorporate different dimensions of RE generation into the algorithm. For a more comprehensive overview, refer to Viethen (2011), Krahmer & van Deemter (2012), and van Deemter (2019).

The studies mentioned thus far are rule-based, algorithmic solutions proposed for attribute selection in the one-shot REG task. In what follows, I present two other approaches that tackle one-shot REG differently: firstly, a machine learning experiment conducted by Jordan & Walker (2005) for automatic attribute selection; and secondly, research focused on the TUNA corpus – a semantically and pragmatically transparent corpus comprising identifying references to objects within visual domains (van Deemter et al. 2006). The TUNA studies emphasize the empirical evaluation of one-shot REG algorithms against human-produced data.

Jordan & Walker (2005) conducted their REG study on the COCONUT corpus using a machine learning approach. This corpus comprises computer-mediated dialogs between two individuals. Each participant had a virtual budget and a list

of furniture inventories. Their task was to collaboratively purchase furniture to furnish two rooms. Jordan & Walker (2005) used RIPPER, a program capable of automatically deriving rules from observations (Cohen 1996), to infer rules. Three sets of features were tested both separately and in combination in their experiment: (1) *contrast set* factors, inspired by the IA (Dale & Reiter 1995), (2) *conceptual pact* factors, drawing on the lexical alignment model of Brennan (1996), and (3) *intentional influence* factors, based on a model by Jordan & Walker (2000). As a baseline for their experiment, the generator simply predicted the most frequent attribute combinations. All systems outperformed the baseline system significantly. Jordan & Walker (2005) concluded that "the choice to use theoretically inspired features is validated, in the sense that every set of cognitive features improves performance over the baseline." Combining the features of all three systems raised the accuracy of their model to 60%. This study is not only among the pioneering machine learning experiments in the REG domain but also underscores the importance of linguistically motivated features in REG studies.

A comprehensive assessment of one-shot REG was conducted in the TUNA project. By applying the above-mentioned algorithms to a semantically and pragmatically transparent corpus known as TUNA, Gatt et al. (2007) investigated whether IA matched speakers' behaviors better than other algorithms. Data were collected through a web experiment in which participants described either singular or plural targets in the presence of six other distractor items. For the sake of generality, objects from two different domains were included, namely constructed images of furniture and actual photographs of people (van Deemter et al. 2006).

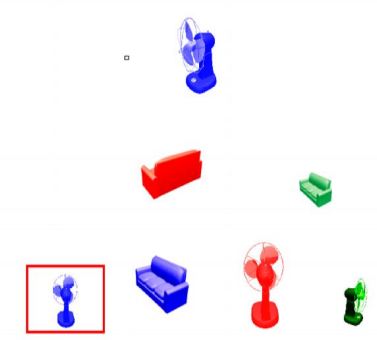

Figure 3.3: An example scene from the TUNA corpus' object domain.

Figure 3.4: An example scene from the TUNA corpus' people domain.

The results of the evaluation demonstrated that the performance of IA was entirely dependent on the preference order established for the attributes (Gatt et al. 2007). TUNA represents one of the initial systematic efforts to evaluate NLG algorithms against human-informed decisions. Moreover, it highlights the significance of corpus-driven approaches in assessing REG algorithms. An additional contribution of TUNA to REG studies was the introduction of the first REG shared tasks, establishing a platform for various research groups to propose solutions to a given problem.

More recent approaches in one-shot RE generation place a greater emphasis on probabilistic methods. Notable examples of such models include *the Probabilistic Referential Overspecification model (PRO)* by van Gompel et al. (2019), the *Rational Speech Act* (RSA) model by Frank & Goodman (2012), which conceptualizes the production and comprehension of REs within a Bayesian framework, and the RSA-based model by Degen et al. (2020) for generating overspecified REs.

As highlighted at the beginning of this section, one-shot REG approaches do not account for discourse context in generating distinguishing descriptions. Considering our daily communication, however, it becomes evident that most REs we produce are embedded within a discourse context. In the remainder of this chapter, I will focus exclusively on the task of REG-in-context.

3.3.2 REG-in-context

As mentioned in Chapter 1, Belz & Varges (2007) defines REG-in-context in this way: "given an intended referent and a discourse context, how do we generate appropriate referential expressions (REs) to refer to the referent at different points in the discourse?" (p. 9). This task can be subdivided into two subtasks: (1) determining the form of the RE, and (2) deciding the content of the RE. The initial step involves determining the form. For example, when referring to Joe Biden at a specific point in a discourse, one must decide whether to use a proper name (*Joe Biden*), a description (*the president of the United States*), a demonstrative form (*this person*), or a pronoun (*he*). The subsequent step involves determining the content, namely, selecting from the various ways a particular RF can be realized. For example, in generating a description of Joe Biden, one must choose whether to mention only his job (for example, *the president entered the Oval Office*) or to include the country as well (for instance, *the president of the United States arrived in Cornwall for the G7 Summit*). These tasks are defined as follows:

Referential Form Selection (RFS): Given a text whose REs are yet to be generated, and given the intended referent for each of these REs, the task of RFS involves developing an algorithm that identifies the appropriate referential

form (RF) from a set of K candidate RFs. RFS presents a classification challenge, where the algorithm's role is to select a referential class from a predefined set of classes. For instance, in a pronominalization task, two classes exist: pronominal and non-pronominal forms (K=2), and the RFS task determines the suitable form to use.

Referential Content Selection (RCS): Given a text whose REs are yet to be generated, and given the intended referent for each of these REs, the RCS task entails building an algorithm that generates all these REs.

The primary focus of this book is the RFS task, although the RCS task also receives attention in Chapter 6. In what follows, I will outline three main approaches that have gained popularity in REG-in-context generation over the last three decades: rule-based, feature-based, and neural network-based models.

In the 1990s and early 2000s, rule-based models were at the forefront of research in REG, utilizing linguistic theories to formulate rules for generating REs (Dale 1989, Passonneau 1996, McCoy & Strube 1999, Henschel et al. 2000). This approach is discussed in detail in §3.3.2.1.

Feature-based ML models have been prevalent since the mid-2000s. The GREC shared tasks (Belz et al. 2010), introduced in Chapter 1, sparked a plethora of feature-based models for REG-in-context (Hendrickx et al. 2008, Bohnet 2008, Greenbacker & McCoy 2009a). This approach is discussed in depth in §3.3.2.2. Rule-based and feature-based models are typically considered two-stage methods, separately addressing RFS and RCS.

More recently, various neural network-based REG models have emerged that are capable of generating REs in an E2E manner, eliminating the need for feature engineering (Castro Ferreira et al. 2018a, Cao & Cheung 2019, Cunha et al. 2020). This method is detailed in §3.3.2.3.

3.3.2.1 Rule-based approach

One of the earliest REG-in-context algorithms was implemented in EPICURE, an NLG pipeline for generating cooking recipes (Dale 1989, 1992). The concepts of *local* and *global focus*, as introduced by Grosz et al. (1983), were fundamental to the REG component of this system. Local focus encompasses the lexical, syntactic, and semantic content of the utterance being generated, whereas global focus pertains to the semantic representation of the recipe as a whole. The discourse center, a concept detailed in Centering Theory (Grosz et al. 1983) and elaborated upon in Chapter 2, is another critical parameter in this system. Within the domain of recipes, the center is defined as the outcome of the preceding operation.

For example, following the utterance *chop the onion*, the center becomes the *onion*. Pronominalization, the algorithm's initial rule, allows the subsequent mention of the center to be pronominalized. Similar to Grosz et al. (1983), Dale (1989) also concluded that other entities presented in local focus can be pronominalized, provided the center itself is a pronoun.

After completing the pronominalization stage, the next step involves selecting appropriate definite descriptions for the remaining referents in the sentence. This process adheres to two principles, akin to Grice's conversational maxims (Grice 1975). The principle of *adequacy* requires that the intended RE should be unambiguous. Conversely, the principle of *efficiency* dictates that REs should not include more information than necessary. Merging these principles yields the shortest distinguishing description, paralleling the use of the Full Brevity Algorithm discussed earlier. Comprehensive details of EPICURE's implementation are available in Dale (1992).

According to Reiter & Dale (2000), two types of errors can occur in such an algorithm: (1) *missed pronouns*, that is, the algorithm decides not to use a pronoun even though it is perfectly acceptable; and (2) *inappropriate pronouns*, that is, when a pronoun is ambiguous in context. To reduce these errors, Passonneau (1996) took the model of Dale (1992) as a baseline and supplemented it with the full focus-structure information of CT. Passonneau suggested that integrating centering constraints could relax the principles outlined in Dale (1992). Consequently, the model could reduce the aforementioned errors. Passonneau (1996) tested the model on a corpus of narratives known as PEAR STORIES (Chafe 1980) to determine whether it could accurately predict the use of pronouns, minimal descriptions, and overspecified descriptions. This study was the first corpus-based investigation of REG-in-context (Viethen 2011). Passonneau discovered that this model was a more effective predictor of minimal and overspecified REs than the model in Dale (1992).

McCoy & Strube (1999) adopted a more critical stance, arguing that backward-looking centers are not typically pronominalized in most natural contexts. They concluded that, in addition to cues such as distance from the antecedent and the presence of competing referents, information about the temporal structure is also crucial in REG accounts. This is because temporal changes in discourse often lead to the use of overspecified NPs. The authors distinguish between four types of temporal cues, driven both by semantic cues in the text (for example, adverbial time phrases) and by changes in verb tense.

According to Henschel et al. (2000), temporal changes are not always critical in all text genres, such as descriptive texts, and are thus not essential in algorithms applied to these texts. Their algorithm is founded on the concept of local

focus, that is, a set of referents eligible for pronominalization. The local focus for each utterance is calculated and defined as the set of referents from the previous utterance that are either discourse-old or realized in the subject position. Theoretically, the local focus set can comprise multiple members; however, in most instances, it consists of a single member, aligning with the backward-looking center as defined in CT. When the target referent does not align with any of the local focus criteria, additional factors such as recency and competition are considered to decide if a pronominal form is appropriate. Algorithm 2 illustrates the implementation proposed by Henschel et al. (2000).

Algorithm 2: Henschel et al.'s (2000) algorithm

Let X be a referent to be generated in utterance ($u2$), and $focus$ be the set of referents of the previous utterance ($u1$) which are
    (a) discourse-old, or
    (b) realized as subject.
X has an antecedent beyond a segment boundary — DESCRIPTION
X has an antecedent two or more utterances distant — DESCRIPTION
X has an antecedent in ($u1$), and
    X occurs in strong parallel context — PRONOUN
    $X \notin focus$ — DESCRIPTION
    $X \in focus$ and
        X has a competing referent $Y \in focus$ — DESCRIPTION
        X has a competing referent Y in ($u1$) amplified with apposition or non-restrictive relative clause — DESCRIPTION
        else — PRONOUN

Henschel et al. (2000) validated their assumptions using the GNOME corpus which comprises texts that describe museum objects and patient information leaflets (Poesio 2004, Poesio et al. 2004, Di Eugenio et al. 1997). Their research primarily addressed the pronominalization problem, where the task is to decide between the use of pronouns and non-pronominal forms. A notable similarity between this algorithm and those previously discussed is the treatment of salience as a *black and white* concept, that is, a referent is deemed either salient or not salient.

Krahmer & Theune (2002) incorporated a graded concept of salience into their incremental algorithm for generating *context-sensitive* REs. Instead of solely focusing on generating unique descriptions that differentiate a referent from distractors, this algorithm also considers the salience of the referent in its choice of

RE. Krahmer & Theune (2002) assigned a salience weight (`sw`) ranging from 0 to 10 to each entity, based on the topic/focus distinction of Hajičová (1993) and the Centering Theory of Grosz et al. (1995).

The algorithm generates context-sensitive expressions by modifying the third line of algorithm 1 (see page 41). It narrows down the distractors to those domain elements whose salience weight is equal to or greater than the salience weight of the target `r`: the third line $C \leftarrow D - \{r\}$ is altered to $C \leftarrow \{x | w(x) \geq sw(r)\} - r$. Figure 3.5 presents a simplified version of (3.2), showcasing attributes such as body color, hair color, and bill color.

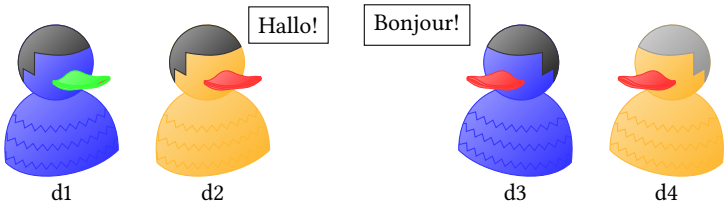

Figure 3.5: A simplified version of Figure 3.2 for generating context-sensitive REs.

Imagine we are at the beginning of a conversation and wish to refer to duck `d2`. As the ducks have not yet been mentioned, they all possess identical salience weights. Therefore, a distinguishing description for `d2` would be *the yellow duck with black hair*, aligning with the classic Incremental Algorithm (IA) prediction.

Now, consider a scenario where based on previous discussions, we know only ducks `d2` and `d3` are capable of speaking. Suppose `sw(d2)=sw(d3)=10` and `sw(d1)=sw(d4)=0`. The salience weights of `d2` and `d3` are highest, rendering the other two ducks non-salient. To refer to the German-speaking duck (`d2`), the contrast set has only one member: `C={d2, d3}-{d2}={d3}`. In this context, selecting the body color attribute is adequate for generating a distinguishing description, making *the yellow duck* an appropriate expression.

According to Krahmer & Theune (2002), integrating CT with the concept of the continuous decrease in salience for unmentioned entities facilitates the generation of context-sensitive REs. Moreover, they noted that pronouns are employed when a referent is the sole salient entity in the discourse context. Hence, this algorithm is capable of generating both context-sensitive descriptions and pronominal forms. The models discussed in this section rely on explicit rules formulated by researchers. While these rules are essential for defining and implementing the models, they might restrict the development of more complex models.

3.3.2.2 Feature-based approach

Feature-based machine learning algorithms deduce rules and learn generalizations from feature-value pairs extracted from corpora. Stoia et al. (2006) developed a dialog agent that leveraged machine learning to provide real-time instructions in a 3D virtual environment. They trained a decision tree to determine the appropriate use of determiners (for example, *the*, *that*, *a*), the type of head noun (for example, pronoun), and whether to include a modifier. The training incorporated features pertinent to dialog history (such as the number of mentions), the referent's position and its relationship with other entities in the visual scene, and its inherent properties (for example, semantic type). This study was among the first to consider both linguistic and non-linguistic factors as predictors of RF.

The initial systematic studies that focused directly on the REG-in-context task were the GREC shared tasks, as documented in (Belz & Kow 2010, Belz et al. 2010). The primary objective of GREC was to explore methods for generating appropriate references to entities in contexts extending beyond a single sentence. The theoretical motivation behind GREC centered on understanding *which types of information* might influence the choice of REs in context. Two distinct corpora, GREC-2.0 and GREC-PEOPLE, were utilized in the GREC shared tasks. Both corpora comprised introductory sections of Wikipedia articles. The GREC shared tasks addressed both RFS and RCS.

3.3.2.2.1 The RFS task in GREC

The GREC RFS task involved a 4-way classification challenge, where participating systems were required to predict the most suitable referential form – be it a pronoun, a zero form, a description, or a proper name – to refer to a given referent within a specific discourse context. The submissions for these shared tasks predominantly employed feature-based algorithms such as *C5.0* decision trees (Greenbacker & McCoy 2009a, Orăsan & Dornescu 2009), *Conditional Random Field* (Bohnet 2008), and *Multilayer Perceptron* (Favre & Bohnet 2009). Additionally, these models leveraged a broad spectrum of features, including the encoding of local context (Hendrickx et al. 2008, Favre & Bohnet 2009), recency (Jamison & Mehay 2008), subjecthood and parallelism (Greenbacker & McCoy 2009a), and competition (Jamison & Mehay 2008).

3.3.2.2.2 The RCS task in GREC

The GREC corpora provide a range of alternative REs, including the original RE found in the corpus, for each reference slot. Table 3.2 displays the initial mention

of James Joyce as it appears in the corpus, represented as *James Augustine Aloysius Joyce*, along with a set of alternative REs corresponding to this reference slot.

Table 3.1: Alternative set of REs for the referent James Joyce

James Augustine Aloysius Joyce himself
James Augustine Aloysius Joyce
James Joyce himself
James Joyce
Joyce himself
Joyce
he himself
he
who himself
who
_ (null RE)

In the shared tasks focused on content selection, the participating systems employed a two-step methodology: (1) determining the RF, and (2) selecting the actual RE from a set of alternatives using various heuristic rules. For instance, Greenbacker & McCoy (2009a) opted for the longest non-emphatic string for the first mention and the shortest non-emphatic string for subsequent mentions when a proper name was predicted as the RF.

Table 3.2 displays the original text alongside output generated by a system named WLV (Orăsan & Dornescu 2009). This model inaccurately predicted null REs in two instances where the use of null references is typically not allowed. As demonstrated in the original sentence of (6a), employing an overt subject (for example, *he*) is required, yet WLV incorrectly predicted a null form.

(6) a. ORIGINAL: After working his way up to production chef at Dean and DeLuca, he worked under James Burns, the acclaimed head chef at Charleston's J. Bistro.

b. WVL: After working his way up to production chef at Dean and DeLuca, - worked under James Burns, the acclaimed head chef at Charleston's J. Bistro.

Table 3.2: An example showing a Wikipedia document on the left and the output generated by WLV on the right, with the target REs highlighted in bold.

ORIGINAL
Sam Talbot (born December 27, 1977) is a Sicilian-American chef from Charlotte, North Carolina, best known as a semi-finalist on Season 2 of Bravo's Top Chef, eventually placing third. **He** also became the fan favorite for the season. **Sam** received **his** education from Johnson & Wales University in Charleston, South Carolina. After working **his** way up to production chef at Dean and DeLuca, **he** worked under **James Burns**, the acclaimed head chef at Charleston's J. Bistro. In Charleston, **he** also met **Sarah Vida who** later became his business partner in Williamsburgh Cafe after **he** moved to New York City. In New York City, **he** has held the position of executive chef at several restaurants, such as the Black Duck, Williamsburgh Cafe, and Punch. Recently, **he** was planning on opening a restaurant named Spitzer's Corner, but when **he** could not agree with **business partners Will and Rob Shamlian** with the direction of the restaurant, **they** separated. **Talbot** is a diabetic and - wears an insulin pump attached to **his** leg. **Talbot** returned to Top Chef for a special "Four Star All Stars" episode along with **Elia Aboumrad**, **Marcel Vigneron**, and **Ilan Hall**.
WLV
Sam Talbot (born December 27, 1977) is a Sicilian-American chef from Charlotte, North Carolina, best known as a semi-finalist on Season 2 of Bravo's Top Chef, eventually placing third. **he** also became the fan favorite for the season. - received **his** education from Johnson Wales University in Charleston, South Carolina. After working **his** way up to production chef at Dean and DeLuca, - worked under **James Burns**, the acclaimed head chef at Charleston's J. Bistro. In Charleston, **he** also met **Sarah Vida she** later became **his** business partner in Williamsburgh Cafe after **he** moved to New York City. In New York City, - has held the position of executive chef at several restaurants, such as the Black Duck, Williamsburgh Cafe, and Punch. Recently, **he** was planning on opening a restaurant named Spitzer's Corner, but when - could not agree with **business partners Will and Rob Shamlian** with the direction of the restaurant, **they** separated. **he** is a diabetic and - wears an insulin pump attached to **his** leg. **Sam** returned to Top Chef for a special "Four Star All Stars" episode along with **Elia Aboumrad**, **Marcel Vigneron**, and **Ilan Hall**.

3.3.2.2.3 Post-GREC studies

In a more recent study, Kibrik et al. (2016) carried out a multifactorial feature-based analysis, incorporating factors from four primary categories: inherent properties of the referent, factors related to the anaphor, factors concerning the antecedent, and the distance between the anaphor and its antecedent. They trained their decision trees on a subset of data from the RST Discourse Treebank (Carlson et al. 2002). A critical observation made by Kibrik et al. (2016) is that human reference production is not entirely categorical or deterministic. Often, more than one RE can be aptly used to refer to an entity at a specific point in the discourse.

The feature-based REG-in-context studies thus far have predominantly relied on written corpora with a single gold standard RE per reference slot. However, this approach presents conceptual challenges, as different authors may opt for varied REs for the same target slot. The non-deterministic generation of REs has been more extensively explored in one-shot REG studies (Gatt et al. 2013, van Deemter 2016, van Gompel et al. 2019), and to a lesser extent in the realm of REG-in-context. Notably, two REG-in-context studies that acknowledged the non-deterministic nature of reference choice are Castro Ferreira et al. (2016a) and Castro Ferreira et al. (2016b).

Castro Ferreira et al. (2016a) developed the VAREG corpus which comprises REs generated by multiple participants in identical contexts. They used the *normalized entropy* metric to analyze the variance in RF choices among different participants, uncovering significant individual differences. For instance, they observed more variation in RF choice when the referent was in the object position. In a subsequent study, Castro Ferreira et al. (2016b) employed the Jensen-Shannon Divergence metric to compare the similarity between distributions produced by humans and those predicted by models. However, the appeal of their non-deterministic approach is offset by the extensive time commitment required to compile a corpus of parallel human judgments. The VAREG corpus, for instance, includes 36 different texts with annotations limited to references to the main topics. For larger-scale projects involving more texts and referential annotations, replicating this experimental approach with numerous human participants to produce REs is often impractical.

3.3.2.3 End-to-end neural network-based approach

Both rule-based and machine learning feature-based studies in REG typically adhere to a two-step generation process. However, the swift advancements in neural methodologies in recent years have enabled the direct mapping of input to

output, bypassing intermediate steps. This end-to-end (E2E) approach offers a significant benefit over feature-based methods as it eliminates the need for extensive feature engineering.

The concept of neural E2E REG was first introduced by Castro Ferreira et al. (2018a). They developed their models, referred to as NEURALREG, using a REG dataset extracted from the WEBNLG corpus. This corpus was established by Gardent et al. (2017) to assess the performance of NLG systems. It was created via a crowd-sourcing experiment where participants were tasked with writing descriptions for given *Resource Description Framework* (RDF) triples. In these experiments, an RDF triple consists of three elements: a subject, a predicate, and an object. The subject and object are either constants or Wikipedia entities, and the predicate elucidates their relationship. Further details about the WEBNLG corpus are provided in Chapter 7.

The NEURALREG models employed an *encoder–decoder* framework, encoding the target referent and its contextual environment into a unified vector representation. This representation was then decoded into an RE specifically tailored to the discourse context. Castro Ferreira et al. (2018a) explored three distinct decoding architectures: a *sequence-to-sequence decoder* (Seq2Seq), a *concatenative attention mechanism* (CAtt), and a *hierarchical attention mechanism* (HierAtt). Their findings indicated that all the neural models surpassed the established baselines in performance, with the CAtt model showing the most effective results, closely followed by the HierAtt model.

While the NEURALREG models demonstrated significant capabilities, they were constrained by their inability to handle unseen entities. To overcome this limitation, Cao & Cheung (2019) introduced the PROFILEREG model, specifically designed to address this challenge. As indicated by its name, the model constructs profiles for entities featured in the WEBNLG dataset. These profiles are composed of the first three sentences from Wikipedia articles about these entities, providing a foundational context from which REs for unseen entities can be generated.

The PROFILEREG model employs three *bidirectional Long Short-Term Memory* (LSTM) encoders to process pre-context, post-context, and entity profiles. Additionally, it uses a unidirectional LSTM decoder for the generation of REs. This approach enables the model to generate REs for both previously encountered (seen) and novel (unseen) entities. The performance of PROFILEREG surpasses that of NEURALREG.

To address the challenge of unseen entities, Cunha et al. (2020) introduced a copy mechanism that enables the transfer of tokens from the input representations of target entities directly to the output. In addition to this mechanism, they incorporated gender and type information into their model's input. This

strategy is grounded in the notion that gender information can aid in generating appropriate pronominal forms, while type information is useful in generating descriptive REs for unseen entities. Notably, their findings revealed a surprising outcome: the simplest baseline model, named ONLYNAMES, which essentially copies a Wikipedia ID (representing the entity name) into the RE slot, performed on par with, or in some cases, even surpassed the more complex neural models. This unexpected result calls for a reevaluation of the effectiveness of end-to-end (E2E) models in this domain. Given the simplicity yet high efficacy of the ONLYNAMES baseline, it raises an important question regarding the potential superiority of well-designed rule-based or feature-based models over the E2E models previously mentioned. This observation suggests that the complexity of a model does not always correlate with its performance in the task of REG.

3.4 Evaluation methods

Van der Lee et al. (2019) assert that the necessity of evaluating the output of NLG systems is undisputed, yet they note, "what is perhaps more contentious is the way in which evaluation should be conducted" (van der Lee et al. 2019: 355). Broadly, there are two primary methods for evaluating a system's performance. The first method is automatic evaluation. Given that most REG-in-context studies are corpus-based, their outputs can be compared against the original texts in the corpus, often referred to as the *gold standard*. An alternative approach is human evaluation, where individuals assess various aspects of the generated text. This section aims to introduce some prevalent automated evaluation metrics and then provide an overview of human evaluation methods, underscoring their significance in REG studies.

3.4.1 Automatic evaluations

A straightforward method for evaluating an algorithm's output compared to the gold standard involves assessing the similarity of the predictions to the original values. *Accuracy*, representing the fraction of correct predictions made by the model, is one such measure. In the GREC shared tasks, two forms of accuracy – *type accuracy* and *string accuracy* – were employed. Type accuracy measures the percentage of correct RF type predictions, while string accuracy calculates the accuracy of the predicted strings (Belz & Kow 2010).

Other commonly utilized metrics include *precision* and *recall*. Precision assesses the proportion of positive identifications that are actually correct, whereas

recall quantifies the proportion of actual positives that were correctly identified. These metrics are largely deterministic, although the choice of RE often is not, as multiple referential options can be valid (see van Gompel et al. 2019 for an extensive discussion on probabilistic modeling and evaluation).

BLEU and *NIST*, initially developed for machine translation, have also been adapted for REG evaluations (Belz & Kow 2010, Belz 2008, van der Lee et al. 2019, Gatt & Belz 2009). BLEU measures the n-gram overlap between strings (Papineni et al. 2002), while NIST, a BLEU variant, places greater emphasis on less frequent n-grams, assuming they carry more information (Krahmer & van Deemter 2019). The *Levenshtein edit distance* is another metric, calculating the minimum number of edits needed to transform a generated string into the original string. A smaller Levenshtein distance is preferable, indicating fewer steps required for the transformation.

While automatic metrics are advantageous for being "fast, cheap, and repeatable" (Reiter & Belz 2009: 555), they are not without drawbacks. In their work, van der Lee et al. (2019) point out the lack of interpretability of these metrics, noting that different types of incorrect outputs can yield similar scores. Additionally, these metrics may assign low scores to correct but uncommon expressions.

String-based metrics, for example, exhibit the issue of scoring the expression *the delicious pie* as being more similar to *the delicious pig* than to *the tasty pie*. As another example, in the context of the second visual scene from Figure 3.1, *the policeman* and *the policewoman* differ only by a Levenshtein distance of two, despite the former being the only appropriate description for target `e4`. Conversely, *the man on the right* and *the policeman* have a Levenshtein distance of 14 but both accurately describe target `e4`. This inconsistency in automatic metrics not aligning with human judgments, as reported by Reiter & Belz (2009) and Belz et al. (2010), has spurred growing interest in human evaluation methods in NLG.

3.4.2 Human evaluations

As a complement or alternative to automatic evaluations, human judges are often employed to assess the texts generated by NLG systems. In many REG studies, the primary evaluation criterion is the quality of the text, including aspects like fluency and naturalness (also known as humanlikeness). However, as van der Lee et al. (2019) point out, measuring these criteria is challenging due to the lack of uniform definitions for terms like fluency or quality. Human evaluations of generated expressions typically employ both *intrinsic* and *extrinsic* methods (Belz & Reiter 2006, Reiter & Belz 2009).

Extrinsic evaluation focuses on the impact of the generated text on other tasks. An example of this can be found in the study by Belz et al. (2010), which involved a comprehension experiment. Participants read the texts and then answered three comprehension questions. The underlying hypothesis was that suboptimal REs could negatively affect comprehension ease, thereby impacting reading speed and comprehension accuracy. This experiment measured reading time, response speed, and answer accuracy. Despite its potential for providing more concrete results, extrinsic evaluation remains rare in the field, with most studies concentrating on intrinsic assessments (van der Lee et al. 2019).

Intrinsic methods, on the other hand, directly evaluate the attributes of the generated outputs. For example, human judges might be asked to assess the fluency and naturalness of an NLG system's output. In one of the GREC shared tasks, participants conducted a rating experiment, evaluating each text in terms of clarity, fluency, and coherence (Belz et al. 2010: 312). The criteria were defined as follows:

Referential clarity: It should be easy to identify who or what the referring expressions in the text are referring to. If a person or other entity is mentioned, it should be clear what their role in the story is. So, a reference would be unclear if an entity is referenced, but their identity or relation to the story remains unclear.

Fluency: A referring expression should *read well*; i.e., it should be written in good, clear English, and the use of titles and names should seem natural.

Structure and coherence: The text should be well structured and well organized. The text should not just be a heap of related information, but build from sentence to sentence to a coherent body of information about a topic.

In the experiment conducted as part of the shared task evaluation, participants were presented with 24 texts, encompassing both generated and original versions. They were tasked with rating these texts on a five-point Likert scale based on specified criteria, viewing each text independently without the need for direct comparison between original and generated texts.

In the NEURALREG study referenced in §3.3.2.3, Castro Ferreira et al. (2018a) implemented a 7-point Likert scale experiment to assess their models. Participants evaluated the texts – including the original, baseline, and experimental models – based on three criteria: *fluency* (assessing the natural flow and readability of the text), *grammaticality* (evaluating the absence of spelling or grammatical errors), and *clarity* (determining the effectiveness of the text in expressing the intended

data). According to the results of this human evaluation, all neural models surpassed the baselines in these criteria. However, only their best neural model, CAtt, achieved a significantly higher performance than the baselines.

Preference judgment tasks represent another approach to human evaluation. This method was employed by Belz et al. (2010) in their GREC-NEG'09 shared task evaluation. Participants were shown a random selection of texts from various systems alongside the original Wikipedia texts. They were asked to indicate their preference between the versions in terms of fluency and clarity. Additionally, the participants quantified the strength of their preference using a slider mechanism. An illustration of this type of task can be seen in Figure 3.6.

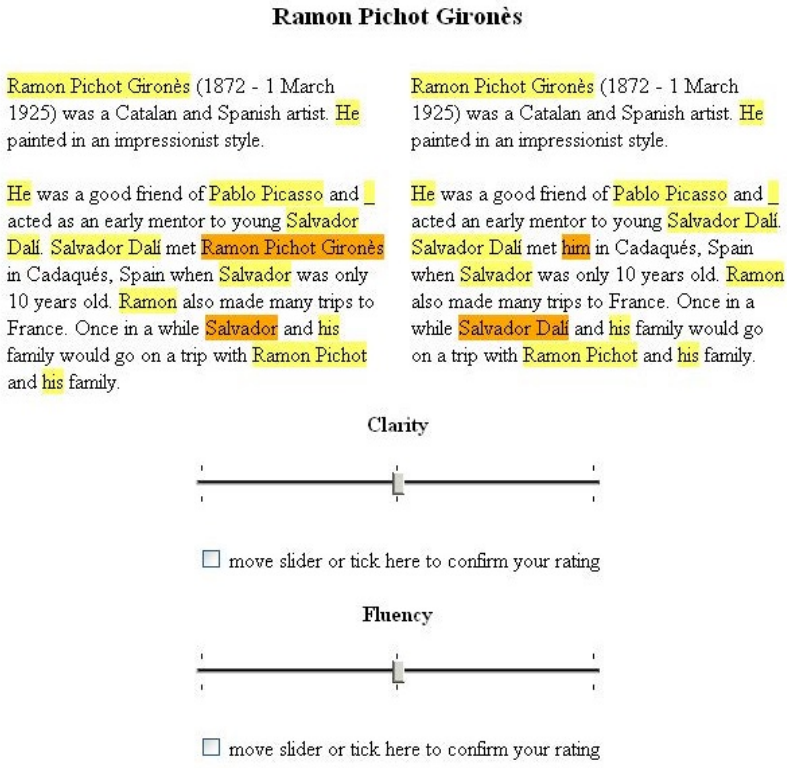

Figure 3.6: Example of a text pair presented in the preference judgment experiment conducted by Belz et al. (2010).

Cao & Cheung (2019) implemented a preference judgment task in their study, where participants were shown outputs from their experimental model alongside a comparison model (either the original text or one of the baselines). The

participants were asked to assess whether the texts were equally good or if they had a preference for one over the other. The evaluation criteria focused on fluency, readability, and grammaticality. In comparing their model's output with the original texts, it was found that 86 out of 100 texts were identical or very similar to the original. While this suggests the models' proficiency in generating high-quality REs, it is important to note, as will be further discussed in Chapter 7, that the WEBNLG dataset used in their experiment predominantly features simplistic texts, potentially limiting referential variability.

This section has provided a concise overview of various automatic and human evaluation methods used in NLG. Howcroft et al. (2020) conducted an analysis of human evaluations in 165 NLG papers and concluded two key points: (1) there is a notable lack of standardized practices in human evaluations within the NLG field, and (2) the information provided in NLG papers regarding human evaluations is often incomplete. These findings underscore the necessity for standardization in human evaluation tasks and improved reporting of human evaluation outcomes in NLG research.

3.5 Summary and discussion

The previous sections chronologically presented REG-in-context studies, starting with rule-based approaches, progressing to data-driven models, which began with feature-based approaches and concluded with E2E neural models. Henceforth, rule-based and feature-based methods are referred to as *classic* approaches.

E2E studies in this chapter suggest that E2E models (1) generally outperform classic REG-in-context models, (2) eliminate the need for feature engineering, and (3) enable direct input-to-output mapping. These claims prompt the question: Why should classic REG-in-context approaches still be considered? The following discussions aim to address this query.

Methodology concerns The automatic evaluation results from Cunha et al. (2020) and human evaluations by Castro Ferreira et al. (2018a) raise doubts about the aforementioned claims. The former study's baseline performance was comparable to that of their neural models. In the latter, while neural models scored higher, the significant difference from the baseline was limited to only one neural model and a single criterion. These mixed results underscore the need for a systematic comparison of the three approaches.

Baseline concerns In their study, Castro Ferreira et al. (2018a) used two baselines, ONLYNAMES and FERREIRA. The uninformed baseline, ONLYNAMES, simply replaced underscores with whitespaces in entity Wikipedia IDs (for example, *Joe_Biden* was converted to *Joe Biden*) to fill reference slots, not generating pronominal REs. The informed baseline, FERREIRA, inspired by Castro Ferreira et al. (2016b), used only three features, namely sentence information status (determining if the RE is new at the sentence level), text information status (determining if the RE is new at the text level), and grammatical role (identifying whether the RE functions as a subject, object, or possessive determiner). The adoption of these baseline models in subsequent studies raises an important consideration: What would be the impact on the outcomes if these baselines were replaced with stronger baseline models? This question underlines the potential for further exploration and refinement in the approach to baseline model selection in REG-in-context studies.

Corpus concerns A significant limitation in the field of E2E REG-in-context studies is the predominant use of a single dataset, specifically WEBNLG. This reliance prompts a critical question: Can the findings derived from this dataset be reliably generalized to other data types, encompassing various genres and complexities? Additionally, the WEBNLG dataset used in the study by Castro Ferreira et al. (2018a) comprised "78,901 referring expressions to 1,501 Wikipedia entities, of which 71.4% (56,321) are proper names, 5.6% (4,467) pronouns, 22.6% (17,795) descriptions, and 0.4% (318) demonstrative REs" (p. 1961). This distribution indicates that a significant majority (over 90%) of the data used for training their models consists of non-pronominal REs.

However, this composition contrasts with common linguistic usage, where pronouns are frequently employed in speech, as discussed in the linguistic studies of Chapter 2. This discrepancy leads to a crucial question: Is the WEBNLG dataset sufficiently representative and versatile to serve as the foundation for E2E REG-in-context model training? The potential mismatch between the dataset's composition and natural language usage poses a challenge to the dataset's suitability for effectively training models that reflect real-world linguistic patterns.

The concerns discussed in the context of end-to-end neural models bring to light broader considerations essential to any REG-in-context study. These key considerations include: (1) the choice of appropriate corpora, (2) the choice of informed rules or features, and (3) the choice of a suitable approach for the task at hand. The subsequent chapters are structured to address these issues.

In Chapter 4, a replication of the GREC systems is applied to three distinct corpora to evaluate their suitability for the REG-in-context task. This chapter

aims to provide insights into the effectiveness of these corpora in capturing the nuances of referring expression generation.

Chapter 5 delves into a comprehensive analysis of the features used in earlier feature-based machine learning studies. The goal is to identify which linguistic features are most influential and effective in contributing to the REG-in-context task.

Chapter 6 shifts the focus to paragraph-related features, an area that has not been extensively explored in previous research. This chapter evaluates the significance of these features in the context of REG-in-context studies.

Finally, Chapter 7 presents a systematic comparison of various REG-in-context approaches, using two markedly different datasets. This comparative analysis is supplemented by both automatic and human assessments of the outputs generated by the models.

Through this structured approach, the upcoming chapters aim to deepen the understanding of the REG-in-context field, addressing the aforementioned concerns and contributing to the ongoing development and refinement of REG-in-context methodologies.

4 The choice of corpora for the REG-in-context task

4.1 Introduction

NLP often relies heavily on the use of language corpora. Employing the terminology of inferential statistics, corpora serve as a *sample* from a broader (data) *population* that NLP researchers aim to understand. The hope is that patterns identified in the sample will be applicable to the wider population (i.e., a specific kind of language use). However, this assumption is only valid if the sample is *representative* of the target population. Otherwise, insights gained from the corpus may not be generalizable to the full spectrum of relevant scenarios. These issues lie at the heart of inferential statistics, yet their applicability to NLP has not been thoroughly explored.

Despite extensive research on the REG-in-context problem, the selection of appropriate corpora for this task remains under-discussed. This chapter aims to assess the suitability of various corpora by presenting a case study focused on REG-in-context models. These models are tasked with selecting the form of REs in context, or as defined earlier in Chapter 3, RFS.

A central question in this chapter is the impact of corpus selection on REG-in-context studies. I hypothesize that the corpora used in prior REG-in-context studies are not adequate for the task, and the lessons learned from these studies may not be generally valid. The term *previous REG-in-context studies* refers specifically to the GREC shared tasks (Belz et al. 2010).[1] In this context, I examine two corpora used in these shared tasks, GREC-2.0 and GREC-PEOPLE, and argue that they are not suitable for the task at hand.

To determine if the findings from GREC are generally valid (i.e., valid for the intended data population) and to provide a wider perspective, a third corpus, the WSJ portion of OntoNotes (Hovy et al. 2006, Weischedel, Ralph et al. 2013), is included in the analysis after additional enrichment. To conduct a comprehensive evaluation of the three corpora (GREC-2.0, GREC-PEOPLE, and WSJ), I replicate the

[1]This chapter does not cover studies such as Kibrik et al. (2016) and Castro Ferreira et al. (2016b).

systems from the GREC shared tasks. I use the same ML algorithm and features as the original systems and train the models on all three corpora.[2]

This chapter is organized as follows: §4.2 provides an overview of the GREC shared tasks and the corpora involved. §4.3 offers a detailed examination of the WSJ corpus and its enrichment for this study. §4.5 details study A, focusing on a systematic reconstruction and evaluation of various RFS models using the three different corpora. Finally, §4.6 summarizes the results and findings.

4.2 GREC corpora and models

As described by Belz et al., "the GREC tasks are about how to generate appropriate references to an entity in the context of a piece of discourse longer than a sentence" (2010: 297). The primary objective of Belz et al. was to explore what information influences the selection of REs in context. The GREC shared tasks delve into both content selection (RCS) and form selection (RFS); this chapter, however, concentrates solely on RFS. The RFS task in GREC is defined as a 4-way classification challenge, where systems are required to determine the most suitable choice among a proper name, a description, a pronoun, or an empty RE for referencing entities. In the upcoming sections, I will introduce the specifics of the GREC corpora in §4.2.1 and examine the systems that were submitted to these shared tasks in §4.2.2.

4.2.1 The corpora used in the GREC shared tasks

The GREC studies consist of four shared tasks that use two distinct corpora: GREC-MSR'08 and GREC-MSR'09 shared tasks employed the GREC-2.0 corpus, while GREC-NEG'09 and GREC-NEG'10 used GREC-PEOPLE. Both corpora were derived from the introductory sections of Wikipedia articles. Specifically, the GREC-2.0 corpus comprises 1941 introductory sections across five domains, including people, rivers, mountains, cities, and countries. The GREC-PEOPLE corpus, on the other hand, includes 1000 introductions from articles about composers, chefs, and inventors. In this study, I focus exclusively on articles from the training sets of these corpora, detailed in Table 4.2. Examples (1a) and (1b) illustrate introductory sentences from the GREC-2.0 and GREC-PEOPLE corpora, respectively:

[2] Methods not considered in GREC, such as E2E neural approaches (Castro Ferreira et al. 2018a, Cao & Cheung 2019, Cunha et al. 2020), are excluded from this discussion as they are not relevant to this specific inquiry.

(1) a. **Berlin** is the capital city and one of the sixteen federal states of Germany. With a population of 3.4 million in **its** city limits, **Berlin** is the country's largest city, and the second most populous city in the European Union. **Berlin** is an influential center in European politics, culture and science.

b. **David Chang** (born 1977) is a noted American chef. **He** is chef/owner of Momofuku Noodle Bar, Momofuku Ko and Momofuku Ssäm Bar in New York City. **Chang** attended Trinity College, where **he** majored in religious studies. In 2003, **Chang** opened **his** first restaurant, Momofuku Noodle Bar, in the East Village.

A key distinction between GREC-2.0 and GREC-PEOPLE is their annotation focus. In GREC-2.0, annotations are limited to references to the main subject of the article, as exemplified in (1a) with references to Berlin. In contrast, GREC-PEOPLE annotations encompass all *human* REs. While (1b) illustrates annotations pertaining solely to the main character, GREC-PEOPLE contains annotations for every human referent mentioned in each document.

Each RE annotation includes details about its form, grammatical role, and semantic type. While the documents are segmented into paragraphs, finer divisions such as sentences or tokens are not provided. To address this limitation, I used *spaCy* for sentence segmentation and tokenization.[3]

4.2.2 The algorithms submitted to the GREC shared tasks

The GREC challenges saw submissions of several feature-based algorithms. For the purpose of model reconstruction in this chapter, the feature sets and machine learning (ML) methods of the following submissions were chosen: Hendrickx et al. (2008) [CNTS], Favre & Bohnet (2009)[ICSI], Bohnet (2008) [IS-G], Jamison & Mehay (2008) [OSU], and Greenbacker & McCoy (2009a) [UDEL]. Table 4.1 summarizes the names, ML methods, and reported accuracies of these models as documented in Belz et al. (2010).

In addition to the variances in the ML algorithms, the feature sets used by each system also significantly differ. The impact of these feature choices on model performance will be thoroughly explored in Chapter 5.

[3] spaCy is a free, open-source Python library for Natural Language Processing (`https://spacy.io/`).

Table 4.1: The first column details the names used for each algorithm in the current study, based on the original names of the systems in Belz et al. (2010). The second column indicates the shared task each algorithm was submitted to. The third column (ML) specifies the machine learning algorithm used, and the fourth column (Acc) reports the accuracy of the models on GREC-2.0 as stated in Belz et al. (2010).

Name	Shared Task	ML	Acc
UDEL	GREC-MSR'09,GREC-NEG'09	*C5.0*	77.71
ICSI	GREC-MSR'09,GREC-NEG'09	*CRF*	75.16
CNTS	GREC-MSR'08	*MBL*	72.61
IS-G	GREC-MSR'08	*MLP*	70.78
OSU	GREC-MSR'08	*MaxEnt*	69.82

4.3 Enriching WSJ for the REG-in-context task

ONTONOTES, as outlined in Weischedel, Ralph et al. (2013), contains a large collection of texts across seven genres, including broadcast conversation, broadcast news, magazine, newswire, pivot text, telephone conversation, and web data. Unlike the GREC corpora, ONTONOTES benefits from comprehensive annotations, including sentence segmentation, tokenization, morpho-syntactic annotation (lemmas, coarse- and fine-grained POS tags), syntactic structure, and shallow semantic details (word sense, coreference). For the purpose of extracting and enriching REs in my study, I use the *ONF* files (OntoNotes Normal File), which contain all levels of annotation.

The WSJ dataset, a part of the newswire genre in ONTONOTES, was selected for several key reasons. Firstly, the predominance of third-person REs in WSJ makes it ideal for studying RF alternations, which are more pronounced in third-person references compared to the typically pronominal first-person and second-person REs. Secondly, the dataset allows for the addition of paragraph segmentation, a crucial feature for the feature-based studies in Chapters 5 and 6. Lastly, various versions of WSJ have been integral to other REG studies (Orita et al. 2015, Kibrik et al. 2016, Rösiger 2018), making it a valuable and relevant choice for comparison and analysis. In the following sections, I detail the additional annotations made to the WSJ dataset in §4.3.1, and discuss the specific cases that were excluded from the dataset in §4.3.2.

4.3.1 Annotating REs of the WSJ dataset

Enhancing the existing annotations in WSJ, I established a set of rules for automatically annotating REs, subsequently verifying and correcting these annotations manually. Below, I explain the specifics of annotating RF, plurality, animacy/entity type, and gender.

Referential form While WSJ inherently includes coreferential chain annotations, it lacks specific annotations for RFs. Using POS tag information from the dataset, REs were automatically labeled as `pronoun`, `description`, or `name`.[4] Despite the high accuracy of this method, relying solely on POS tags can occasionally result in misannotations. For instance, the phrase *the United States* might be mistakenly labeled as a description based on its initial determiner, despite being a proper name. To correct such inaccuracies, I conducted a manual review of these tags.[5] (2) from document WSJ-0990 illustrates various REs with their corresponding annotations.

(2) **Their**$_{\texttt{pronoun}}$ tone was good-natured, with **Mr. Packwood**$_{\texttt{name}}$ saying **he**$_{\texttt{pronoun}}$ intended to offer **the proposal**$_{\texttt{description}}$ again and again on future legislation and **Sen. Mitchell**$_{\texttt{name}}$ saying **he**$_{\texttt{pronoun}}$ intended to use procedural means to block **it**$_{\texttt{pronoun}}$ again and again.

Plurality The *plurality* feature in the annotation process identifies whether a referent is `plural` or `singular`. This classification is based on the content of the REs. When a coreferential chain contained a singular pronominal form, such as *he* or *she*, the entire chain was tagged as `singular`. In contrast, if the chain included a plural pronoun, then `plural` was assigned to all its members. Cases that lacked explicit pronominal forms within the coreferential chains were annotated manually.

Animacy/entity type For the animacy or entity type annotation, the name entity annotations already present in the corpus, as documented in (Weischedel et al. 2012: 21), were initially used. These annotations were then reviewed and confirmed manually. The category of animacy/entity type includes a variety of values, such as `person`, `organization`, `location`, among others.

[4] Here, the term *name* is used synonymously with *proper name*.

[5] These tags are further categorized into more detailed classifications, taking into account factors such as expression length and modification, as discussed in (Belz & Varges 2007). However, this book does not delve into these finer distinctions.

Gender The feature *gender* indicates the gender of individual human referents. Referential chains with at least one male pronominal form were assigned the tag `male`, and those with female pronominal forms were given the tag `female`. I tagged all other cases as `other`. I checked and manually corrected the annotations if necessary.

4.3.2 REs excluded from the WSJ dataset

This section outlines the criteria used to exclude certain REs from the WSJ dataset.

First-person and second-person REs Given the minimal RF alternation in first-person and second-person REs, these have been excluded from the dataset. The exclusion process involved first converting the REs to lowercase and then omitting the following pronominal forms, amounting to 1847 REs in total:

> First-person and second-person REs= `{"i", "my", "mine", "me", "you", "your", "yours", "we", "us", "ours", "our"}`.

In addition to first-person and second-person REs, certain REs were also excluded from the set of 42,032 third-person REs. The criteria for their exclusion include:

Appositives The ONTONOTES corpus annotates two distinct types of referential relations, namely *IDENT* and *APPOS*, as described in (Weischedel et al. 2012):

- IDENT: Identity relation includes the annotation of REs with anaphoric mentions (coreference).
- APPOS: Apposition relation contains the annotation of the head of an appositive phrase along with one or more attributes associated with the head.

Figure 4.1 provides an example of both identity and appositive chains. In the upper portion of the figure, an IDENT chain is depicted where the REs *Gary Hoffman, a Washington lawyer specializing in intellectual-property cases*, *he*, and *He* are in a coreferential relationship. Conversely, the lower part showcases an appositive chain where *Gary Hoffman* is the head, and *a Washington lawyer specializing in intellectual-property cases* functions as the attribute, providing additional information about the head. For the purposes of this study, appositive chains were excluded from the dataset.

IDENT CHAIN
RE 1: Gary Hoffman, a Washington lawyer specializing in intellectual - property cases,
RE 2: he
RE 3: He

APPOS CHAIN
HEAD: Gary Hoffman
ATTRIB: a Washington lawyer specializing in intellectual - property cases

Figure 4.1: An example of IDENT and APPOS chains in the WSJ corpus.

Non-coreferential chains Consider the scenario illustrated in (3), where *a criminal defendant* and *a criminal defendant who chooses not to testify* are treated as separate but nested REs with a coreferential relationship. In such instances, only the maximal span, namely *a criminal defendant who chooses not to testify*, is considered as an RE, while the shorter expression *a criminal defendant* is excluded from the analysis.

(3) This privilege against self-incrimination precludes the drawing of an adverse inference against **[[a criminal defendant] who chooses not to testify]**.

Verbal and coordinated expressions In addition to excluding appositives and non-coreferential instances, I also omitted other specific cases from the analysis. These include verbal expressions and coordinated expressions where the NPs do not align in terms of their `person` attribute. For instance, in (4), the verbal expression *applied* is coreferential with the noun phrase *its application*. However, due to its verbal nature, *applied* is not included in the dataset. Similarly, in (5), there is a coordination between *Mr. Apple*, a third-person REs, and *I*, a first-person RE. The difference in `person` information between these coordinated NPs led to their exclusion from the analysis.

(4) An official of the Palestinian Olympic Committee said the committee first **applied** for membership in 1979 and renewed its application in August of this year.

(5) All of which has enabled those of us in Washington who enjoy wallowing in such things to go into high public dudgeon, as **Mr. Apple and I** did the other night on ABC's "Nightline."

With these exclusions applied, the dataset retains 30,471 REs. The following summary provides a comprehensive overview of the WSJ dataset alongside the other two corpora discussed earlier.

4.4 Interim summary

Sections §4.2 and §4.3 have introduced the three corpora – WSJ, GREC-2.0, and GREC-PEOPLE – that will be pivotal in study A. Table 4.2 offers an in-depth comparison of these corpora, encompassing various characteristics and referential expression distributions.

Table 4.2: A comparison of GREC-2.0, GREC-PEOPLE, and WSJ in terms of their length-related features and RF distributions.

Characteristics	GREC-2.0	GREC-PEOPLE	WSJ
Genre	Wikipedia	Wikipedia	Newspaper
Number of documents	1655	808	585
Word/doc (mean)	148.3	129.3	530.7
Sentence/doc (mean)	7.2	5.8	25
Paragraph/doc (mean)	2.3	2.2	11
Referent/doc (mean)	1	2.6	15
RE/doc (mean)	7.1	10.9	52.1
Total number of REs	11,705	8378	30,471
`Description` n(%)	1620(13.84%)	335(4%)	12,020(39.45%)
`Name` n(%)	4459(38.09%)	3417(40.79%)	11,164(36.64%)
`Pronoun` n(%)	4891(41.79%)	4084(48.75%)	7287(23.91%)
`Empty` n(%)	735(6.28%)	542(6.47%)	-

As the table shows, there are considerable differences in length-related features and RE distribution between GREC-2.0 and GREC-PEOPLE, on the one hand, and WSJ, on the other. For instance, while the occurrence of (proper) names is relatively consistent across all corpora, descriptions show a marked variance, being most frequent in WSJ (39.45% of its total REs) compared to their presence in GREC-2.0 (~14%) and GREC-PEOPLE (~4%). The next section will delve into the

methodology and implications of reconstructing the GREC systems with these diverse corpora, aiming to provide insights into how these differences influence the outcomes and interpretations in the study.

4.5 Study A: Reconstruction of the GREC RFS models

Study A is dedicated to the meticulous reconstruction of several RFS models.[6] The primary objective of this study is to assess whether the original corpora employed in these models were adequately suited for their intended tasks. This examination is crucial for understanding the effectiveness and generalizability of the findings derived from these models. The section unfolds as follows: §4.5.1 delves into the necessity of thoroughly inspecting the corpora used in REG-in-context studies. It aims to evaluate their representativeness and appropriateness in capturing the nuances of the RFS task. §4.5.2 describes the various RF categories employed in RFS studies. §4.5.3 explains the methodology behind the reconstruction of the RFS models, detailing the steps taken to replicate the original models using the selected corpora. Lastly, §4.5.4 presents a comprehensive evaluation of the reconstructed models.

4.5.1 The need for a closer inspection of REG-in-context corpora

As previously discussed, the GREC tasks have introduced several feature-based REG-in-context models into the realm of REG (Bohnet 2008, Greenbacker & McCoy 2009a, Orăsan & Dornescu 2009). However, certain aspects of the GREC corpora raise concerns about their suitability for these tasks. This section elaborates on these issues and explains the rationale for incorporating a third corpus in this study.

Representativity Despite the notable differences in annotation practices between the GREC-2.0 and GREC-PEOPLE corpora, both draw from the same genre of text, namely, Wikipedia articles. Given that one of the principal objectives of this chapter is to assess the representativeness of these resources, I have incorporated a third corpus into the study. This additional corpus contrasts with GREC-2.0 and GREC-PEOPLE not only in genre but also in terms of length and various other features. The inclusion of this divergent corpus will allow for a more comprehensive analysis, shedding light on how genre and other textual characteristics impact the outcome of the models.

[6] Refer to Same et al. (2023) for an updated version of study A, which also integrates pretrained Language Models.

The GREC task definition According to the GREC task definition cited in §4.2, the minimum requirement is to generate REs in a piece of text longer than *a sentence*. Upon evaluating the two corpora, it was observed that 80 documents in GREC-2.0, representing 4.8% of the training set, and 114 documents in GREC-PEOPLE, accounting for 14.1% of the training set, comprise only a single sentence. Consequently, these documents fail to meet the minimum task requirement. In contrast, such instances are notably rare in the WSJ corpus, occurring in merely two documents (0.3%). This disparity underscores the significance of conducting an initial evaluation of corpora to confirm their alignment with the primary requirements of the study's targeted task.

The distribution of RFs As illustrated in Table 4.2, the two predominant RF types in the GREC corpora, specifically pronouns and proper names, collectively constitute a significant portion of the referential instances: 80% in GREC-2.0 and 89.5% in GREC-PEOPLE. The remaining referential forms, which include descriptions and empty references, contribute to approximately 20% of the instances in GREC-2.0 and around 10% in GREC-PEOPLE. This distribution presents a crucial question for the study: Does this imbalance in the frequency of different RF types impact the effectiveness of the RFS algorithms? More specifically, the study seeks to investigate whether RFs that appear less frequently are accurately predicted by these algorithms.

Incorporating the WSJ corpus, with its distinct distribution of RF types, provides an opportunity to explore these questions in a varied context. The diverse distribution of RFs in WSJ may offer insights into how algorithms perform across a more balanced range of referential forms.

4.5.2 RF categories considered

REG-in-context studies have explored various RF classifications. These include:

- 2-Way (binary) classification (McCoy & Strube 1999, Henschel et al. 2000, Poesio et al. 2004): This classification distinguishes between pronominal and non-pronominal REs.
- 3-Way classification (Kibrik et al. 2016): Here, REs are categorized as either a pronoun, a description, or a proper name.
- 4-Way classification (Belz et al. 2010): This approach includes an additional category of empty forms, along with pronouns, descriptions, and proper names.

- 5-Way classification (Castro Ferreira et al. 2016b): This classification adds demonstrative forms to the four categories mentioned above.

In the context of this study, a faithful reconstruction would ideally adhere to the 4-way classification model employed in the GREC tasks. However, a notable limitation is the absence of annotations for empty references in the WSJ corpus. To accommodate this, I modify the approach to focus on a 3-way classification task, considering the labels `pronoun`, `description`, and `(proper) name`. This adaptation, though necessary due to the dataset constraints, might impact the study's outcomes, as it may affect the generalizability and comparability of the findings with those from the original GREC tasks.

4.5.3 Architecture of the models

In assessing the performance of the algorithms across different corpora, two approaches were considered:

- Construction of GREC-PEOPLE and WSJ algorithms: This method involves constructing the algorithms for these two corpora and comparing their accuracy with the reported accuracy of the GREC-2.0 algorithms in Belz et al. (2010), as shown in Table 4.1.
- Reconstruction of algorithms for all three corpora: This approach entails reconstructing the algorithms for GREC-2.0, GREC-PEOPLE, and WSJ using the feature sets and ML methods detailed in Belz et al. (2010).

I have opted for the second solution based on two key considerations. Firstly, there is a lack of clarity regarding the specific parameters employed in the algorithms submitted to GREC.[7] Given this uncertainty, constructing algorithms for the WSJ and GREC-PEOPLE corpora using new parameters and then comparing their performance to the original GREC-2.0 algorithms' accuracy might yield inaccurate or misleading results.

Secondly, a comprehensive evaluation of the algorithms' effectiveness in predicting *each* class necessitates a per-class analysis of the predictions. Such a detailed level of evaluation is not provided in the GREC report (Belz et al. 2010).

As indicated in §4.2, this study involves the reconstruction of five systems originally submitted to the GREC tasks. These systems are based on five distinct

[7] I contacted the developers of the models submitted to the GREC shared tasks to obtain information about their models' parameters, but unfortunately, the responses received did not provide conclusive details.

ML algorithms, each implemented using specific software packages and configurations:

1. Conditional Random Field [CRF]: The *crfsuite* package (`https://cran.r-project.org/web/packages/crfsuite/`) is used for training CRF models. Settings include 3000 iterations and Stochastic Gradient Descent (SGD) with L2-regularization (`l2sgd`) as the learning method.

2. C5.0 Decision Tree [C5.0]: The *C5.0* R package (Kuhn et al. 2018) is employed to construct decision trees. The number of boosting iterations (trials) is set to 3, using information gain (entropy) as the splitting criterion.

3. Memory-Based Learning [KNN]: While Hendrickx et al. (2008) referenced the TiMBL package (Daelemans et al. 2007) for Memory-Based Learning, I implement the k-Nearest Neighbors (KNN) algorithm, a direct descendant of Memory-Based Learning (Daelemans et al. 2007). The *caret* package, with the method *knn*, is used for this implementation.

4. Maximum Entropy Classifier [MaxEnt]: The *multinom* algorithm from the *nnet* R package is used, given the theoretical similarity of the Maximum Entropy classifier to multinomial logistic regression. This choice was made as the original *maxent* package has since been deprecated.

5. Multilayer Perceptron [MLP]: The MLP is implemented using the *Keras* package. The model includes two hidden layers with 16 and 8 units respectively, employing the rectified linear activation function (ReLU) for the hidden layers and the Sigmoid activation function for the output layer. The model is trained over 50 epochs, with a batch size of 50 samples.

I have developed feature sets for each of the three corpora – GREC-2.0, GREC-PEOPLE, and WSJ – on which the respective algorithms are trained. Figure 4.2 illustrates the reconstruction process for each algorithm. As an example, the first dashed box in the figure indicates that the UDEL feature set, coupled with its ML method C5.0, is employed to train decision trees for all three corpora.[8] Applying these feature sets and ML methods across the three corpora results in a total of 15 distinct classifiers. For the purpose of training and evaluating these models, the data in each corpus is partitioned into two segments: 70% is allocated for the training and development sets, and the remaining 30% forms the test set.

[8] A comprehensive description of the features used in these systems is detailed in Chapter 5.

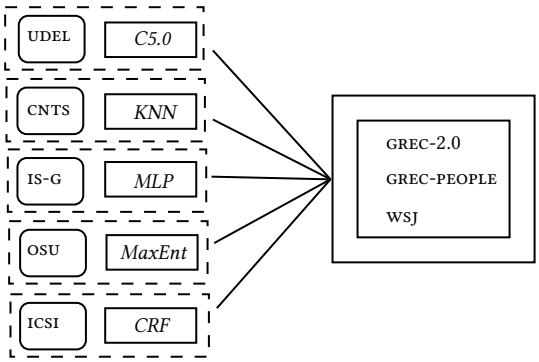

Figure 4.2: The feature sets and the ML methods of the five GREC algorithms used to train classifiers on the three corpora.

4.5.4 Evaluation of the algorithms

This section is dedicated to a thorough evaluation of the reconstructed algorithms. The primary focus is on assessing the accuracy of the models, consistent with the metrics used in the GREC-MSR shared tasks. As documented in Table 4.1, accuracy was the sole measure reported for all of these systems, while precision and recall were detailed only in the context of the GREC-NEG shared tasks (refer to Belz et al. 2010 for more on this decision). Thus, the evaluation in §4.5.4.1 will primarily report on the accuracy of each model across the three corpora.

Following the accuracy assessment, a Bayes Factor (BF) analysis is conducted in §4.5.4.2. This analysis will provide statistical evidence to determine whether the observed accuracy rates across different models and corpora stem from similar or different distributions.

Finally, a per-class evaluation of the predictions is carried out in §4.5.4.3, which aims to dissect the performance of the algorithms for each referential form class (e.g., pronoun, description, proper name). This step is essential to evaluate the true effectiveness of the algorithms, as it reveals how well they perform for each specific category of RF.

4.5.4.1 Accuracy of the models

Table 4.3 presents the overall accuracy achieved by each model in the 3-way classification task.

The results indicate varied performance across the corpora. For instance, the ICSI model shows consistent performance across all three corpora, while other models like CNTS and UDEL exhibit notable variations. Such discrepancies could

Table 4.3: Overall accuracy of the algorithms illustrated in Figure 4.2.

	UDEL	CNTS	IS-G	ICSI	OSU
GREC-2.0	0.67	0.68	0.69	0.72	0.70
GREC-PEOPLE	0.80	0.75	0.78	0.78	0.79
WSJ	0.63	0.59	0.69	0.70	0.70

be attributed to differences in the corpora's textual characteristics and the distribution of referential forms.

Table 4.4 presents two specific predictions made by the ICSI, UDEL, and OSU models on the GREC-2.0 corpus. The first instance pertains to the bold RE mentioned in the first sentence, while the second instance relates to the bold RE in the third sentence.[9] In the evaluation of the first instance, all algorithms – ICSI, UDEL, and OSU – successfully predicted the correct RF. However, in the second instance (third sentence), discrepancies arise in the models' performance: both UDEL and OSU models inaccurately predicted the RF, whereas the ICSI model accurately classified the label as `pronoun`.

Ranking of the models The data from Table 4.3 reveals intriguing trends in the performance of the various models across the corpora. In the GREC-2.0 models, ICSI stands out with the highest accuracy, closely followed by OSU. On the other hand, UDEL demonstrates a lower performance with an accuracy rate of 0.67, positioning it at the lower end of the spectrum. A similar pattern is observed in the WSJ models, where ICSI and OSU again emerge as top performers, while UDEL and CNTS show the least accuracy. Interestingly, this trend is reversed in the GREC-PEOPLE corpus, where UDEL achieves the highest accuracy rate at 0.80, outperforming the other models.

Overall performance In order to complement the individual assessments of each model, an aggregated analysis was conducted to examine the average accuracy rates across all algorithms and corpora, offering a comprehensive view of their overall effectiveness. As depicted in Table 4.5, this analysis calculates the mean accuracy of the algorithms across the three corpora. According to the combinations outlined in Figure 4.2, the ICSI model demonstrates the highest overall performance, closely followed by the OSU and IS-G models.

[9] It is noteworthy that the occurrence of *Chicago* in the second sentence does not establish coreference with the REs in the first and third sentences.

Table 4.4: Two predictions from the GREC-2.0 corpus. The term *name* stands for `proper name` and *pro* stands for `pronoun`.

Num	Sentence	ORIGINAL	ICSI	UDEL	OSU
1	**Chicago** is the largest city in Illinois and the third-most populous city in the United States, with approximately 2.9 million people.	`name`	`name`	`name`	`name`
2	"Chicago" can also refer to the Chicago Metropolitan area, known as Chicagoland, with a population of 9.4 million in three states.	-	-	-	-
3	**It** is located along the southwestern shore of Lake Michigan.	`pro`	`pro`	`name`	`name`

Table 4.5: Average accuracy of each algorithm across the three corpora.

UDEL	CNTS	IS-G	ICSI	OSU
0.70	0.673	0.72	0.733	0.73

Furthermore, the mean accuracy of all algorithm-corpus combinations has been calculated to provide further insights into the performance of the algorithms within each specific corpus. As indicated in Table 4.6, the algorithms applied to the GREC-PEOPLE corpus exhibit, on average, the highest accuracy among the three.

Table 4.6: Mean accuracy rates of the algorithms within each corpus.

GREC-2.0	GREC-PEOPLE	WSJ
0.692	0.78	0.662

Based solely on the accuracy metrics of the models, it can be inferred that those models which incorporate the ICSI specifications, specifically the Condi-

tional Random Field (CRF) method and the feature set outlined in Favre & Bohnet (2009), demonstrate a higher proficiency in predicting RFs. Additionally, it is noteworthy that models applied to the GREC-PEOPLE corpus exhibit the best overall performance. However, to gain a more comprehensive understanding of these results, two pivotal questions remain:

1. Are the accuracy rates reported in Table 4.3 for each corpus evidentially different from one another?

2. What factors contribute to the variance in rankings and the disparity in accuracy rates of models trained on the GREC-PEOPLE corpus compared to those trained on the other two corpora?

To address the first question, a BF analysis is employed, providing a statistical measure of the evidence for differences between the corpora. The second question is tackled through a per-class evaluation, aiming to dissect and understand the specific causes behind the observed differences in model performance across the corpora.

4.5.4.2 Bayes factor analysis

The BF is employed as a statistical tool to assess whether two distinct accuracy rates (for example, those of the best-performing and worst-performing algorithms) are likely to originate from distributions with either *similar* or *different* probability parameters. This analysis involves determining whether there is substantial evidence to conclude that the difference in accuracy rates of the models exceeds 0.03, suggesting they are derived from different distributions, or whether the difference is less than 0.03, indicating a potential similarity in distribution.

Furthermore, the implementation of BF analysis in this study includes descriptive statements about the strength of the evidence. By interpreting BF scores, we can discern whether there exists *light*, *positive*, *strong*, or *very strong* evidence to support or refute the hypothesis of similar or different distributions in the ratio of probabilities. The categorization of these evidence strengths is based on the guidelines set forth by Kass & Raftery (1995).

GREC-2.0 The BF analysis comparing the accuracy rates of the highest performing model, ICSI, and the lowest, UDEL, yields a BF score of 3.55. This score falls into the category of positive evidence according to Kass and Raftery's scale, suggesting that the accuracy rates of these two algorithms are statistically different

Table 4.7: Interpretation of Bayes Factors according to Kass & Raftery (1995: 777).

BF score	Meaning
>150	Very strong
20 to 150	Strong
3 to 20	Positive
1 to 3	Worth of a bare mention

from each other within the GREC-2.0 corpus. Therefore, it can be inferred that the efficacy of ICSI in predicting RFs significantly differs from that of UDEL in this specific context.

GREC-PEOPLE In the case of the GREC-PEOPLE corpus, the analysis shows a BF score of 11.9 when comparing the best-performing model, UDEL, with the least accurate, CNTS. This score, indicating positive evidence, implies that these two models likely originate from distributions with different probability parameters. Conversely, the comparison of UDEL with the other intermediate models does not show a statistically significant difference, suggesting that their performances are relatively similar.

WSJ For the WSJ corpus, the analysis comparing OSU, ICSI, and ISG (the top three algorithms) indicates that their accuracy rates are likely from similar distributions. However, there is very strong evidence to suggest that the top two models, ICSI and OSU, differ significantly in performance from the lower-ranked models, UDEL and CNTS. This finding highlights a clear distinction in effectiveness between the higher and lower-ranked algorithms within this corpus.

Note that the results for the GREC-2.0 and WSJ corpora exhibit closer similarities to each other than either does to the GREC-PEOPLE corpus. This observation is evidenced by the nearly identical ranking of algorithms in both GREC-2.0 and WSJ, along with their comparable performance metrics. This similarity raises a pertinent question regarding the source of the discrepancy observed in the GREC-PEOPLE corpus compared to the other two. One plausible explanation for this variation could be attributed to the non-uniform distribution of RF classes within the GREC-PEOPLE corpus, where approximately 90% of the expressions are concentrated in the two dominant classes. This distribution pattern suggests that standard accuracy metrics might not fully capture the effectiveness of an algorithm,

as a model could potentially achieve high accuracy by predominantly predicting these dominant classes.

To explore this hypothesis further, the subsequent section will delve into a per-class evaluation. This analysis aims to discern whether the observed high accuracy in the GREC-PEOPLE corpus stems primarily from the overprediction of the dominant classes or if it reflects genuinely robust performance across all RF classes.

4.5.4.3 Per-class evaluation

While the GREC-2.0 and WSJ corpora exhibit similar patterns in overall accuracy, the GREC-PEOPLE corpus presents distinct trends. This section aims to determine whether the accuracy rates reported in Table 4.3 truly reflect the algorithms' success or are skewed due to the overprediction of dominant classes.

The per-class evaluation, detailed in Table 4.8, offers insights into this matter, offering precision, recall, and F1 scores for each model. The F1 score, a weighted average of precision and recall, is emphasized as it presents a balanced measure of performance for each class.

A comparative analysis of the F1 scores for the class `description` across the three corpora reveals that the WSJ corpus achieves the highest scores, with all algorithms surpassing an F1 score of 0.50. In contrast, the F1 scores for `description` in both GREC-2.0 and GREC-PEOPLE are significantly lower. Particularly noteworthy are the results from GREC-PEOPLE, where UDEL and OSU display F1 scores near zero, and IS-G fails to correctly predict this class at all. This suggests that the GREC-PEOPLE models might struggle with predicting descriptions due to a lack of sufficient instances in the training dataset.

Another notable finding is the high recall for pronoun prediction in the GREC-PEOPLE models. For instance, NEG-ISG exhibits a recall of 0.96, indicating that 96% of pronoun cases are accurately predicted, possibly hinting at an overprediction tendency.

The per-class evaluation indicates that, with the exception of ICSI, the GREC-PEOPLE models perform poorly in predicting the class `description`. While GREC-2.0 algorithms also show weak performance in this area, they are somewhat more successful than those from GREC-PEOPLE. In contrast, WSJ models consistently predict this class with more than 50% accuracy. These observations suggest that feature-based classification models can achieve reliable predictions across all classes only when trained with a suitably diverse and balanced dataset. The corpus-dependent nature of these models' performance is evident, highlighting the need for further investigation into whether neural end-to-end (E2E) models

Table 4.8: Per-class precision, recall, and F-1 score of each label. The results report on training five different algorithms on three corpora for predicting three labels, namely `proper name`, `description`, and `pronoun`.

Model	Label	GREC-2.0			GREC-PEOPLE			WSJ		
		prec	recall	f1	prec	recall	f1	prec	recall	f1
UDEL	description	0.46	0.14	0.21	1.00	0.01	0.02	0.61	0.65	0.63
	name	0.77	0.57	0.66	0.81	0.73	0.77	0.62	0.51	0.56
	pronoun	0.63	0.93	0.75	0.80	0.93	0.86	0.70	0.82	0.76
ICSI	description	0.57	0.34	0.43	0.70	0.30	0.42	0.74	0.64	0.69
	name	0.79	0.64	0.71	0.77	0.71	0.74	0.66	0.73	0.69
	pronoun	0.71	0.92	0.80	0.78	0.87	0.82	0.69	0.75	0.72
CNTS	description	0.49	0.28	0.36	0.33	0.10	0.15	0.67	0.45	0.54
	name	0.75	0.58	0.65	0.72	0.73	0.72	0.53	0.64	0.58
	pronoun	0.66	0.90	0.76	0.79	0.82	0.80	0.59	0.72	0.65
OSU	description	0.53	0.28	0.37	1.00	0.03	0.06	0.78	0.55	0.65
	name	0.70	0.68	0.69	0.81	0.69	0.75	0.63	0.80	0.70
	pronoun	0.71	0.85	0.77	0.78	0.93	0.85	0.74	0.80	0.77
ISG	description	0.57	0.20	0.30	0.00	0.00	0.00	0.67	0.78	0.72
	name	0.67	0.73	0.70	0.84	0.64	0.73	0.69	0.60	0.64
	pronoun	0.71	0.80	0.75	0.75	0.96	0.84	0.74	0.68	0.71

exhibit similar corpus-dependent characteristics. This exploration is continued in Chapter 7, where the generalizability of these findings will be further assessed.

4.6 Summary and discussion of study A

This chapter has critically examined the impact of corpus selection on the performance of algorithms in the REG-in-context task. Drawing on the systematic assessment provided by GREC (Belz et al. 2010), this study sought to evaluate the suitability of the corpora used for the REG-in-context task. The key findings are summarized as follows:

4.6.1 Different corpora favor different algorithms

The Bayes Factor (BF) analyses focusing on the overall accuracy rates indicate a clear pattern: Different corpora favor different algorithms. A prime example of

this is UDEL. While its performance is markedly different (and inferior) compared to the top-performing models in the GREC-2.0 and WSJ corpora, it emerges as the best-performing algorithm in the GREC-PEOPLE corpus, achieving an accuracy of 0.8.

This discrepancy is particularly notable: Despite the similar genre and length characteristics of GREC-PEOPLE and GREC-2.0, their model performances diverge significantly. In contrast, despite genre- and length-related differences between GREC-2.0 and WSJ, these corpora show a preference for almost identical algorithms, suggesting a complex interplay between corpus characteristics and algorithm efficiency.

4.6.2 Explaining the corpus differences

Study A reveals that the results for GREC-2.0 and WSJ are more closely aligned with each other than with GREC-PEOPLE. A crucial distinction lies in the scope of annotated referents: GREC-PEOPLE exclusively annotates references to human referents, whereas GREC-2.0 and WSJ include a broader spectrum, such as mountains, cities, countries, rivers (in GREC-2.0), and non-human entities like organizations, places, and objects (in WSJ). It appears that the class *description* is predominantly employed for non-human referents. The distribution of RFs in WSJ, comprising approximately 31,000 instances, indicates that there are 12,000 descriptions, of which 80% are non-human. This pattern suggests that a corpus limited to human referents might not be optimally suited for a 3-way prediction task of this nature.

4.6.3 Explaining the performance differences

To elucidate the performance disparities observed between models trained on the GREC-PEOPLE corpus and those trained on other corpora, a detailed per-class evaluation was conducted. This assessment revealed that, in the case of GREC-PEOPLE, all algorithms except for ICSI either entirely fail to predict the description class or demonstrate markedly poor performance in doing so (for instance, the CNTS model shows an F1 score of only 0.15 for this class). This pattern suggests a significant limitation of the GREC-PEOPLE corpus for the current classification task, particularly since one of the key classes is infrequently predicted.

Furthermore, a notable proportion of the GREC-PEOPLE documents, approximately 14.1%, fail to meet the minimal criteria of the GREC task, which mandates the generation of referring expressions in texts exceeding a single sentence in length. These findings collectively indicate that the GREC-PEOPLE corpus, as it stands, is not ideally suited for the objectives of this task.

4.6.4 Limitations of the accuracy metric

While reporting the accuracy of classification models is a standard practice, it is crucial to approach overall accuracy with caution. The addition of a per-class evaluation in this study brought in an extra layer of complexity and insight. Notably, it was observed that algorithms achieving high accuracy in the GREC-PEOPLE corpus had a tendency to overpredict pronouns. This finding is particularly significant in situations where the distribution of referential form classes is imbalanced, suggesting that evaluation measures other than overall accuracy should be considered. This point can be argued from several angles.

From a theoretical linguistics viewpoint, any algorithm that completely *overlooks* some RF classes is inadequate, just like it would be unforgivable if an anthropological study described the population of the USA as consisting exclusively of English speakers, overlooking significant minorities such as speakers of Spanish. Linguists, in particular, would be ill-served if the algorithms did not address some of the linguistic classes that they are interested in.

In light of these considerations, there are two ways of looking at accuracy. One is to view accuracy as only part of the story, to be complemented by additional information (such as the information offered in §4.5.4.3, which provided an analysis per category). Another perspective is to *replace* accuracy by a new metric that measures the extent to which the distribution predicted by a REG algorithm matches the distribution found in a corpus (van Gompel et al. 2019).

We do not yet know how well the lessons drawn here generalize to other NLG and NLP tasks. The focus here was on the GREC algorithms; however, far from being limited to these particular algorithms, the conclusions give reason to suspect substantial *corpus dependence* for any REG algorithm, including neural models. This chapter does suggest that when a language corpus is employed for training and testing a CL algorithm – whether this is a conventional rule-based algorithm or, for example, an algorithm based on deep learning, the question must always be asked whether the corpus is representative of the type of language use in which the researchers are interested and about which they are making claims. The issue of representativeness and its implications will be revisited in Chapter 7, where a systematic comparison of different REG approaches is undertaken.

5 The choice of features in feature-based REG-in-context models

5.1 Introduction

Building on the work in Chapter 4, where the adequacy and representativeness of corpora for the REG-in-context task were assessed through the reconstruction of five systems from the GREC shared tasks, this chapter aims to provide a comprehensive overview of the features used in these systems.[1]

The choice of features is a crucial aspect of the success of any feature-based REG-in-context model. As discussed in Chapter 3, the process of feature engineering demands significant resources. Focusing only on those features whose contribution to the task is certain can make the feature selection process more efficient. To this end, study B, detailed in §5.2, focuses on analyzing the features employed in feature-based systems to identify those most beneficial for the REG-in-context task.

The complexity of feature engineering is further compounded by the variety of ways a single feature can be implemented. Taking the measurement of recency as an example, the myriad approaches to calculating the distance between the target RE and its antecedent (ANTE) can lead to substantial variations in model performance. Study C, discussed in §5.3, investigates different implementations of the recency feature to evaluate their respective impacts on the REG-in-context task. This study also includes an analysis of recency across two different corpora, aiming to uncover any corpus-dependent traits.

[1]The studies presented in this chapter are based on two published articles: [study B] Fahime Same & Kees van Deemter. 2020a. A linguistic perspective on reference: Choosing a feature set for generating referring expressions in context. In *Proceedings of the 28th International Conference on Computational Linguistics*, 4575–4586. Barcelona: International Committee on Computational Linguistics. DOI: 10.18653/v1/2020.coling-main.403. [study C] Fahime Same & Kees van Deemter. 2020b. Computational interpretations of recency for the choice of referring expressions in discourse. In *Proceedings of the First Workshop on Computational Approaches to Discourse*, 113–123. Association for Computational Linguistics. DOI: 10.18653/v1/2020.codi-1.12.

5.2 Study B: Choosing a consensus set of features for the RFS task

Chapter 2 highlighted various theories of reference, such as those proposed by Gundel et al. (1993), Ariel (2001), Grosz et al. (1995), and von Heusinger & Schumacher (2019), which seek to explain the RF choices speakers make. A common thread in these theories – referred to here as the linguistic tradition – is the correlation between the form used for referring to a referent and its prominence status at a given point in the discourse. Short anaphoric forms like pronouns are typically sufficient for prominent referents, while longer, more semantically rich forms are employed for less prominent ones. The prominence of a referent is shaped by various prominence-lending cues, including recency and frequency of mention (Ariel 1990), grammatical function (Brennan 1995), animacy (Fukumura & van Gompel 2011), and competition (Arnold & Griffin 2007).

Study B aims to scrutinize feature-based RFS models through a linguistic lens. The features in these models vary significantly, with some aligning with the linguistic tradition and others being more abstract and less linguistically interpretable. This study systematically evaluates previous feature-based REG-in-context models to determine the relative contribution of different features within these models.

We hypothesize that not all features are equally contributory and that a reduced set of features from each feature set could perform comparably to the full feature set. To test this, we analyze the features employed in various REG-in-context models for their impact on the task. Our second hypothesis posits that a small set of features drawn from previously published datasets can form a model substantially as accurate as the best-performing existing model. Through several feature selection experiments, we aim to identify a *consensus* set of the most effective features for the REG-in-context task and compare these features against those prioritized in the linguistic tradition.

The structure of this study is as follows: §5.2.1 introduces relevant studies focusing on feature selection in the REG-in-context task. In §5.2.2, we provide a comprehensive overview of the features considered for our feature selection studies and detail their application to the WSJ corpus. §5.2.3 discusses our feature selection experiments and the resulting consensus feature set. In §5.2.4, we examine the consensus set from a linguistic standpoint, assessing the interpretability of these features within a linguistic framework. Finally, §5.2.5 summarizes our key findings.

5.2.1 The importance of feature selection

Data-driven, feature-based models rely on carefully selected features to approximate human decision-making processes. In a study grounded in psycholinguistic insights, Greenbacker & McCoy (2009b) emphasized the critical role of feature selection for the REG-in-context task. They developed several models informed by linguistic insights and analyzed misclassifications made by these models. Their aim was to determine if psycholinguistic research could shed light on the observed misclassification patterns. Additionally, they incorporated features from Hendrickx et al. (2008) and compiled a comprehensive list of features believed to influence RF choice. Training C5.0 decision trees on various subsets of these features, they discovered that using the maximum number of features did not always result in optimal performance. However, their study had limitations, such as the lack of clarity in their feature subset selection process and a somewhat subjective approach to feature selection, primarily focused on features they deemed important rather than a broader range of features used in other models. Moreover, they did not offer a linguistic explanation for the performance of their best model.

Kibrik et al. (2016) also touched upon feature selection, specifically highlighting the significance of recency-related features. Although their study was linguistically informed, the annotation effort behind it was very intense. This raises a question: Could a more concise set of features achieve similar results? Our study seeks to systematically evaluate the features utilized in various REG-in-context studies, building upon and extending the work of Greenbacker & McCoy (2009b) and Kibrik et al. (2016). We aim to identify a potentially smaller yet effective set of features that can yield comparable outcomes in REG-in-context tasks.

5.2.2 Prerequisites for a systematic evaluation

In conducting a systematic evaluation, a fundamental aspect is the selection of the objects of study, which, in our case, are the feature sets of RFS algorithms. Our approach to selecting these feature sets is as follows:

- We begin by selecting all RFS algorithms submitted to the GREC shared tasks, as documented in Belz et al. (2010), and extract their feature sets. GREC is chosen as the starting point due to its aggregation of major RFS algorithms existing at that time.
- We extend our scope to include two additional feature sets from papers in the ACL anthology. The criteria and methods used for this selection are detailed in §5.2.2.1.

- The features identified through this process are then reimplemented following the methodology outlined in §5.2.2.2.

It is important to note that our systematic evaluation excludes E2E models, such as those proposed by Castro Ferreira et al. (2018a), Cao & Cheung (2019), and Cunha et al. (2020). The reason for this exclusion is that, in their current state, these models do not yet offer many possibilities for linguistic interpretation. Therefore, our focus is on linguistically interpretable features, providing a clearer insight into the mechanisms driving effective RFS algorithms.

5.2.2.1 Feature sets used

In selecting feature sets for our study, we established the following criteria: we targeted studies that (1) primarily focus on RFS, (2) employ a machine learning (ML) method, (3) use an English dataset, and (4) incorporate interpretable features.

Our initial step involved selecting feature sets from the RFS algorithms submitted to the GREC challenges. We excluded the JUNLG set (Gupta & Bandopadhyay 2009) due to its rule-based approach, and the WLV feature set (Orăsan & Dornescu 2009) because we were unable to interpret some of their features.

Given that the GREC challenges were conducted several years ago, it was imperative to also consider more recent research in order to include contemporary RFS feature sets meeting our criteria. To achieve this, we downloaded the complete BibTeX file of the ACL anthology.[2] We then used regular expressions to manually search for specific terms (listed in Table 5.1) in the titles and abstracts of the articles. This approach helped in identifying relevant studies that have contributed to the field of RFS after the GREC challenges, ensuring a comprehensive and current selection of feature sets for our evaluation.

Based on our search results and adherence to the predefined criteria, we selected the feature sets from the studies by Castro Ferreira et al. (2016b) and Kibrik et al. (2016). These two sets, in conjunction with the feature sets from the GREC challenges, constitute the seven sets used in our feature selection experiments.[3] The datasets employed in this study are detailed in Table 5.2. The GREC feature sets are referenced as named in Belz et al. (2010), while the other two feature sets are identified by the last names of their primary authors.

[2] `http://www.aclweb.org/anthology/`

[3] Several studies identified in our search were excluded as they did not meet our criteria, including Zarrieß & Kuhn (2013), Siddharthan et al. (2011), Stent (2011), and Castro Ferreira & Paraboni (2017).

Table 5.1: Terms used to search for RFS studies.

Regular expressions
[R|r]eferring [E|e]xpression.*[M|m]achine [L|l]earning
title =.*[R|r]eferring [E|e]xpression
[G|g]enerat[ion|ing].*[R|r]eferring [E|e]xpression.*discourse
title =.*[R|r]efer.* [G|g]eneration
[D|d]ata-driven.*[E|e]xpression

Table 5.2: The feature sets used in study B. The first two columns specify the numerical identifier and the name assigned to each dataset, facilitating easy reference throughout the study.

Number	Dataset	Reference	Number of features
1	IS-G	Bohnet (2008)	5
2	FERREIRA	Castro Ferreira et al. (2016b)	5
3	OSU	Jamison & Mehay (2008)	8
4	ICSI	Favre & Bohnet (2009)	14
5	KIBRIK	Kibrik et al. (2016)	17
6	U-DEL	Greenbacker & McCoy (2009a)	18
7	CNTS	Hendrickx et al. (2008)	21

To facilitate a structured overview, we have classified the features into nine broad categories for analysis: *grammatical role*, *inherent features*, *referential status*, *recency*, *competition*, *antecedent form*, *surrounding patterns*, *position*, and *protagonism*. These categories are further elaborated in the subsequent sections. Within the context of the datasets discussed in this chapter, the term REF denotes the current referent, and ANTE refers to its coreferential antecedent.

In Table 5.3 to Table 5.8, the first column, labeled `Feature`, provides a description of each feature. The `Type` column categorizes the value of each feature as either *numeric* (`num`), *categorical* (`cat`), *boolean* (`bool`), or *character* (`char`). Additionally, the notation `[N]` next to the `Type` attribute indicates the number of distinct features encoded. For example, a feature like *grammatical role of the* 2^{nd} *and* 3^{rd} *ANTE* with the type attribute `cat[2]` signifies two categorical features: *grammatical role of the* 2^{nd} *ANTE* and *grammatical role of the* 3^{rd} *ANTE*. The `DT` column specifies which datasets include each feature, corresponding to the values in the `Number` column of Table 5.2. Lastly, the `Symbol` column presents the nomenclature used to refer to the features in our analysis.

Grammatical role This category encompasses features related to the syntactic properties of both the referent (REF) and its antecedent (ANTE), as detailed in Table 5.3.

Table 5.3: Grammatical features encoded in different feature sets.

Feature	Type[N]	DT	Symbol
Grammatical role of REF	`cat[1]`	1–7	`gm`
Grammatical role of ANTE	`cat[1]`	5,6	`gm_p1`
Grammatical role of the 2^{nd} and 3^{rd} ANTE	`cat[2]`	6	`gm_p2`, `gm_p3`
Trigram grammatical roles of the three antecedents	`cat[1]`	7	`gm_tri`
Is REF the subject of this & the two previous sentences?	`bool[3]`	6	`subj_S`, `subj_prevS`, `subj_prev2S`
Is ANTE in the subject position?	`bool[1]`	6	`ante_subj`
Are REF and ANTE prepositional phrases?	`bool[2]`	5	`ref_pp`, `ante_pp`

Inherent features of a referent This category encompasses features that describe the intrinsic semantic properties inherent to referents. These features provide insight into the essential characteristics of the referents themselves, independent of their contextual usage or syntactic roles in discourse.

Table 5.4: Inherent features encoded in different feature sets.

Feature	Type	DT	Symbol
Animacy/semantic category	`cat[1]`	3,4,5,7	`anim`
Gender	`cat[1]`	5	`gender`
Plurality	`cat[1]`	5	`plur`

Position This category, as outlined in Table 5.5, includes features that provide information about the position of the referent (REF) within the text.

Recency The recency category, detailed in Table 5.6, focuses on features that quantify the distance between the REF and its ANTE. This distance is measured in various units such as words, noun phrases (NPs), markables, sentences, and paragraphs.

Competition This category, as outlined in Table 5.7, includes features that capture the competition between REF and other potential referents in the discourse. In this context, a competitor refers to any other entity in the text, regardless of its gender or type.

Table 5.5: Positional features of different feature sets.

Feature	Type[N]	DT	Symbol
Sentence Number	num[1]	6,7	sent_num
NP number	num[1]	7	np_num
Mention number	num[1]	1,5,6	ment_num
Referent number	num[1]	6	ref_num
How many times has REF occurred since the beginning? (1,2,3,4+)	cat[1]	4	count_bef
How many times does REF occur since the last change? (1,2,3,4+)	cat[1]	4	count_aft
Mention order (first, second, middle, last)	cat[1]	3	ment_ord
Does REF appear in the first sentence?	bool[1]	7	first_sent
Does REF appear at the beginning of a paragraph?	bool[1]	4	firstS_par

Table 5.6: Recency features of different feature sets.

Feature	Type[N]	DT	Symbol
Distance in number of words	num[1]	1,5	dist_w
Distance in number of NPs	num[1]	7	dist_np
Distance in number of markables	num[1]	5	dist_mark
Distance in number of sentences	num[1]	5,7	dist_sent
Distance in number of paragraphs	num[1]	5	dist_par
Distance to the nearest non-pronominal antecedent	num[1]	5	dist_full
Word distance (5 bins of 0–10, 11–20, 21–30, 31–40 and 40+)	cat[1]	2	bin5_w
Word distance (3 bins of 0–5, 6–12 and 13+)	cat[1]	3	bin3_w
Sentence distance (+/-2 sentences)	cat[1]	6	bin2_sent
Sentence distance (3 bins of 0, 1, 2+ sentences)	cat[1]	3	bin3_sent

Table 5.7: Competition features of different feature sets.

Feature	Type[N]	DT	Symbol
Does the previous RE refer to the same entity?	bool[1]	4	same_ante
Does REF have a competitor in the whole text?	bool[1]	3	compet_txt
Does REF have a competitor since the beginning of the text?	bool[1]	3	compet_beg
Is there a competitor between REF and ANTE?	bool[1]	3,6	compet_prev
Are there other referents in the same sentence?	bool[1]	6	compet_sent
Does the previous sentence contain another referent?	bool[1]	7	compet_prevS

Surrounding patterns This category, as detailed in Table 5.8, encompasses features related to the lexical content and Part-Of-Speech (POS) tags of the tokens surrounding the target RE.

Table 5.8: Surrounding pattern features of different feature sets.

Feature	Type[N]	DT	Symbol
Word unigram and bigram before and after REF	`char[4]`	4,7	`w_(uni|bi)_(bef|aft)`
Word trigram before and after REF	`char[2]`	7	`w_tri_(bef|aft)`
Three POS tags before and after REF	`char[6]`	7	`pos_(1|2|3)_(bef|aft)`
Punctuation type before and after REF	`cat[2]`	4	`punct_(bef|aft)`
Morphology of the previous and next words (-ed, -ing, -s, -)	`cat[2]`	4	`morph_(bef|aft)`
Is REF immediately followed by `and`, `but`, `then`?	`bool[3]`	6	`w_(and|but|then)`
Is REF between a comma and "and"?	`bool[1]`	6	`w_command`

Antecedent form This feature, known as `ante_form`, specifically addresses the form of the antecedent (ANTE). As noted by Bohnet (2008), in most prediction tasks, the input to this feature is typically a predicted referential form rather than one produced by a human. Therefore, the content of this feature might contain elements of uncertainty. It is used in datasets 1 and 5.

Referential status This category includes features that determine the newness of the referent within various textual scopes. For instance, whether REF is new in the sentence (`same_sent`) [datasets 1 & 2], new in the paragraph (`new_in_par`) [dataset 2], or new in the text (`new_in_text`) [dataset 2]. Some of these features can also be considered under the category of recency features.[4]

Protagonism In their study, Kibrik et al. (2016) introduced two measures of protagonism. The first, `protagonism1`, calculates the ratio of REF's chain length to the maximum chain length in the text. The second, `protagonism2`, compares REF's chain length to the sum of all REs in the text. These measures provide insights into the relative occurrence of REF within the overall text.

An overview of each feature set Table 5.9 provides a summary of the types of features used in each model, categorized according to the nine categories previously discussed. This table highlights the predominant type of features present

[4]The symbol `same_sent` is used to indicate that if a referent is not new in the sentence, it implies that its antecedent is in the *same* sentence.

in each feature set, offering an insight into the primary focus of each model's feature selection.

For instance, in the ICSI feature set, eight out of 14 features are dedicated to describing the surrounding patterns of REF. Similarly, the KIBRIK feature set includes four features that emphasize recency, and another four that provide information about the grammatical position of REF. Such an overview allows for a comparative analysis of the different models, revealing the varied emphases placed on certain feature categories and their potential impact on the models' performance in the REG-in-context task.

Table 5.9: General classes of the features used in the feature sets.

General classes	ISG	FERREIRA	OSU	ICSI	KIBRIK	UDEL	CNTS
Grammatical role	1	1	1	1	4	8	2
Inherent	0	0	1	1	3	0	1
Referential status / Givenness	0	3	0	0	0	0	0
Distance / Recency	2	1	2	0	4	1	2
Competition	0	0	3	1	0	2	1
Antecedent form	1	0	0	0	1	0	0
Pattern	0	0	0	8	0	4	12
Position	1	0	1	3	1	3	3
Protagonism	0	0	0	0	2	0	0
Total number of features	5	5	8	14	15	18	21

5.2.2.2 Applying feature sets to the WSJ dataset

In light of the preceding discussion, the WSJ corpus emerges as a particularly apt choice for our feature selection experiments. By applying the various feature sets to WSJ, we identify a total of 65 distinct features. However, some features could not be implemented due to the limitations of the existing annotations within the corpus. Notable examples include the elementary discourse unit (EDU) and rhetorical distance (RhD) measurements from Kibrik et al. (2016).

Furthermore, WSJ does not inherently have the paragraph-related information. To incorporate these features, we used paragraph information from the PDTB parser.[5]

[5] https://github.com/WING-NUS/pdtb-parser/tree/master/external/aux_data/paragraphs

Implementing these features presents its own set of challenges. For instance, the recency features, particularly the calculation of word distance between mentions, require a decision on whether to include or exclude punctuation in the count. In this study, punctuation is considered in calculating word distance.

The WSJ corpus, encompassing approximately 30,500 REs, is partitioned into 60% for training, 10% for validation, and 30% for testing. The RFS task is structured as a 3-way classification, with the classes being `pronoun`, `proper name`, and `description`. Detailed methodologies and results of the feature selection experiments are discussed in §5.2.3.

5.2.3 Feature selection experiments for assessing the features

To evaluate the efficacy of the proposed feature sets, our initial step involves constructing classifiers using these features applied to the WSJ corpus. Following this, we conduct two distinct feature selection experiments, named Experiment 1 and Experiment 2. These experiments aim to scrutinize the significance and impact of the individual features.

Upon analyzing the outcomes of experiments 1 and 2, we proceed with Experiment 3, in which the models are rerun with various subsets of features. The objective here is to determine whether models with these feature subsets can match or even surpass the performance of models employing the full array of features. This approach allows us to assess not only the individual contribution of each feature but also the optimal combination of features for efficient and effective RFS modeling.

5.2.3.1 Building models for predicting RF

In this study, we employ the *Random Forest* algorithm, a well-known ensemble learning method that operates by creating a multitude of decision trees during training and aggregating their results for classification (Nayak & Natarajan 2016, Biau 2012). One of the key advantages of using Random Forest is its ability to assess the *permutation importance* of variables, which is crucial for understanding the contribution of each feature in the classification models.

For the implementation of Random Forest, we use *ranger* (Wright & Ziegler 2017), a fast implementation of Random Forest in the R programming language. The performance results of the models, each employing the original features from their respective feature sets, are presented in Table 5.10.

Table 5.10 reveals that the model trained with the KIBRIK feature set achieves the highest accuracy, making it the best-performing model among those tested.

Table 5.10: Accuracy rates of the RFS models with their original features.

Dataset	IS-G	FERREIRA	OSU	ICSI	KIBRIK	UDEL	CNTS
Accuracy	0.68	0.601	0.697	0.69	0.793	0.624	0.723

It is followed by the models using the CNTS and OSU feature sets. In the subsequent section, we evaluate the features from each set to identify those that most significantly contribute to the predictive success of the models.

5.2.3.2 Experiment 1: Evaluating the importance of the features using RFI

In this experiment, we employ the built-in permutation importance feature of the Random Forest algorithm, known as RFI (Random Forest Importance), which ranks features based on their *importance* in the model (Breiman 2001). As outlined by Strobl et al. (2008), the process to determine the importance of a specific feature X_i involves several steps. Initially, the model is built, and its accuracy is assessed using *Out-of-bag* (OOB) observations. Then, the connection between the values of X_i and the model's outcome is disrupted by permuting all values of X_i. The model's accuracy is recalculated using these permuted values.

The permutation importance of X_i is quantified as the difference between the accuracy of the model with permuted values and the original model's accuracy. A feature with little to no impact on the model's predictions would show negligible change in accuracy upon permutation. Conversely, a substantial change in accuracy indicates the feature's significant role in the prediction task. Figure 5.1 illustrates the importance of various features across the seven models, with the *Mean Decrease in Accuracy* represented on the x-axis indicating the relative importance of each feature. The higher this score on the x-axis, the more important the feature.

In addition to assessing feature importance through RFI, we also calculate the p-values for the variables using the method of Altmann et al. (2010) . The test is based on the null hypothesis that permuting the values of the variables has no effect on the model's accuracy.

Out of the 65 distinct features analyzed, the null hypothesis was confirmed for four features from the UDEL dataset (`w_and`, `w_but`, `w_then`, and `w_command`) and one feature from the OSU dataset (`compet_txt`). This indicates that these particular features did not have a statistically significant impact on the accuracy of

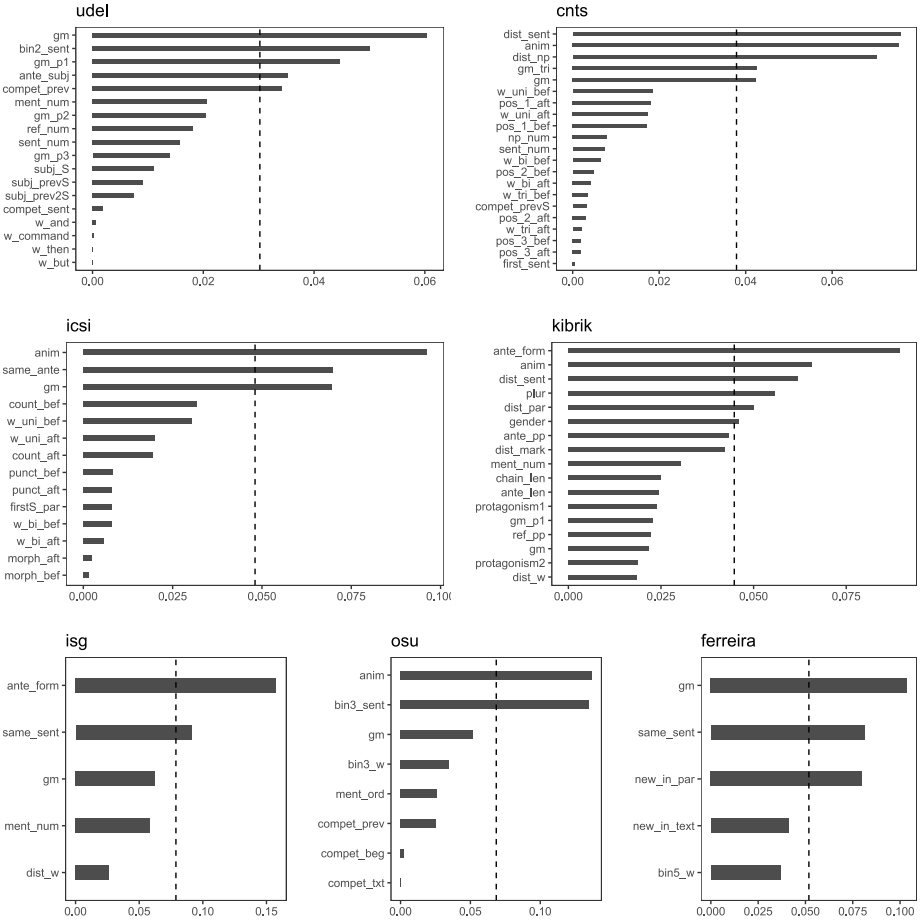

Figure 5.1: Variable importance plot of the RFS models. The y-axis lists the features of each model; the x-axis shows the permutation importance (Mean Decrease in Accuracy) of each feature.

the models. On the other hand, the remaining features demonstrated varying degrees of contribution to the models, with their effects on model accuracy being statistically significant.

5.2.3.3 Experiment 2: Evaluating the importance of the features using SFS

The second technique employed in our feature selection process is the Sequential Forward Search (SFS) algorithm. SFS begins with an empty set and incrementally adds features, continuing this process until there is no substantial improvement in accuracy. We have set a threshold of alpha=0.01 as the minimum required improvement for the algorithm to continue adding features. Once the improvement in accuracy falls below this threshold, the algorithm terminates.

For the implementation of SFS, we used the R package *mlr* (Bischl et al. 2016). In this framework, the learner specified for our model is `classif.randomForest`, and the chosen resampling strategy is *Holdout*. Table 5.11 provides an overview of the features selected from each feature set using the SFS algorithm.

5.2.3.4 Experiment 3: Exploring different feature subsets based on their importance

In this experiment, we first analyze the accuracy of various subsets within each feature set, as determined by the results of the RFI and SFS experiments. Subsequently, we investigate combinations of all features to identify an optimal consensus set for the REG-in-context task.

Subsets of each feature set Table 5.12 provides a comparative overview of model accuracies. The `original` column displays the initial accuracy of each model, as previously shown in Table 5.10. The columns `top1`, `top2`, and `top3` report the Random Forest performance using the one, two, and three most important features from each feature set, as identified by their permutation importance in Figure 5.1. For instance, in the OSU model, the top two features are `{anim, bin3_sent}`. The `top 50%` column presents the performance when applying the Random Forest to the top 50% of features in each set, as indicated by the dashed line in Figure 5.1. For KIBRIK, this includes features like `{ante_form, anim, dist_sent, plur, dist_par, gender}`. Finally, the `sfs` column reflects the Random Forest results using the feature subsets selected in the SFS experiment, as outlined in Table 5.11. An example of this is `{gm, gm_p1, bin2_sent, ref_num}` from the UDEL feature set.

Table 5.11: An overview of features selected from each set using SFS.

Model	Feature 1	Feature 2	Feature 3	Feature 4	Feature 5
IS-G	grammatical role of REF [gm]	RF of ANTE [ante_form]	whether REF and ANTE are in the same sentence [same_sent]		
FERREIRA	grammatical role of REF [gm]	whether REF and ANTE are in the same sentence [same_sent]	word distance in five bins [bin5_w]	new in paragraph [new_in_par]	
OSU	grammatical role of REF [gm]	animacy or entity type [anim]	sentence distance in three bins [bin3_sent]	word distance in three bins [bin3_w]	
ICSI	grammatical role of REF [gm]	animacy or entity type [anim]	whether previous RE refer to the same REF [same_ante]	number of times REF occurs since the last change [count_aft]	punctuation before REF [punct_bef]
KIBRIK	animacy or entity type [anim]	plurality [plur]	sentence distance [dist_sent]	paragraph distance [dist_par]	RF of ANTE [ante_form]
UDEL	grammatical role of REF [gm]	grammatical role of ANTE [gm_p1]	sentence distance in two bins [bin2_sent]	referent number [ref_num]	
CNTS	grammatical role of REF [gm]	animacy or entity type [anim]	sentence distance [dist_sent]	NP distance [dist_np]	POS tag of preceding word [pos_1_bef]

Table 5.12: The accuracy of various subset models based on the RFI (columns `top1`, `top2`, `top3`, and `top50%`) and SFS (column `sfs`) experiments. For each model, the best result is **boldfaced**, the second best result is *italicized*.

Feature sets	`original`	`top1`	`top2`	`top3`	`top50%`	`sfs`
IS-G	*0.680*	0.560	0.660	**0.684**	0.660	**0.684**
FERREIRA	**0.601**	0.492	0.577	*0.593*	*0.593*	**0.601**
OSU	*0.697*	0.556	0.665	0.694	0.665	**0.701**
ICSI	**0.690**	0.556	0.593	0.661	0.659	*0.681*
KIBRIK	**0.793**	0.560	0.632	0.712	*0.761*	*0.761*
UDEL	**0.624**	0.492	0.560	0.585	0.597	*0.608*
CNTS	**0.723**	0.556	0.661	0.670	0.696	*0.701*

An interesting observation is that for models like IS-G and OSU, the accuracy with SFS-selected features slightly exceeds that of the original feature sets. In the case of the FERREIRA model, the accuracies of the original and SFS-based models are identical.

Subsets of all features After examining subsets within individual feature sets, we now turn our attention to exploring various combinations of *all* features. This approach is guided by the insights gained from the RFI and SFS experiments, aiming to determine the most effective feature combinations for the REG-in-context task.

1. Initially, we apply the Random Forest algorithm to the set of features with the highest permutation importance from each model, as identified in Figure 5.1. This set includes {`gm`, `anim`, `dist_sent`, `ante_form`}, resulting in a model accuracy of 0.728.

2. Next, we test the algorithm on the union of the two top features from each feature set: {`gm`, `bin2_sent`, `dist_sent`, `bin3_sent`, `same_ante`, `ante_form`, `anim`, `same_sent`}. Surprisingly, this model yields a slightly lower accuracy of 0.723, contrary to our expectations.

3. We then train the algorithm on the union of all the SFS features, as detailed in §5.2.3.3. This combined set, comprising 19 distinct features, achieves an accuracy of 0.779 – the highest among the subsets tested.

4. Building on the success of the subset in item 3, we further apply the *SFS* algorithm to this set of 19 features. Our aim is to find an optimal balance between a minimal number of features and maximum performance. The SFS-selected subset consists of {`gm`, `ante_form`, `bin3_sent`, `anim`, `plur`, `dist_par`} and achieves an accuracy of 0.776. We then employ BF analysis with a beta distribution, setting a threshold of 0.01, to test whether there is evidence of a difference greater than or less than 0.01 between the best performing model, KIBRIK, with an accuracy of 0.793, and the new model. The evidence positively suggests that the two rates come from similar distributions, so they are not significantly different.[6]

5.2.4 The consensus set of features from a linguistic perspective

The consensus set of features, as identified in item 4 of the previous section, encompasses elements from four of the nine feature classes outlined in §5.2.2.1. Table 5.13 presents these features alongside their respective categories. In the following discussion, we aim to link the relevance of these categories to REG-in-context in light of various linguistic theories.

Table 5.13: The consensus set of features for the REG-in-context task.

Category	Feature
Inherent Feature	Animacy Plurality
Recency	Sentence distance (3 bins of 0, 1, 2+ sentences) Distance in number of paragraphs
Antecedent form	Antecedent form
Grammatical role	Grammatical role of the current RE

In the remainder of this section, we frequently refer to *row-wise* and *column-wise* data distributions. The term row-wise refers to the distribution within rows where the sum equals 1 (or 100%), and similarly, the term column-wise pertains to distributions within columns that also sum up to 1 (or 100%).

[6]While the original paper (Same & van Deemter 2020a) reports the BF results with a threshold of 0.05, we have adjusted it to 0.01 for this analysis to capture very small differences. The evidence still indicates that the two models' accuracy rates are derived from similar distributions.

5.2.4.1 Inherent features

Animacy and plurality are two inherent features of a referent that play an important role in predicting RF.

5.2.4.1.1 Animacy

The significance of animacy in RF selection has been well-noted in linguistic studies. As mentioned in Chapter 2, research has shown that pronouns are more frequently used for animate than inanimate referents (Dahl & Fraurud 1996, Fukumura & van Gompel 2011, Vogels 2014). Within the WSJ corpus, 32.2% of REs (n=9817) pertain to humans, while 67.8% (n=20654) are non-human. The row-wise distribution, marked by the sign `row%`, in Table 5.14 indicates that pronouns are used 40.8% of the time for human REs, compared to only 15.9% for non-human REs. This pattern suggests that models are more likely to predict non-pronominal forms (descriptions or proper names) for non-human referents.

Table 5.14: Cross-tabulated distribution of animacy (human vs. other) by RF (description, name, pronoun). For each animacy value, the first row shows the raw number of cases in the entire dataset (e.g., 2297 human REs with the form *description*). The second row (row%) shows the row-wise distributions of each value. For example, 23.4%, 35%, and 40.8% of all human REs are descriptions, proper names, and pronouns, respectively.

Animacy	RF		
	description	name	pronoun
human	2297	3510	4010
row %	23.4%	35.8%	**40.8%**
other	9723	7654	3277
row %	47.1%	37%	**15.9%**

However, distinguishing between descriptions and proper names can be challenging, as the distribution is relatively balanced between the two (47.1% for descriptions versus 37% for proper names). In what follows, we investigate whether the inclusion of more fine-grained animacy values contributes to the choice of descriptions vs. proper names.

Figure 5.2 provides a more detailed breakdown of animacy values in the WSJ corpus, illustrating how different animacy values (or entity types) correlate with RF choices. This mosaic plot reveals distinct trends: For example, geopolitical

entities (gpe) such as cities and countries are commonly referred to by proper names, whereas the *other* category, i.e., something other than a *city*, *country*, *person*, or *organization*, tends to be described using descriptions. This tendency might be attributable to the fact that many entities outside typical categories (like *objects*) lack proper names and are therefore referred to using descriptions or pronouns.

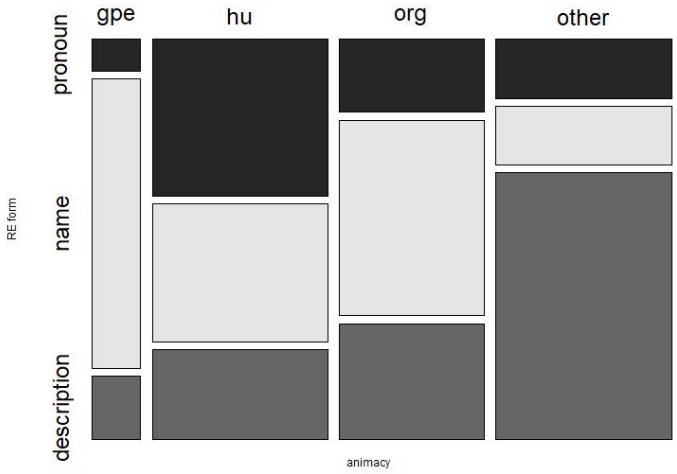

Figure 5.2: Mosaic plot of the animacy values (x-axis) and their RFs (y-axis). The animacy values are: gpe – geopolitical locations such as cities and countries, hu – human, org – organization, other – referents such as objects, time, other locations, etc. The colors black, light gray, and dark gray correspond to the classes pronoun, name, and description, respectively.

Tables 5.14 and Figure 5.2 collectively highlight how different animacy values influence the distribution of RFs. For a 2-way RFS task, a broad distinction between human and non-human referents suffices, while a more detailed classification, as illustrated in Figure 5.2, is beneficial for a 3-way classification task to effectively differentiate between descriptions and proper names.

5.2.4.1.2 Plurality

Plurality is another inherent feature that contributes significantly to feature-based REG-in-context models. Within the WSJ corpus, the majority of REs (82%, $n = 24963$) pertain to singular entities, while a smaller proportion (18%, $n = 5508$)

refers to plural entities. The corpus exhibits various types of plural REs, including conjoined noun phrases (NPs) (*South Korea, Taiwan, and Saudi Arabia*), plural definite descriptions (*the sponsors*), numerically quantified NPs (*the five senators*), and inherently plural NPs (*the Senate*). Table 5.15 presents the distribution of RFs for plural referents in WSJ, highlighting that proper names are used in only 10.53% of cases referring to plural referents.

Table 5.15: Frequency of the RFs of plural referents in WSJ.

	Frequency	Percent%
`description`	3038	55.16
`name`	580	10.53
`pronoun`	1890	34.31
Total	5508	100.00

This distribution suggests that the plurality feature plays a crucial role in guiding the choice between using a proper name and a description. Given that plural referents are more frequently represented as descriptions rather than proper names, the plurality feature can be a key determinant in the REG-in-context decision-making process. Further research is needed to fully understand the extent and nuances of this feature's contribution to the task of RFS.

5.2.4.2 Antecedent form

The referential form of the antecedent plays a pivotal role in the RFS task. As previously mentioned, in the practical implementation of REG algorithms, the RF of an antecedent is typically predicted, rendering it a piece of uncertain information (Bohnet 2008). Nonetheless, for this study, we had access to accurate antecedent data from the corpus and included these values in our evaluation. The significant impact of this factor on the RFS task aligns with findings from linguistic research (Kaiser 2003, Gundel 2008, Brilmayer & Schumacher 2021).

Bohnet (2008) suggests that the importance of antecedent form is partly due to a tendency to avoid repetitive expression use. While this may be particularly true for consecutive uses of pronouns, our models are not designed to confirm this hypothesis, as we only considered short antecedent–target chains, encoding merely the RF of the immediate antecedent.

Table 5.16 shows the likelihood of encountering different types of antecedents for each RF. The antecedents can be categorized as pronouns (`ante_pronoun`),

Table 5.16: The probability of different antecedent types (`ante-description`, `ante-name`, `ante-pronoun`) for each referential form. For instance, 69.4%, 17.7%, and 12.9% of all the REs with the RF type `description` have an antecedent of the type `description`, `proper name`, and `pronoun`, respectively.

Current RF	Antecedent Type		
	`ante-description`	`ante-name`	`ante-pronoun`
description %	69.4%	17.7%	12.9%
name %	13.1%	68.5%	18.4%
pronoun %	37.2%	31.8%	31.0%

proper names (`ante_name`), or descriptions (`ante_description`). The percentages in Table 5.16 indicate the likelihood that an RE has an antecedent of the same type. According to this table, descriptions have a 69.4% chance, proper names 68.5%, and pronouns 31% of sharing the same RF type with their antecedent. Notably, the distribution of antecedent types for pronouns is relatively balanced (37.2% for description antecedents, 31.8% for name antecedents, and 31% for pronoun antecedents), while non-pronominal REs predominantly share the same RF type as their antecedent. This finding is in line with Brilmayer & Schumacher (2021)'s observation in an ERP study, where they noted that pronoun anaphors, unlike noun anaphors, exhibit less dependence on the RF of their antecedent.

5.2.4.3 Grammatical role

Numerous studies, particularly those based on Centering Theory, have established that referents in the subject position are more likely to be pronominalized in subsequent sentences (Brennan et al. 1987, Brennan 1995, Kaiser 2010). These research efforts have typically emphasized the subjecthood of the antecedent rather than the subjecthood of the current mention. However, our analysis indicates that the grammatical role of the current mention may play a more pivotal role than that of the antecedent in predicting RF. This observation is visually represented in Figure 5.3, which displays a mosaic plot illustrating the correlation between various grammatical roles and their respective RFs.

An intriguing aspect of our findings is the divergence from traditional linguistic studies, which predominantly emphasize the role of the subject position. Our analysis reveals a rather equal distribution of RF types across instances where the referent is in the subject position. In contrast, the other two categories – pos-

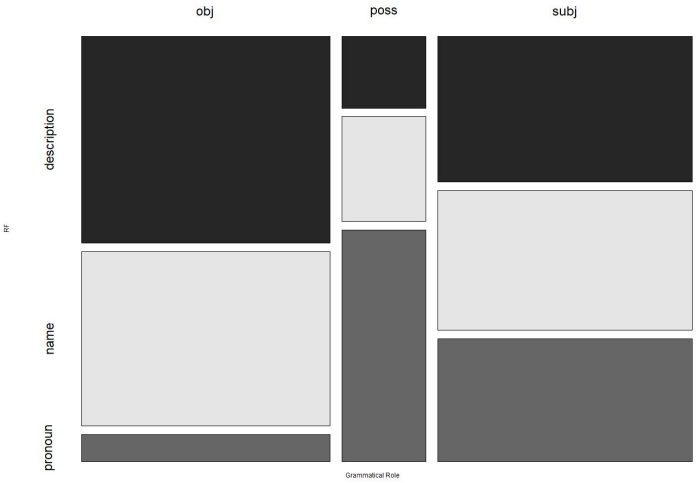

Figure 5.3: Mosaic plot of the RE's grammatical roles (x-axis) and their RFs (y-axis). The grammatical role values are: `subj` – subject, `poss` – possessive determiner, `obj` – other grammatical roles. The colors black, light gray, and dark gray corresponds to the classes `description`, `name`, and `pronoun`, respectively.

sessive modifiers and objects – demonstrate more distinct patterns in their RF usage.

Specifically, possessive modifiers predominantly take the form of pronouns. Conversely, referents in the object position are infrequently realized as pronouns. This distinct pattern suggests that the grammatical roles of possessive modifiers and objects are particularly influential in determining whether a referential expression will be pronominal or non-pronominal. These findings highlight the nuanced ways in which different grammatical roles can guide the selection of RFs, underscoring the importance of considering a range of grammatical positions beyond just the subject in the study of RF.

5.2.4.4 Recency

Recency is a concept frequently emphasized in linguistic research, though its definition often lacks specificity. In our study, we propose a more concrete definition of recency, suggesting it should be primarily quantified by the number of sentences separating the antecedent and the current RE. A secondary, yet effective, measure is the number of paragraphs between them. This operational

definition aligns with the linguistic tradition's inclination towards *higher-level* measures of discourse structure, as highlighted in works by Fox (1987b), Tomlin (1987a), Henschel et al. (2000), and Arnold et al. (2009).

Further exploration and analysis of the recency concept will be undertaken in study C of this chapter, where we aim to delve deeper into its implications and applications within the context of RFS.

5.2.5 Summary and discussion of study B

The primary objective of this study was to conduct a systematic examination of feature-based RFS models, with the aim of identifying the most effective features driving their success. By analyzing the feature sets of computational RFS studies from a linguistic perspective, we endeavored to bridge the gap between the computational application of these features and the linguistic theories underpinning them.

Following two feature selection experiments conducted across seven distinct feature sets, as well as the methodology detailed in §5.2.3.4, we identified a consensus set comprising six key features. The significance and functionality of these features were further explored and elucidated in §5.2.4.

This comparative analysis between the consensus set and previously established feature sets not only sheds light on the mechanics of feature-based RFS models but also offers insights that are relevant to both computational RFS research and linguistic studies. We will discuss the implications of this systematic analysis for both these fields.

5.2.5.1 Implications for feature-based RFS studies

Our findings suggest that a model equipped with a limited yet well-chosen set of features can achieve satisfactory performance. Notably, our proposed model, using merely six features, demonstrates performance comparable to the best-performing model in the REG-in-context literature, which employs 2.5 times as many features.

One key insight from our research is that the inclusion of a large number of features in a model does not necessarily guarantee optimal performance. The results presented in Table 5.12 indicate that models with a carefully selected subset of features often perform as well as those with a more extensive feature set. An interesting observation from Table 5.12 is that two models actually exhibited slightly improved performance when utilizing a subset of their original features, compared to their full feature sets. This finding warrants further investigation to

understand why the exclusion of certain features from these models resulted in enhanced performance.

Table 5.6 reveals that all systems, except ICSI, incorporate some form of recency measurement. These encodings vary, with some systems using lower-level units such as word, NP, and intervening RE counts, while others employ higher-level units like sentences and paragraphs. In both feature selection experiments, metrics at the higher level are consistently deemed more important than their lower-level counterparts. This disparity underscores the notion that different operationalizations of a concept contribute unevenly to a model's success. The nuances of these encoding strategies and their impact on RFS models will be further explored in study C.

5.2.5.2 Implications for linguistics

In §5.2.2.1, we categorized the features employed by earlier REG-in-context models into nine broad categories. Several of these, such as grammatical role, recency, and referential status, align with the prominence-lending cues frequently highlighted in linguistic research (Kaiser & Trueswell 2011, Gundel 2003, von Heusinger & Schumacher 2019). Meanwhile, other categories like surrounding lexical patterns are more commonly found in computational studies. Intriguingly, our feature selection experiments revealed that the most contributive features for the REG-in-context task tend to fall into the categories emphasized in linguistic tradition (namely, inherent features, recency, grammatical role, and antecedent form). This observation suggests a convergence of mechanisms in both the production and generation of referential expressions in context.

However, notable differences also emerge. For instance, as discussed in §5.2.4, linguistic studies typically focus on the grammatical role of the antecedent, particularly linking the prominence of a referent to the subjecthood of its antecedent. Contrastingly, only two of the feature sets in §5.2.2.1 incorporate the grammatical role of the antecedent, while all seven feature sets consider the grammatical role of the current RE. Furthermore, our feature selection experiments indicate that it is the grammatical role of the current RE, rather than that of the antecedent, which significantly impacts model performance.

This discrepancy raises an intriguing question: If the mechanisms underlying reference production and generation are similar, why do we observe this divergence in practice? Understanding the roots of this discrepancy could provide valuable insights into the interplay between linguistic theory and computational modeling in the domain of RFS.

While most features in the consensus set align with established themes in linguistic studies of reference in context, the role of plurality has not been as extensively explored. In the realm of computational studies, the generation of plural referents is addressed in several one-shot REG studies (van Deemter 2002, Gatt & van Deemter 2007a), but its exploration in REG-in-context studies remains limited. A notable exception is the work of Gatt & van Deemter (2009), who examined the behavior of plural REs in discourse using the GNOME corpus (Poesio 2000).

The GNOME corpus, with its detailed annotations, allowed Gatt & van Deemter (2009) to identify three main classes of plural anaphors: *identity*, *element*, and *split*. Their examples 1a–1c illustrate these classes. The study found that for identity anaphors, there is an equal tendency to use either pronominal forms or *same-head* non-pronominal forms. Conversely, the use of pronouns to refer to the other two types of anaphors is considerably less common.

(1) a. IDENTITY: [Precious metals such as silver and gold]$_i$, have been widely used from antiquity to the present day. [Their]$_i$ use is due, at least in part, to [their]$_i$ essential physical properties. (GNOME:TEXT3:34–35)

b. SPLIT: [[Caffieri]$_i$'s [wife]$_j$] bought a royal privilege . . . which allowed [them]$_{i+j}$ to gild bronze as well as cast it . . . (GNOME:GETTY:49)

c. ELEMENT: [[The Swiss artist Verena Sieber Fuchs]$_i$ and the [German-born but Irish-based artist Brigitte Turba]$_j$]$_{i+j}$ use discarded or waste materials as a source for their work. For [Sieber Fuchs]$_i$, old pill packaging, wrapper or film create possibilities. . . (GNOME:TEXT3:22–26)

Conducting an analysis of plural REs in the WSJ corpus, akin to the approach of Gatt & van Deemter (2009), could offer deeper insights into the use of plural REs within this corpus. Such an analysis would potentially enhance our understanding of the distributions observed in Table 5.15.

5.3 Study C: Computational interpretations of recency

Study B, presented in §5.2, underscored the significance of recency-based metrics in feature-based REG-in-context research. However, this study left unexplored the nuances of how different implementations of recency might impact the effectiveness of these metrics. The primary objective of our current study is to find out the most effective conceptualization of recency for the RFS task. Drawing on insights from Study B and established methodologies in linguistic research, we

hypothesize that recency metrics that encode *higher-level* distances contribute more to RFS than those based on *lower-level* distances. Additionally, we postulate that the effectiveness of recency metrics varies depending on corpus-specific characteristics, such as the genre and structure of texts.

To test these hypotheses, our approach involves two key steps: firstly, we will develop a comprehensive taxonomy of the various computational operationalizations of recency, providing a clear overview of the spectrum of metrics employed. Subsequently, we will conduct an evaluation of these recency metrics across two distinct corpora, GREC-2.0 and WSJ, which differ in terms of their genre and structural attributes. This comparative analysis aims to shed light on the relative effectiveness of different recency measures in varying textual contexts.

The structure of this section is as follows: §5.3.1 offers a concise overview of the concept of recency in both linguistic and computational studies, setting the stage for the development of a taxonomy of recency metrics in §5.3.2. An in-depth evaluation of these metrics is conducted in §5.3.3, followed by a summary and review of the study's findings in §5.3.4.

5.3.1 Recency in linguistic and computational linguistic studies

The concept of recency posits a direct relationship between an RE's form and the referent's distance from its antecedent (Vonk et al. 1992, Givón 1992, Arnold 2010). Specifically, the greater the distance between the referent and its antecedent, the more likely it is that richer RFs will be employed, and vice versa. As detailed in Chapter 2, linguistic literature interprets recency in three primary ways. The first two interpretations focus on measuring distance in terms of the number of sentences or clauses. Immediate context considers the antecedent's presence within the same or preceding utterance (or clause) (Hobbs 1978, Ariel 1990, Hitzeman & Poesio 1998, Henschel et al. 2000, Poesio et al. 2004), while non-local context typically encompasses a broader range of sentences (McCoy & Strube 1999, Arnold et al. 2009), with studies like Givón (1983) measuring up to 20 clauses back. The third interpretation, unit boundary, extends beyond sentence level to consider paragraphs (Fox 1987b, Tomlin 1987a). The question at hand is which of these interpretations is most effective for predicting RF in discourse.

Most feature-based REG-in-context models integrate various interpretations of recency. For instance, the binary feature of Bohnet (2008), indicating if the antecedent appears in the same sentence, aligns with the immediate context interpretation. Other models measure recency differently, such as counting the intervening words between an RE and its antecedent (Bohnet 2008, Jamison & Mehay 2008).

As shown in Table 5.6 from study B, the metrics in feature-based studies vary in their units of measurement (e.g., word distance vs. sentence distance) and encoding strategies. Some distances are quantified using natural numbers, while others are grouped into broader *bins*. For instance, in the following example, from GREC-2.0 (Belz et al. 2010), one could measure the distance between the expression *its* and its antecedent *Berlin* as 21 words (a natural number). Alternatively, following the approach of Castro Ferreira et al. (2016b), distances can be categorized into bins, placing the distance between *its* and *Berlin* in the 21–30 word bin.

(2) **Berlin**$_{(1)}$ is$_{(2)}$ the$_{(3)}$ capital$_{(4)}$ city$_{(5)}$ and$_{(6)}$ one$_{(7)}$ of$_{(8)}$ the$_{(9)}$ sixteen$_{(10)}$ federal$_{(11)}$ states$_{(12)}$ of$_{(13)}$ Germany$_{(14)}$ ·$_{(15)}$ With$_{(16)}$ a$_{(17)}$ population$_{(18)}$ of$_{(19)}$ 3.4$_{(20)}$ million$_{(21)}$ in$_{(22)}$ **its**$_{(23)}$ city$_{(24)}$ limits$_{(25)}$,...

5.3.2 Taxonomy of recency metrics in computational studies

To develop a comprehensive understanding of recency metrics used in Machine Learning (ML) literature, we compiled a wide range of metrics and constructed a taxonomy, as shown in Table 5.17. These metrics exhibit significant variation, particularly in three key aspects: *the type of antecedent*, *the unit of measurement*, and *the type of encoding*.[7]

5.3.2.1 The type of antecedent

Most metrics identify the antecedent as the closest preceding mention of the same entity. However, one metric (metric 14 in Table 5.17) measures the distance to the nearest full NP antecedent, rather than the nearest mention.

5.3.2.2 The unit of measurement

The metrics vary in their chosen unit for measuring distance. In Table 5.17, the units include: (1) *words* [metrics 1–3], (2) *sentences* [metrics 4–11], (3) *NPs* [metric 12], (4) *markables* (textual expressions between which coreferential relations are established) (Chiarcos & Krasavina 2005) [metrics 13–14], and (5) *paragraphs* [metric 15].

[7] Greenbacker & McCoy defined their recency metric as follows: "Referring expressions which were separated from the most recent reference by more than two sentences were marked as long distance references" (2009a: 101). This definition is interpreted in two ways in metrics 5 and 6 of our taxonomy.

Table 5.17: List of recency metrics collected from different ML studies.

Metric	Type of encoding & description	Meas unit	Reference
1	Numerical distance	word	Bohnet (2008)
2	Categorical distance (five bins of 0-10, 11-20, 21-30, 31-40 and 40+ words)	word	Castro Ferreira et al. (2016b)
3	Categorical distance (three bins of 0-5, 6-12 and 13+ words)	word	Jamison & Mehay (2008)
4	Numerical distance	sentence	Orăsan & Dornescu (2009), Hendrickx et al. (2008), Kibrik et al. (2016), Saha et al. (2011)
5	Categorical distance [1^{st} interp] (+/-2 sentences)	sentence	Greenbacker & McCoy (2009a)
6	Categorical distance [2^{nd} interp] (four bins of 0,1,2,+ 2 sentences)	sentence	Greenbacker & McCoy (2009a)
7	Categorical distance (three bins of 0, 1, 2+ sentences)	sentence	Jamison (2008), Saha et al. (2011)
8	Log distance	sentence	Saha et al. (2011)
9	Exponential distance	sentence	Modi et al. (2017)
10	Antecedent in the same sentence	sentence	Bohnet (2008)
11	Normalized distance	sentence	Newly implemented
12	Numerical distance	NP	Hendrickx et al. (2008)
13	Numerical distance	markable	Kibrik et al. (2016), Saha et al. (2011)
14	Numerical distance to the nearest non-pronominal antecedent	markable	Kibrik et al. (2016)
15	Numerical distance	paragraph	Kibrik et al. (2016)

5.3.2.3 The type of encoding

A key distinction between metrics, as exemplified in (2), is whether the distance is represented as a numeric value or categorized into bins. In Table 5.17, metrics 2, 3, 5, 6, 7, and 10 are categorical, while the others are numeric. The numeric values themselves are encoded differently across metrics: 1, 4, and 12–15 use natural numbers (including 0), metric 8 applies the natural logarithm, metric 9 uses an exponential form, and metric 11 involves normalized distance, as detailed below.[8]

Scaled/normalized sentence distance. To address disparities in sentence distance measurements, we normalize these values between $[0, 1]$ using the formula below. This metric, along with the other 14, will be further discussed in §5.3.3.

$$x_{\text{norm}} = \frac{x_i - x_{\text{min}}}{x_{\text{max}} - x_{\text{min}}}$$

5.3.3 Assessing recency metrics

In this section, we undertake a comprehensive evaluation of the recency metrics discussed previously. Our primary objective is to determine which metrics contribute most to the success of REG-in-context models. As a first step, we first outline the necessary prerequisites for our experimental approach in §5.3.3.1. Subsequently, in §5.3.3.2, we delve into the classifiers used for the assessment, examining their performance and the implications of the findings. Following this, two distinct methodologies for evaluating the recency metrics are explored: a Bayesian approach detailed in §5.3.3.3, and a sequential forward search (SFS) method presented in §5.3.3.4.

5.3.3.1 Prerequisites of the studies

This section outlines the foundational elements necessary for our investigation of recency metrics in REG-in-context models. We begin by detailing the datasets used for this assessment, followed by a description of the baseline model and the machine learning methodology employed to construct the RFS classifiers. The performance of each model, measured in terms of accuracy, is also presented.

Corpora used in this study A critical aspect of our study is determining how the choice of recency metrics might vary depending on the characteristics of the

[8] The exponential distance is not included for WSJ in this study.

corpus. Since corpora can differ significantly in size, genre (e.g., Wikipedia articles, newspapers, and medical reports), and document structure (such as length and sentence structure), we selected two corpora examined in Chapter 4: WSJ and GREC-2.0. These corpora differ notably in text genre and length-related attributes, providing a diverse basis for our analysis. We omitted GREC-PEOPLE due to its limited suitability for a 3-way classification task, primarily because only 4% of its REs are descriptions. Table 5.18 compares these two corpora, highlighting differences in document length, sentence and paragraph counts, and distribution of RE types. Based on this table, the WSJ documents are on average 3.5 times longer than the GREC-2.0 documents, with a mean length of 530.7 words for WSJ compared to 148.3 words for GREC-2.0. Additionally, WSJ documents contain notably more sentences and paragraphs.

Table 5.18: Comparison of the GREC-2.0 and WSJ corpora, focusing on length-related features and RF distributions.

Corpus features	GREC-2.0	WSJ
Genre	Wikipedia	Newspaper
Number of documents	1655	589
Average number of words per document	148.3	530.7
Average number of sentences per document	7.2	25
Average number of paragraphs per document	2.3	11
Average number of referents per document	1	15
Average number of REs per document	7.1	52.1
Average length of sentences	25.8	29.5
Number of descriptions	1613	6917
Number of proper names	2813	7695
Number of pronouns	4880	6953

For the 3-way RFS classification task, models must choose between `pronoun`, `proper name`, and `description`. First-mentioned referents, which lack an antecedent and therefore have no recency value, are excluded from our analysis. The total number of REs in GREC-2.0 and WSJ amount to 9306 and 21565, respectively, with 70% allocated to a training set and 30% to a test set.

Baseline algorithms and ML method To effectively assess the impact of recency metrics, we establish a baseline algorithm that excludes recency metrics, serving as a comparison point for the experimental algorithms incorporating

these metrics. The baseline algorithm features the grammatical role of the current and preceding mentions. This choice ensures consistency across both corpora, eliminating discrepancies that might arise from differing annotations.

The study uses a Multilayer Perceptron (MLP), a type of artificial neural network algorithm, for model training. This network is *feedforward*, meaning information flows in one direction from input to output. It includes two hidden layers with 16 and 8 units, respectively, which are internal layers that help process the data. These layers use the Rectified Linear Unit (ReLU) activation function to determine the output of each unit, effectively allowing the network to learn nonlinear patterns. The output layer employs the softmax function, converting the network's output into probabilities for classification.

The MLP undergoes training in 50 epochs, where each epoch is a complete pass of the entire dataset through the network. Training is conducted in batches of 50 samples to optimize learning efficiency. Since MLPs cannot directly handle categorical data, a transformation technique known as one-hot encoding is used. In one-hot encoding, each categorical value within the dataset is converted into a new, separate categorical column. These columns are then filled with binary values: a `1` is assigned to the column corresponding to the data point's actual category, and `0`s are assigned to all other columns. This process effectively converts each categorical value into a distinct binary vector. As a result, every integer value, previously categorical, is represented in a format that the MLP can process.

5.3.3.2 Building classifiers using MLP

Baseline algorithms In our study, we trained MLP algorithms on both the GREC-2.0 and WSJ corpora. These algorithms use the grammatical roles of the current RE and its antecedent to form the baseline models. The accuracies achieved by the baseline models are 0.585 for GREC-2.0 and 0.55 for WSJ, respectively.

Assessing recency metrics Each experimental algorithm, in addition to the two baseline features, incorporates a single recency metric. For instance, MODEL 4 includes the grammatical role of an RE and its antecedent, coupled with metric 4, which quantifies the numerical distance in the number of sentences. In total, we tested 28 experimental algorithms across the two corpora, encompassing 14 distinct recency metrics.[9] The rationale behind testing each recency metric separately is to isolate its individual contribution to the algorithm's success, avoiding confounding effects that could arise from combining multiple metrics. The

[9] Metrics 9 and 13 were not applicable to WSJ and GREC-2.0, respectively.

accuracies achieved by these experimental algorithms, which integrate varying recency metrics, are summarized in Table 5.19.

Table 5.19: Accuracy of experimental algorithms. The first column indicates the measurement unit of each metric, as detailed in §5.3.2.2.

Measurement Unit	**Name**	GREC-2.0	WSJ
–	baseline	0.585	0.55
Word	MODEL 1	0.60	0.576
	MODEL 2	0.594	0.551
	MODEL 3	0.592	0.572
Sentence	MODEL 4	0.607	0.62
	MODEL 5	0.588	0.582
	MODEL 6	0.608	0.622
	MODEL 7	0.602	0.622
	MODEL 8	0.607	0.611
	MODEL 9	0.609	-
	MODEL 10	0.589	0.597
	MODEL 11	0.602	0.604
NP	MODEL 12	0.59	0.623
Markable	MODEL 13	-	0.577
	MODEL 14	0.594	0.561
Paragraph	MODEL 15	0.625	0.616

While all experimental algorithms outperform the baseline in terms of accuracy, it remains to be determined whether the inclusion of recency metrics significantly enhances their performance. Notably, for the WSJ corpus, seven models (encompassing six sentence metrics and one NP metric) achieved higher accuracy than their GREC-2.0 counterparts with corresponding metrics. The significance of these findings and the performance of the experimental algorithms will be further evaluated in §5.3.3.3 and §5.3.3.4.

5.3.3.3 Bayes factor analysis

To investigate whether the experimental and baseline algorithms derive from distributions with similar or different underlying probability parameters, we con-

duct a Bayes Factor (BF) analysis using a beta distribution. Specifically, BF analysis is employed to determine if the differences in accuracy rates between models are less than or exceed a predefined threshold of 0.01, chosen to detect very small differences. The evidence is in favor of similar distributions if the difference in accuracy is below the threshold. If it is above the threshold, there is good evidence that the outcome comes from different distributions. A result suggesting different distributions implies that the inclusion of recency metrics substantively improves the performance of experimental algorithms. We assess the strength of the evidence for each experimental model compared to the baseline following the scale by Kass & Raftery (1995), as presented in Table 4.7 in Chapter 4. Here, we report only those results where experimental algorithms and baselines appear to originate from *different* distributions.

BF analysis of the GREC-2.0 models Comparing each experimental model's correct prediction rates with the baseline reveals that only MODEL 15, incorporating paragraph distance as a recency metric, shows positive evidence (BF = 3.286) of differing from the baseline. Other models outperform the baseline, but lack sufficient evidence to assert that they differ from the baseline.

BF analysis of the WSJ models For WSJ, eight models exhibit accuracy rates distinct from the baseline. Similar to GREC-2.0, MODEL 15 significantly differs. Additionally, six out of seven sentence-based recency metrics and MODEL 12 (using NP distance) also show very strong evidence of improved performance compared to the baseline. The exception is MODEL 5, which does not demonstrate a significant impact.

In the WSJ model assessments, it is notable that all sentence-based recency metrics, with the exception of metric 5, significantly enhance model performance. This discrepancy invites a deeper examination of why metric 5 is an outlier. A key distinction of metric 5 is in its categorical binning approach for measuring immediate context. Specifically, it combines instances where the antecedent and the referent are either in the same sentence (distance = 0) or adjacent sentences (distance = 1) into a single category. This binning strategy contrasts with the methods employed by other sentence-based metrics, which typically assign these scenarios to distinct categories. In essence, metrics other than metric 5 provide a finer-grained distinction by separately categorizing instances where referents share the same sentence from those where they are separated by a single sentence.

Table 5.20: The BF analysis provides the ratio of probabilities, assessing whether the underlying accuracy rates are within a 1% margin of each other or not. According to the scale by Kass & Raftery (1995), as presented in Table 4.7, there is very strong evidence suggesting that the accuracy rates of all these models differ significantly from the baseline. In the column labeled `Def`, a brief definition of the metrics is provided, in accordance with Table 5.17. For example, `cat(4)` refers to the categorical distance across four bins.

Name	Measurement unit	Definition	BF
MODEL 4	sentence	num	54×10^{8}
MODEL 6	sentence	cat (4)	19×10^{9}
MODEL 7	sentence	cat (3)	37×10^{9}
MODEL 8	sentence	log	78×10^{5}
MODEL 10	sentence	binary	14×10^{2}
MODEL 11	sentence	norm	12×10^{4}
MODEL 12	NP	num	56×10^{9}
MODEL 15	paragraph	num	16×10^{7}

BF analysis of the best performing models We further compare the models with best performance across different measurement units. The only difference between the models is the recency metric they use. If the accuracy difference exceeds the threshold, we attribute it to the use of distinct recency metrics. Table 5.21 lists the top algorithms for each measurement unit.

Table 5.21: Best-performing algorithms of each measurement unit.

Meas Unit	GREC-2.0	WSJ
Word	MODEL 1	MODEL 1
Sentence	MODEL 9	MODEL 7
NP	MODEL 12	MODEL 12
Markable	MODEL 14	MODEL 13
Paragraph	MODEL 15	MODEL 15

I. GREC-2.0 models Upon making one-to-one comparisons between the best-performing models for each measurement unit in the GREC-2.0 corpus, we observe that these models do not exhibit statistically significant differences in performance.

II. WSJ models For the WSJ corpus, a different picture emerges. The models employing sentence (model 7), NP (model 12), and paragraph (model 15) recency metrics do not seem to evidentially differ from each other, suggesting these metrics are equally effective. However, when these three models are compared with the best-performing models using word and markable metrics, there is a significant shift in accuracy rates exceeding the 0.01 threshold. This indicates that models 7, 12, and 15 are statistically different from those using word and markable metrics.

Upon contrasting GREC-2.0 and WSJ, it becomes evident that the incorporation of recency metrics yields more substantial improvements in WSJ. While only one GREC-2.0 model significantly outperforms its baseline, eight WSJ models exhibit statistically superior performance compared to their baseline. Notably, metrics based on sentences, paragraphs, and NPs significantly enhance the performance of the WSJ algorithms.

While this section has focused on individually assessing recency metrics, it is important to note that many REG-in-context models incorporate multiple recency metrics. In the following section, we will explore a feature selection study aimed at identifying the most effective combinations of recency metrics for REG-in-context models.

5.3.3.4 Sequential forward search

To assess the combined impact of various recency metrics in REG-in-context models, we employ SFS with the learner `classif.mlp`. We also use 5-fold cross-validation for resampling. This approach follows the methodology detailed in §5.2.3.3. The aim is to identify which combinations of recency metrics result in the most effective predictive models.

GREC-2.0 experiment For the GREC-2.0 corpus, SFS identifies two recency metrics as particularly influential: metric 15 (distance measured in paragraphs) and metric 9 (exponential distance in sentences). Incorporating these metrics into our model yields an accuracy of 0.637. Subsequent BF analysis confirms that the outcome of this model is statistically different from the baseline.

WSJ experiment In the WSJ experiment, SFS selects metric 15 (distance in paragraphs) and metric 8 (logarithmic distance in sentences) as the most impactful combination. The model trained with these metrics achieves an accuracy of 0.631. BF analysis provides very strong evidence that the performance of this model differs significantly from the baseline, reinforcing the importance of paragraph distance in the context of WSJ.

The results from both corpora consistently point to the relevance of paragraph-based distance as a key factor in REG-in-context studies. This insight aligns with the broader theme of emphasizing higher-level structural elements in text for effective reference generation. A more in-depth exploration of the role of paragraph structure in REG-in-context models is presented in the next chapter.

5.3.4 Summary and discussion of study C

In study C, we delved into the diverse interpretations of recency to identify the most effective metrics for predicting the form of referring expressions in context. Additionally, we examined how corpus-specific characteristics, such as text genre and structure, influence the choice of recency metrics. This study's findings are of interest to both theoretical linguists and computational linguists, who have explored the relationship between recency and RF.

The concept of recency has often been explored in the linguistic tradition without a clear definition being offered. In the computational tradition, on the other hand, researchers have dwelt less on theoretical justifications but have had to provide precise definitions to ensure that their algorithms can handle a wide range of inputs. For example, Kibrik et al. (2016) proposed seven distinct implementations of recency based on various units of measurement, while Saha et al. (2011) explored different sentence-related metrics.

Interestingly, computational research has ventured beyond conventional sentence or paragraph-level metrics, incorporating unconventional measures like word count, markables, and noun phrases (NPs). This expansion of recency metrics in computational studies potentially paves the way for novel insights and could prompt a reevaluation of recency concepts within linguistic theory. In many computational works, however, there is no explanation as to why a particular metric or encoding method was chosen over another. Our results contribute the following to the literature:

Creating a taxonomy of recency metrics Our study not only gives an overview of recency interpretations in the linguistic tradition but also, for the first time to our knowledge, establishes a comprehensive taxonomy of recency metrics used

in feature-based ML studies. This taxonomy clarifies the nuances between these metrics, providing a foundational step for analyzing various aspects of recency and for developing new, refined metrics.

Assessing a wide range of recency metrics Using an MLP algorithm, we constructed classifiers based on individual recency metrics. Subsequent Bayes factor analysis assessed whether models with recency metrics diverged significantly from baseline models. Additionally, a comparative BF analysis among the top-performing models of each measurement unit was conducted to verify if noticeable differences existed in their outcomes.

As indicated in Table 5.20, the WSJ models integrating NP, paragraph, and sentence metrics showed a substantial difference (>0.01) from the baseline. There is also strong evidence that these models significantly differ from those incorporating word and markable distance measures. Sequential Forward Search experiments demonstrated that a combination of paragraph and sentence metrics yielded the best results for both corpora. This aligns with study B's findings, where sentence- and paragraph-based distances were key components in the consensus feature set.

The combined results from BF analysis and SFS suggest that higher-level metrics (paragraph and sentence) may enhance algorithm performance more effectively than lower-level metrics (word and markable). This observation leads to a pertinent question: Why is a higher-level measurement, like the distance in the number of sentences, more effective than a lower-level measurement, such as the distance in the number of words? This consideration is particularly significant given that word-based distance measures might more accurately reflect the physical proximity between mentions, considering the considerable variability in sentence lengths.

One potential explanation is that the physical distance between referents does not influence their prominence status. Instead, what may be more critical are the transitions between distinct units. This reasoning can explain the effectiveness of sentence and paragraph metrics, since the former involve a transition between sentences, and the latter a transition between paragraphs. However, this explanation does not sufficiently account for the observed success of the NP-based metric.

Another notable finding is the varying effectiveness of different solutions. For example, in one metric, sentence distance categorized into two bins yielded a marginal performance improvement, while another metric with four bins significantly enhanced accuracy. This underscores the importance of how recency is operationalized in computational models.

Significance of paragraph-based distance Both BF and SFS analyses underscore the importance of paragraph-based distance in the success of algorithms. This finding corroborates the discussion in Chapter 2, where transitions across episode boundaries were shown to influence referent accessibility and form. Including paragraph distance can thus markedly improve algorithms' ability to predict referential form. Despite its significance, paragraph distance has been underutilized in computational studies, with only one featured study from Table 5.17 incorporating it. The impact of paragraph boundaries on REG-in-context tasks will be further explored in Chapter 6.

Importance of the choice of corpus Our study reveals that the impact of recency metrics in REG-in-context models is significantly influenced by the characteristics of the chosen corpus. This was particularly evident when comparing the effects of recency measures in the WSJ and GREC-2.0 corpora.

In GREC-2.0, only the metric measuring distance in the number of paragraphs yielded a distribution significantly different from the baseline when considered independently. In contrast, eight distinct recency metrics in WSJ led to significant divergences. This discrepancy can be attributed to the distinct structural features of these corpora. As shown in Table 5.18, WSJ texts are substantially longer, with nearly four times as many words, sentences, and paragraphs compared to GREC-2.0. This difference in length-related features might account for the varying importance of recency metrics in the respective models.

Furthermore, the genre of each corpus could play a role. GREC-2.0 documents, typically introductory sections of Wikipedia articles, often focus on a single main topic. This format likely results in repeated mentions of the referent across sentences, diminishing the relevance of metrics like sentential distance.

Table 5.22 presents the distribution of sentential distances in GREC-2.0. Approximately 88% of REs are found within the immediate context of their antecedents (distance < 2 sentences). This concentration in immediate context may reduce the effectiveness of sentential recency metrics in this corpus.

Our findings suggest that the complexity and diversity of a text's discourse structure significantly influence the efficacy of recency metrics. Therefore, un-

Table 5.22: Sentence-based recency distributions in GREC-2.0.

distance (num)	0	1	2	3	4	5	<5
total	30.98	56.86	7.54	2.57	1.03	0.48	0.55
total_cumulative	30.98	87.83	95.37	97.94	98.98	99.45	100.00

derstanding the genre and structural features of the textual source is imperative when selecting recency metrics for computational studies.

5.4 Discussion and final remarks

The studies presented in this chapter both dealt with the choice of features for the feature-based REG-in-context task.

Study B functioned as a *survival of the fittest* challenge among a variety of features, culminating in the selection of six key features. These features spanned four primary categories: grammatical role, antecedent form, inherent characteristics, and recency. The principal aim of this study was to propose a concise set of features as a robust foundation for constructing effective feature-based REG-in-context models.

A unique aspect of this study was its endeavor to not just identify but also elucidate the significance of these features in the context of REG-in-context models. Although computational models do not rely on explicative frameworks, understanding the underlying reasons for a model's performance is crucial for further improvements. However, one limitation of study B was its relative lack of focus on the specific operationalization of these features, a crucial aspect for comprehending their full impact.

In contrast, study C delved into the intricate details of one feature category, providing a nuanced understanding of how different dimensions of a feature (such as the unit of measurement and encoding method) can influence the effectiveness of REG-in-context models. Focused exclusively on recency-based metrics, this study revealed key insights into the multifaceted nature of this feature class.

Importantly, study C extended the findings of study B by highlighting that while recency is a pivotal feature in reference studies, the choice of what and how to encode this feature is critical. For instance, the study demonstrated that sentence-based recency metrics were more impactful for models using the WSJ corpus compared to those using GREC-2.0, emphasizing the need to consider corpus-specific characteristics in feature selection. Additionally, the study illuminated that not all operationalizations of sentence-related recency metrics contributed equally to the models' success, underscoring the necessity of a strategic approach in feature encoding.

The insights from study C have broader implications, extending beyond recency metrics to other feature-based studies in reference. They underscore the importance of meticulous feature analysis and selection in developing sophisticated and effective computational models for studying reference in context.

6 The effect of paragraph structure on the choice of referring expressions

6.1 Introduction

The preceding chapters discussed the selection of corpora and features for the task of RFS. While the two studies in Chapter 5 have demonstrated the significance of a paragraph-based recency feature, the exact manner in which this feature influences the choice remains unclear. Moreover, those studies highlighted the relevance of only a single paragraph-related factor. It is uncertain whether additional aspects of paragraph structure contribute to its relevance for this task. This chapter will examine various paragraph-related factors that might be relevant to the choice of RF.

To better understand the potential importance of including paragraph structure in REG-in-context, consider (1). This example demonstrates the paragraph structure in an excerpt about Walter White, a character from the television series "Breaking Bad". [1] To illustrate the realization of the REs, the first paragraph is presented in full; the content of the subsequent paragraphs is summarized up to the point where the character Walter White is first mentioned in the subject position.

(1) a. **Walter Hartwell "Walt" White Sr.**, also known by **his** clandestine pseudonym and business moniker Heisenberg, was an American drug kingpin. A former chemist and high school chemistry teacher in Albuquerque, New Mexico, **he** started manufacturing crystal methamphetamine after being diagnosed with terminal lung cancer. **He** initially does this in order to pay for **his** treatments and secure the financial future of **his** family: wife Skyler, son Walter Jr., and infant daughter Holly, but confesses before **his** death that **he** actually did it for **himself**, due to being good at it and feeling alive.

b. In the 1980s, **Walt** co-founded the company Gray Matter Technologies ...

[1] `https://breakingbad.fandom.com/wiki/Walter_White`

c. After joining **his** brother-in-law and accomplished DEA agent Hank Schrader on a drug bust and hearing from Hank about the lucrative profits that drug manufacturing and dealing could produce, **Walt** decided to use **his** knowledge of chemistry to become involved in the drug trade ...
d. While initially heavily reluctant to use violence, **Walt** gradually came to see it as a necessity ...
e. After accumulating over $80 million USD from **his** involvement in the drug trade, and following a resurgence in **his** cancer, **Walt** retired from the drug business permanently ...
f. **Walt** returned home and attempted to escape with **his** family ...
g. **Walt** went on to confront and then ...

As the first paragraph (1a) demonstrates, Walter White is introduced for the first time using his full name, *Walter Hartwell "Walt" White Sr.* It is noteworthy that in the following paragraphs, the initial mention of Walter White in the subject position is *consistently* non-pronominal. Despite the proximity of just one sentence from its antecedent in the second paragraph (1b), Walter White is referred to in a non-pronominal form. Does this non-pronominal usage stem from the transition between paragraphs?

Additionally, the excerpt features frequent use of pronominal REs within the first paragraph. If queried about the primary subject of the first paragraph, one would likely respond that it centers on Walter White. Is this pattern merely coincidental, or does the prominence of the character within the paragraph contribute to the exclusive use of pronouns?

In all paragraphs except the third (1c), Walter White is first mentioned in the subject position. In the third paragraph, the possessive determiner *his* serves as a cataphoric reference to reintroduce him. Does the grammatical role of the RE, being a possessive determiner, influence the preference for a pronominal form at the beginning of this paragraph?

The primary objective of this chapter is to explore whether paragraph structure influences the choice of RF. The observations mentioned earlier hint at several aspects of paragraph structure that could affect this decision: (1) transitions between paragraphs, (2) the prominence of referents within a paragraph, and (3) the grammatical roles of REs that reintroduce a referent into a paragraph. This chapter comprises two studies, study D and study E, which respectively assess the impact of paragraph structure on the choice of RF in a corpus and through a feature-based REG-in-context analysis.

In study D, I analyze various factors within the WSJ corpus, hypothesizing that (1) *paragraph-prominent* entities are substantially more likely to be pronominalized, (2) *paragraph-new* and *paragraph-initial* REs are substantially more likely to be non-pronominal, and (3) paragraph-new REs are more likely to be pronominal if the referent is prominent in the current (P_i) and the previous (P_{i-1}) paragraph.

Study E integrates several paragraph-related features into a feature-based pronominalization model, hypothesizing that the inclusion of paragraph-related information substantially improves the performance of feature-based REG-in-context models. This study also aims to understand *why* the inclusion of paragraph-related features is critical for the task. To augment the explainability of its findings, the study employs two methods: a *SHapley Additive exPlanations* (SHAP) analysis and an *error analysis*.

The chapter is structured as follows: §6.2 discusses the concept of a paragraph and reviews studies that (1) underscore the significance of paragraphs or episodic information more broadly, and (2) delve into the interactions between paragraph boundaries and RF choices. A corpus analysis of the WSJ corpus is conducted in Study D, as detailed in §6.3, to gain deeper insights into paragraph structure. Study E, described in §6.4, introduces various paragraph-related features. It then proceeds with a series of REG-in-context model evaluations and an error analysis to evaluate the contribution of paragraph-related information to feature-based REG-in-context models.

6.2 Paragraph boundary: Linguistic theories

In the process of writing, authors typically possess an intuitive sense about where to conclude a paragraph and where to initiate a new one. Moreover, poor paragraphing can hinder readers' comprehension of the text, as noted by Hofmann (1989). Despite this, there are no universally agreed-upon characteristics that define a paragraph (Hofmann 1989, Filippova & Strube 2006). Linguistic theories offer limited insight about what paragraph breaks signify or on the criteria for dividing texts into paragraphs. Complicating matters further, there is often more than one acceptable method for structuring paragraphs.

Notwithstanding the existence of various approaches to paragraph formation, Hofmann (1989) points out that certain instances clearly warrant the start of a new paragraph, making any other choice incorrect. Similarly, there are circumstances where extending a current paragraph is inappropriate. Nonetheless, the focus of this section is not to critically analyze the definitions of paragraph boundaries. Instead, it concentrates on examining studies that have, in one form or another, incorporated the concept of paragraph boundaries.

6.2.1 Paragraph boundary: Its detection, importance, and applications

This subsection delves into various studies that have explored the significance of paragraph boundaries from diverse perspectives, such as their impact on information processing. In her research, Stark (1988) investigated how paragraph boundaries influence reading time and perceived importance of ideas. Her findings suggest that the presence of a paragraph boundary can heighten the reader's attention to the opening sentence of the paragraph, leading to a higher perceived importance of that sentence.

Moreover, Stark (1988) conducted an experiment with 21 participants who were presented with three essays stripped of their paragraph markers. The participants were tasked with identifying where they believed the paragraph boundaries should be, denoting them with a slash between sentences. The aim was to determine the degree of consensus on the placement of paragraph boundaries. Stark (1988) observed varying levels of agreement among participants (with a minimum agreement rate of 0.25 and a maximum of 0.47) and noted the highest accuracy in identifying paragraph boundaries as 0.6 for one of the texts. Participants, according to Stark, generally concurred with each other and with the original authors of the essays to a greater extent than would be expected by random chance. However, Stark did not specify the method used to calculate this "chance level".

One key finding of Stark (1988)'s study is the influence of *over-reference* – the use of full forms when a pronoun would suffice – on the detection of paragraph boundaries in unparagraphed texts. For instance, Stark argues that the perception of a paragraph boundary in (2b) "is consistent with evidence that over-reference is used by speakers at episode boundaries" (1988: 291).

(2) a. **It** [spring] comes seeping in everywhere like one of those new poison gases which pass through all filters.

b. **The spring** is commonly referred to as "a miracle" and during the past five or six years this worn-out figure of speech has taken on a new lease on life. (Orwell, 1945, p. 143)

Stark (1988) also discovered that individuals do not consistently divide texts into paragraphs of uniform length. If paragraph boundaries were merely aesthetic elements, one might expect paragraphs of similar lengths. However, Stark (1988) found that paragraph lengths varied significantly. Interestingly, people demonstrated a notable ability to identify the boundaries of paragraphs, even when their lengths deviated considerably from the average.

In summary, her experiments revealed a number of key insights: (1) paragraph boundaries are not arbitrarily placed, as evidenced by people's agreement on their detection exceeding chance levels; (2) paragraphs are not solely aesthetic constructs, as indicated by the successful identification of paragraphs of varying lengths; and (3) Over-reference is frequently employed by individuals to discern paragraph boundaries.

From a computational standpoint, paragraph boundaries hold significant importance in various applications, including document summarization and the creation of layouts for generated texts. However, as Sporleder & Lapata (2006) notes, paragraph boundary detection has received less attention than the closely related task of topic segmentation. This is partly because paragraph boundaries are often explicitly marked in texts by a new line and additional space. Yet, in newly generated texts from text-to-text or speech-to-text applications, a clearly defined paragraph structure is usually absent (Sporleder & Lapata 2006). Only a handful of studies (Bolshakov & Gelbukh 2001, Sporleder & Lapata 2004, 2006, Filippova & Strube 2006) have proposed models for detecting paragraph boundaries using linguistic cues.

Bolshakov & Gelbukh (2001) employed text cohesion as an indicator for paragraph boundary detection, using collocation networks and semantic links between words to assess cohesion. They posited that the connection between the first sentence of a paragraph and its preceding sentence is generally weaker than the links between sentences within a paragraph.

Similarly, Filippova & Strube (2006) used cohesive features based on discourse cues, pronominalization, and information structure to identify paragraph boundaries. Their research, involving 970 texts from the German Wikipedia, demonstrated that pronominalization and information structure are pivotal in detecting paragraph boundaries. Echoing Stark (1988), they argued that over-reference is indicative of a new paragraph's onset. Hence, if a sentence employs a non-pronominal form where a pronominal reference would be suitable, it likely signals the start of a new paragraph.

Contrasting with Filippova & Strube (2006), who concentrated on using RF to identify paragraph boundaries, the studies in this chapter assess how paragraph boundaries influence the choice of RF. Before delving into the corpus- and machine learning feature-based experiments in §6.3 and §6.4, §6.2.2 will explore studies that emphasize the impact of paragraph or episode boundaries on the choice of RF.

6.2.2 Paragraph boundaries as determinants of RF

In the previously discussed Walter White example, it was observed that all subject REs at the beginning of each paragraph were non-pronominal. This pattern of using non-pronominal forms at paragraph openings is posited to be correlated with the presence of paragraph boundaries, as suggested in the works of Hinds (1977) and Hofmann (1989). Notably, Hofmann (1989) views paragraph breaks as *barriers* to anaphora, suggesting that these breaks significantly influence the choice of RF.

> A pronoun or other anaphoric element cannot be used if its nearest antecedent is embedded in a preceding paragraph. Even in the cases that the pronouns are sufficient and a non-pronominal expression is redundant, when there is a paragraph break, a non-pronominal form is being used. Paragraph boundary can be seen as machinery that is deactivating most of what precedes (1989: 241).

Hofmann illustrates the inaccessibility of pronouns across paragraph boundaries with an analogy: Consider a teacher using a blackboard to present various pieces of information. As long as the information remains on the board, the teacher can refer to it using pronouns. However, once erased, the teacher must resort to more detailed expressions or rewrite the information to reference it again. Similarly, Hofmann argues, transitioning to a new paragraph necessitates the use of more elaborate REs (Hofmann 1989). Despite taking a firm stance on the barrier to cross-paragraph anaphora, Hofmann acknowledges the occasional use of pronouns at the start of a paragraph. These instances act as a bridge between the preceding and current paragraphs, aiming to "unite them into larger functional units" (Hofmann 1989: 245).

Ariel's Accessibility Theory (1990, 2004) posits that the more accessible a referent is, the more likely it will be referred to by a reduced form, such as a pronoun. Accessibility is determined by two factors: (1) the intrinsic salience of the referent (e.g., being topical), and (2) its relational accessibility to its antecedent. Under this framework, Ariel introduced the *unity* criterion, which assesses whether the antecedent and its reference share the same frame, world, point of view, segment, or paragraph. References spanning different paragraphs are deemed not coherently close.

In her analysis of the distribution of REs across various text positions in a corpus, Ariel (1990) observed that over 80% of definite descriptions were used either within the same paragraph but beyond the immediate preceding sentence,

or across paragraph boundaries. Additionally, when examining REs from the perspective of textual positions, Ariel found that across paragraph boundaries, 58.9% of REs were definite descriptions, and notably, 26.7% were pronouns. However, a limitation of this study is its scope, encompassing only 775 REs, with 70% being pronouns.

Vonk et al. (1992), expanding on Ariel's work, investigated the distribution of pronouns and definite descriptions in text fragments. Their findings suggest that overspecified REs play a role in structuring discourse, whereas pronouns indicate continuity. In this context, a full NP serves as a marker for establishing a new informational chain within the discourse representation (Vonk et al. 1992, Smith 2003).

Tomlin (1987b) critiqued linear models of recency in discourse, highlighting their inadequacy in explaining two phenomena: (1) the use of a definite description when its antecedent is merely a single clause away without any ambiguity, and (2) the maintenance of pronominal expressions over extended distances. To address these inconsistencies, Tomlin proposed an episode/paragraph model. He theorized, "the alternation between noun and pronoun to be a function of the limited capacity of working memory, which is manifested in the text artifact primarily through its paragraph, or episodic organization" (1987b: 456). Each episode in this model contains a thematic macroproposition that remains the focus of attention until a shift occurs. In narrative discourse, paragraph boundaries signify such shifts in attention. Tomlin elucidates this relationship between episode or paragraph boundaries and attention allocation:

> The alternative use of a noun or pronoun in discourse production is a function of attention allocation by the speaker. During the online process of discourse production, the speaker uses a pronoun to maintain reference as long as attention is sustained on that referent. Whenever attention focus is disrupted, the speaker reinstates reference with a full noun, no matter how few clauses intervene between subsequent references (p. 458).

This perspective links the hierarchical structural organization of discourse with the cognitive mechanism of attention, offering a more nuanced understanding of attention dynamics in discourse. According to Tomlin (1987b), this episodic theory explains 84% of referential expression choices in his study. The other 16% are categorized as (1) *intra-episode nominals* (non-pronominal REs within an episode) and (2) *inter-event pronominals* (rare pronouns used at the start of a paragraph). Intra-episode nominals are often employed to resolve ambiguity, whereas no specific linguistic explanation is provided for inter-event pronominals due

to their scarcity in the study. Tomlin supports an attention-driven episodic approach over a linear recency model, arguing that the latter fails to account for these exceptions.

Similar to Tomlin (1987b), Fox (1987a) also challenged linear reference theories, questioning their implications on text structure and attention flow. She argued that if only the distance from the antecedent mattered, all clauses would contribute equally to this measurement, implying a linear model where discourse is perceived as "an undifferentiated string of clauses which follow one another in time but do not form larger units that could perform communicative functions in relation to one another" (p. 158). In such a model, attention is treated as a uniform concept, disregarding the need to signal new developments about the same referent or interruptions of previous information.

Contrarily, Fox (1987a) posited that the use of a full NP where a pronoun would suffice marks the hierarchical structure of the narrative. A full NP signals the start of a new development unit within the text. This does not imply all new units begin with full NPs, but rather, full NPs are used where pronouns are typically admissible. While Fox (1987a) does not explicitly define these development units, Huang (2000) suggests they may manifest as turns, paragraphs, episodes, events, or themes in a hierarchical anaphora framework.

6.2.3 Interim summary

In §6.2, I discussed the detection and applications of paragraph boundaries from cognitive and computational perspectives. The focus then shifted to the relationship between paragraph boundaries and the choice of RF. The theories examined here offer insights into why full NPs often appear at the start of new paragraphs, attributing this pattern to factors like working memory constraints, the initiation of new informational chains, and the dynamics of attention allocation.

Furthermore, this section delineates a crucial distinction between linear and hierarchical models of reference. While linear models focus on the immediate proximity of antecedents, hierarchical models integrate larger textual units, such as paragraphs and episodes, into the analysis. This distinction underscores the limitations of linear models, particularly in explaining the complexities of cross-boundary transitions. The upcoming corpus analysis aims to delve deeper into this issue, exploring how paragraph structure intricately influences the choice of referential forms in a more empirical and data-driven manner.

6.3 Study D: A corpus analysis of the impact of paragraph structure on RF

Study D undertakes a thorough analysis of paragraph-related factors within the WSJ corpus, aiming to decipher how paragraph structure influences RF choices. This section is structured as follows: §6.3.1 presents basic statistics of the WSJ corpus's paragraphs. Subsequently, paragraphs are examined from two perspectives in §6.3.2 and §6.3.3: (1) *intra-paragraph*, focusing on the internal structure of paragraphs, and (2) *inter-paragraph*, concerning transitions between paragraphs. Lastly, §6.3.4 offers a concise summary and discussion.

6.3.1 Basic overview of paragraphs in WSJ

The WSJ corpus, known for its relatively lengthy newspaper articles (average 25 sentences per article), initially lacks explicit paragraph structure information. This information was later integrated from an external source and assigned to the articles, as detailed in Chapter 5.[2]

Our analysis uses 5561 paragraphs from the WSJ corpus.[3] The corpus exhibits an average of 11.01 paragraphs per document (ranging from 1 to 53 paragraphs). Table 6.1 provides insights into the number of referents, REs, sentences, and words per paragraph.

Table 6.1: WSJ paragraph statistics. Note: the minimum word count of 1 in the dataset can occur in instances of direct short answers such as *Yes* or *No*, which are examples of one-word paragraphs. This happens when a document quotes a dialogue.

Per paragraph	Mean	Std. Dev	Min	Median	Max
Number of referents	3.24	1.89	1	3	13
Number of REs	7.21	5.08	1	6	40
Number of sentences	2.13	1.18	1	2	11
Number of words	48.43	26.24	1[4]	44	270

[2] https://github.com/WING-NUS/pdtb-parser/tree/master/external/aux_data/paragraphs

[3] Two articles (WSJ-0591 and WSJ-1482) are excluded from the analysis due to the absence of paragraph segmentation.

6.3.2 Intra-paragraph factor: Paragraph-prominent referents

This section examines the hypothesis that referents prominent within a paragraph are more frequently referred to using pronominal REs. Consider (3), which repeats the opening paragraph of (1). This example illustrates a pattern where, following the first explicit mention, subsequent references to the prominent character, Walter White, are predominantly pronominal. In this example, except for the initial explicit mention, Walter White is subsequently referred to via pronominal REs such as *he* and *his*. This pattern suggests a tendency towards pronominalization for a paragraph's prominent character.

(3) **Walter Hartwell "Walt" White Sr.**, also known by **his** clandestine pseudonym and business moniker Heisenberg, was an American drug kingpin. A former chemist and high school chemistry teacher in Albuquerque, New Mexico, **he** started manufacturing crystal methamphetamine after being diagnosed with terminal lung cancer. **He** initially does this in order to pay for **his** treatments and secure the financial future of **his** family: wife Skyler, son Walter Jr., and infant daughter Holly, but confesses before **his** death that **he** actually did it for **himself**, due to being good at it and feeling alive.

Prominent referents in a paragraph To investigate the impact of paragraph-prominence on RF, this study adopts the frequency of mention as a measure of referents' prominence within a paragraph. This concept, inspired by Siddharthan et al.'s assertion that frequency features can effectively indicate a referent's global salience within a document (2011), is applied at the paragraph level to assess the prominence of entities within specific discourse segments.

In this approach, referents that receive the most mentions within a paragraph are marked as prominent. When multiple referents share the highest frequency of mentions, each is considered equally prominent. For instance, in (4), "Anne Volokh" is identified as the prominent referent in the paragraph. Therefore, all references to her within this paragraph, which are shown in bold in (4), are tagged as prominent.

(4) WSJ-1367

a. [Paragraph 1] When **Anne Volokh** and **her** family immigrated to the U.S. 14 years ago, they started life in Los Angeles with only $ 400. They 'd actually left the Soviet Union with $ 480, but during a stop in Italy **Ms. Volokh** dropped $ 80 on a black velvet suit. Not surprisingly, **she** quickly adapted to the American way. Three months after **she** arrived

> in L.A. **she** spent $ 120 **she** did n't have for a hat. "A turban," **she** specifies, "though it was n't the time for that 14 years ago. But I loved turbans."

In the analyzed dataset, a total of 2,472 paragraphs feature only one prominent referent, while 3,089 paragraphs have multiple prominent referents. Table 6.2 details the distribution of RFs, distinguishing between mentions of prominent and non-prominent referents. According to the table, there are 18,121 REs classified as `prominent` and 12,350 as `non-prominent`, as indicated in the `Total` row of Table 6.2.

Table 6.2: Cross-tabulated distribution of RFs (`description`, `name`, `pronoun`) by prominence of referents within a paragraph (`non-prominent`, `prominent`). The percentages reflect the proportion of each RF type relative to the prominence status in the paragraph. For instance, 19.3% of the REs with the RF `description` are `non-prominent`.

RF	Prominence in paragraph	
	`non-prominent`	`prominent`
`description`	19.3%	20.1%
`name`	16.6%	20.1%
`pronoun`	4.6%	19.3%
Total	12,350	18,121

Analysis of Table 6.2 reveals that the distribution of non-pronominal forms (i.e., proper names and descriptions) is relatively similar for both prominent and non-prominent referents. For instance, descriptions account for 20.1% of references to prominent entities and 19.3% to non-prominent entities. However, the pattern diverges for pronominal REs: 19.3% of pronominal references are made to prominent referents, compared to only 4.6% for non-prominent referents. This discrepancy indicates a significantly higher likelihood of using pronominal REs for prominent referents than for non-prominent ones.

6.3.3 Inter-paragraph factors: Cross-boundary transitions

Prior research, as discussed in §6.2.2, indicates a propensity for referents to be expressed in non-pronominal forms when their antecedents are located in a preceding paragraph. This pattern may arise from stylistic choices, where authors prefer non-pronominal forms for the initial REs in a new paragraph due to their

prominent position. Alternatively, it could be attributed to a decrease in the referents' prominence status due to paragraph transitions, necessitating their reintroduction via non-pronominal REs. The following sections, 6.3.3.1 and 6.3.3.2, explore these hypotheses, while §6.3.3.3 examines a distinct aspect: cross-boundary pronominalization.

6.3.3.1 Paragraph-initial position

This analysis focuses on the first RE in each paragraph, termed the *paragraph-initial* slot. Since the first mention of an entity in a text is usually non-pronominal, first-mention REs – also known as *discourse-new* REs – are excluded from this analysis. This approach helps to avoid conflating the effects of discourse-new REs with those of paragraph-initial references. Therefore, the criteria for selecting paragraph-initial referents are:

1. Select only the first RE in each paragraph.
2. Exclude discourse-new REs.

Following these criteria, a total of 3,257 discourse-old, paragraph-initial REs were identified. As Table 6.3 shows, over 90% of these REs are non-pronominal, aligning with findings from previous studies (Tomlin 1987b, Pu 2019). This rate of non-pronominalization surpasses that reported by Ariel (1990) and mirrors Pu (2019)'s findings in a Chinese text corpus, where a similar percentage of cross-boundary REs were non-pronominal. A critical question remains: Is the tendency for non-pronominalization specific to paragraph-initial REs, or is it a broader phenomenon related to reintroducing referents across paragraph boundaries? The subsequent section, §6.3.3.2, addresses this question by examining the reintroduction of REs within a paragraph.

Table 6.3: Distribution of non-new REs in the paragraph-initial position.

RF	Frequency	Percent (%)
`description`	1281	39.33
`name`	1658	50.91
`pronoun`	318	9.76
Total	3257	100.00

6.3.3.2 Reintroduction of entities at the paragraph level

The focus here shifts from paragraph-initial REs (as discussed in §6.3.3.1) to the first reintroduction of each referent within a new paragraph, which I will refer to as *paragraph-new REs*. Unlike the previous analysis, this does not solely concentrate on the first reference slot of each paragraph; instead, it focuses on the initial occurrence of a referent being *reactivated* within the paragraph. This approach is similar to the blackboard analogy in §6.2.2, where a teacher's use of pronouns is acceptable as long as the referential context remains visible on the board. Once erased, fuller forms are necessary for reactivation. A similar dynamic is expected here: Pronouns remain viable within the same paragraph, but the transition to a new paragraph necessitates fuller forms for referent reactivation, presumably due to a reduction in their prominence. The subset for this analysis adheres to the following criteria:

1. Focus on the first mention of *each referent* within each paragraph.
2. Exclude discourse-new REs in order to avoid conflating first mentions with paragraph-new reintroductions.

Applying these criteria yields 9,784 discourse-old, paragraph-new REs. Table 6.4 presents their distribution, revealing a striking pattern: Approximately 94% of paragraph-new REs are realized in non-pronominal forms. This significant tendency strongly suggests that non-pronominal forms are preferred for reintroducing referents in new paragraphs.

Table 6.4: Distribution of discourse-old paragraph-new REs.

RF	Frequency	Percent (%)
`description`	4156	42.48
`name`	5033	51.44
`pronoun`	595	6.08
Total	9784	100.00

6.3.3.3 Cross-boundary pronominalization

While a predominant portion of cross-boundary REs are non-pronominal (over 90%), as noted in previous analyses, there remains a notable subset of pronominal instances. This section delves into three potential factors that might influence

this occurrence of cross-boundary pronominalization: (1) the length of the preceding paragraph, (2) the grammatical role of the paragraph-initial RE, and (3) the prominence of the referents involved.

Length of the preceding paragraph The hypothesis here is that pronominal REs at the start of a paragraph might be linked to stylistic choices, particularly following short, one-sentence paragraphs. To investigate this, we examine the 318 paragraph-initial pronominal REs identified in Table 6.3, focusing on the length of the preceding paragraph (P_{i-1}) in terms of sentences. Table 6.5 shows the sentence-wise length of paragraph P_{i-1} that precedes paragraph P_i containing a paragraph-initial pronominal RE.

Table 6.5: The sentence-wise length of paragraph P_{i-1} that precedes the paragraph-initial pronominal REs in P_i.

P_{i-1} length	1	2	3	4	5	6	7	8	Total
Frequency	81	100	56	46	22	6	5	2	318
Percent (%)	25	31.45	17.61	14.47	6.92	1.89	1.57	0.63	100.00

Table 6.5 displays the sentence length of P_{i-1}. The data reveals a range from one to eight sentences, with 25% of cases having only one sentence in the preceding paragraph. However, since the majority have two or more sentences, we can conclude that the length of P_{i-1} is not a sole determinant for the use of pronominal REs at the beginning of a paragraph.

The subsequent parts of this section will further explore paragraph-initial pronominal REs from two perspectives: examining the grammatical role of these REs and investigating their relation to the prominence status of the referents.

The impact of grammatical role The example presented at the beginning of this chapter highlights a unique instance of a paragraph-initial pronominal RE, specifically a cataphoric possessive determiner (1c). This case is intriguing as the use of the possessive determiner *his* is essential for sentence coherence, and alternative RFs would render the sentence structurally inappropriate.

A broader examination of the WSJ dataset reveals that possessive modifiers constitute 14% of the REs, distributed across different RFs as follows: 17.6% `descriptions`, 25.7% `proper names`, and a significant 56.7% `pronouns`. This distribution pattern suggests that possessive modifiers in the dataset predominantly take

pronominal forms. Consequently, it is reasonable to hypothesize that a significant proportion of paragraph-initial pronominal REs are possessive modifiers.

The following example from the WSJ corpus illustrates the transition between paragraphs 17 and 18, where a pronominal possessive determiner is used as the first RE in paragraph 18. In this context, other methods of referring to Judge Ramirez, as shown in (5b), are deemed unacceptable. The possessive pronominal form appears to be the only viable option.

(5) WSJ-0049

a. [Paragraph 17] Judge Ramirez, 44, said it is unjust for judges to make what they do. "Judges are not getting what they deserve. You look around at professional ballplayers or accountants... and nobody blinks an eye. When you become a federal judge, all of a sudden you are relegated to a paltry sum."

b. [Paragraph 18] At **his** new job, as partner in charge of federal litigation in the Sacramento office of Orrick, Herrington & Sutcliffe, he will make out much better.

However, the notion of obligatoriness does not always dictate the choice of RFs, and in many instances, alternative RFs are equally appropriate. For example, consider the RE "Mr. Greenspan's" in (6b), which appears as the first RE of paragraph 4. While a pronominal form such as *his* could have been used and would have been contextually acceptable, the text opts for a proper name instead.

(6) WSJ-0598

a. [Paragraph 3] Such caution was evident after the recent Friday - the - 13th stock market plunge. Some Bush administration officials urged Mr. Greenspan to make an immediate public announcement of his plans to provide ample credit to the markets. But he refused, claiming that he wanted to see what happened Monday morning before making any public statement.

b. [Paragraph 4] **Mr. Greenspan's** decision to keep quiet also prompted a near-mutiny within the Fed's ranks. [...]

Given the substantial proportion of pronominal possessive modifiers in the dataset, which constitute 56.7% of all possessive modifiers, and the instances where such forms are obligatory, I propose that paragraph-initial pronominal REs are substantially more likely to be possessive modifiers than other grammatical roles. This hypothesis is tested using the paragraph-initial pronominal

REs from Table 6.3, although similar patterns are observable in the dataset represented in Table 6.4.

Table 6.6 details the distribution of these REs across different grammatical roles. In the first row, labeled *count*, the raw frequency of each grammatical role is presented. Notably, the data indicates that out of the paragraph-initial pronominal REs, 231 are subjects, while 69 are possessive modifiers.

Table 6.6: Distribution of paragraph-initial pronominal REs by grammatical roles (obj, poss, subj). The first line (`count`) shows the raw count of each grammatical role category. The second line (row%) shows the row-wise distribution of paragraph-initial pronominal REs across different grammatical roles (e.g., 72.6% of paragraph-initial pronouns are in the subject position). The third line (col%) shows the conditional relative frequency of paragraph-initial pronominal REs given the grammatical roles. For instance, according to the information in this row, only 18.6% of the possessive modifiers appearing paragraph-initially are pronominal.

Paragraph-initial pronominal REs	Grammatical role			Total
	obj	poss	subj	
count	18	69	231	318
row %	5.7%	21.7%	72.6%	
col %	2.1%	18.6%	11.5%	

The row-wise distributions in Table 6.6, (`row%`), offer valuable insights into the roles of paragraph-initial pronominal REs. These percentages reveal that a smaller proportion of paragraph-initial pronominal REs are objects (5.7%) and possessive modifiers (21.7%), while a significant majority, 72.6%, function as subjects. This finding is intriguing as it challenges the initial hypothesis that possessive determiners would predominantly characterize paragraph-initial pronominal REs. (7b) shows a pronominal subject as the first RE at the beginning of paragraph 4.

(7) WSJ-0121

a. [Paragraph 3] "I think program trading is basically unfair to the individual investor," says Leo Fields, a Dallas investor. He notes that program traders have a commission cost advantage because of the quantity of their trades, that they have a smaller margin requirement than individual investors do and that they often can figure out earlier where the market is heading.

b. [Paragraph 4] But **he** blames program trading for only some of the market's volatility.

The third row (`col%`) of Table 6.6 shows the conditional relative frequency of paragraph-initial pronominal REs given the grammatical roles. According to the information in this row, only 18.6% of the possessive modifiers appearing paragraph-initially are pronominal. Therefore, even in the case of possessive modifiers, non-pronominal REs are more common than pronominal ones in the initial-paragraph position. Regarding the REs in the paragraph-initial subject position, we see that 231 instances (11.5% of all paragraph-initial subject REs) appear pronominally.

The third row (`col%`) in Table 6.6 details the conditional relative frequency of paragraph-initial pronominal REs based on grammatical roles. From this data, it is evident that only 18.6% of possessive modifiers at the start of a paragraph are realized pronominally. This shows a predominant preference for non-pronominal forms over pronominal ones as paragraph-initial possessive modifiers. Additionally, the analysis of subject REs in paragraph-initial positions reveals that 231 instances, accounting for 11.5% of all paragraph-initial subject REs, are expressed using pronominal forms.

In summary, Table 6.6 provides information about the grammatical role and RF of discourse-old paragraph-initial pronominal REs. The table shows that these pronouns are not confined to possessives but frequently occur in subject positions. Furthermore, despite the general tendency for possessive modifiers to be pronominalized in the WSJ corpus, they are more likely to be realized as non-pronominal REs in paragraph-initial positions. In the next section, I will delve deeper into paragraph-initial pronominal instances, examining their occurrence in relation to the prominence of referents.

The prominence status of referents §6.3.2 revealed that within paragraphs, prominent referents are more frequently pronominalized compared to their non-prominent counterparts. Additionally, we noted that cross-boundary pronominals extend beyond obligatory possessive determiners. These findings suggest a potential link between cross-boundary pronominalization and the prominence status of referents. Building on this, I propose the hypothesis that a referent newly introduced in a paragraph (paragraph-new referent) is significantly more likely to be pronominalized if it is considered prominent in both the current paragraph (P_i) and the preceding one (P_{i-1}). To investigate this hypothesis, I focus exclusively on REs that reference paragraph-prominent entities. Consequently, the criteria for selecting this subset are:

1. The dataset is narrowed to include only those REs classified as prominent in Table 6.2.
2. Focus is placed on the initial mention of each referent within a paragraph.
3. Only discourse-old REs are considered.[5]

Out of the 4291 instances meeting these conditions, approximately 38% are categorized under the `prominent` condition (the referent is prominent in both the current and preceding paragraph), and around 62% under the `non-prominent` condition (the referent lacks prominence in both paragraphs). Interestingly, only 331 of these REs are pronominal, representing less than 8% of this dataset. However, a noteworthy aspect is that 58% of these pronominal REs are associated with the prominent condition, while 42% are linked to the non-prominent condition. Another point of interest is that merely 5.3% of the REs in the non-prominent condition are pronominal, as opposed to 11.7% in the prominent condition. These findings suggest that (1) the likelihood of cross-boundary pronominalization remains low even when the referent is prominent in both paragraphs, and (2) although the limited data points inhibit definitive conclusions, it appears that cross-boundary pronominalization more frequently involves prominent referents. For example, in (8), paragraph 6 introduces the referent *Betty Raptopoulos* and frequently mentions her within the paragraph. Paragraph 7 then re-introduces her with the pronoun *She*. The conceptual proximity of these paragraphs and the use of a pronominal RE at the start of paragraph 7 might be interpreted as an attempt to weave them into larger, cohesive units, aligning with the perspectives of Hofmann (1989).

(8) WSJ-1203

a. [Paragraph 6] **Betty Raptopoulos, senior metals analyst at Prudential-Bache Securities in New York**, agreed that most of the selling was of a technical nature. **She** said the market hit the $ 1.18 level at around 10 a.m. EDT where it encountered a large number of stop-loss orders. More stop-loss orders were touched off all the way down to below $ 1.14, where modest buying was attracted. **Ms. Raptopoulos** said the settling of strikes in Canada and Mexico will have little effect on supplies of copper until early next year. **She** thinks the next area of support for copper is in the $ 1.09 to $ 1.10 range. "I believe that as soon as the

[5]The premise of the hypothesis says that the referent should be prominent in both the current and preceding paragraphs, thus these REs must have been introduced earlier in the text.

selling abates somewhat we could see a rally back to the $ 1.20 region," **she** added.

b. [Paragraph 7] **She** thinks a recovery in the stock market would help copper rebound as well. **She** noted that the preliminary estimate of the third-quarter gross national product is due out tomorrow and is expected to be up about 2.5% to 3%.

The observations regarding cross-boundary pronominals offer a platform for more detailed investigations. However, before conducting a comprehensive analysis, a thorough preliminary study is needed to identify potentially challenging cases. For instance, in (9), the initial RE of paragraph 9 occurs within a quotation. This indicates that the RE was expressed by an individual other than the document's author. Such instances represent a distinct usage of obligatory REs at the beginning of a paragraph, diverging from the assumptions discussed so far. Consequently, these cases necessitate separate consideration, as they operate under different dynamics from the ones outlined in the previous analyses.

(9) WSJ-1474

a. [Paragraph 9] "Unless **it** gets more help, the U.S. industry won't have a chance," says Peter Friedman, Photonics's executive vice president.

6.3.4 Summary and discussion of study D

The corpus study in this section investigated the influence of paragraph-related features on the selection of REs, focusing on both intra-paragraph and inter-paragraph effects. The key findings from this study are summarized below:

Intra-paragraph effects The study revealed that entities that are prominent within a paragraph are significantly more likely to be referred to using pronominal forms. This observation aligns with the concept that the prominence of a referent within a discourse segment increases the likelihood of its pronominalization.

Inter-paragraph effects Consistent with prior research, the study found that most REs crossing paragraph boundaries are non-pronominal. This supports the idea that paragraph transitions often require the use of fuller, non-pronominal forms to effectively reintroduce or reactivate referents.

Initial position vs. reintroduction of referents The study distinguished between referents appearing in the initial position of a paragraph and those being reintroduced in a paragraph. Over 90% of paragraph-initial REs are non-pronominal, indicating a preference for fuller forms in this position. However, this study alone could not fully clarify whether the initial position inherently favors non-pronominal REs or if the paragraph transition demotes the prominence status of referents, necessitating their reactivation with fuller forms. Regardless, the analysis in §6.3.3.2 showed that when referents are reintroduced in a paragraph, whether in paragraph-initial position or elsewhere in the paragraph, they are usually realized as non-pronominal REs.

Grammatical role and pronominalization The investigation into pronominal paragraph-initial cases revealed that these are not limited to possessive modifiers. Subject REs also appear in pronominal form, though pronominalization of object REs is rare.

Prominence and transition Preliminary findings suggest that prominent referents are more likely to be pronominalized across paragraph boundaries. However, this conclusion requires further data and in-depth analysis for validation.

In conclusion, this study contributes to a better understanding of how paragraph structure and transitions influence the choice of REs in text. In the next section, some of the findings of this corpus analysis will be put into practice in a REG-in-context task.

6.4 Study E: REG-in-context models incorporating paragraph-related features

Study E is a computational analysis aimed at understanding the impact of paragraph structure on the selection of RFs in text. The primary hypothesis of this study is that incorporating paragraph-related information will significantly enhance the performance of feature-based REG-in-context models. Building on insights gained from previous studies, especially the positive contribution of the distance in the number of paragraphs highlighted in Chapter 5, study E seeks to thoroughly investigate which paragraph-related features are most influential and how they improve model performance. Below are the key aspects of study E:

Focus on pronominalization task The literature review and corpus analysis in §6.2 and §6.3 primarily examined the pronominalization task, i.e., the choice between pronominal and non-pronominal REs. Study E continues this focus, specifically exploring how paragraph structure influences the choice between pronominal and non-pronominal REs.

Exclusion of first Mentions The study recognizes that the initial mentions of entities in a text are typically non-pronominal. Therefore, to ensure a more precise analysis, first mentions are excluded from the study's scope. This exclusion allows for a clearer examination of the impact of paragraph transitions on subsequent mentions of entities.

In summary, the studies in this section share two key characteristics: (1) they address a 2-way RFS task, specifically distinguishing between `pronominal` and `non-pronominal` forms, and (2) they consider only discourse-old REs. The structure of this section is as follows: In §6.4.1, I explore various paragraph-related features to determine their appropriateness for the studies outlined in §6.4.2. Subsequently, in §6.4.2, I evaluate diverse feature-based models to ascertain the efficacy of paragraph-related features in addressing the REG-in-context RFS task. Finally, in §6.4.3, I conduct an error analysis of the models' results to identify where most mispredictions occur and to understand the conditions under which incorporating paragraph-related features enhances performance.

6.4.1 Introducing paragraph-related features for REG

Drawing on insights from study C in Chapter 5, I experimented with various numeric and categorical implementations of paragraph recency in the validation set. The selected paragraph-based recency metric is as follows:

- `dist_par`: This numeric metric measures the distance in the number of paragraphs between a target RE and its antecedent.

Previous research, such as the works of Tomlin (1987b) and Fox (1987a), has questioned linear models of recency, highlighting the significance of the hierarchical and organizational structure in narrative texts over mere linear order. Aligning with this perspective, the corpus studies in §6.3 prioritized examining cross-boundary transitions. In this context, I hypothesize that the contribution of the paragraph distance metric to the REG-in-context task lies not just in quantifying the distance between mentions but in implicitly indicating whether the RE and its antecedent are within the same paragraph. To put it simply, a distance

of zero suggests that both the RE and its antecedent are located in the same paragraph, while any distance greater than zero indicates a paragraph transition. To explore this hypothesis, I introduce an additional feature, `par_givenness`, which categorizes the relationship between the RE and its antecedent into two states: (1) `new` – the target RE and its antecedent are at least one paragraph away, that is, marking the first mention of the target RE following a paragraph transition, and (2) `given` – indicating that the RE and its antecedent appear within the same paragraph.

- `par_givenness`: whether the RE is paragraph-new or paragraph-old.

The recency feature, `dist_par`, not only implicitly encodes the paragraph-givenness of referents but also provides insight into the linear distance between a target RE and its antecedent. In contrast, the `par_givenness` feature focuses solely on identifying whether the target RE and its antecedent are situated within the same paragraph or are separated by a paragraph transition. This distinction raises a relevant question: Which aspect – linear distance or paragraph givenness – plays a more substantial role in influencing the performance of REG-in-context models?[6]

To address this query, two distinct models are constructed: one incorporating the `dist_par` feature and the other employing the `par_givenness` feature. Should these models exhibit similar performance levels, it would suggest that paragraph givenness is the primary factor influencing the choice of RF. Conversely, if the model featuring `dist_par` demonstrates superior performance, it would imply that either both paragraph givenness and linear distance or solely the latter are influential in determining RF selection.

I employed *XGBoost*, a technique from the Gradient Boosting Decision Trees family, to train the classifiers for this study (Chen & Guestrin 2016). The evaluation of the classifiers' performance, as indicated by the confusion matrices shown in Figure 6.1, reveals that both models exhibit identical performance metrics. This finding underscores that paragraph givenness is the key factor in determining the choice of RF. Consequently, based on these insights, the feature `par_givenness` is selected as the primary focus for the models discussed in §6.4.2.

In the preceding discussions, we have explored paragraph-related factors that encode recency and givenness. In what follows, I introduce two additional features: `par_prom` and `par_subj_1`. These are defined as:

[6]The term *paragraph givenness* is used here to indicate the referential status of referents at the paragraph level, which also implicitly signals the presence of paragraph transitions.

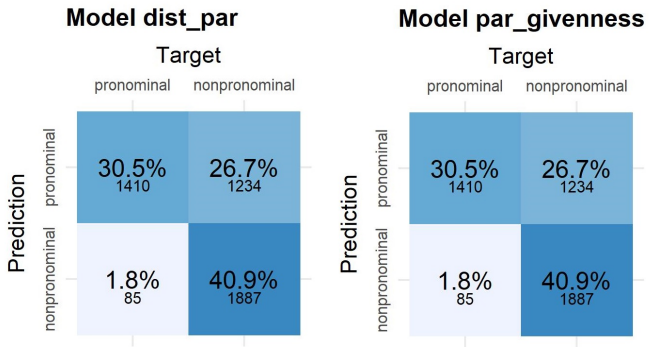

Figure 6.1: Confusion matrices of the `dist_par` and `par_givenness` models.

- `par_prom`: This feature assesses whether the target RE refer to a referent that is prominent within its paragraph. This is informed by the observation in §6.3.2 that prominent referents within a paragraph are more likely to be referenced pronominally.
- `par_subj_1`: This feature identifies whether the target RE is the first subject RE in the paragraph.

The aim of this section is to determine an effective set of paragraph-related features for the comprehensive study outlined in §6.4.2. Building on the previously established significance of paragraph givenness, the forthcoming model will integrate `par_givenness` along with `par_prom` and `par_subj_1`. The study has the following specifications:

1. Task: binary (pronominal vs. non-pronominal) classification
2. Features: `par_givenness`, `par_prom`, `par_subj_1`
3. Model: XGBoost with the parameters outlined in Table 6.7.

To assess the impact of the features `par_givenness`, `par_prom`, and `par_subj_1` on the REG-in-context task, a SHapley Additive exPlanations (SHAP) analysis is employed. This analysis, originating from coalitional game theory, effectively deconstructs the model's predictions into individual contributions attributable to various variables. As mentioned by Molnar (2019), "a prediction can be explained by assuming that each feature value of the instance is a 'player' in a game where the prediction is the payout" (p. 177). In this context, SHAP values act as a fair

Table 6.7: The parameters used in the XGBoost model.

parameters	value
nrounds	500
max_depth	5
eta	0.05
gamma	0.01
colsample_bytree	0.75
min_child_weight	0
subsample	0.5
objective	multi:softprob

means of allocating the "payout" – in this case, the prediction – amongst different feature values of the instance.

Figure 6.2 demonstrates the outcomes of the SHAP analysis for this model, highlighting how each feature–value influences the model's predictions. This approach allows for an understanding of the individual and collective impact of these paragraph-level features on the model's performance. The analysis is divided into two parts: one focusing on non-pronominal REs (top graph) and the other on pronominal REs (bottom graph). In this framework, green bars indicate a feature–value's positive contribution towards a particular prediction, while red bars signify a negative contribution.

For the prediction of non-pronominal REs, the top graph reveals that two primary factors increase the likelihood of choosing non-pronominal forms and discourage the use of pronouns. These are: (1) the referent is newly introduced in the paragraph (`par_givenness = par_new`), and (2) the RE is the first subject to appear in the paragraph (`par_subj_1 = yes`). Conversely, if the referent is marked as prominent within the paragraph (`par_prom = prominent`), there is a decreased likelihood of opting for non-pronominal forms.

Notably, the `par_givenness` feature has a more substantial impact on the task than the other two features. Nonetheless, given that all these features contribute in varying degrees to the model's performance, they are all incorporated into the main study outlined in §6.4.2.

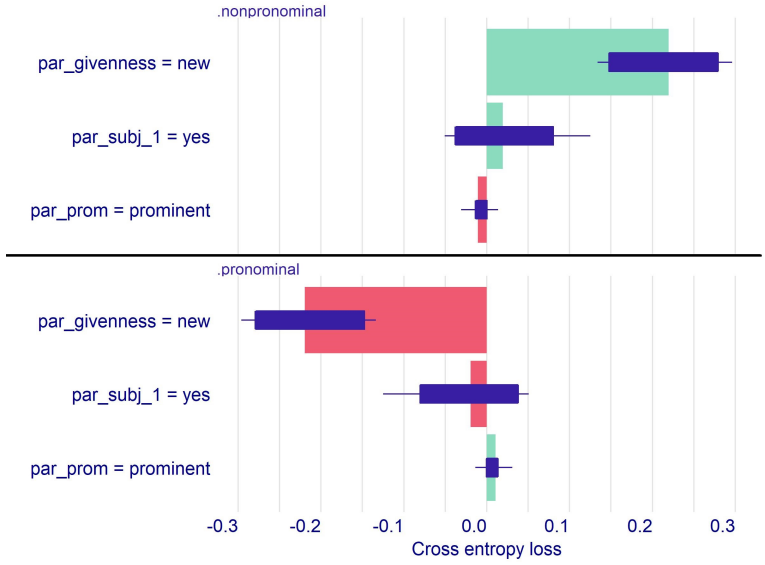

Figure 6.2: Shapley values with box plots for ten random orderings of explanatory variables in the paragraph-related model. The top graph shows the contribution of the factors to the prediction of non-pronominal REs, and the bottom graph shows the contributions to pronominal REs. The green and red bars represent positive and negative contributions, respectively.

6.4.2 A comparison of REG models with and without paragraph features

In the study presented in this section, I investigate the impact of paragraph-related features on the performance of REG-in-context models. The goal is to assess whether incorporating these paragraph-level features significantly enhances the model's ability to choose between pronominal and non-pronominal referential forms. For this purpose, I compare the models with added paragraph features against three baseline models:

- RANDOM: This baseline model assigns a pronominal or non-pronominal value to each instance in the test dataset randomly. This approach serves as a basic comparison point, representing a scenario where no specific features or logic are used in the decision-making process.
- MINIMUM: The MINIMUM baseline incorporates only the local features of a referent, excluding any features related to the antecedent. The features included in this model are: grammatical role (`gm`), animacy, and plurality.

- INFORMED: The INFORMED baseline builds upon the MINIMUM baseline by adding a categorical measure of sentential distance (`dist_s`). This feature classifies the distance between the referent and its antecedent into three categories: `same sentence` (the referent and its antecedent are in the same sentence), `one sentence away` (the referent and its antecedent are separated by one sentence), and `plus-one sentence away` (the referent and its antecedent are separated by more than one sentence).

The EXPERIMENTAL model is constructed by combining features from both the INFORMED model and the paragraph-related features previously introduced in §6.4.1. The features included in the EXPERIMENTAL model are as follows:

- EXPERIMENTAL: grammatical role (gm), animacy, plurality, sentence distance (`dist_s`), paragraph givenness (`par_givenness`), prominence in paragraph (`par_prom`), paragraph subjecthood (`par_subj_1`).

The models are trained using the XGBoost algorithm, with the training process involving 5-fold cross-validation. The specific parameters used for training are detailed in Table 6.7.

The performance of the models is assessed using a variety of metrics. These include the overall accuracy of the models, as well as their macro-averaged precision, recall, and F1 scores. Accuracy provides a measure of overall correctness, while the macro-averaged precision, recall, and F1 scores address potential class imbalances. These macro-averaged scores are calculated by taking the arithmetic mean, also known as the unweighted mean, of the scores for each class. Precision assesses the model's accuracy in identifying relevant instances, recall evaluates its ability to capture all relevant cases, and the F1 score, being the harmonic mean of precision and recall, offers a balanced measure of the model's sensitivity and specificity. Table 6.8 presents the overall performance statistics of the models.

Table 6.8: Overall statistics of the models.

Model name	accuracy	macro-precision	macro-recall	macro-F1
RANDOM	0.495	0.495	0.494	0.478
MINIMUM	0.745	0.753	0.632	0.638
INFORMED	0.853	0.831	0.84	0.835
EXPERIMENTAL	0.869	0.85	0.85	0.85

All three models outperform the RANDOM baseline, as shown in Table 6.8. The MINIMUM and INFORMED models differ in only one feature, yet their performance varies significantly (MINIMUM macroF1 = 0.638 vs. INFORMED macroF1

= 0.835). The EXPERIMENTAL model shows improved performance over the INFORMED model, although the margin of this improvement is not very large. Figure 6.3 presents the confusion matrices for the INFORMED and EXPERIMENTAL models, illustrating their respective performances in terms of correct and incorrect predictions.

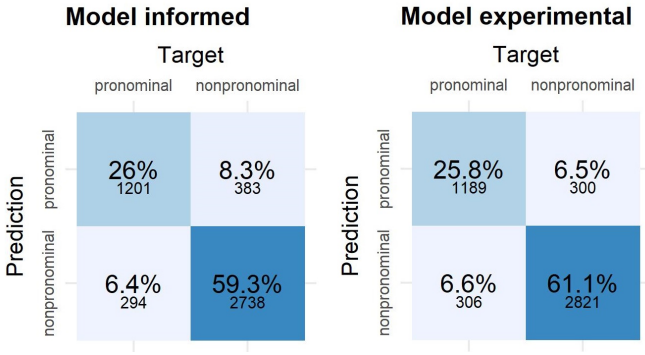

Figure 6.3: Confusion matrices of the INFORMED and EXPERIMENTAL models.

Pronominal cases Both models have very similar performance in correct and incorrect prediction of pronominal cases. Regarding the correct predictions, both models show comparable accuracy, with the INFORMED model at 26% and the EXPERIMENTAL model slightly lower at 25.8%. Regarding the incorrect predictions, again, the performance is similar, with the INFORMED model at 6.4% and the EXPERIMENTAL model at 6.6%.

Non-pronominal cases When comparing the non-pronominal cases, the difference between the two models is more pronounced. Regarding the correct predictions, the EXPERIMENTAL model shows a marked improvement, correctly identifying 61.1% of non-pronominal cases compared to 59.3% for the INFORMED model. Regarding the incorrect predictions, the EXPERIMENTAL model also performs better in reducing incorrect predictions of non-pronominal cases, with a rate of 6.5% compared to 8.3% for the INFORMED model. These results suggest that the EXPERIMENTAL model, with its additional paragraph-related features, is particularly more effective at identifying non-pronominal cases.

The SHAP analysis, as depicted in figures 6.4, 6.5, and 6.6, provides insights into how each model uses its features to arrive at predictions. These figures illus-

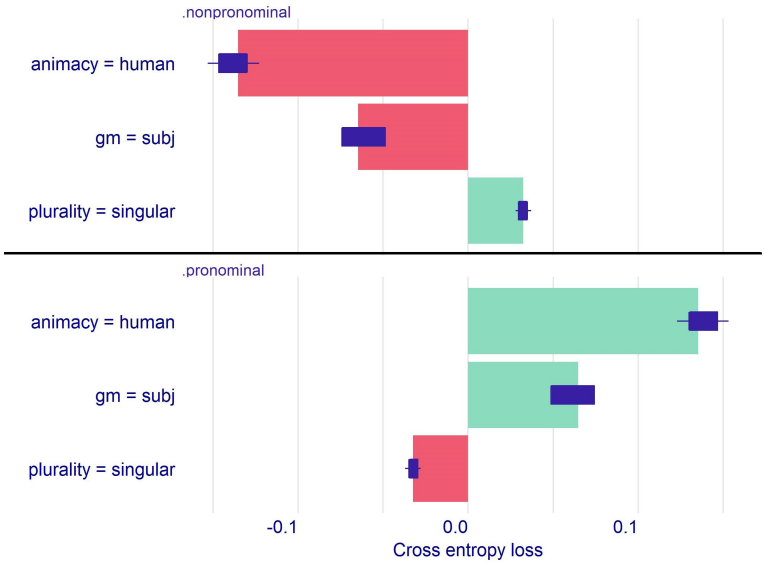

Figure 6.4: Shapley values with box plots for ten random orderings of explanatory variables in the MINIMUM model.

trate the average contribution of each feature across all observations, highlighting the importance and the order in which they impact the model's predictions. However, it is important to note that the specific contributions of features can vary for each individual observation within the data set.

As shown in Figure 6.4, animacy contributes the most to the predictions of the MINIMUM model, followed by grammatical role and plurality. The figure demonstrates that both animacy (value: `human`) and grammatical role (value: `subj`) negatively impact non-pronominalization. In other words, REs that are human and in subject position have a higher likelihood of being pronominal. Conversely, singular referents (plurality value: `singular`) tend to favor non-pronominalization.

Transitioning to the INFORMED model, as illustrated in Figure 6.5, the significance of animacy decreases, making way for sentence distance (`dist_s` = `plus_one`) to become the dominant feature. This model demonstrates a tendency towards non-pronominalization when the distance between a referent and its antecedent spans more than one sentence. This shift in feature importance suggests a more nuanced approach by the INFORMED model in predicting referential forms, taking into account the extended context beyond immediate sentence boundaries.

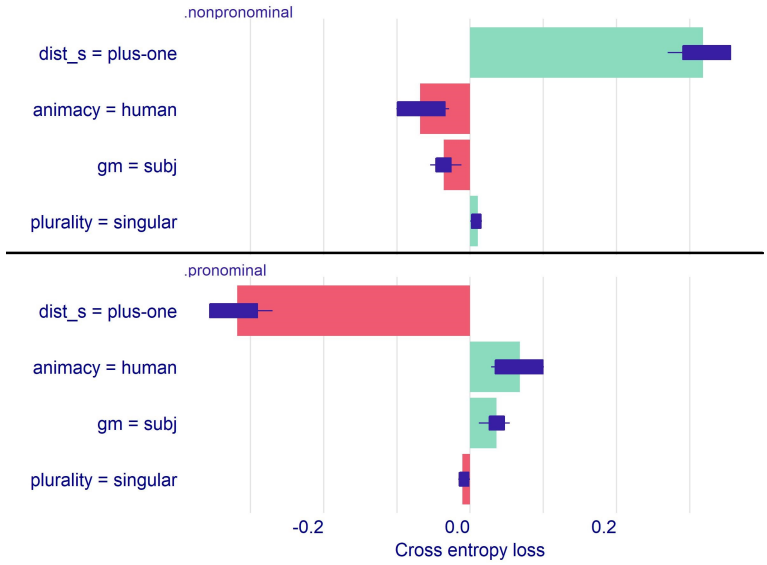

Figure 6.5: Shapley values with box plots for ten random orderings of explanatory variables in the INFORMED model.

The EXPERIMENTAL model, as delineated in Figure 6.6, continues to prioritize sentential recency, particularly when the value is `plus-one`, as a crucial factor in predicting non-pronominal forms. Additionally, paragraph givenness (`par-givenness`), especially when the RE and its antecedent are in separate paragraphs (`par_givenness = new`), emerges as the second most significant contributor to opting for non-pronominal forms. The feature `par_subj_1`, a composite metric reflecting both the first-mention status and subjecthood within a paragraph, also plays a key role in the model's decisions. Specifically, when `par_subj_1` equals `yes`, indicating the RE is the paragraph's initial subject mention, there is a tendency towards non-pronominal forms. The importance of the animacy feature, however, has dropped to the fourth place in the EXPERIMENTAL model.

In evaluating the EXPERIMENTAL model, which integrates paragraph-based information, it is noteworthy that it outperforms both the RANDOM and MINIMUM baseline models in terms of assessment metrics. However, its performance is only marginally superior to that of the INFORMED model, the most robust of the baseline models. This marginal difference in performance between the EXPERIMENTAL and INFORMED models necessitates a deeper error analysis to discern the specific areas and conditions under which paragraph-related features enhance model performance.

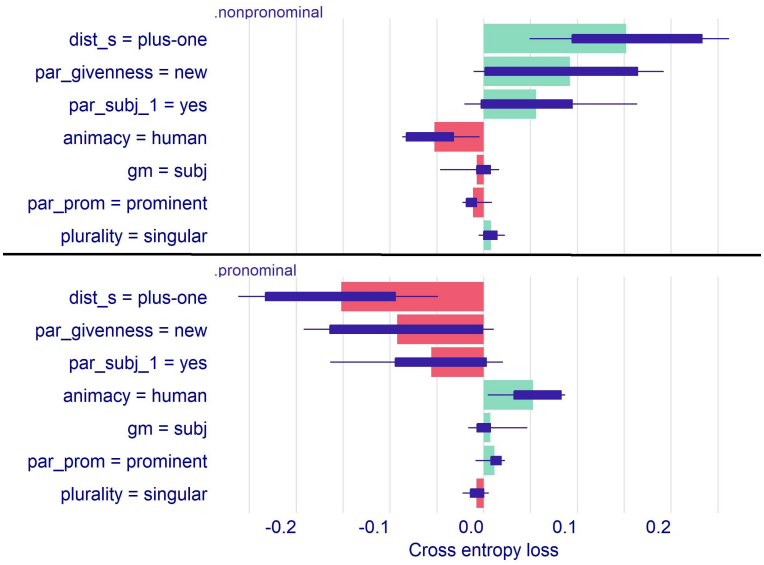

Figure 6.6: Shapley values with box plots for ten random orderings of explanatory variables in the EXPERIMENTAL model.

6.4.3 Error analysis of the INFORMED and EXPERIMENTAL models

The INFORMED and EXPERIMENTAL models cumulatively make 748 incorrect predictions. A notable observation is that 535 of these errors are shared between both models, while 213 are unique to either one. These errors manifest in two distinct forms: Firstly, *pronoun errors*, wherein a non-pronominal (-p) RE is inaccurately predicted as a pronoun; and secondly, *non-pronoun errors*, where a pronominal (+p) RE is predicted to be non-pronominal. To examine these inaccuracies more closely, sections 6.4.3.1 and 6.4.3.2 will investigate the shared errors across both models and the individual, model-specific errors, respectively.

6.4.3.1 Errors made by both models

The 535 errors shared by both models comprise 282 pronoun errors (where a non-pronominal (-p) RE is predicted as a pronoun (+p)) and 253 non-pronoun errors (where a pronominal (+p) RE is predicted as non-pronominal (-p)). This breakdown reveals a tendency for pronoun errors to be slightly more prevalent than non-pronoun errors. When considering the entire spectrum of predictions, Table 6.9 indicates that only 9% of non-pronominals were incorrectly predicted by both models, while the error rate for pronominals predicted as non-pronominals

stands at 17%. Thus, it appears that these models demonstrate a more robust accuracy in predicting non-pronominal forms. A deeper look into each type of prediction error can be gained through tables 6.10 and 6.11, which detail the most frequent feature combinations found in misclassified instances.

Table 6.9: The percentage of wrong predictions by both models. The column total_freq shows the frequency of each RF in the test set. The columns wrong_pred and wrong_pred_perc show the frequency and percentage of the RFs which are predicted wrongly.

original	prediction	total_freq	wrong_pred	wrong_pred_perc
non-pronominal	pronominal	3121	282	9%
pronominal	non-pronominal	1495	253	16.9%

Non-pronoun errors Table 6.10 presents the features and their corresponding values for cases where REs are pronouns, but are incorrectly classified as non-pronominal. The INFORMED model relies on the first four features for its predictions: grammatical role (gm), animacy, plurality, and sentence distance (distr_s). The EXPERIMENTAL model, on the other hand, uses all the listed features.

Table 6.10: Top three feature combinations of the pronominal cases predicted as non-pronominals.

pred	gm	animacy	plurality	dist_s	par_givenness	par_subj_1	par_prom	N
-p	subj	other	singular	one	given	no	prominent	49
-p	subj	other	plural	one	given	no	prominent	17
-p	subj	other	singular	one	given	no	not-prominent	16

For example, the most recurrent feature–value combination leading to 49 misclassifications in Table 6.10 is:

{ grammatical role (gm): `subject (subj)`
animacy: `not human (other)`
plurality: `singular`
sentence distance (dist_s): `one`
paragraph givenness (par_givenness): `given`
paragraph's first-mention subject RE (par_subj_1): `no`
within-paragraph prominence (par_prom): `prominent` }

This particular combination of features seems to create a conflict in the prediction process. While some factors, such as `gm:subj` and `dist_s:one`, generally lean towards pronominalization, others like `animacy:other` tend to favor non-pronominal forms. Notably, the non-pronoun errors in Table 6.10 do not include any human referents in the animacy column, suggesting that animacy plays a significant role in these misclassifications. The dominant presence of non-human referents in these errors, coupled with other feature combinations that typically support pronominalization, indicates that the models struggle particularly in cases where a non-human subject is mentioned just one sentence away from its antecedent.

Pronoun errors Table 6.11 illustrates the top three feature–value combinations of the non-pronominal REs that were incorrectly predicted to be pronominal by both models.

Table 6.11: Top three feature combination of the non-pronominal cases, predicted as pronominals.

pred	gm	animacy	plurality	dist_s	par_givenness	par_subj_1	par_prom	N
+p	subj	human	singular	one	given	no	prominent	66
+p	poss	other	singular	same	given	no	prominent	41
+p	subj	other	singular	same	given	no	prominent	30

The first row of Table 6.11 details 66 instances where the REs in the corpus are non-pronominal but were predicted as pronominal by the models. The feature–value combinations in this row include paragraph-given and prominent human referents in the subject position. According to the SHAP analyses presented earlier, these feature–value combinations strongly favor pronominalization. Non-pronominal forms in such contexts may be used for clarity, to resolve ambiguity, or to avoid repetitive pronouns. For example, in the WSJ excerpt shown in (10a), the RE *Mr. Boren* is incorrectly predicted to be pronominal by the models. In this case, it seems that a non-pronominal form is employed to prevent excessive repetition of pronouns.

(10) WSJ-0771

a. **He** points to a letter on **his** desk, **his** second in a week from President Bush, saying that they "do n't disagree." More broadly, **Mr. Boren** hopes that Panama will shock Washington out of its fear of using military power.

The other two rows in Table 6.11 reveal misclassifications mainly due to the RE and its antecedent being in the same sentence. In the WSJ corpus, 75% of REs with their antecedent in the same sentence are typically realized as pronouns, contributing to this prediction error.

6.4.3.2 Errors made by individual models

In the analysis of unique errors by each model, it was found that the INFORMED model made 142 incorrect predictions, while the EXPERIMENTAL model made 71.

Table 6.12 presents the incorrect predictions of the INFORMED model. This model uses only the first four features for its predictions: grammatical role (gm), animacy, plurality, and sentence distance (dist_s). Thus, in the INFORMED model, the REs with the feature–value combinations in the first and third rows are treated identically. These REs refer to singular human referents that are only one sentence away from their antecedent. The INFORMED model predicted the REs in these two rows to be pronominal, though they are actually non-pronominal cases.

Table 6.12: Top three feature combinations of the wrong predictions of the informed model. The INFORMED model uses only the first four features, namely gm, animacy, plurality, and dist_s.

orig	inf	gm	animacy	plurality	dist_s	par_given	par_subj_1	par_prom	N
-p	+p	subj	human	singular	one	new	yes	prominent	36
+p	-p	poss	human	singular	one	given	no	prominent	32
-p	+p	subj	human	singular	one	new	yes	not-prominent	12

Figure 6.7 demonstrates a breakdown plot for a single observation from the INFORMED model, with feature–value combinations as shown in the first row of Table 6.12. The breakdown plot decomposes the model's prediction into contributions from different variables, with green and red bars indicating positive and negative changes, respectively, in the model's mean predictions. This breakdown plot helps in understanding how different features contribute to a specific prediction, revealing insights into why certain predictions may be erroneous.

Figure 6.7 demonstrates that for the INFORMED model, the main factors influencing the prediction to be pronominal are the human referent and the subject position. The model relies on these two features to predict these cases as pronominal. However, this model lacks the critical feature of `par_givenness`, which indicates whether an RE and its antecedent are in the same paragraph. This missing feature leads to inaccuracies, as the model predicts these cases to be pronominal

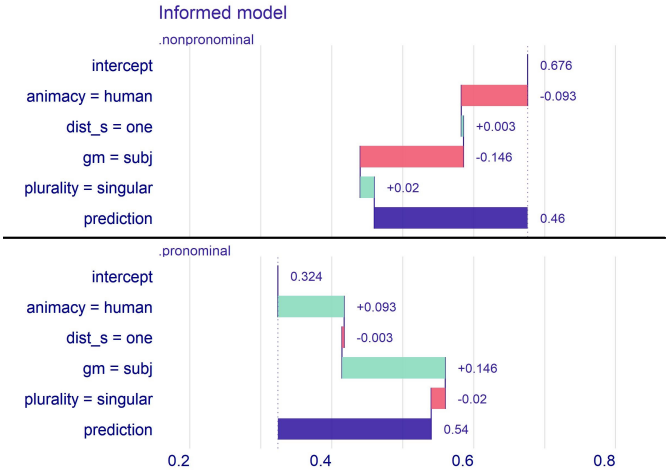

Figure 6.7: Breakdown plot for a single observation from the INFORMED model.

when they are actually non-pronominal, as a result of being the first mention of the referent in the new paragraph.

In contrast, the EXPERIMENTAL model, which includes the `par_givenness` feature, correctly predicts these cases. Figure 6.8 shows that `par_givenness` is the dominant feature in the prediction for the EXPERIMENTAL model. Therefore, one of the occasions in which the EXPERIMENTAL model performs better than the INFORMED model is in predicting REs that are only one sentence away from their antecedent, but across a paragraph boundary. This indicates that the model effectively uses the paragraph structure information to improve its predictions.

The error analysis of the EXPERIMENTAL model, as illustrated in Table 6.13, shows certain cases where the model diverges in its predictions from the actual data. Notably, these errors are fewer in number compared to those made by the INFORMED model alone.

Table 6.13: Top 3 feature combination of the wrong predictions of the experimental model.

orig	exper	gm	animacy	plurality	dist_s	par_given	par_subj1	par_prom	N
+p	-p	subj	human	singular	one	new	yes	prominent	12
+p	-p	dobj	other	singular	same	given	no	not-prominent	11
-p	+p	poss	human	singular	one	given	no	prominent	11

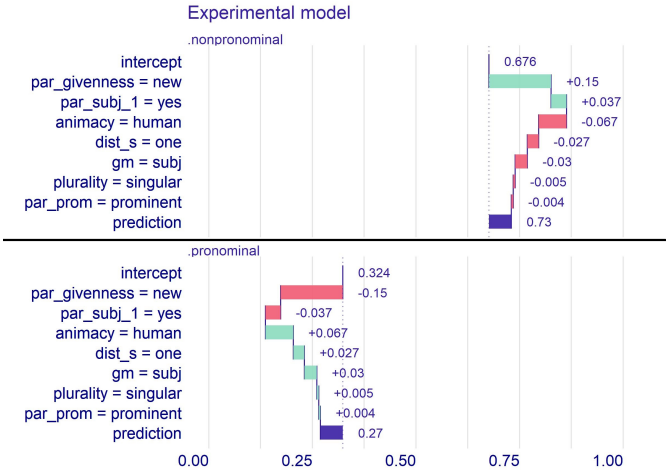

Figure 6.8: Breakdown plot for a single observation from the EXPERIMENTAL model.

The predominant type of misclassification involves paragraph-new human referents, which the model incorrectly predicts as non-pronominal, although they are represented as pronouns in the dataset. This discrepancy suggests that the model may not be adequately capturing the nuances of pronominalization across paragraph boundaries, for instance, in cases where pronominalization is used to unite two paragraphs and retain the continuity of the narrative. Here is an example from the corpus:

(11) WSJ-1102

a. Paragraph 9: [Tom Trettien], a vice president with Banque Paribas in New York, sees a break in the dollar's long-term upward trend, a trend that began in January 1988.

b. Paragraph 10: [**He**] argues that the dollar is now "moving sideways", adding that "the next leg could be the beginning of a longerterm bearish phase."

The overall error analysis, encompassing both shared and model-specific errors, demonstrates that the inclusion of paragraph transition information in the EXPERIMENTAL model particularly improves its performance in cases where there is a short linear distance but a paragraph boundary between the target RE and its

antecedent.[7] While both the EXPERIMENTAL and INFORMED models show similar accuracy levels, the nuanced differences in performance, especially in handling paragraph transitions, become evident through this detailed error analysis. This insight underscores the importance of paragraph structure as a significant factor in referential form selection.

6.5 Discussion and final remarks

The chapter's meticulous exploration of paragraph-related features in the WSJ corpus, and their impact on the choice of referential form (RF), offers significant insights and contributions to both linguistic and computational fields. What follows is a summary of the key findings and their implications.

6.5.1 Comprehensive corpus analysis of paragraph attributes

The chapter's corpus analysis (study D) delved into pronominalization and non-pronominalization in relation to paragraph structure, aligning with previous literature findings (Tomlin 1987a, Fox 1987b, Hofmann 1989). The majority (over 90%) of REs were found to be non-pronominal across paragraph boundaries. The study probed whether this is due to the initial position in a new paragraph or the transition itself. This distinction, while challenging to make due to overlapping concepts, revealed that over 90% of the paragraph-initial REs and over 93% of the discourse-old paragraph-new REs in WSJ are non-pronominal. This study also looked more closely at a few pronominal cases across paragraph boundaries, showing that pronominal paragraph-new REs were twice as frequent when referring to prominent referents than to non-prominent referents. However, since the data points were insufficient (only 331 cross-boundary pronominal cases), the results are inconclusive. In addition to examining the role of paragraph transitions, this study also examined the internal structure of paragraphs and showed that prominent referents within a paragraph have a greater likelihood of being pronominalized.

6.5.2 Impact of paragraph-related features on REG-in-context models

With a few exceptions, including Kibrik et al. (2016) and Castro Ferreira et al. (2016b), the majority of machine learning feature-based REG-in-context models

[7] As mentioned earlier, reference production is considered a non-deterministic task; therefore, misclassified cases are not necessarily incorrect or implausible.

use local features of the referent or linear recency-based concepts. Study E presented in §6.4 brought to light the role of paragraph transitions in REG-in-context models. It introduced three paragraph-related features based on corpus analysis, distinguishing between linear and hierarchical representations. This study highlighted that the transition to a new paragraph is a crucial factor, improving model performance, but only modestly compared to strong baselines. The limited improvement might be partly due to the specific characteristics of the WSJ corpus, where only 8% of REs have an antecedent one sentence away, but in a different paragraph. The contribution of paragraph-related transitions might be more pronounced if applied, for example, to a Wikipedia corpus (e.g., GREC-2.0 or GREC-PEOPLE mentioned earlier) in which the majority of sentences revolve around the main topic of the document.

6.5.3 Advocacy for model explainability

Rather than focusing exclusively on improving the performance of the models, study E sought to offer explanations for the predictions made by the models in three ways: (1) the SHAP analysis provided information on the extent and direction of the contribution of each explanatory variable to the models' predictions, (2) the error analysis provided information on the cases where the models did not perform well, and (3) the breakdown method made it possible to compare the performance of two models by looking at the individual decisions each model makes. This approach is vital as it bridges the gap between computational predictions and linguistic theories, offering a pathway to refine models based on linguistic insights.

The findings presented in this chapter highlight the significance of paragraph structure in the linguistic analysis and computational modeling of reference. By emphasizing explainability and the nuanced roles that paragraph-related features play, this chapter contributes positively to both fields. It paves the way for future research that integrates linguistic theory with computational techniques.

7 A systematic evaluation of REG-in-context approaches

7.1 Introduction

As outlined in Chapter 3, three aspects of REG-in-context warrant particular attention. The first two aspects, the choice of corpora and features, have been extensively covered in Chapters 4, 5, and 6. In this current chapter, we shift our focus to the third aspect: exploring the diverse approaches for addressing the REG-in-context task.[1]

In study F (§7.2), we aim to systematically compare three distinct methodologies: rule-based, feature-based, and E2E neural network approaches. This comparison is conducted across two significantly different datasets, WEBNLG and WSJ, and involves evaluations based on both automated metrics and human judgments.

Beyond performance metrics, it is also crucial to consider how transparently different models can explain their decision-making processes. Rule-based and feature-based models typically excel in this aspect, offering greater clarity in their operational mechanisms, in contrast to DL models, which tend to be more opaque or *black-box* in nature. To delve deeper into this disparity, study G in §7.3 conducts a series of probing experiments. These experiments are designed to enhance our understanding of neural REG-in-context RFS models.

[1]The studies presented in this chapter are based on two published articles: [study F] Fahime Same et al. 2022. Non-neural models matter: A re-evaluation of neural referring expression generation systems. In Smaranda Muresan et al. (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 5554–5567. Dublin: Association for Computational Linguistics. DOI: 10.18653/v1/2022.acl-long.380. [study G] Guanyi Chen et al. 2021. What can neural referential form selectors learn? In *Proceedings of the 14th International Conference on Natural Language Generation*, 154–166. Aberdeen: Association for Computational Linguistics. DOI: 10.18653/v1/2021.inlg-1.15.

7.2 Study F: A systematic comparison of REG-in-context approaches

Historically, REG-in-context studies have often adopted a two-step approach to REG, as noted by Henschel et al. (2000) and Krahmer & Theune (2002). The initial step involves determining the form of an RE – whether it should be a proper name (e.g., *Marie Skłodowska-Curie*), a description (*the physicist*), or a pronoun (*she*). Subsequently, the focus shifts to content selection, where the REG system decides on the different ways in which an RF can be realized. For instance, in referencing Marie Curie, should the REG system opt for a simple description involving only her profession like *the physicist* or include additional details like *a Polish-French physicist*? This dichotomy of form and content selection has been a hallmark of most rule-based and feature-based models. In contrast, E2E models appear to integrate these steps, simultaneously addressing both aspects (Castro Ferreira et al. 2018a, Cao & Cheung 2019, Cunha et al. 2020).

In Chapter 3, we discussed Castro Ferreira et al. (2018a)'s pioneering work in E2E REG-in-context. For their study, they derived the WEBNLG dataset specifically for the REG-in-context task from the original WEBNLG corpus (Gardent et al. 2017). They developed a neural REG system leveraging a sequence-to-sequence model with attention mechanisms. Despite automatic and human evaluations showing neural REG systems' superiority over rule-based and feature-based baselines, the strength of the baseline models used for comparison, notably ONLYNAME and FERREIRA, was modest. ONLYNAME is a rule-based system that always generates a proper name given an entity, and FERREIRA is a Naive Bayes model with only three simple features.[2]

This context sets the stage for study F in this chapter, which seeks to reassess the relative efficacy of state-of-the-art (SOTA) neural REG models, hypothesizing that neural REG models are not always better than rule-based and feature-based models. Our hypothesis challenges the prevailing notion, questioning whether these advanced neural models indeed outperform more traditional rule-based and feature-based systems in the REG-in-context task. This study aims to provide a more comprehensive comparison by employing stronger baselines than those previously used.

The current study strategically introduces a variety of rule-based and feature-based models to assess how neural models compare against *well-designed* non-

[2]It is worth noting that Cunha et al. (2020)'s human evaluation presented mixed results, with the ONLYNAME model performing comparably to neural REG models in terms of fluency, grammaticality, and adequacy. However, due to the limited number of evaluators, these findings should be approached cautiously.

neural counterparts. It is crucial to note that the efficacy of a model is not solely contingent on its complexity. In fact, a simple, yet well-formulated rule-based system with one or two rules can be highly effective. Given that neural E2E models are favored for their minimal need for feature engineering, our comparison incorporates two types of baselines: (1) models necessitating minimal expert effort, and (2) more resource-intensive models, which leverage linguistically well-established rules and features, as discussed in Chapter 3 (§3.3.2.1) and study B of Chapter 5.

Our analysis aims to not just compare the performance of these models but also to consider the amount of resources each requires. This holistic view is vital to fully comprehend the trade-offs involved. While neural models benefit from requiring less linguistic expertise and annotation effort, they demand significant computational resources and expertise in deep learning. In contrast, rule-based and feature-based models might necessitate more intensive linguistic input but are generally less demanding in terms of computational power. This study endeavors to shed light on these varying demands, offering a nuanced perspective on the practicalities of implementing different REG-in-context models.

To align this study with previous E2E REG-in-context investigations, we use the WEBNLG dataset.[3] A notable constraint of this corpus is its high rate of entity recurrence: approximately 99.34% of entities in the test set also appear in the training set. This aspect limits the dataset's ability to assess performance on unseen entities. Additionally, since many sentences in WEBNLG are paraphrases of one another, evaluating neural models on this corpus alone may overestimate their performance. Recognizing these limitations, Castro Ferreira et al. (2019) expanded WEBNLG to include unseen domains with numerous unseen entities. Similarly, Cunha et al. (2020) have developed models tailored to these new challenges, dividing their test set into two subsets: one comprising documents with 99.34% *seen* entities, and the other with 92.81% *unseen* entities. However, such composition could render the dataset somewhat unrealistic (for an in-depth discussion, see §7.2.1). To counter this, we create a dataset we consider more representative of real-world scenarios, derived from the WSJ corpus (Hovy et al. 2006, Weischedel, Ralph et al. 2013).

The structure of the study is as follows: §7.2.1 and §7.2.2 detail the datasets and REG models employed. In §7.2.3, we elaborate on our methodologies for both human and automatic evaluations. Finally, §7.2.4 will present a comparative analysis of the results, offering insights and recommendations for future

[3]We use version 1.5 of the WEBNLG dataset available at `https://github.com/ThiagoCF05/webnlg`.

research endeavors. This comprehensive approach aims to provide a more accurate assessment of REG models' capabilities in handling diverse and realistic data scenarios.

7.2.1 Task and datasets

This section explains the REG-in-context task and describes the two English datasets used to conduct the experiments.

7.2.1.1 The REG-in-context task in study F

In study F, we define the task of REG-in-context as follows: given a text whose REs have not yet been generated, and given the intended referent for each of these REs, the REG-in-context task is to build an algorithm that generates all of these REs.

The rule-based and feature-based models in this study adopt a two-step approach to this task, encompassing both RFS (Referential Form Selection) and RCS (Referential Content Selection). In contrast, the E2E models are designed to address both steps simultaneously. To illustrate, consider the delexicalized text in Table 7.1, featuring the entity AWH_Engineering_College. Given the entity AWH_Engineering_College, REG selects an RE based on the *entity* and its *pre-context* (*AWH_Engineering_College is in "Kuttikkattoor", India in the state of Kerala.*), and its *post-context* (*has 250 employees and Kerala is ruled by Kochi. The Ganges River is also found in India.*).

Table 7.1: An entry from the WEBNLG corpus. In the delexicalized text, every entity is **in bold**.

Triples: (AWH_Engineering_College, country, India), (Kerala, leaderName, Kochi), (AWH_Engineering_College, academicStaffSize, 250), (AWH_Engineering_College, state, Kerala), (AWH_Engineering_College, city, "Kuttikkattoor"), (India, river, Ganges)
Text: AWH Engineering College is in Kuttikkattoor, India in the state of Kerala. The school has 250 employees and Kerala is ruled by Kochi. The Ganges River is also found in India.
Delexicialized Text: **Pre-context**: **AWH_Engineering_College** is in "**Kuttikkattoor**" , **India** in the state of **Kerala** . **Target Entity**: **AWH_Engineering_College** **Post-context**: has 250 employees and **Kerala** is ruled by **Kochi** . The **Ganges** River is also found in **India** .

7.2.1.2 The WEBNLG dataset

The WEBNLG corpus, introduced by Gardent et al. (2017), serves as a benchmark for evaluating NLG systems. This dataset originated from a crowd-sourcing experiment where participants were tasked to write textual descriptions for given Resource Description Framework (RDF) triples (as exemplified in Table 7.1), with each entry comprising 1 to 7 triples. Later, Castro Ferreira et al. (2018a) and Castro Ferreira et al. (2018b) enriched and adapted the corpus, particularly delexicalizing it, to align with the REG-in-context task. Castro Ferreira et al. (2019) further extended WEBNLG and divided the documents into the test sets *seen* (where all entities appear in the training data) and *unseen* (where none appear in the training data). This division was intended to help assess models' ability to handle seen and unseen entities. Since the maximum number of triples in the unseen set is five, we would expect the unseen data to be less complex than the seen data. In our study, we used version 1.5 of WEBNLG, featuring 67,027, 8278, and 19,210 REs in the training, development, and test sets, respectively (see Table 7.1 for an example).

Despite its utility, WEBNLG exhibits certain limitations. Primarily, it consists of relatively formal texts with simplistic syntactic structures, which may not accurately represent the complexity and variability of everyday language use. Additionally, the texts in WEBNLG are notably brief, averaging only 1.4 sentences each. There is also a significant imbalance in the types of REs used, with a predominance (71%) of proper names, and a high proportion (85%) of first-mention REs. Furthermore, in the test samples, entities are usually either entirely seen or unseen, lacking a realistic mix of both. Given these constraints, we decided to complement our analysis with a second corpus, aiming to provide a more robust and comprehensive evaluation of the algorithms across different linguistic contexts.

7.2.1.3 The WSJ dataset

In an effort to complement the WEBNLG dataset and to introduce a more diverse linguistic environment, we followed the approach introduced by Castro Ferreira et al. (2018a) and developed a new English REG dataset based on the Wall Street Journal (WSJ). As in Chapter 4, we use ONF format files of the OntoNotes corpus (Weischedel, Ralph et al. 2013) for the creation of this dataset. The dataset excludes first and second person REs and assumes a linear presentation order, thereby omitting complex cases like union REs. For instance, in (1), while individual REs ([*Mary*], [*John*], [*David*]) are included, their union form (*Mary, John, and David*), is excluded.

(1) [*Mary*], [*John*] and [*David*] got their booster shots yesterday.

The resulting WSJ dataset includes 582 newspaper articles, containing 20,186, 2362, and 2781 REs across training, development, and test sets, respectively. Each document in this dataset is substantially longer than those in WEBNLG, averaging about 25 sentences. Furthermore, this dataset exhibits a more balanced distribution of first mentions (23%) and subsequent mentions.

To prepare the dataset, we first delexicalize the REs. This dataset comprises nearly 8000 coreferential chains. In each chain, the REs are replaced with corresponding delexicalized expressions, similar to those illustrated in Table 7.1.

The delexicalization process involves three key steps: (1) using the information contained within the content of each RE, (2) leveraging the fine-grained annotations of the RFs, and (3) considering the entity type of each referent. For instance, in the delexicalization of human REs, we start by identifying concise yet informative RFs, such as the combination of first and last names (e.g., *Barack Obama*). When such an expression appears in a coreferential chain, its delexicalized form (with tokens separated by underscores, e.g., *Barack_Obama*) is assigned to all REs in that chain. This structured approach ensures that each referent is represented in a consistent and informative manner across the dataset. Below is the order in which the human referents are searched and delexicalized:

- [firstname-lastname]
- [title-firstname-lastname]
- [modified firstname-lastname]
- [title-lastname]
- [lastname]
- [modified-lastname]
- [firstname]

After delexicalizing all REs across various entity categories, we then define the context for each RE. This includes not only its pre-context and post-context at the local sentence level but also an extended context comprising K preceding and following sentences, where K is referred to as the *context length*. This extended context provides a more comprehensive background for each RE. An example from the WSJ dataset, illustrating this approach with a context length of K=2, is presented in Table 7.2.

7.2.2 REG models

This section presents the rule-based, feature-based, and neural SOTA REG models used in this study.

Table 7.2: An entry from the WSJ corpus. In the delexicalized text, every entity is **in bold**.

Text: Ronald B. Koenig , 55 years old , was named a senior managing director of the Gruntal & Co. brokerage subsidiary of this insurance and financial - services firm . Mr. Koenig will build the corporate - finance and investment - banking business of Gruntal , which has primarily been a retail - based firm . He was chairman and co-chief executive officer of Ladenburg , Thalmann & Co. until July , when he was named co-chairman of the investment-banking firm along with Howard L. Blum Jr. , who then became the sole chief executive . Yesterday , Mr. Blum , 41 , said he was n't aware of plans at Ladenburg to name a co-chairman to succeed Mr. Koenig and said the board would need to approve any appointments or title changes .
Delexicalized Text:
Pre-context: **Mr._Koenig** was named a senior managing director of the **Gruntal** brokerage subsidiary of this insurance and financial-services firm .
Target Entity: **Mr._Koenig**
Post-context: will build the corporate-finance and investment-banking business of **Gruntal**. **Mr._Koenig** was chairman and co-chief executive officer of **Ladenburg** until July , when **Mr._Koenig** was named co-chairman of **Ladenburg** along with **Mr._Blum**. Yesterday, **Mr._Blum** said **Mr._Blum** was n't aware of plans at **Ladenburg** to name a co-chairman to succeed **Mr._Koenig** and said the board would need to approve any appointments or title changes.

7.2.2.1 Rule-based REG

Rule-based models have been widely used for generating REs in context (McCoy & Strube 1999, Henschel et al. 2000). Here, we build rule-based systems for binary classification of REs into two classes, namely `pronominal` and `non-pronominal`.

Simple rule-based model (RREG-S) Our first rule-based algorithm, RREG-S, is outlined in Algorithm 3. This model operates under two primary rules:

1. An entity *r* is classified as a pronoun if it meets two conditions:
   a) *r* is *discourse-old*, meaning it has been mentioned in the preceding context.
   b) *r* has no *competitor* in the current or the previous sentence. A competitor is defined as another entity that could be referred to using the same pronoun as *r*.
2. In all other cases, *r* is realized as a non-pronominal RE.

Algorithm 3: Simple rule-based system (RREG-S).

1 The target entity r is realized as a pronominal RE if:
2 r is *discourse-old*;
3 r has no *competitor* in the current sentence and the previous sentence,
4 Otherwise, r is realized as a non-pronominal RE.

For the realization of pronouns, we have developed a dictionary that stores the pronouns associated with each entity. For entities already seen in the training data, their pronouns are extracted directly from this data. In cases where an entity has multiple pronominal forms, the most frequently occurring form is selected. For unseen entities, we determine the appropriate pronoun based on their meta-information (which is also used in E2E systems of Cunha et al. 2020). Assuming an entity in WEBNLG has the meta-information `PERSON` and gender `FEMALE`, we assign the pronoun *she* to that entity. Following Castro Ferreira et al. (2016b), pronouns are further tailored to align with the grammatical role of the entity in a sentence. For example, in the object position, the pronoun *he* is changed to *him*. For non-pronominal REs, realization is achieved by converting underscores in the entity label to whitespaces. This process transforms *Adenan_Satem* to *Adenan Satem*, as previously described by Castro Ferreira et al. (2018a).

Linguistically-informed rule-based model (RREG-L) This model draws from the pronominalization rules outlined by Henschel et al. (2000). This model is founded on the principles of *local focus*, a concept based on a simplified implementation of Centering Theory (Grosz et al. 1995), and *parallelism*, which examines if the target entity r and its antecedent share the same grammatical role (Henschel et al. 2000). Algorithm 4 details the process of generating REs using RREG-L.

As input, the system takes the target entity r together with the current sentence (u_2) and the previous sentence (u_1). The output is the surface form of r.

The algorithm initiates by checking for the presence of an antecedent of r in u_1 (line 3). In the absence of an antecedent, r is realized non-pronominally (line 13).

If an antecedent exists, the model checks for parallelism (line 4), verifying if r shares the same grammatical role (subject or object) with its antecedent. If parallelism is established, r is realized as a pronoun (line 5).

In the absence of parallelism (line 6), the local focus theory of Henschel et al. (2000) is applied. Here, a referent becomes the local focus if it is either discourse-

Algorithm 4: Linguistically-informed rule-based algorithm (RREG-L).

```
1  Input The target entity r, the sentence u_2 that r is in, and its previous
     sentence u_1.
2  Output The surface form of r.
3      if r has an antecedent in u_1 then
4          if r occurs in parallel context then
5              RealizeProRE(r)
6          else
7              F := FocusSetConst(u_1)
8              if r's antecedent ∈ F and r has no competitor r' ∈ F then
9                  RealizeProRE(r)
10             else
11                 RealizeNONProRE(r)
12     else
13         RealizeNONProRE(r)
```

old or occupies the subject position. The `FocusSetConst` function (line 7) constructs a set $\mathcal{F}$ of local focus entities from u_1. If r's antecedent is in $\mathcal{F}$ and r has no competitor that is an element of $\mathcal{F}$ (line 8), it is then realized pronominally (lines 9). If not, r is realized non-pronominally (line 11).

The surface realization functions, `RealizeProRE` and `RealizeNONProRE`, are similar to those in RREG-S and responsible for generating pronominal and non-pronominal REs, respectively.

7.2.2.2 Feature-based REG

For building our feature-based REG models, we use *CatBoost* (Prokhorenkova et al. 2018), a powerful machine learning algorithm known for handling categorical features effectively. These models are designed to predict whether a reference should be realized as a `pronoun`, a `proper name`, or a `description`. Once the RF is predicted, the model's next task is to select the specific content for the RE.

In the content selection phase, the most frequent variant of RE in the training corpus is chosen, matching both the predicted RF class and the referent, considering the complete set of features. This approach ensures that the model's predictions are grounded in the most typical usage patterns observed in the training data.

However, there may be instances where no exact match for the predicted RF and referent is found in the training corpus. In such cases, we employ a back-

off method (Castro Ferreira et al. 2018a). This method progressively removes one feature at a time from the set, in order of increasing importance, until a matching RE is found. This means we start by removing the least important feature first and then proceed to remove features that are increasingly more important. The importance order of the features is determined using CatBoost's inherent feature importance evaluation method, which ranks features based on their contribution to the model's predictive power.

To explore different dimensions of feature-based REG, we have developed two models: ML-S and ML-L. The ML-S model is designed to be simpler, using a limited set of features, while ML-L is more complex and incorporates a broader range of linguistic features. The distinction between these models allows us to assess how varying levels of feature complexity influence the effectiveness of REG in different contexts.

Simple feature-based model (ML-S) In order to determine the upper-bound performance of a system that operates without requiring additional linguistic information or extensive annotation efforts, we have developed a model named ML-S. To achieve this, ML-S deliberately omits more complex linguistic features that would require external processing tools like a syntactic parser. Instead, it focuses on features that can be directly and automatically extracted from the text. Key features in this model include recency indicators, referential status, and positional information. Table 7.3 shows the specific features employed in ML-S and offers a brief description of each.

Table 7.3: Features used in the ML-S models. Each feature is defined and computed for each target entity *r*. Antecedent (ANTE) refers to the first coreferential RE preceding the current *r*.

Feature class	Definition
Referential status	Is *r* the first mention of the entity in the text?
Referential status	Is ANTE in the same sentence?
Recency	Categorical distance between *r* and ANTE in number of sentences
Recency	Categorical distance between *r* and ANTE in number of words
Competition	Is there any other RE between *r* and ANTE?
Position	Is *r* the first, second, middle, or last mention?

Linguistically-informed feature-based model (ML-L) ML-L is constructed to explore the upper limits of performance achievable with a linguistically informed

feature-based model. Drawing primarily from the findings of study B in Chapter 5, ML-L incorporates a set of carefully selected features that are expected to have significant effect on the generation of contextually appropriate REs. Table 7.4 presents the features used in ML-L.

Table 7.4: Features used in the WEBGNLG and WSJ ML-L systems. Each feature is defined and calculated for each target entity *r*. `Antecedent` (*ANTE*) refers to the first co-referential RE preceding the current *r*.

Feature class	Definition
Grammatical Role	Grammatical Role of *r*
Grammatical Role	Grammatical Role of ANTE
Meta information	Entity type (e.g., human, city, country, organization)
Meta information	Plurality: Is *r* plural or singular? (only WSJ)
Meta information	Gender
Recency	Categorical distance between *r* and ANTE in number of words
Recency	Categorical distance between *r* and ANTE in number of *sentences*
Recency	Categorical distance between *r* and ANTE in number of paragraphs (only WSJ)

One of the key adaptations in both feature-based models is the conversion of numeric features into categorical values, a decision made to facilitate their use in a back-off method for content selection of REs. Notably, the consensus set from study B does not include *distance measured in number of words* as a feature. However, models trained on the WEBNLG corpus demonstrate a marked preference for word recency over sentence recency. This preference is likely attributed to the brief length of WEBNLG documents, underscoring the significant impact of corpus characteristics on feature selection. Below is a detailed breakdown of each recency feature, focusing on measuring the distance between the referent *r* and its antecedent:

Word distance: This feature is categorized into five quantile groups. It is used in all feature-based models on both the WEBNLG and WSJ datasets.

Sentence distance: This is measured by (1) two quantile groups in the WEBNLG ML-R, ML-S and WSJ ML-S models, and (2) three distinct bins that categorize if the antecedent (ANTE) is in the same sentence, one sentence away, or more than one sentence away, as used in the WSJ ML-L model.

Paragraph distance: Here, the distance is segmented into four bins, depending on whether the referent (*r*) and its ANTE are in the same paragraph, one paragraph away, two paragraphs away, or more than two paragraphs away. This feature is specifically employed in the WSJ ML-L model.

The ML-L models are designed to fully leverage syntactic information. This syntactic data is sourced from the annotations available in the ONTONOTES corpus. For parsing the WEBNLG documents, we use the Python library *spaCy*. In addition to syntactic features, these models also incorporate meta-information such as *plurality* (applicable only in the WSJ models), *gender*, and *person* as input features. It is worth noting that the rule-based and the E2E models also use this meta-information for RE realization.

7.2.2.3 Neural REG

One of the challenges with rule-based and feature-based models is their limited capacity to handle cases where an RF, such as a proper name, can have multiple realizations (for example, *Lady Gaga* or *Stefani Germanotta*). This is where End2End neural REG models demonstrate their strength. These models are adept at generating REs from scratch, accommodating various possible realizations of an entity. In this study, we focus on three distinct neural REG systems, all of which have been designed to deal with unseen entities. Each of these systems is built upon the foundation of the sequence-to-sequence model with attention, as introduced by Bahdanau et al. (2014).

ATT+COPY The ATT+COPY model, proposed by Cunha et al. (2020), employs three bidirectional Long Short-Term Memory networks (LSTMs) (Hochreiter & Schmidhuber 1997). These LSTMs encode three key components: the pre-context, the post-context, and the proper name of an entity, where the entity labels are modified by replacing underscores with whitespaces. The outputs of these encoders are three distinct hidden vectors: $h^{(\text{pre})}$, $h^{(\text{post})}$, and $h^{(r)}$. During each step of decoding, designated as t, the model employs three separate attention mechanisms. These mechanisms focus on the encoded contexts and compute attention vectors, which are then merged into a unified context vector, denoted as $\mathbf{v}_t^{(c)}$. The model uses an autoregressive LSTM-based decoder that generates the REs relying on these context vectors.

To effectively address the challenge of unseen entities, Cunha et al. (2020) incorporated a copy mechanism within the decoder. This mechanism grants the decoder the ability to directly copy words from the given contexts into its output, enhancing the model's flexibility and adaptability in handling unseen entities.

ATT+META The ATT+META model, as developed by Cunha et al. (2020), integrates meta-information about each entity to refine the generation of REs. During each decoding step, denoted as t, the model combines the context vector $\mathbf{v}_t^{(c)}$

with embeddings representing the meta-information, prior to inputting it into the decoder. In the case of the WEBNLG dataset, this meta-information includes the type of the entity, represented as $\mathbf{v}^{(type)}$, and gender, denoted as $\mathbf{v}^{(gender)}$. For the WSJ dataset, the model additionally incorporates plurality, signified by $\mathbf{v}^{(pl)}$, alongside the entity type and gender embeddings.

PROFILEREG The PROFILEREG model, developed by Cao & Cheung (2019), leverage the content of entity profiles extracted from Wikipedia to generate REs. In a departure from the above-mentioned approaches that focus on the encoding of the proper names of entities, PROFILEREG instead asks the entity encoder to encode the entirety of an entity's profile to obtain the hidden vector $h^{(r)}$. However, it is important to note that this model is specifically tailored for the WEBNLG dataset, as the complete profiles for entities in the WSJ dataset are not readily accessible. Consequently, our evaluation of PROFILEREG is exclusively focused on its performance with the WEBNLG dataset.

For the implementation of the models on the WEBNLG dataset, we adhere to the original parameter settings as defined in their original implementations.[4] For the WSJ dataset, a key aspect of our training is to determine the optimal context length, denoted as K. To achieve this, we experiment with varying K, ranging from 1 to 5 sentences, both preceding and following the target sentence. These variations are tested using the ATT+META model on the WSJ development set. Our findings indicate that the model achieves its peak performance when the context length K is set to 2 sentences. This specific configuration of K is therefore selected for the subsequent implementations on the WSJ dataset.

7.2.3 Evaluation

In this section, we conduct a thorough evaluation of all systems outlined in §7.2.2, applying them to both the WEBNLG and WSJ datasets. Our evaluation process contains both automatic and human assessments to ensure a comprehensive analysis of each system's performance.

7.2.3.1 Automatic evaluation

Metrics In our evaluation of REG systems, we use a comprehensive set of metrics, following the approach of Cunha et al. (2020). The assessment is conducted from three distinct perspectives:

[4]The original implementations for ATT+COPY and ATT+META can be found at `https://github.com/rossanacunha/NeuralREG`, and for PROFILEREG at `https://github.com/mcao610/ProfileREG`.

- **RE Accuracy and String Edit Distance (SED)**: These metrics are employed to evaluate the quality of each generated RE. SED, as defined by Levenshtein (1966), measures the minimum number of edits needed to transform one string into another, providing a clear indication of the similarity between the generated and the expected REs.
- **BLEU and Text Accuracy**: Upon inserting the REs into the original documents, we use the BLEU score (Papineni et al. 2002) and Text Accuracy as evaluative tools. The BLEU score measures the correspondence between the machine-generated text and the human-generated reference text, while Text Accuracy assesses the overall accuracy of the text after RE insertion.
- **Precision, Recall, and F1 score for pronominalization**: These standard metrics are used to evaluate the effectiveness of pronominalization within the generated texts. Precision assesses the accuracy of the pronominal REs, recall measures the ability to identify all relevant instances of pronominalization, and the F1 score provides a harmonic balance between precision and recall.

Results for WEBNLG In the evaluation of the WEBNLG dataset, as detailed in Table 7.5, a notable pattern emerged: Both rule-based and feature-based models perform better than the neural models overall. However, it was observed that neural models perform better in pronominalization tasks. The feature-based model ML-L demonstrated superior performance in generating REs, as evidenced by its highest scores in RE Accuracy and BLEU, along with the second best results in SED and Text Accuracy. On the other hand, PROFILEREG performs best in pronominalization, closely followed by the simpler rule-based system RREG-S.

Interestingly, RREG-S, the simplest of the rule-based models, exhibited exceptional performance. It not only surpassed its linguistically informed counterpart, RREG-L, but also exceeded the neural models ATT+COPY and ATT+META in both RE generation and pronominalization. This outcome underscores the potential effectiveness of simpler, rule-based approaches in certain contexts. Additionally, Table 7.6 provides a breakdown of these results into seen and unseen subsets.

As shown in Table 7.6, the neural models – ATT+COPY, ATT+META, and PROFILEREG – showed a distinct pattern in their performance. They ranked as the top three models for data they had seen before (seen data) but exhibited the weakest performance for unseen data, as evidenced by lower scores in RE Accuracy, SED, BLEU, and Text Accuracy. Conversely, the feature-based models, while performing slightly below their rankings for seen data (ranked fourth and fifth), displayed

Table 7.5: Results of the automatic evaluation of WEBNLG. The best results are **boldfaced**, while the second best are underlined. ↑ means the higher the metric the better, while ↓ means the lower the better.

Model	RE Acc.↑	SED↓	BLEU↑	Text Acc.↑	Precision↑	Recall↑	F1↑
RREG-S	54.60	**3.65**	72.05	16.28	**89.52**	77.57	82.28
RREG-L	53.43	3.77	71.27	15.49	73.94	73.96	73.95
ML-S	54.35	3.70	70.89	15.43	71.70	63.52	66.39
ML-L	**56.69**	3.66	**72.25**	16.36	81.66	63.62	68.36
ATT+COPY	48.75	4.46	68.48	14.88	85.33	75.74	79.63
ATT+META	53.34	4.22	70.82	**16.54**	86.32	75.56	79.81
PROFILEREG	40.96	7.40	61.04	11.39	86.44	**87.40**	**86.91**

Table 7.6: Results of the automatic evaluation of WEBNLG for seen and unseen entities, respectively.

Model	RE Acc.↑	SED↓	BLEU↑	Text Acc.↑	Precision↑	Recall↑	F1↑
RREG-S	54.76/**54.44**	3.94/**3.35**	73.98/**69.90**	18.85/**13.56**	85.50/**91.51**	71.21/81.61	76.29/**85.73**
RREG-L	54.10/52.75	3.86/3.68	73.56/68.72	18.49/12.41	69.66/77.65	73.34/74.43	71.30/75.90
ML-S	58.61/50.06	3.38/4.02	73.73/67.67	18.65/12.12	74.15/56.99	85.05/50.26	78.35/48.24
ML-L	63.01/50.32	3.30/4.03	76.10/67.91	20.30/12.33	83.68/73.90	85.00/50.21	84.32/47.98
ATT-COPY	71.47/25.84	2.64/6.28	**80.46**/54.50	26.39/3.08	**86.90**/83.66	87.75/68.12	87.32/72.97
ATT-META	**71.64**/34.88	**2.62**/5.82	80.26/60.00	**27.88**/4.93	86.48/86.66	87.97/67.72	87.21/73.14
PROFILEREG	68.26/13.43	3.27/11.55	78.24/41.13	21.82/0.7	86.79/86.04	**94.80**/**82.66**	**90.33**/84.24

a modest decline in their effectiveness on unseen data, a drop less pronounced than that experienced by the neural models.

This discrepancy in performance, particularly the marked drop in the neural models' effectiveness on unseen data, is likely attributable to their limitations in handling unseen entities, for instance, because the models fail to conduct domain transfer. This issue seems especially acute given that the unseen data in WEBNLG comprises different domains than the seen data. It indicates a potential overfitting by the neural models to the training data, which compromises their ability to generalize and adapt to new domains. Consequently, this raises concerns about the neural models' reliance on extensive parameterization, which, while beneficial in certain contexts, may impede their ability to generalize effectively across diverse datasets and domains.

The fact that the unseen entities come from different domains could also explain the minimal difference in RE accuracy between the ML-L and ML-S models (50.32% vs. 50.06%, respectively). As indicated in Table 7.4, the ML-L model incorporates the entity type (analogous to the domain label in WEBNLG) as one of its

features. However, since the training and test sets have different domain labels, this feature does not contribute to the predictions. Consequently, the anticipated advantage of ML-L over ML-S in handling domain-specific information diminishes when encountering entities from previously unseen domains.

Rule-based systems, which do not depend on training data, are not susceptible to the same performance variations as seen in feature-based models when encountering unseen data. This accounts for their relatively stable, and in some cases, enhanced performance in pronominalization tasks with unseen entities. As noted in §7.2.1, the unseen data in WEBNLG comprises fewer triples, potentially contributing to this improved performance.

Conversely, feature-based models, such as ML-L and ML-S, demonstrate weaker results in pronominalizing unseen entities. This can be attributed to their complex task of making a three-way distinction between pronouns, proper names, and descriptions, as opposed to the simpler binary categorization of pronominal vs. non-pronominal forms employed by rule-based models. The complexity of distinguishing among these three referential forms in feature-based models may lead to their reduced effectiveness in handling unseen entities, particularly when such entities diverge significantly from those in the training set.

The annotation practices employed in WEBNLG may also undermine the performance of feature-based models. Given their reliance on the quality of data, the quality of annotations can significantly impact these models' efficacy. In WEBNLG, a non-pronominal RE is marked as a `description` if it includes a determiner, and as a `proper name` otherwise. This approach leads to some inaccuracies: For instance, *United States* is classified as a proper name, while *The United States* is incorrectly labeled as a description. While we have retained these original annotations to maintain consistency with prior studies, it is important to recognize that these labeling practices may have inadvertently contributed to the suboptimal performance observed in feature-based models. This aspect should be considered when interpreting their results and in future modifications to the dataset.

Results for WSJ The results for the WSJ dataset, as shown in Table 7.7, reveal that the ML-L model stands out as the best model in both RE generation and pronominalization, surpassing all other models by a significant margin. In the context of rule-based models, RREG-L demonstrates better performance over RREG-S across all evaluation metrics. This observation aligns with our earlier insights, suggesting that the WSJ dataset, unlike WEBNLG, includes more varied and naturalistic texts, as discussed in §7.2.1.2. This complexity in the WSJ texts appears to favor the more sophisticated, linguistically informed approach used in the RREG-L model.

Table 7.7: Automatic evaluation results for WSJ. Note that the text accuracy for a WSJ document is always 0, since it is extremely unlikely that every RE will be generated correctly. We therefore report the sentence-level accuracy instead.

Model	RE Acc.↑	SED↓	BLEU↑	Text Acc.↑	Precision↑	Recall↑	F1↑
RREG-S	35.89	12.54	81.71	34.78	56.34	53.00	51.73
RREG-L	37.22	12.37	82.06	36.07	67.11	54.31	52.08
ML-S	37.18	12.56	82.28	36.07	77.93	56.70	55.52
ML-L	**56.60**	**9.23**	**85.64**	**50.03**	**85.12**	**85.75**	**85.43**
ATT+COPY	29.27	15.19	79.01	32.55	76.33	54.16	51.10
ATT+META	35.56	12.11	81.07	36.83	72.72	70.50	71.42

Turning to the neural models, the data shows that including meta-information significantly enhances RE prediction accuracy. A comparison of ATT+META with ATT+COPY reveals that meta-information substantially improves the recall in pronominalization tasks.

To illustrate the outputs produced by these models, Table 7.8 presents examples from one of the WSJ documents, showcasing the generated outputs by RREG-S, ML-L, and ATT-META.

7.2.3.2 Human evaluation on WEBNLG

Materials From our test set of WEBNLG seen entities, we drew a random sample of four instances from each group of triples, with sizes ranging from 2 to 7. For the unseen entities, we randomly selected six instances from groups ranging in size from 2 to 5. This resulted in a total of 48 original instances, evenly divided into 24 seen and 24 unseen instances. In addition to the original instances, we included seven different versions generated by the models, comprising three neural, two feature-based, and two rule-based models. Consequently, we compiled a total of 384 items (48 instances multiplied by 8 versions each).

Design The 384 items were randomly allocated into 12 lists, each containing 32 items. Each list was rated by ten participants. The participants were instructed to evaluate each text on a 7-point Likert scale, where 1 represented *very bad* and 7 represented *very good*. The evaluation criteria included *fluency*, *grammaticality*, and *clarity*, defined according to Castro Ferreira et al. (2018a).

Fluency: Does the text flow in a natural, easy-to-read manner?

Grammaticality: Is the text grammatical (no spelling or grammatical errors)?

Clarity: Does the text clearly express the data in the table?

Participants To conduct the human evaluation, we used Amazon Mechanical Turk (MTurk). We restricted MTurk workers to U.S. residents who had an ap-

Table 7.8: Examples of an original text from the WSJ dataset together with the outputs generated by the models RREG-S, ML-L, and ATT-META. References to *MGM Grand Inc.* are in bold.

ORIGINAL: **MGM Grand Inc.** said **it** filed a registration statement with the Securities and Exchange Commission for a public offering of six million common shares. **The Beverly Hills , Calif.-based company** said **it** would have 26.9 million common shares outstanding after the offering. **The hotel and Gaming company** said Merrill Lynch Capital Markets will lead the underwriters. Proceeds from the sale will be used for remodeling and refurbishing projects , as well as for the planned MGM Grand hotel / casino and theme park.
RREG-S: **MGM Grand Inc.** said **MGM Grand Inc .** filed a registration statement with the Securities and Exchange Commission for the offering. **MGM Grand Inc.** said **MGM Grand Inc.** would have 26.9 million common shares outstanding after the offering. **MGM Grand Inc.** said Merrill Lynch Capital Markets will lead the underwriters . Proceeds from the offering will be used for remodeling and refurbishing projects , as well as for the planned MGM Grand hotel / casino and theme park .
ML-L: **MGM Grand Inc.** said **it** filed a registration statement with the Securities and Exchange Commission for an offering of common shares. **MGM Grand Inc.** said **it** would have 26.9 million common shares outstanding after the offering. **MGM Grand Inc.** said Merrill Lynch Capital Markets will lead the underwriters . Proceeds from the offering will be used for remodeling and refurbishing projects , as well as for the planned MGM Grand hotel / casino and theme park.
ATT-META: **MGM Grand Inc.** said **MGM Grand Inc.** filed a registration statement with the Securities and Exchange Commission for the offering of the company of the market. **MGM Grand Inc.** said **it** would have 26.9 million common shares outstanding after the offering. **MGM Grand Inc.** said Merrill Lynch Capital Markets will lead the underwriters. Proceeds from the offering will be used for remodeling and refurbishing projects , as well as for the planned MGM Grand hotel / casino and theme park .

proval rate of at least 95% and had completed over 1,000 HITs. We established specific criteria for rejecting a worker's contribution, which included: (1) assigning a score lower than two to human-produced (original) descriptions more than three times, and (2) providing scores with a standard deviation of less than 0.5.

A total of 120 workers participated in the study, covering 12 lists with 10 workers each. This resulted in 11,520 judgments (384 items evaluated across 3 criteria by 10 participants each). The demographic breakdown of the participants was as follows: 80 men, 36 women, and four who identified as other or did not disclose their gender. The average age of the participants was 37 years.

Results Table 7.9 presents the results of the human evaluation for the WEBNLG dataset. Notably, only a few differences reach significance.[5] This was determined by Wilcoxon's signed-rank test with Bonferroni correction[6]. This suggests that WEBNLG might not be the most suitable for distinguishing between different REG models.

The significant differences are observed when comparing RREG-S with ATT+META and PROFILEREG, particularly in terms of grammaticality for unseen data. These results position RREG-S as the top-performing model in generating REs on WEBNLG, showing equivalent performance to neural models on seen data and surpassing them on unseen data. Interestingly, in contrast to the results from our automatic evaluation, ATT+COPY slightly outperforms ATT+META in the human evaluation, as indicated by their rankings.

Table 7.9: Human evaluation results on WEBNLG. Ranking is determined by significance testing. Per column, results that have *no* superscript letters in common are significantly different from each other.

	All			Seen			Unseen		
Model	Fluency	Grammar	Clarity	Fluency	Grammar	Clarity	Fluency	Grammar	Clarity
RREG-S	5.76^{A}	**5.73**^{A}	**5.71**^{A}	5.73^{A}	5.62^{A}	5.68^{A}	5.79^{A}	**5.83**^{A}	**5.75**^{A}
RREG-L	5.68^{A}	5.52^{A}	5.67^{A}	5.63^{A}	5.45^{A}	5.64^{A}	5.74^{A}	5.60^{A,B}	5.70^{A}
ML-S	5.73^{A}	5.65^{A}	**5.71**^{A}	5.65^{A}	5.64^{A}	5.70^{A}	**5.82**^{A}	5.65^{A,B}	5.72^{A}
ML-L	**5.78**^{A}	5.63^{A}	5.67^{A}	5.73^{A}	5.63^{A}	5.62^{A}	**5.82**^{A}	5.62^{A,B}	5.72^{A}
ATT+COPY	5.65^{A}	5.62^{A}	5.68^{A}	5.71^{A}	5.64^{A}	5.76^{A}	5.58^{A}	5.60^{A,B}	5.59^{A}
ATT+META	5.68^{A}	5.56^{A}	5.66^{A}	5.69^{A}	**5.68**^{A}	5.65^{A}	5.67^{A}	5.43^{B}	5.66^{A}
PROFILEREG	5.70^{A}	5.56^{A}	5.61^{A}	**5.81**^{A}	**5.68**^{A}	**5.77**^{A}	5.58^{A}	5.43^{B}	5.44^{A}
HUMAN	5.81^{A}	5.69^{A}	5.82^{A}	5.77^{A}	5.69^{A}	5.83^{A}	5.84^{A}	5.70^{A}	5.80^{A}

[5]We define a difference as significant only if $p < 0.01$. Interestingly, all differences that are not significant in tables 7.9 and 7.10 have p-values greater than 0.1.

[6]This implies that p-values were adjusted by multiplying them with the number of comparisons.

7.2.3.3 Human evaluation on WSJ

7.2.3.3.1 Materials

From the WSJ test set, we randomly selected 30 documents (*reference texts*). Additionally, we gathered the six versions of each reference text generated by each WSJ model, referred to as *target texts*. This process resulted in 180 reference-target pairs.

7.2.3.3.2 Design

Given that the average length of WSJ documents is 25 sentences (as noted in §7.2.1.3) and there are no RDF input representations, we decided to use a *Magnitude Estimation* (ME) experiment (Bard et al. 1996) for our study design. This approach involved participants viewing both the reference and target texts side by side for comparison. To facilitate manageability, the texts were shortened to a maximum of 20 sentences each. These 180 reference-target pairs were then divided across 12 lists, with each list containing 15 items. In total, ten participants were assigned to rate each list, focusing on fluency, grammaticality, and clarity. For fluency and grammaticality, we adhered to the definitions established in the WEBNLG task (§7.2.3.2). The criterion for clarity was defined as follows:

Clarity: How clearly does the target text allow you to understand the situation described in the standard text?[7]

In the evaluation process, participants were presented with the following question for each of the three criteria: *Assuming that standard text has a score of 100, how do you rate the fluency|grammaticality|clarity of target text?* This format allowed participants to assign any positive number as their rating, providing a flexible and intuitive way to quantify their assessment of the texts.

7.2.3.3.3 Participants

In the experiment, a total of 120 individuals participated, comprising 65 men, 54 women, and one person who either responded as "other" or did not disclose their gender. The average age of the participants was 38 years. This resulted in a total of 5400 judgments, calculated from 180 items, each evaluated on 3 criteria by 10 different judges. For quality control, we rejected responses from crowdworkers whose scores were less than 5 standard deviations from the mean.

[7] In the experiment, we referred to the reference text as *standard text*.

7.2.3.3.4 Results

In the results analysis, we accounted for possible response errors in Magnitude Estimation, such as a worker entering 600 instead of 60. To manage this, outliers were excluded, defined as values falling below 3 standard deviations from the median or exceeding 3 standard deviations above the median for a given item. For significance testing, the scores were then down-sampled. The results, as illustrated in Table 7.10, show noticeable differences.

Table 7.10: Human evaluation results on WSJ.

Model	Fluency	Grammar	Clarity
RREG-S	$76.13^{A,B}$	$75.74^{A,B,C}$	78.03^{A}
RREG-L	$72.56^{A,B}$	73.38^{C}	74.76^{A}
ML-S	77.48^{A}	$78.39^{A,B}$	78.76^{A}
ML-L	77.52^{A}	78.45^{A}	79.45^{A}
ATT+COPY	$74.43^{A,B}$	$74.57^{A,B,C}$	75.63^{A}
ATT+META	71.94^{B}	$72.95^{B,C}$	73.95^{A}

Contrasting with the WEBNLG outcomes, significant differences were more frequently observed in the WSJ dataset. Specifically, for fluency, ML-S and ML-L demonstrated the highest performance, while ATT+META exhibited the worst performance. In terms of grammaticality, ML-L emerged as the top-performing model, significantly outperforming RREG-L and ATT+META. Interestingly, RREG-L was among the lower-scoring models for grammaticality, a finding that merits further investigation. Clarity did not show any significant differences, which might be attributed to the challenge participants faced in comparing longer documents. Overall, the ML-L model exhibited the best performance, closely followed by the simpler ML-S and RREG-S models.

7.2.3.4 Ethical considerations

We collected our human evaluations using Amazon Mechanical Turk. The payment for each task was set at $7.5/hour (slightly above the US minimum wage, i.e., $7.25/hour). We expected the amount to be a fair remuneration, but given the actual time some participants needed, their remuneration turned out to be on the low side. In future crowd-sourcing experiments, we will base our remuneration on a more generous estimate of the duration per experimental task.

During the study, we collected demographic data such as age, gender, and English proficiency level. We made it clear to participants that their information would be used solely for research purposes and assured them of anonymity. Moreover, these demographic fields were optional, not mandatory. It is important to note that this demographic data will be kept confidential and will not be publicly disclosed.

7.2.4 Summary and discussion of study F

7.2.4.1 Why neural REG does not defeat classic REG

Received wisdom suggests that while neural-based models may fall short in interpretability, they excel in performance. Recent neural models, such as Large Language Models, might perform better than those presented here, but our results challenge this received wisdom. One plausible reason is that neural NLG systems typically show superior performance in surface realization tasks but struggle with tasks demanding deep semantic understanding. As noted by Reiter (2018), these systems sometimes generate hallucinated content in Data2Text generation tasks. Given that REG heavily relies on semantic content, this could be a critical factor affecting the performance of neural networks in this area.

7.2.4.2 Role of linguistically-informed features

In our study, rule-based models demonstrated remarkable effectiveness, particularly with the WEBNLG dataset, outperforming other approaches. However, in the case of WSJ, the model equipped with linguistically informed features (ML-L) surpassed all others. This suggests that the nature and complexity of the text greatly influence the success of different REG methods. While simpler texts appear to be well-handled by rule-based models, linguistically informed features become increasingly vital for processing more complex texts. Therefore, the choice of an appropriate REG method should be guided by the complexity and type of the text at hand.

7.2.4.3 Resources use

The acceptable performance of RREG-S and ML-S on WSJ and WEBNLG becomes even more significant when considering resource efficiency. RREG-S stands out for its minimal demands on human resources, context, and computing power, and its

independence from training data.[8] Unlike RREG-S, ML-S requires training data and possibly more human effort for feature engineering and model training. However, it does not rely on external tools or meta-information. When building any model, various types of resources should be factored into the decision-making process:

1. *The amount of context*: The neural models in WEBNLG access complete pre- and post-contexts, while in WSJ, they are limited to two sentences around the target entity. The feature-based models extract features based on the current sentence and its entire preceding context, whereas the rule-based models consider only the current and the preceding sentence.

2. *External tools*: The neural models do not require external tools, while the rule-based and ML-L models require a syntactic parser.

3. *Meta-information*: Both rule-based models (RREG-S and RREG-L), the linguistically informed feature-based model (ML-L), and the neural model (ATT+META) rely on entity meta-information. Furthermore, PROFILEREG requires the profile description of each entity, which is difficult to obtain for most REG tasks.

4. *Computing resources*: Neural models typically need GPUs, while other models can be developed on standard personal computers.

5. *Training data volume*: Rule-based models operate without the need for training data, whereas feature-based and neural models demand substantial training datasets, which are not always easily available for REG tasks.

In conclusion, the ultimate choice between models is contingent upon a nuanced assessment of various resource-related considerations. This choice is crucial and should be made by weighing the specific characteristics and limitations of each approach, ensuring an optimal balance of efficiency and effectiveness.

7.2.4.4 Generalizability

While our study used neural REG to underscore the significance of incorporating non-neural baselines, it is important to recognize that our conclusions may not generalize to more complex E2E NLG systems. However, as indicated by Dušek et al. (2018), well-designed template-based NLG systems can still demonstrate

[8]The pronominal content (e.g., *he* or *she*) for an entity can be determined either from training data or using the entity's meta-information.

competitive performance. This suggests that our insights might be applicable to other components of the NLG pipeline, such as content determination, aggregation, and lexicalization. Notably, pipeline NLG systems, which handle these tasks sequentially, are occasionally deemed capable of outperforming comprehensive E2E NLG systems, as discussed by Castro Ferreira et al. (2019).

7.3 Study G: Neural REG and explainability

The outcomes of neural models, as demonstrated in study F and analogous research efforts (Castro Ferreira et al. 2018a, Cao & Cheung 2019, Cunha et al. 2020), leave open questions about the extent to which these models learn to encode linguistic features. Addressing this gap, study G introduces a sequence of probing tasks aimed at inspecting the internal structure of neural models.[9] This investigation seeks to discern which linguistic features are effectively encoded within neural models designed for the RFS task. Moreover, this study aims to explore whether models focused solely on pronominalization learn different contextual features compared to those trained for more fine-grained classification challenges. Additionally, we aim to understand how RFS performance is influenced by varying neural network architectures.

An established method to determine the nature of information encoded in the latent representations of neural models is through probing tasks. This approach has been widely adopted in various fields, such as machine translation (Belinkov et al. 2017), language modeling (Giulianelli et al. 2018), and relation extraction (Alt et al. 2020). Additionally, probing studies have been conducted in areas closely related to our study's focus, like bridging anaphora and coreference resolution (Sorodoc et al. 2020, Pandit & Hou 2021), which also seek to address the complexities of reference phenomena.

In a typical probing task, a diagnostic classifier is trained using the representations generated by the neural models. The classifier's performance is indicative of the extent to which these representations encode the information pertinent to the probing task. In this study, we aim to understand the linguistic features that are encoded by neural models in the context of choosing the appropriate RF.

As discussed in Chapters 2 and 5, the majority of research in the linguistic tradition on reference production primarily concentrates on the choice of the referring expression form, rather than on the detailed realization of its content. Consistently aligning with this tradition, study G focuses on the task of RFS. In

[9] Refer to Chen et al. (2023) for an updated version of study G, which includes probing experiments on WEBNLG and on WSJ, the latter being conducted in both English and Chinese.

this study, we use the REG model proposed by Castro Ferreira et al. (2018a) for addressing RFS. For a comprehensive analysis, we also incorporate: (1) a robust baseline model that uses a single encoder, in contrast to the multiple encoders used by Castro Ferreira et al. (2018a), and (2) pre-trained word embeddings (such as GloVe) and language models (like BERT), to gauge their impact on model performance.

We begin this section by outlining the task of RFS using the WEBNLG corpus and proceed to develop several neural models tailored for this task. To understand the linguistic features influencing the choice of RF, we introduce eight probing tasks. These tasks are informed by the prominence-lending cues discussed in Chapter 2 and the features examined in Chapter 5. Using these probing tasks, we aim to delve into the inner workings of our RFS models to interpret and explain their behavior.[10] As this study represents the first probing experiment focused on REG-in-context models, we conduct it using WEBNLG, the de facto standard corpus in NLG research. Future works will broaden the scope of this study to include more complex corpora, such as WSJ.

7.3.1 Neural referential form selection

This section is dedicated to introducing the task of RFS within the context of the WEBNLG dataset. Following this introduction, we will delve into the specifics of the *NeuralRFS* models.

7.3.1.1 The RFS task in study G (NeuralRFS)

In this study, the RFS task involves identifying the appropriate RF for a given context. This context comprises a pre-context $x^{(\text{pre})} = \{w_1, w_2, ..., w_{i-1}\}$ (where w denotes a word or a delexicalized entity label), the target referent $w^{(r)} = \{w_i\}$, and a post-context $w^{(\text{post})} = \{w_{i+1}, ..., w_n\}$. The algorithm's objective is to select the most suitable RF $\hat{f}$ from a set of K possible RFs $\mathcal{F} = \{f_k\}_{k=1}^{K}$.

In the WEBNLG dataset, RFs are categorized into four classes: `proper name`, `description`, `demonstrative`, and `pronoun`. However, due to the rarity of demonstrative NPs in the dataset, we also opt for a 3-way classification combining descriptions and demonstratives into a single category. Additionally, considering the focus of most linguistic studies on pronominalization, we also conduct a 2-way classification task. The classification categories used in our study are outlined in Table 7.11.

[10] The code for the RFS models and the probing classifier is accessible at: `github.com/a-quei/probe-neuralreg`.

Table 7.11: Three different types of RF classification.

Type	Classes
4-way	`Demonstrative`, `Description`, `Proper Name`, `Pronoun`
3-way	`Description`, `Proper Name`, `Pronoun`
2-way	`Non-pronominal`, `Pronominal`

The primary aim of study G is to uncover the specific linguistic features encoded by neural models in the RFS task. A key aspect of our investigation is to determine if neural models trained on less complex tasks (e.g., 2-way classification) capture different contextual features compared to those trained on more demanding tasks (e.g., 3-way and 4-way classification). Additionally, we are curious to explore whether models incorporating an attention mechanism encode a broader range of linguistic features in their latent representations compared to simpler neural architectures. As this is the first series of probing experiments for the NeuralRFS task, study G adopts an exploratory approach.

7.3.1.2 NeuralRFS models

In the context of our study, we have developed NeuralRFS models by employing two key strategies. First, we adopt the top-performing neural REG model from Castro Ferreira et al. (2018a). This step allows us to build upon a proven foundation, leveraging the strengths of an existing, effective model. Secondly, we introduce an alternative model that is inherently simpler in design and more adaptable in incorporating pre-trained representations.

con-att In our study, we use the CAtt model from Castro Ferreira et al. (2018a), identified as the most effective for the REG-in-context task among the models they evaluated. To process the inputs, we first use a *Bidirectional GRU* (*BiGRU*), as detailed by Cho et al. (2014), to encode both the pre-context $x^{(\text{pre})}$ and the post-context $x^{(\text{post})}$. Specifically, for each $k \in [pre, post]$, we transform $x^{(k)}$ into an encoded form $h^{(k)}$ using the BiGRU: $h^{(k)} = \text{BiGRU}(x^{(k)})$.

Differing from the approach of Castro Ferreira et al. (2018a), we then apply self-attention, following the method by Yang et al. (2016), to encode $h^{(k)}$ into a context representation $c^{(k)}$. This process involves computing attention weights $\alpha^{(k)}j$ at each step j. The attention weight is computed by first determining

$$(2) \quad e_j^{(k)} = v_a^{(k)T} \tanh(W_a^{(k)} h_j^{(k)})$$

and then normalizing these scores across all N steps in $h^{(k)}$ using

$$\alpha_j^{(k)} = \frac{\exp(e_j^{(k)})}{\sum_{n=1}^{N} \exp(e_n^{(k)})}, \tag{3}$$

$$e_j^{(k)} = v_a^{(k)T} \tanh(W_a^{(k)} h_j^{(k)}), \tag{4}$$

$$\alpha_j^{(k)} = \frac{\exp(e_j^{(k)})}{\sum_{n=1}^{N} \exp(e_n^{(k)})}, \tag{5}$$

where v_a is the attention vector and W_a is the weight in the attention layer.

In this model, the context representation for each context $x^{(k)}$ is obtained through a weighted sum of its encoded form $h^{(k)}$. Specifically, we calculate the context representation $c^{(k)}$ as follows:

$$c^{(k)} = \sum_{j=1}^{N} \alpha_j^{(k)} h^{(k)}. \tag{6}$$

After acquiring the context representations $c^{(\text{pre})}$ and $c^{(\text{post})}$ for the pre-context and post-context, we concatenate these with the embedding of the target entity $x^{(r)}$. This concatenated vector is then fed into a feed-forward neural network layer. The output of this layer, denoted as R, is obtained using the ReLU activation function:

$$R = \text{ReLU}(W_f[c^{(\text{pre})}, x^{(r)}, c^{(\text{post})}]), \tag{7}$$

Here, W_f represents the weights of the feed-forward layer. This process effectively combines the context information with the target entity information to produce a final representation, R. R, the input to the probing classifiers (discussed in §7.3.2), provides a rich representation encapsulating the context and entity information, which is then used to make informed predictions:

$$P(f|x^{(\text{pre})}, x^{(r)}, x^{(\text{post})}) = \text{Softmax}(W_c R), \tag{8}$$

In this equation, W_c represents the weights of the output layer, and the function Softmax is used to calculate the probability distribution over the possible referential forms (f) for the target entity $x^{(r)}$, given the pre-context $x^{(\text{pre})}$ and post-context $x^{(\text{post})}$. The Softmax function ensures that the predicted probabilities are normalized, meaning that they sum up to one. This makes it possible to interpret these probabilities in a meaningful way, with higher values indicating a greater likelihood of a particular referential form being the correct choice in the given context.

C-RNN In addition to the CON-ATT model, we explore a streamlined yet effective structure, termed *centered recurrent neural networks* (C-RNN). This model diverges from the CON-ATT approach by employing a single BiGRU for encoding the context and the target entity. The process involves concatenating the pre-context $x^{(\text{pre})}$, the target referent $x^{(r)}$, and the post-context $x^{(\text{post})}$ into a single sequence, which is then collectively encoded:

$$h = \text{BiGRU}([x^{(\text{pre})}, x^{(r)}, x^{(\text{post})}]). \quad (9)$$

Here, h represents the encoded sequence. For the target entity positioned at index i within this sequence, we specifically extract the corresponding i-th encoded representation h_i. This representation is then processed through a ReLU activation function in the feed-forward layer to obtain the final representation, denoted as $R = \text{ReLU}(W_f h_i)$, where W_f signifies the weights of the feed-forward layer.

Once R is obtained, the subsequent steps are similar to those of the CON-ATT model. The final representation R serves as the input for the same classification procedure, using the Softmax function to predict the probability distribution of the RFs.

Pre-training In this study, we also aim to explore the potential benefits of pre-trained word embeddings and language models in the RFS task, a domain where their efficacy remains largely unexplored. While Cao & Cheung (2019) have previously used pre-trained embeddings, the specific contributions of these embeddings have not been thoroughly examined through an ablation study. Consequently, we experiment with both the C-RNN and CON-ATT models, integrating them with GloVe embeddings (Pennington et al. 2014) to determine whether these pre-trained embeddings significantly influence the choice of RF.

Furthermore, we extend our exploration to the utilization of the pre-trained language model BERT (Devlin et al. 2019). To enhance BERT's effectiveness in encoding delexicalized entity labels, we first train it as a masked language model using the training data from WEBNLG. Subsequently, we freeze BERT's parameters and employ the model for input encoding. This encoded input is then fed into C-RNN.[11]

[11]Other approaches involving BERT were also tested, including the combination of BERT with a feed-forward layer for deriving h, as well as training scenarios where BERT's parameters were not frozen. However, these variations did not yield high-performance models.

Feature-based model In the construction of our feature-based RFS classifiers, we use the XGBoost algorithm (Chen & Guestrin 2016), employing a 5-fold cross-validation approach for training.[12] Initially, a comprehensive range of features, amounting to 16 in total, are extracted from the WEBNLG corpus to train the classifiers. After conducting a variable importance analysis, we select a subset of the most significant features for inclusion in the final models. The chosen features, which are deemed to have the greatest impact on the performance of the classifiers, are detailed in Table 7.12.

Table 7.12: Features used in the `XGBoost` models. The description of the features can be found in §7.3.2.1. The features `SenStat` and `DistAnt_W` were not implemented in the 2-way and 3-way classification models, respectively.

Feature	Definition
`Syn`	Description is provided in §7.3.2
`Entity`	Values: person, organization, location, number, other
`Gender`	Values: male/female/other
`DisStat`	Description is provided in §7.3.2
`SenStat`	Description is provided in §7.3.2
`DistAnt_S`	Description is provided in §7.3.2 (DistAnt)
`DistAnt_W`	Distance in number of words (5 quantiles)
`Sent_1`	Does RE appear in the first sentence?
`MetaPro`	Description is provided in §7.3.2
`GloPro`	Description is provided in §7.3.2

7.3.1.3 Evaluation

7.3.1.3.1 Implementation details

For the evaluation of our models, we fine-tune their hyper-parameters on the development set, selecting the configuration that yields the highest macro F1 score. In the case of the BERT model, to circumvent issues with tokenization, we use the cased version of *bert-base* and augment its vocabulary with all entity

[12] Although the CatBoost algorithm was also considered, as it was used in study F, the performance difference between CatBoost and XGBoost was marginal. We opted for XGBoost due to the ease of conducting SHAP analysis with this algorithm.

labels from our dataset.[13] This specific BERT model is trained on the WEBNLG corpus for 25 epochs, with a masking probability set at 0.15.

As for the XGBOOST models, the learning rate is set at 0.05, and the minimum split loss is calibrated at 0.01. We limit the maximum depth of a tree to 5 and set the subsampling ratio for the training instances at 0.5.

To assess the performance of each model, we focus on the macro-averaged precision, recall, and F1 score, calculated on the test set. We repeat each model's run five times to ensure consistency and report the averaged outcomes. It is important to note that the dataset used for this evaluation is composed exclusively of seen entities from the WEBNLG corpus.

7.3.1.3.2 Results

The results of the various classification tasks are presented in Table 7.13. Across the board, all neural model variants demonstrate superior performance compared to the feature-based baseline. In the binary classification task, the performance gap between the neural models and the feature-based model is relatively narrow. However, this gap is much larger in the more complex 3-way and 4-way classification tasks. It is noteworthy that the 2-way classification task is considerably less complex than its 3-way and 4-way counterparts, which is likely why the feature-based model achieves results nearly on par with the neural models in this scenario.

Table 7.13: Evaluation results of our RFS systems on WEBNLG. Best results are **boldfaced**, whereas the second best results are underlined.

	4-way			3-way			2-way		
Model	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
XGBOOST	53.77	51.98	51.55	71.27	69.24	68.34	86.64	82.76	84.57
C-RNN	68.79	62.95	64.96	84.49	82.52	**83.63**	90.31	88.01	89.09
+GLOVE	**69.10**	**63.90**	**65.40**	84.29	**82.55**	83.30	89.33	88.02	88.63
+BERT	62.63	61.80	62.15	83.02	81.44	82.15	**90.98**	88.00	**89.42**
CONATT	67.42	62.39	64.07	**85.04**	82.21	83.53	89.30	**89.19**	89.23
+GLOVE	65.98	62.49	63.67	83.62	81.41	82.45	89.60	88.06	88.80

When examining the performance of the neural variants, it becomes apparent that the more straightforward C-RNN model outperforms CON-ATT in the 4-way classification task, while achieving comparable results in both the 3-way and

[13] https://huggingface.co/bert-base-cased

2-way classifications. A plausible reason for this could be the approach CON-ATT adopts, which involves dividing the input into three distinct components – the target entity, the pre-context, and the post-context, encoding these elements separately, and then merging the encoded representations back together before making predictions. This *divide and merge* procedure might impede the model's ability to learn certain useful information.

In considering the impact of pre-trained models, we observe that the GloVe embeddings enhance the performance of C-RNN exclusively in the context of 4-way classification tasks. However, for 2-way and 3-way classifications, GloVe embeddings do not demonstrate any significant contribution. Interestingly, the integration of GloVe embeddings into the CON-ATT model seems to have a negative effect, as evidenced by a decrease in overall performance when GloVe is employed. Moreover, it is quite surprising to note that BERT, contrary to expectations, negatively impacts C-RNN in both 4-way and 3-way predictions. The F1 scores show a decline from 64.86 to 62.15 and from 83.63 to 82.15, respectively, when BERT is used, indicating that its inclusion might not always be beneficial for certain classification complexities.

Regarding pronominalization performance, the use of BERT demonstrates a marginal improvement, elevating the F1 score from 89.09 to 89.42. This increment, albeit slight, suggests some level of effectiveness of BERT in pronominalization tasks. However, this improvement is notably less pronounced compared to the substantial gains often observed in other NLP tasks when employing BERT. One plausible reason for this limited enhancement could be the nature of the WEBNLG dataset. Although BERT was retrained using delexicalized sentences from WEBNLG, it appears that the entity labels may introduce a form of noise that impedes optimal learning, thus restricting the full potential of BERT in this specific context.

The confusion matrices depicted in Figure 7.1 for XGBOOST and the best performing neural model, C-RNN+GLOVE, provide a deeper understanding of how each model behaves in the 4-way classification scenario. It is notable that both models perform well in identifying pronouns and proper names, which explains the minor performance discrepancies observed in the 2-way classification task. However, they struggle with predicting demonstratives, likely due to the scarcity of such references in the WEBNLG dataset.

A critical distinction between the two models emerges in their ability to differentiate between proper names and descriptions. The XGBOOST model frequently misclassifies descriptions as proper names, doing so in 62.58% of cases. In contrast, the C-RNN+GLOVE model shows a markedly improved distinction, with a misclassification rate of only 20.18%. This disparity suggests that neural models

like C-RNN+GLOVE may be capturing useful discourse-related features that are not readily captured through the feature engineering process employed in the XGBOOST model.

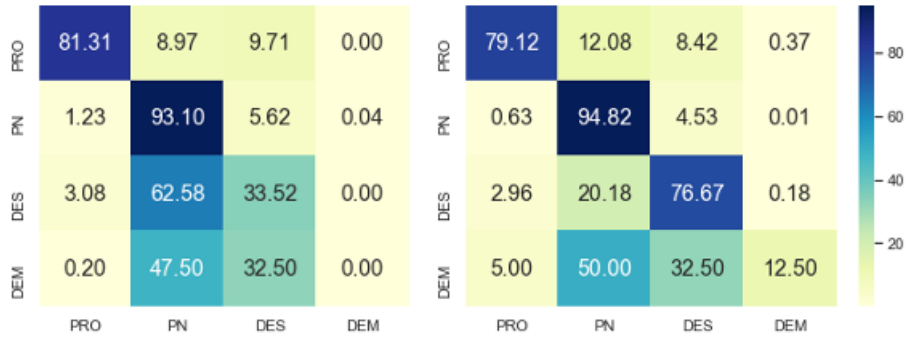

Figure 7.1: Confusion matrices for the 4-way classification results of XGBOOST (left) and C-RNN+GLOVE (right), where *PRO*, *PN*, *DES*, and *DEM* stand for `pronoun`, `proper name`, `description`, and `demonstrative`, respectively. The *y*-axis shows the true labels and the *x*-axis shows the predicted labels.

Furthermore, it is crucial to acknowledge certain annotation inconsistencies present in the WEBNLG dataset, which may impact the performance of both the XGBOOST and C-RNN+GLOVE models. An example is the inconsistent annotation of the entity "United States". While "United States" is annotated as a `proper name`, the phrase "the United States" is labeled as a `description`. Such irregularities in the dataset can lead to confusion in the classification of proper names and descriptions by both models. This factor must be taken into account when interpreting their performance and the differences in their ability to distinguish between these two referential forms.

7.3.2 Probing NeuralRFS models

In our approach, we use a logistic regression classifier for probing. This methodology involves generating the representation R using the models outlined in §7.3.1 for each input instance. As previously noted in §7.3.1, each model undergoes five runs, with their performance averages subsequently reported. When implementing the probing tasks, we specifically leverage the representations derived from those models which exhibited the most effective RFS performance during their evaluation on the development set.

7.3.2.1 Probing tasks

Drawing on the insights from Chapters 2 and 5, we have developed eight probing tasks pertaining to four classes of features, namely referential status (DisStat and SenStat), syntactic position (Syn), recency (DistAnt and IntRef), and discourse structure prominence (LocPro, GloPro, and MetaPro).

7.3.2.1.1 Referential status

Referential status is a key factor influencing the choice of referential form, as explored from both linguistic and computational perspectives (Chafe 1976, Gundel et al. 1993, Castro Ferreira et al. 2016b). In this study, we define referential status at two distinct levels: discourse-level and sentence-level.

DisStat. This feature has two possible values: (a) `discourse-old` indicates that the entity has previously appeared in the discourse, and (b) `discourse-new` denotes that the entity is new to the discourse.

SenStat. Similarly, this feature also encompasses two values: (a) `sentence-new` means that the RE is the first mention of the entity in the sentence, and (b) `sentence-old` denotes that the RE is not the first mention of the entity within the sentence.

7.3.2.1.2 Syntactic position

The syntactic role of entities in sentences significantly influences their likelihood of being pronominalized. Research by Brennan (1995) and Arnold (2010) suggests a higher tendency for subjects to be pronominalized compared to objects.

Syn. The syntax probing task focuses on this aspect through a binary classification: (a) `subject`: if the referent functions as a subject in the sentence, and (b) `other`: any non-subject referent in the sentence.

7.3.2.1.3 Recency

The proximity of a referent to its antecedent is a crucial factor in determining the referential form. We introduce two probing tasks focusing on recency:

DistAnt. This feature measures the distance between the target entity and its antecedent, which can contain four values: the entity and its antecedent are (a) in the `same` sentence, (b) `one` sentence apart, (c) `more than one`

sentence apart, and (d) the referent is the `first mention` of its entity in the discourse.

IntRef. This task assesses whether another referent intervenes between the target entity and its nearest antecedent, potentially indicating competitor referents. The values for this feature are: (a) the target entity is newly introduced in the discourse (`first-mention`), (b) the immediately preceding RE refers to the `same entity`, and (c) the immediately preceding RE refers to a `different entity`. The presence of intervening referents, especially those sharing animacy and gender with the target, can influence the referential form selection due to potential competition.

7.3.2.1.4 Discourse structure prominence

Discourse structure prominence is influenced by organizational elements within a discourse. Studies based on Centering Theory, such as those by Grosz et al. (1995) and Henschel et al. (2000), often account for pronominalization through the concept of local focus. *Local focus* considers both the current and preceding utterance. In contrast, *Global focus* identifies a referent within a broader context, encompassing either the entire text or a specific discourse segment (Hinterwimmer 2019). This approach links concepts like the familiarity and importance of a referent to the global prominence status of entities (Siddharthan et al. 2011). Our work introduces three probing tasks, each designed to capture distinct properties of discourse.

LocPro. The notion of local prominence originates from Centering Theory, which highlights the idea of local focus (Grosz et al. 1995). The implementation of local prominence follows Henschel et al. (2000): an entity is considered locally prominent if it is both *discourse-old* and *realized as a subject*. Local prominence is a binary feature, characterized by two distinct values: (a) `locally prominent`, and (b) `not locally prominent`. This probing task is a hybrid function of the two previously-mentioned probing tasks, DisStat and Syn.

GloPro. The concept of *global prominence* draws upon the idea of *global salience*, as explored by Siddharthan et al. (2011). This concept pertains to determining whether an entity is a major or minor referent within the text. Siddharthan et al. (2011) posits that "the frequency features are likely to give a good indication of the global salience of a referent in the document" (p. 820). In the GloPro probing task, the entity that appears most frequently

in a text is designated as globally prominent. This task allows for two outcomes: (a) `globally prominent`, and (b) `not globally prominent`.

MetaPro. In line with the concept of global prominence, we aim to explore how *prominence beyond a single text* – for instance, at the level of a collection of texts – affects RF. The underlying principle of this exploratory feature is that people tend to use fewer semantic details when referring to entities known outside the text. To implement this probing task, each RE is categorized based on the frequency of mentions of the target entity within the entire WEBNLG corpus, with four potential values representing different intervals: (a) `[1,50)`, (b) `[50,150)`, (c) `[150,290)`, and (d) `[290,∞)`. For instance, the category $[1, 50)$ includes entities mentioned fewer than 50 times in the corpus.

7.3.2.2 Importance analysis

To understand the relative importance of features used in the probing tasks for feature-based models, we conduct a feature importance analysis. This analysis acts as a validation step, ensuring that the learned representations significantly incorporate features that are crucial for the RFS task.

We train an XGBoost model, employing only the features discussed in §7.3.2.1. To evaluate the importance of each feature, we compute the model-agnostic, permutation-based variable importance for each model (Biecek & Burzykowski 2021). This method assesses how the model's performance is impacted when a specific feature is excluded. Essentially, it measures the degree of change in performance resulting from the removal of each feature. The results, illustrated in Figure 7.2, display the change in performance associated with each feature in the context of the 4-way classification task.

As depicted in Figure 7.2, the features DisStat (Discourse status) and Syn (Syntactic position) make the most substantial contributions to the predictions in the 4-way classification. An interesting observation is the lower importance of LocPro (Local Prominence), which can be attributed to its composition as a hybrid feature, combining aspects of both Syn and DisStat. Therefore, we also expect LocPro to exhibit a high performance in the probing experiment.

In addition to the 4-way classification results, Figure 7.3 presents the findings for the variable importance in the 2-way and 3-way classification tasks. Notably, there is a consistent pattern in the ranking of variables across all three classification models. This consistency indicates that certain linguistic features, particularly DisStat and Syn, consistently play a pivotal role, regardless of the complexity of the classification task.

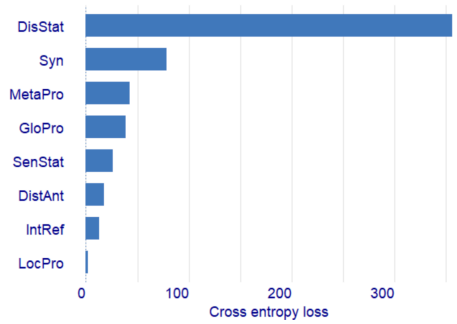

Figure 7.2: Feature importance of the XGBOOST classifiers for 4-way predictions. A higher loss indicates greater importance of a feature.

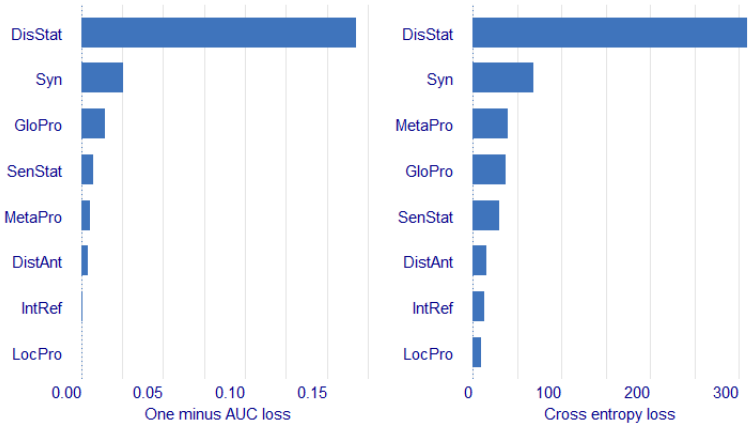

Figure 7.3: Feature Importance of the XGBOOST 2-way (left figure) and 3-way (right figure) predictions.

7.3.2.3 Probing results

We carry out a series of probing tasks in order to assess whether the latent representations of the NeuralRFS models effectively encode the linguistic features outlined in §7.3.2.1. These tasks aim to evaluate the extent to which these features are captured within the model's representations. We determine the performance of these tasks using accuracy and macro-averaged F1 scores, as depicted in Table 7.14. To ensure reliability, each classifier is trained five times and the average score of these iterations is reported. Moreover, we employ two baseline methods for comparison: RANDOM, which randomly assigns a label to each input, and MAJORITY, where the most frequently occurring label in each probing task is assigned to all inputs.

Table 7.14: Results of each probing task. Results are reported in the format of A(B), where A is accuracy and B is macro F1.

Model	Type	DisStat	SenStat	Syn	DistAnt	IntRef	LocPro	GloPro	MetaPro
RANDOM	-	49.57 (41.83)	33.11 (22.87)	49.65 (48.99)	25.19 (14.90)	33.30 (22.92)	50.05 (49.84)	49.75 (48.02)	25.24 (25.20)
MAJORITY	-	86.91 (46.50)	86.91 (31.00)	61.27 (37.99)	86.91 (23.25)	86.91 (31.00)	56.28 (36.01)	68.49 (40.65)	28.12 (10.97)
C-RNN	4-way	85.16 (84.06)	93.28 (73.72)	94.16 (85.34)	92.84 (53.84)	91.71 (55.43)	83.37 (82.92)	70.62 (56.00)	44.76 (42.32)
	3-way	84.78 (83.72)	92.59 (72.60)	93.50 (83.60)	92.58 (54.78)	91.24 (53.21)	82.17 (81.67)	70.87 (56.70)	45.42 (41.79)
	2-way	88.84 (88.04)	92.77 (73.84)	93.49 (84.00)	92.53 (54.93)	91.01 (52.31)	86.08 (85.69)	71.24 (59.98)	44.32 (41.65)
C-RNN +GLOVE	4-way	85.84 (84.85)	93.58 (74.59)	94.56 (87.04)	93.30 (55.67)	92.06 (55.93)	83.71 (83.20)	70.55 (53.53)	44.23 (41.71)
	3-way	85.09 (83.89)	91.89 (67.24)	93.23 (82.48)	91.72 (50.94)	90.92 (51.17)	82.08 (81.44)	70.20 (52.49)	45.58 (42.34)
	2-way	88.88 (88.02)	92.38 (71.25)	93.32 (82.67)	92.25 (53.67)	90.94 (51.43)	85.81 (85.22)	71.78 (63.17)	44.92 (41.03)
C-RNN +BERT	4-way	95.85 (90.64)	94.41 (78.04)	84.05 (82.71)	93.60 (56.91)	92.27 (54.30)	82.03 (81.67)	71.04 (54.24)	45.27 (43.07)
	3-way	94.00 (84.80)	92.74 (72.29)	85.12 (84.08)	92.57 (54.21)	91.28 (53.25)	82.92 (82.53)	71.69 (57.31)	43.64 (42.80)
	2-way	94.59 (87.28)	92.94 (69.69)	85.75 (84.74)	92.50 (54.19)	92.06 (54.88)	83.27 (82.77)	73.80 (63.07)	41.05 (40.75)
CONATT	4-way	94.86 (87.81)	94.12 (77.11)	88.64 (88.00)	93.69 (57.09)	92.11 (55.88)	86.93 (86.34)	72.22 (60.15)	48.37 (46.14)
	3-way	93.91 (84.39)	93.15 (74.19)	87.43 (86.66)	92.93 (55.26)	91.35 (54.09)	85.32 (84.56)	72.61 (60.61)	49.35 (47.47)
	2-way	93.74 (84.20)	92.78 (73.18)	89.01 (88.44)	92.50 (53.98)	91.19 (53.64)	87.05 (86.75)	70.65 (56.39)	44.24 (41.81)
CONATT +GLOVE	4-way	94.86 (87.82)	94.10 (77.70)	87.98 (87.24)	93.66 (57.52)	92.10 (55.22)	86.06 (85.69)	71.94 (58.54)	53.19 (49.94)
	3-way	93.79 (84.35)	92.78 (72.83)	89.54 (88.91)	92.59 (54.23)	91.39 (51.96)	87.09 (86.80)	71.91 (59.05)	49.27 (46.36)
	2-way	93.81 (84.38)	92.86 (73.21)	87.69 (86.96)	92.84 (56.14)	91.50 (53.33)	85.61 (85.27)	72.48 (62.46)	44.47 (39.63)

Results of each probing task The performance of all probing classifiers surpasses the random baseline across all tasks. This indicates a notable degree of learning and encoding of specific linguistic features by the models. In what follows, we detail the probing results of different tasks:

Referential status and syntactic position: The high performance on DisStat, SenStat, and Syn tasks across all models suggests that they effectively learn referential status and syntactic position from the WEBNLG corpus.

Recency (DistAnt and IntRef): Compared to referential status and syntactic position, the models exhibit a lower performance in recency-related tasks. The F1 scores for DistAnt and IntRef are not only lower than those for DisStat, SenStat, and Syn but also closer to the baseline values. This aligns with the findings from §7.3.2.2 where DistAnt and IntRef were deemed less important. The lesser importance of recency in WEBNLG could be attributed to a large portion of the documents being single-sentence, reducing the impact of long-distance dependencies. Alternatively, consistent with previous probing studies on coreference and bridging anaphora (Sorodoc et al. 2020, Pandit & Hou 2021), the models might inherently struggle with capturing long-distance dependencies.

Discourse structure prominence: The models handle LocPro effectively, likely due to its derivation from DisStat and Syn. However, they perform poorly on GloPro and MetaPro, as indicated by the low F1 scores in these tasks. This is contrary to the importance analysis results, which highlighted the relative importance of both GloPro and MetaPro (ranked three and four in Figure 7.2). This discrepancy suggests that neural models may lack a comprehensive understanding of the broader document or corpus context, a crucial aspect for accurately handling GloPro and MetaPro tasks.

Comparing C-RNN and CON-ATT In §7.3.1, we established that the C-RNN model outperformed CON-ATT in the 4-way RF classification task. However, it is intriguing to note that CON-ATT demonstrates superior performance in several probing tasks, particularly in DisStat, LocPro, GloPro, and MetaPro. This enhanced performance may stem from CON-ATT's use of self-attention, a feature that potentially improve the model's ability to grasp long-distance dependencies within text.

It is crucial to note that the WEBNLG corpus, upon which these models were evaluated, has distinct characteristics that might not align perfectly with more general or diverse text types. For instance, a notable majority (over 85%) of REs in

WEBNLG are categorized as first-mentions, and a single entity (*United_States*) appears in 21% of its documents. These specifics point to a usage of REs in WEBNLG that could substantially differ from more varied and realistic text sources.

Given these corpus-specific peculiarities, it is possible that the true capabilities of the more complex CON-ATT model, particularly its proficiency in encoding a broader range of contextual features, are not entirely used within the limits of the RFS tasks conducted with the WEBNLG corpus. This suggests a need for further exploration and testing of these models in more varied and representative text environments to fully leverage their respective strengths.

The effect of pre-training As mentioned earlier, we are interested to know whether RFS benefits from different architectures such as pre-trained word embeddings and language models. Notably, the inclusion of GloVe embeddings does not yield a significant impact on either C-RNN or CON-ATT.

In contrast, the incorporation of BERT shows a more notable effect, particularly in improving the models' ability in learning about DisStat. This improvement is likely attributable to BERT's self-attention mechanism, which enhances its ability to discern and encode the discourse-old or discourse-new status of entities. However, it is important to consider the composition of the WEBNLG corpus, where a large proportion of REs are first mentions. This characteristic of the corpus means that while accuracy in the DisStat task is improved with BERT, this does not translate into a corresponding substantial increase in the overall RFS performance. The specificity of WEBNLG's content thus limits the extent to which the benefits of BERT can be observed and leveraged in the broader context of RFS.

Comparing different RF classifications The probing results intriguingly suggest that the type and degree of information encoded in neural models' latent representations are influenced by the nature of the RF classification tasks they are trained on. Notably, the simpler 2-way classification task seems to enhance the ability of C-RNN to effectively learn about referential status.

Conversely, in the context of more complex 4-way classification tasks, models equipped with attention mechanisms, namely CON-ATT+GLOVE, CON-ATT, and C-RNN+BERT, exhibit superior proficiency in encoding referential status. This enhanced ability can be attributed to the attention mechanism's capacity to focus on and integrate relevant contextual information more effectively. Moreover, in the case of CON-ATT+GLOVE, we see that more fine-grained classifications help the model learn more about meta-prominence (MetaPro).

7.3.3 Summary and discussion of study G

Study G was dedicated to exploring the capacity of neural models to encode features pertinent to referential form selection (RFS). This investigation was structured around eight probing tasks, each targeting critical linguistic aspects such as referential status, syntactic position, recency, and discourse structure prominence.

Firstly, the performance of the probing classifiers consistently surpassed that of the RANDOM and MAJORITY baselines across all tasks. This outcome is indicative of the models' ability to learn and represent these linguistic features beyond mere chance or frequency-based guessing.

Notably, the probing tasks associated with referential status, syntactic position, and local prominence yielded particularly strong performances. This suggests that neural models are especially proficient in understanding and encoding information related to these aspects. However, the models demonstrated weaker performance in tasks assessing recency-related features. This could be attributed to the inherent complexity of encoding long-distance dependencies and contextual relations in the neural architecture. Additionally, tasks probing global prominence and meta-prominence proved challenging for the models, indicating a potential area for further development in capturing more abstract, discourse-level features.

While probing tasks in study G have provided valuable insights into the abilities of neural models in RFS, there are inherent limitations to this approach that warrant a cautious interpretation of the results.

One of the key criticisms of probing is that low performance does not necessarily indicate a lack of encoding of a feature; rather, it could imply that the feature is not pivotal to RFS. This aspect was somewhat addressed by a complementary variable importance analysis, where discourse status and syntactic position were identified as important. These features showed good predictability in probing classifiers. However, it is important to note that this importance analysis was based on a feature-based model, not the neural models in question. Consequently, the same level of importance may not hold across different model types.

Moreover, the overall validity of probing as a method has been debated in various studies, such as those by Hewitt & Liang (2019) and Kunz & Kuhlmann (2020). These critiques argue that it is challenging to discern whether probing classifiers are simply learning the probing task at hand or genuinely extracting encoded linguistic information. This ambiguity means that a probing classifier's high performance does not unequivocally indicate that a model has effectively encoded

the linguistic information being tested. Consequently, we cannot directly quantify *how well* the linguistic information was learned based on the performance of probing classifiers. Therefore, we must be careful in how we interpret the levels of linguistic understanding based solely on probing results. In essence, although probing offers valuable insights, its findings should be viewed as one part of a broader analysis, rather than definitive proof of a model's capabilities.

Based on the results of our probing experiments, we draw several conclusions:

1. All neural models have indeed learned some information about the features pertinent to the probing tasks. However, the extent to which this information has been effectively learned remains uncertain.

2. The widespread usage of the WEBNLG corpus in studying discourse REG is problematic for analyzing discourse-related aspects of RFS. The texts in this corpus are predominantly short, and the majority of REs are first mentions.

3. The choice of neural architecture and the specificity of the label set employed are crucial in determining how well a model can learn a particular feature. The implementation of an attention mechanism and a more granular label set appears to aid models in learning more information.

4. Models consistently struggle with features that derive from a broader contextual understanding, such as GloPro (global prominence) and MetaPro (meta prominence). This suggests a gap in the models' ability to interpret and use wider contextual cues, indicating an area for future improvement and research in neural model development.

7.4 Discussion and final remarks

In studies F and G, we examined neural models, focusing on performance in the former and on interpretability in the latter.

In study F, our approach involved reevaluating the latest neural REG systems using a combination of well-designed rule-based and feature-based models. This analysis led us to recognize that traditional REG models hold their ground quite effectively against neural models in terms of performance.

Recognizing the lack of interpretability inherent in neural models, study G introduced several probing tasks as a method to inspect the latent representations of these models. The objective of this study was to determine whether neural

REG-in-context models could learn some of the linguistic features associated with the choice of referential form. The following sections outline the lessons learned regarding various facets of the REG-in-context task, along with suggestions for future research.

7.4.1 The choice of corpora

A need for corpus diversification. The discussions in this chapter highlight the crucial role of selecting an appropriate corpus for each task. Study F demonstrated that model performance assessments vary significantly based on the training corpora. To make well-rounded generalizations about various approaches, it is essential to diversify the corpora used in NLG tasks. In pursuit of this, we created a new dataset for the REG-in-context task using the WSJ corpus, proposing that it might be more suitable for this particular task. Nevertheless, this dataset requires further refinement, including enhancements in delexicalization.

Inclusion of more languages. The REG-in-context task exhibits some degree of language specificity, as languages differ in their referring expression form inventories. Yet, if the impact of features hinges on their prominence in context, it is plausible that the core underlying referential mechanisms are consistent across languages. To develop a broader understanding of the generation of REs in various contexts, it is vital to extend these studies to include languages beyond English. Specifically, languages that predominantly use zero pronouns, like Chinese, pose unique challenges for the REG-in-context task and thus are of particular interest. Extending our research to include such languages will provide deeper insights into the nuances of REG in different linguistic contexts.

7.4.2 The choice of REG-in-context approaches

7.4.2.1 The importance of classic REG models

Our findings in study F underscore the relevance of traditional REG-in-context methods. These approaches are still widely practiced, particularly in commercial applications (Reiter 2017). Study F revealed that carefully crafted classic models are capable of competing with their neural counterparts in terms of efficacy and reliability.

7.4.2.2 Pre-trained language models

In study G, we took the initiative to integrate pre-trained language models into our NeuralRFS frameworks. However, it is premature to draw definitive conclu-

sions about their effectiveness. This caution stems from two primary factors: First, these models were tailored specifically for the RFS task, which is notably less complex than the end-to-end REG-in-context challenge; second, their training was conducted using the WEBNLG corpus. As indicated in our research, this corpus may not be the most appropriate for comprehensively evaluating the capabilities of these models in the REG-in-context domain.

7.4.3 The choice of features

7.4.3.1 Varied efficacy of features

In study G, it was observed that probing classifiers did not perform optimally on certain tasks. This could be attributed to the models' challenges in effectively encoding the associated features. It raises the question of whether the difficulty lies in the feature itself or if a different model architecture might yield better results. Another consideration is the relevance of the feature to the task or to the specific corpus. For instance, the feature *distance in number of sentences*, discussed in study C, was significant in feature-based models applied to the WSJ corpus, but less so in models trained on the GREC-2.0. This highlights the necessity of evaluating the feature's relevance to the corpus in question.

7.4.3.2 Significance of feature-based models

The superior performance of the linguistically informed feature-based model on the WSJ corpus underlines the critical role of features in the REG-in-context task. An area for further exploration is understanding why certain features are so influential. It might be that some linguistic features represent bottlenecks for neural models, limiting their effectiveness.

7.4.3.3 Bottleneck features in REG-in-context

Features can be seen as tools for encoding characteristics not immediately discernible from text. One such feature is recency, which considers longer text spans. Study G highlighted that neural models struggle with capturing long-distance dependencies. Another intriguing feature is transitions, as suggested by the findings in studies C and E. These studies proposed that transitions, along with distance, significantly impact RF choice. Recognizing and understanding these bottleneck features is essential, as they are pivotal to the task but may not be adequately addressed by current end-to-end models. Enhanced comprehension of these features could lead to the development of more effective neural models.

7.4.4 Better human evaluation methods

The use of the Likert scale in evaluating WEBNLG predictions seems to align well with the requirements of this task. However, our perspective on the Magnitude Estimation approach used for the WSJ data is more cautious. The human evaluation results did not reveal any significant distinctions in clarity among the models. This outcome suggests that either the evaluation task was not adequately defined or the experimental design was not suitable for discerning differences. It is noteworthy that most human evaluation experiments in this domain have been limited to shorter text segments. Even those performed on the GREC systems (Belz et al. 2010) involved narratives that were simpler and shorter than those in this study. Consequently, this signifies the need for further research and additional experiments to develop and refine methods for effectively evaluating longer texts in future studies.

8 Conclusion

8.1 Introduction

As highlighted in Chapter 3, the concept of REG involves two distinct tasks: one-shot REG and REG-in-context. The one-shot REG task focuses on identifying a set of attributes to distinguish a referent from a group of distractors. It aims to single out a referent in a single shot, without considering the surrounding linguistic context (Krahmer & van Deemter 2012). In contrast, REG-in-context extends beyond this by incorporating the linguistic context of an expression. It entails producing suitable referring expressions (REs) to denote a referent at different stages in a discourse (Belz & Varges 2007).

This book has delved into various studies related to the REG-in-context task. In Chapters 2 and 3, we explored both linguistic and computational perspectives on the selection of REs in context. Chapter 4 focused on comparing three corpora for predicting the form of REs in context, while also addressing the limitations of accuracy-based metrics for evaluating model performance. The exploration continued in Chapters 5 and 6, which focused on the linguistic features critical to feature-based REG-in-context studies. Chapter 5 revisited features from previous studies, while Chapter 6 shed light on paragraph-related concepts, a relatively less explored area in computational research. Finally, Chapter 7 provided a comprehensive evaluation of various approaches to the task. This evaluation included rule-based, feature-based, and E2E neural network models, supplemented by a series of experiments aimed at interpreting the outputs of neural RFS models.

This chapter is organized as follows. In §8.2, a critical overview of the various aspects of the REG-in-context task, as discussed in previous chapters, is provided. This section not only delineates the strengths and weaknesses of the conducted studies but also proposes directions for future research. The studies presented in this book have been exclusively focused on the REG-in-context task. To provide a broader perspective, in section §8.3, a concise comparison between REG-in-context and the other principal task, one-shot REG, is drawn. This comparison intends to shed light on their similarities and differences to offer a comprehensive understanding of the field.

While the primary focus of this work has been on computational solutions to enhance the performance of REG-in-context models, it also recognizes the value of integrating insights from linguistics. In §8.4, the discussion revolves around how insights from the linguistic approach can augment the computational generation of REs, and conversely, how computational methods can inform linguistic understanding.

Finally, §8.5 presents a summary of the preceding chapters, highlighting the main findings that emerged from this research.

8.2 Different aspects of the REG-in-context task

Addressing the REG-in-context task involves making a series of critical decisions, such as the choice of corpus, approach, features, and evaluation methods. Each of these choices introduces its own complexities and limitations. In the following subsections, I will discuss some of the most significant choices and their implications. The organization of this section is as follows:

In §8.2.1, the focus lies on the selection of corpora and their inherent limitations. This discussion aims to shed light on how the choice of corpus can impact the outcomes and interpretations in REG-in-context studies.

§8.2.2 explores the features employed in feature-based REG-in-context models, as well as various methods for assessing these features. This subsection will delve into the nuances of feature selection and evaluation, highlighting their critical role in the effectiveness of the models.

The methodologies adopted in this work are discussed in §8.2.3. This includes an examination of the three primary REG-in-context approaches: feature-based, rule-based, and end-to-end neural network-based methods. The unique strengths and challenges of each approach will be outlined, providing a comprehensive overview of the techniques employed.

Closely linked to the choice of approach is the issue of interpretability, which is discussed in §8.2.4. This subsection will explore the balance between model performance and the ability to interpret the model's decision-making processes, a critical consideration in the development of these models.

Finally, in §8.2.5, the limitations of the evaluation methods used in this research are discussed. This will involve a critical assessment of the metrics and procedures employed in evaluating the effectiveness of the REG-in-context models.

8.2.1 Limitations of the corpora that were used

The selection of a corpus for linguistic research can significantly influence the derived outcomes and insights. In this book, I used four different corpora to investigate various aspects of the REG-in-context task. The appropriateness and implications of these corpus choices were primarily addressed in Chapter 4. However, the relevance of this choice resurfaced in Chapters 5 and 7, particularly during discussions on the selection of features and approaches for the REG-in-context task. This recurrence underscores the pivotal role the choice of corpus plays in shaping the research's trajectory and its outcomes. Table 8.1 provides a comprehensive overview of the corpora used, alongside a breakdown of each corpus' contribution to the respective chapters.

Table 8.1: Different corpora used in this book.

	GREC-2.0	GREC-PEOPLE	WSJ	WEBNLG
Genre	Wikipedia	Wikipedia	Newspaper articles	crowdsourced texts from RDFs
Task	RFS	RFS	RFS & RCS	RFS, RCS, NeuralRFS
Chapter	4 & 5	4	4–7	7
Num of doc	1655	808	585	25373 (triples)
Num of RE	11,705	8378	30,471	94,515
Length (sent)	7.2	5.8	25	1.4

Appropriateness of corpora for the task Whenever a task involves using corpora, the primary question should be the suitability of the selected corpus for the intended task. In Chapter 4, the suitability of the chosen corpora for the REG-in-context task was discussed. For instance, the GREC-PEOPLE corpus, as highlighted, was found to be unsuitable for a 3-way classification task (involving `pronoun`, `description`, and `name`). Its utility was limited primarily to addressing the pronominalization task. This limitation stems from a significant imbalance in label distribution within the GREC-PEOPLE corpus, where the `description` class represents a mere 4% of the data. Such imbalances can skew classification algorithms towards the majority class, undermining the learning of minority classes.

This situation raises an important question: Is the scarcity of the `description` class a general phenomenon, or is it specific to this dataset? The prevalence of this class as the majority in the WSJ dataset suggests the latter. Furthermore, some classes might be inherently less common in specific linguistic contexts. For example, empty REs are a relatively rare class in English, constituting only 6.28% and 6.47% of REs in the GREC-2.0 and GREC-PEOPLE corpora, respectively.

To address such imbalances, researchers can employ resampling techniques to alter the distribution of classes in the dataset and force the algorithm to learn the less-represented classes (Weiss & Provost 2003, Mountassir et al. 2012, Branco et al. 2016, Padurariu & Breaban 2019). Common resampling methods include random *undersampling*, which involves removing instances from the majority class, and *oversampling*, which adds copies of the minority class. Each method has its drawbacks: Undersampling risks losing valuable information for the majority class, while oversampling increases the likelihood of overfitting. More sophisticated resampling strategies are reviewed in Branco et al. (2016), offering insights into how to balance datasets more effectively without compromising the integrity of the data.

Corpus-dependent choice of features in feature-based models In the realm of feature-based REG-in-context models, the selection of features is intrinsically linked to the characteristics of the corpora used. This dependency was evident in study C of Chapter 5. Our findings revealed that while the distance in the number of sentences is a significant factor in models trained on the WSJ corpus, it holds less importance in models based on the GREC-2.0 dataset. This observation prompts an important question: How crucial is sentential distance in feature-based models? The answer is not straightforward but is contingent on the specific corpus used for training the models.

This scenario underscores a central lesson for researchers: When developing feature-based models, or when defining probing tasks based on certain features, it is imperative to ensure that these features are relevant and meaningful within the context of the chosen corpora. This relevance is crucial not only for the accuracy and effectiveness of the models but also for the validity of the research findings. A feature that is influential in one corpus may not necessarily hold the same level of significance in another, highlighting the need for a careful, corpus-specific approach to feature selection.

Representativeness of WSJ for the REG-in-context task In chapters 4 and 7, we discussed the limitations of the GREC (comprising GREC-2.0 and GREC-PEOPLE) and WEBNLG corpora, highlighting their specific constraints for the REG-in-context task. For instance, GREC-PEOPLE, focusing solely on human referents, is unsuitable for a 3-way classification involving descriptions, which are predominantly used for non-human entities. GREC-2.0, despite its broader range of documented entities (such as cities, countries, and mountains), is limited by its annotation of only the main topic of each document, thereby failing to capture the

dynamics of competition mechanisms and the presence of other entities. The WEBNLG data, examined in Chapter 7, comprises very short documents with simple syntactic structures, averaging 1.4 sentences per document. This brevity restricts our understanding of REs in longer, more complex texts.

In contrast, the WSJ corpus, used throughout this study, overcomes many of these limitations. It offers a balanced representation of referential classes, with annotations of multiple referents per document, and features an average document length of 25 sentences. This corpus, therefore, provides a more comprehensive landscape for studying REs. However, it is not without its own limitations. For instance, null cases are not annotated in the WSJ corpus. Additionally, the syntactic complexity of the WSJ documents, primarily comprising Wall Street Journal articles with a focus on finance, may surpass the complexity found in other text sources. This higher complexity is exemplified in (1), which features an introductory sentence with various nested clauses from a WSJ document. While this example alone does not quantify the overall complexity of the WSJ corpus, it serves as an indicator of the types of sentences prevalent in this corpus.

(1) The U.S., claiming some success in its trade diplomacy, removed South Korea, Taiwan and Saudi Arabia from a list of countries it is closely watching for allegedly failing to honor U.S. patents, copyrights and other intellectual-property rights.

These insights lead to the pivotal question: What kind of language use does the WSJ corpus represent, and what can we learn from it? The corpus's lack of the previously mentioned limitations enriches our study by providing a more robust and diverse dataset. However, the unique features of the WSJ corpus, such as its syntactic complexity and financial focus, may also influence the generalizability of our findings.

Another notable aspect regarding the structure of the WSJ documents is their distinct paragraph format, which differs significantly from other written genres like essays and commentaries. As highlighted in Chapter 6, the average number of sentences per paragraph in the WSJ is 2.13, with approximately 25% of paragraphs consisting of only a single sentence. This brevity is atypical when compared to the average paragraph length in other written sources.

This structural difference has important implications for the generalizability of models trained on the WSJ corpus. In their study, Hendrickx & Hoste (2009) found that the performance of their coreference resolver declined markedly when applied to less structured data sources, such as blog texts and reader comments, after being trained on structured news texts. Their findings align with my belief

that to enhance the generalizability of results in corpus-based studies, a more diverse dataset is essential. The reliance on a single type of text structure, as is the case with the WSJ corpus, can limit the applicability of models to other genres and formats of text.

8.2.2 The importance of linguistic features for REG-in-context

The crucial role of linguistically informed features in REG-in-context models has been a recurring theme in this book, particularly emphasized in Chapters 5, 6, and in study F of Chapter 7. This section is dedicated to discussing three key aspects: (1) the significance of thoroughly understanding linguistic features, (2) the formation of an effective set of features in feature-based models, and (3) the various methods of evaluating linguistic features.

The importance of a thorough understanding of linguistic features The selection of features is a pivotal step in constructing feature-based REG-in-context models. Appropriate feature inclusion can markedly enhance model performance. A deep understanding of the features that influence the choice of REs is equally vital for crafting rules in rule-based models. Although feature selection is not directly applicable to E2E models, a comprehensive grasp of linguistic features is essential for designing probing experiments to inspect these models.

Features used in feature-based REG-in-context models In study B of Chapter 5, we undertook a comprehensive review, gathering 65 features from prior feature-based REG-in-context studies and categorizing them into nine distinct groups. These groups were grammatical role, inherent features, referential status, recency, competition, antecedent form, surrounding patterns, position, and protagonism. According to the findings in Chapter 2, the characterization of linguistic features can be based on their inherent or derived accessibility. The consensus set of features in study B contains both categories, including crucial aspects like the grammatical role of the RE, the antecedent's form, animacy, plurality, sentence distance, and paragraph distance. Features such as animacy reflect the inherent accessibility of a referent, while concepts like recency are derived.

Various ways of assessing linguistic information In feature-engineered models, linguistic features are more transparent and interpretable, as demonstrated in studies B, C, and E. These studies involved integrating various features into

models and observing their impact on predictive capabilities, representing a direct method of assessment. The goal is to determine whether incorporating a specific feature enhances model performance.

E2E models, such as those employed in study F, present a more opaque scenario, where the internal workings are less immediately apparent. It is theorized that these models' effectiveness stems from their capacity to encode a continuous analogue of linguistic structures (Torroba Hennigen et al. 2020). To decode these encoded features, study G employed probing experiments. These experiments involve training neural network-based models on the RFS task to produce representations, which are then used in probing classifiers. These classifiers, designed around features thought to influence RF choice (like recency and grammatical role), test if the model has learned information pertinent to the classifiers' tasks. Good performance by these classifiers suggests successful encoding of the relevant features, offering insights into the neural models' latent representations. This method, considered an extrinsic evaluation (Torroba Hennigen et al. 2020), allows us to view a neural model's representations through the lens of linguistic features.

8.2.3 The approaches used for tackling REG-in-context

In this book, three distinct approaches were employed to study the REG-in-context task: rule-based models, feature-based machine learning models, and end-to-end neural models. Each approach presents unique characteristics and methodologies.

Rule-based and feature-based models follow a modular pipeline architecture. In these models, two sequential subtasks are performed: Referential Form Selection (RFS) and Referential Content Selection (RCS). In contrast, E2E architectures simplify this process by generating referring expressions in a single integrated step. This approach contrasts markedly with the segmented process of modular architectures. Study F, in agreement with Rudin (2019) and Castro Ferreira et al. (2019), suggested that more complex architectures do not always equate to superior performance.

In the following sections, I discuss the specifics of RFS and RCS, the two integral components of the modular approach. Subsequently, I will examine each of the three methodologies used in this book in greater detail, highlighting their respective advantages and shortcomings.

RFS and RCS: Two subtasks of the modular architecture RFS is often considered the initial module in a modular REG architecture. However, it can also

be viewed as a testing ground for linguistic assumptions. This is because most linguistic theories concentrate on the choice of RF, as opposed to the specific content of an RE. The reason for this is that, while the classifications of RF, such as the Accessibility Hierarchy by Ariel (2001), can be categorized into a finite list of classes, the actual content of an RE has infinite possibilities. In this book, alongside classic rule-based and feature-based models, neuralRFS models were used in study G for class prediction. Interestingly, for the WEBNLG corpus, these neural classifiers outperformed feature-based classifiers across all classification tasks. In contrast, with a corpus like WSJ, the performance gap between neural and feature-based models may not be as pronounced. Should neural classifiers surpass feature-based models in effectiveness, a potential alternative architecture could involve combining a neural form selector with a data-driven content selector in a modular REG model.

RCS is the subsequent module in the modular architecture. In this book, the rule-based models described in Chapter 7 adhere to a strict realization protocol. If the predicted form is not pronominal, the model refers to the entity using a delexicalized phrase, converting underscores in the phrase to whitespaces. Feature-based models take a different approach: They determine the most frequent RE for each combination of features. If no corresponding content is available, they resort to the backoff method detailed in Chapter 7. While this strategy allows for a broader range of realizations, it remains deterministic, relying on the frequency of REs in the training data. For example, when referring to *Arria NLG*, the most common reference in the dataset, *Arria*, is selected. However, the more natural approach might be to use the full company name on its first mention. To address such nuances, a hybrid approach combining rules and data-driven methods could potentially yield more natural results. Nevertheless, this integration poses the risk of overly complicating the content realization process.

Rule-based models A key advantage of rule-based models lies in their transparency. These models operate with a clear set of predefined rules, making their decision-making process fully comprehensible and traceable. This transparency is particularly valuable when the aim is to understand the underlying logic of the model's decisions. However, designing these systems can be challenging, especially when attempting to incorporate a wide array of conditions and their interplay. As demonstrated in study F, a straightforward but thoughtfully designed rule-based system like RREG-S performs well with simpler corpora such as WEBNLG. Yet, when dealing with more intricate data sets, this simplicity may not suffice. In these cases, the development of more complex rules and the careful consideration of their interactions become imperative.

Feature-based machine learning models These models distinguish themselves by learning rules automatically from input data, as opposed to relying on manually encoded knowledge (Bishop 2006). These models exhibit transparency at two distinct levels: (1) model or algorithmic transparency (Barredo Arrieta et al. 2020), and (2) transparency in the choice of features.

Model or algorithmic transparency is about the transparency of a model's decision-making process. For example, Figure 8.1 illustrates a decision node from the application of the C5.0 decision tree algorithm to the OSU dataset in study B.

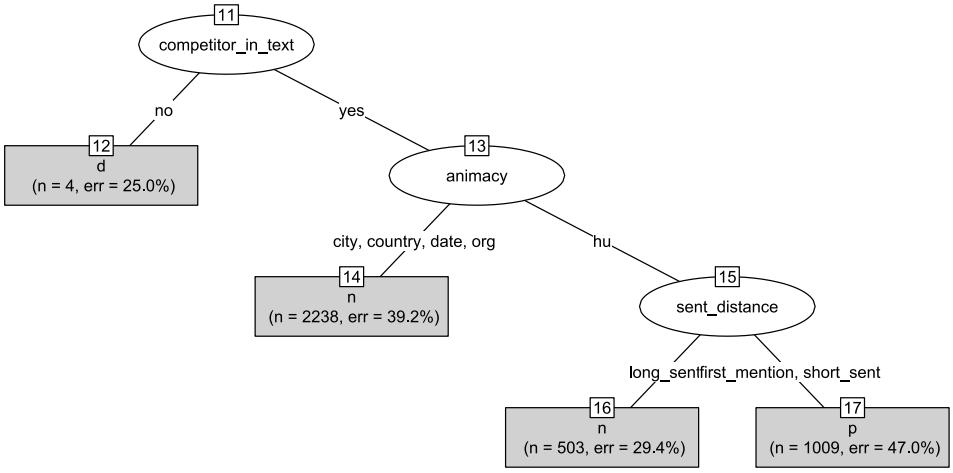

Figure 8.1: One of the branches of the OSU decision tree for a 3-way distinction. In this image, `n` is `proper name`, `d` is `description`, and `p` is `pronoun`.

This figure exemplifies a clear decision-making process: Each node in the decision tree splits the data based on certain criteria, leading to a prediction. The decision node number 13, for instance, divides into a termination leaf node (node 14) and a further decision node (node 15). Each termination leaf node provides the number of instances examined (`n`) and the error rate (`err`). Node 15's detailed breakdown is shown in Figure 8.2, where we observe that node 16 correctly predicts 70% of cases, with the remaining 30% falling into other categories.

The second level of transparency pertains to the choice of features in these models. In feature-based ML models, the features, their types, and values are explicitly known. This transparency is particularly evident in models like decision trees, where the decision-making process and the role of each feature are clearly visible. However, this level of transparency varies across different ML algorithms. For instance, while decision trees, as shown above, provide a clear insight into both decision steps and features, other algorithms like Random Forest and gradient boosting methods do not offer the same degree of transparency.

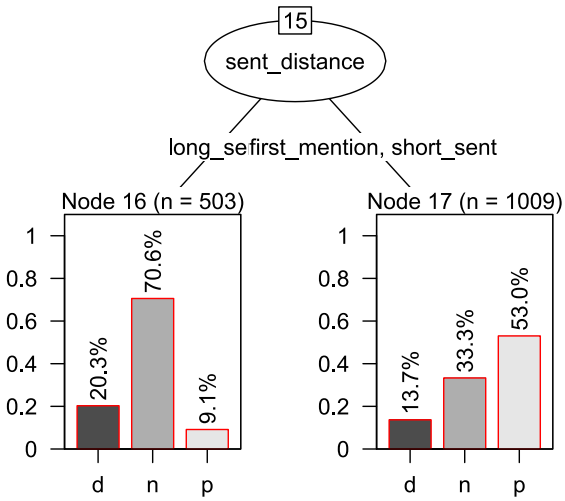

Figure 8.2: Branch 15 and its termination leaves.

Neural E2E models Neural end-to-end (E2E) models present a stark contrast to rule- and feature-based methods, particularly in terms of transparency. Firstly, these models do not use predefined features in their decision-making processes. This absence of explicit features means that the basis on which these models make decisions remains largely unknown. Secondly, the internal decision-making steps of neural E2E models are not accessible, obscuring the path by which they arrive at predictions. This opacity is a significant departure from the more interpretable nature of rule- and feature-based models.

Despite this lack of transparency, the success of neural E2E models is undeniable. Their effectiveness is largely attributed to the combination of sophisticated learning algorithms and their extensive parametric space, as noted by Castelvecchi (2016), West (2018), and Barredo Arrieta et al. (2020). The rapid evolution of deep learning approaches, along with their continually improving learning, reasoning, and adaptation capabilities, has enabled these models to achieve remarkable performance levels in complex computational tasks (West 2018). This advancement has established them as a central and increasingly dominant force in NLP.

However, with the growing reliance on and applicability of these models, the importance of interpretability and explainability has come to the forefront. In recent years, the field of *eXplainable AI* (XAI) has gained significant momentum, with numerous research initiatives aimed at elucidating the inner workings of neural systems. These efforts reflect a growing recognition in the AI community

of the need to bridge the gap between the high performance of neural E2E models and our understanding of their decision-making processes (Dosilovic et al. 2018, Barredo Arrieta et al. 2020).

Accuracy vs. model transparency Figure 8.3 illustrates a fundamental trade-off in model architectures between accuracy and transparency, as discussed by Dosilovic et al. (2018) and Barredo Arrieta et al. (2020). This trade-off is a key consideration in the development and selection of models for computational tasks.

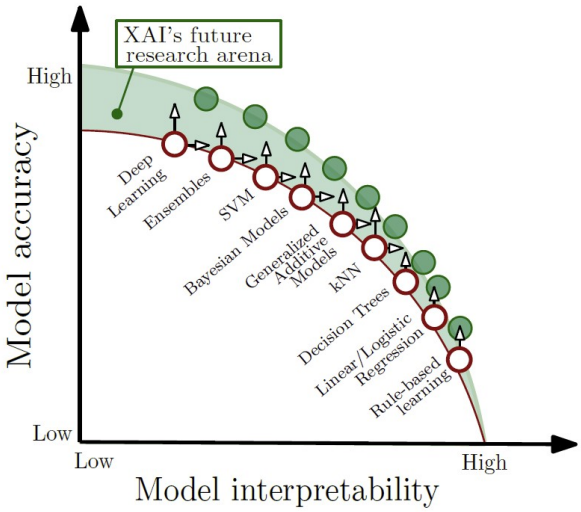

Figure 8.3: Trade-off between model interpretability and performance taken from Barredo Arrieta et al. (2020).

In the scope of study F, from a methodological perspective, rule-based models emerge as the most interpretable owing to their explicit decision-making rules. However, when it comes to model accuracy, particularly in handling complex data, feature-based models, especially those employing XGBoost (a gradient boosting ensemble algorithm), demonstrate superior performance. This increased accuracy may be attributed to various factors, including the use of linguistic features or the inherent strengths of the model architecture itself. The observed dichotomy between interpretability and accuracy presents a significant challenge. On the one hand, the need for transparency in models is paramount for

understanding the models. On the other hand, the need for high accuracy, particularly in complex data scenarios, often necessitates more sophisticated and less transparent models.

8.2.4 Interpretability and explainability of the outcome of REG-in-context models

The previous section primarily focused on the transparency of different model architectures, categorizing them into transparent models, like rule-based models and decision trees, and more opaque models, such as CatBoost, XGBoost, and neural models. This book also explores several *post-hoc explainability* methods (Barredo Arrieta et al. 2020) to enhance the explainability of these models. Post-hoc explainability refers to external techniques used to make the predictions of models more interpretable. This concept is akin to how individuals justify their decisions without fully understanding their own decision-making processes (Dosilovic et al. 2018).

In this work, various methods were used to elucidate the findings, including model-agnostic variable importance, Sequential Forward Selection (SFS), global and local SHapley Additive exPlanations (SHAP), probing experiments, and error analysis. The following discussion provides an overview of these methods: (1) feature relevance analysis, which assesses the importance of features in feature-based model predictions, (2) probing experiments, which add interpretability to neural models, and (3) error analysis, which helps interpret the predictions made by REG-in-context models.

Feature relevance explanation In this work, both global and local feature relevance analyses were conducted to shed light on how different features influence model predictions. The global feature analysis aimed to identify the features with the most significant contribution to a model's overall predictive performance. This analysis was crucial for several reasons. For instance, in study B, it was instrumental in determining a consensus set of features critical for the REG-in-context task. Study G used feature relevance analysis to hypothesize which features would be encoded in the latent representations of neural models.

The methodologies for assessing feature relevance vary, and each offers unique insights. Techniques like model-agnostic variable importance and Sequential Forward Selection assess each feature's overall contribution to a model's performance. In contrast, SHAP analysis goes a step further by illustrating both the positive and negative impacts of each feature on the prediction of individual

classes. This dual perspective provided by SHAP analysis not only identifies important features for the task but also clarifies their role in predicting specific classes.

On a more granular level, local feature relevance analysis, as exemplified in the error analysis of study E, calculates the contribution of features to individual predictions. This local approach to explainability is particularly valuable for diagnosing the causes of specific incorrect predictions. It allows for a detailed examination of why a model might have made an error in a particular instance, enhancing our understanding of model behavior at a micro-level.

Probing tasks Probing tasks offer a unique approach to understanding deep learning models, especially in cases where traditional post-hoc explainability techniques used in feature-based models are not applicable. However, the application of probing tasks comes with certain limitations, as highlighted in study G. One key limitation is the difficulty in generalizing the results, partly due to the unsuitability of the underlying corpus used for these tasks. Additionally, there is a critical distinction to be made between "extracting the linguistic structure encoded in the representations" and "learning the probing task" itself, as noted by Kunz & Kuhlmann (2020: 5136). This distinction raises caution in interpreting probing results. If a model appears to encode a feature, it might not necessarily mean that the feature is truly encoded in the model's latent representations. Instead, it could indicate that the probing classifier has learned to perform the task independently of the model's actual encoding capabilities (Kunz & Kuhlmann 2020).

Moreover, the absence of a feature in the model's latent representation does not necessarily imply that the feature is irrelevant to the task. It could suggest a limitation in the model's ability to encode that particular feature. Given that probing is a relatively new field in NLP research, the methodologies employed still lack a robust theoretical foundation, making cautious interpretation essential. Despite these challenges, probing tasks represent a significant step forward in enhancing the interpretability of neural models.

Error analysis Error analysis, while being a vital method, is frequently overlooked in the natural language generation (NLG) community. As pointed out by van Miltenburg et al. (2021), there is a notable trend of *underreporting* errors in NLG research. This underreporting, as they define it, occurs when "authors neither include any error analysis nor provide any examples of errors made by the system, and they do not make reference to different kinds of errors that may appear in the output" (p. 140).

Recognizing this gap, in study E, I undertook a comprehensive error analysis to assess the impact of incorporating paragraph-related attributes in REG-in-context models. Initially, the analysis involved reporting the errors made by both models to identify their general weaknesses. Subsequently, I examined the errors made by each model individually. Out of the 213 individual errors identified, 142 were attributed to the strong baseline model, while 71 errors were associated with the model that included paragraph-related information. Notably, the strong baseline model made nearly twice as many incorrect predictions as the experimental model. Despite the minimal difference in overall accuracy between the two models, this significant disparity in error rates would have remained undetected without a thorough error analysis.

Further, the study examined how individual features contributed to specific individual errors. It revealed that errors in the strong baseline model were often related to the lack of encoded paragraph transitions. For example, errors frequently occurred when a human subject referent was situated just one sentence away from its antecedent but located in a different paragraph. While not a common scenario in the WSJ dataset, this pattern could be more prevalent in other document types, like Wikipedia articles where sentences typically center around the main subject of the article. Here, paragraph boundaries might significantly affect model performance.

Similarly, study F from Chapter 7, which compared different REG-in-context approaches, could greatly benefit from detailed error analysis. As noted by van Miltenburg et al. (2021), error reporting is particularly valuable when assessing various implementation paradigms, such as pipeline-based data-to-text systems versus neural end-to-end systems. Each system has unique weaknesses, and identifying where they falter can provide valuable insights. For instance, study F found that feature-based models with well-designed features are better suited for complex data. An in-depth error analysis can illuminate why these models outperform others and identify the types of predictions that are challenging for different approaches. Furthermore, by examining the errors, we can determine whether the primary difficulties lie in the step of selecting the appropriate form of a referring expression, or in the content realization step.

8.2.5 Evaluation methods used

This work presented a comprehensive evaluation of the REG-in-context models, using both automatic and human evaluation methods. The following sections delve into the specifics of each approach. Additionally, the necessity for in-depth qualitative evaluations of model outcomes is explored.

Automatic evaluations In this book, we explored a range of automatic evaluations conducted to assess the performance of both the RFS and RCS tasks. This section will offer an overview of the key metrics used in these evaluations.

In study A, as detailed in Chapter 4, it was demonstrated that overall model accuracy is not always a reliable indicator of performance, particularly in cases of imbalanced class distributions. This observation aligns with the insights provided by Padurariu & Breaban (2019), who argue for a shift in focus from optimizing overall classification accuracy to a more nuanced approach that balances precision and recall, especially in imbalanced datasets. This approach emphasizes the importance of evaluating the precision, recall, and F1 score for each individual class. For instance, the analysis in study A revealed that a model might achieve high overall accuracy but still overpredict certain classes, like pronouns, while underperforming in others, such as descriptions, as evidenced by very low F1 scores.

In addition to these class-specific metrics, we also used weighted-averaged scores. These scores provide a comprehensive perspective on the performance of RFS classification models by taking into account the prevalence of each class in the dataset. This method ensures a more balanced evaluation which is especially important in datasets with uneven class distributions.

In study F, which concentrated on the actual realization of REs, we employed two distinct classes of automatic metrics to evaluate the outcomes: (1) RE Accuracy and String Edit Distance (SED, Levenshtein 1966) for measuring the quality of the generated REs, and (2) Text Accuracy and BLEU (Papineni et al. 2002) for assessing the quality of the documents after the REs were inserted. These metrics combined offer a comprehensive evaluation of the REG-in-context models. While RE Accuracy and SED focus on the precision of individual REs, Text Accuracy and BLEU provide a broader assessment of how these REs integrate into and affect the overall quality of the generated documents.

Human evaluations In study F, we conducted two human evaluation experiments to assess the fluency, grammaticality, and clarity of the outputs generated by the models. Human evaluations involve participants reviewing and rating the generated text based on these criteria. However, this process can present certain challenges, as not all aspects of these criteria are immediately apparent. Consider the following example generated by the neural model ATT-META:

(2) William Anders was born in British Hong Kong but is an American . He was a member of Apollo 8 which is operated by NASA . His backup pilot was Buzz Aldrin . **They** retired in 1969-09-01 .

At first glance, this example appears grammatically correct. However, a closer look at the corresponding RDF data, as shown in Table 8.2, reveals an inconsistency. The retirement date is associated only with *William Anders*, not with both *William Anders* and *Buzz Aldrin*. Therefore, the italicized *They* in the example is incorrect.

Table 8.2: The RFD of (2).

William_Anders	dateOfRetirement	1969-09-01
William_Anders	was a crew member of	Apollo_8
William_Anders	nationality	United_States
William_Anders	birthPlace	British_Hong_Kong
Apollo_8	backup pilot	Buzz_Aldrin
Apollo_8	crewMembers	Frank_Borman
Apollo_8	operator	NASA

This discrepancy poses a challenge for human evaluators. Some may recognize the error based on the RDF information and deem the sentence ungrammatical, while others, focusing solely on the text, might interpret a plural subject and consider the sentence grammatically correct.

Human evaluations play a critical role in understanding the nuances of natural language generation, beyond what automated metrics can capture. However, as this example illustrates, ensuring accuracy in human evaluations requires careful consideration of the evaluators' perspectives and the provision of complete information to guide their judgments.

Another significant challenge in evaluating NLG models lies in reconciling the results of automatic and human evaluations. Research by Novikova et al. (2017) and Castro Ferreira (2018) has highlighted that there is not always a clear correlation between the results of these two types of evaluation. For instance, consider a scenario where a model scores highly in automatic metrics but receives low ratings in human evaluations. This discrepancy raises important questions about how to accurately interpret the performance of the model. Are the automatic metrics not capturing certain qualitative aspects that human evaluators are sensitive to, or are there other factors at play?

Moreover, it is important to note that both automatic and human evaluations discussed in this work are forms of intrinsic evaluation metrics. As pointed out by van Miltenburg et al. (2021), these intrinsic metrics, despite their usefulness, may be too coarse-grained to capture all relevant information. They provide general evaluations of system performance, estimating an average case performance

across a limited set of abstract dimensions. However, this approach might not fully reflect the real-world effectiveness and applicability of the models.

Given these limitations, there is a growing recognition of the value of extrinsic methods in measuring performance. While these methods are more expensive and time-consuming (Gkatzia & Mahamood 2015), they offer a more comprehensive view of a model's performance by evaluating it in the context of its intended use. Thus, despite the challenges, incorporating extrinsic evaluation methods alongside intrinsic ones could offer a more complete picture of a REG-in-context model's capabilities.

A need for in-depth qualitative evaluations As previously discussed in Chapter 7, one of the significant challenges faced by E2E generation systems is the occurrence of *hallucinations*. In the context of natural language generation, a hallucination refers to instances where the system generates content that is either untrue or not present in the input data (Rohrbach et al. 0010, Reiter 2018, Nie et al. 2019). This phenomenon raises serious concerns about the reliability and accuracy of these systems. An illustrative example of this can be seen in (3), which compares the original text (3a) with the output generated by the model PROFILEREG (3b) in study F.

(3) a. ORIGINAL: **The Aston Martin V8** is assembled in the United Kingdom, the leader of which, is Elizabeth II.

b. PROFILEREG: **Simon Martin** is assembled in the United Kingdom, the leader of which, is Elizabeth II.

In this instance, the model replaces *The Aston Martin V8* with *Simon Martin*, which is not only incorrect but also not found in the input data. Such errors highlight the necessity for in-depth qualitative evaluations.

In addition to hallucinations, another type of error that may not be adequately captured by automatic or human evaluations is the issue of incomplete information. This occurs when the generated text, while not incorrect per se, fails to provide comprehensive and informative content. An illustrative case of this can be seen in (4) generated by the ATT-META system:

(4) AIDAstella is operated by Rostock based AIDA, Cruises. It was built by **Meyer** and is owned by Costa Crociere.

The corresponding RDF data for this example is shown in Table 8.3. Here, the text mentions *Meyer* as the builder of *AIDAstella*. While this is not incorrect, it is

not entirely complete either. The full name of the builder, as indicated in the RDF, is *Meyer Werft*. Since this is the first mention of the referent *Meyer Werft* in the text, the omission of *Werft* results in incomplete information being presented.

Table 8.3: The RFD of (4).

AIDA_Cruises	location	Rostock
AIDAstella	operator	AIDA_Cruises
AIDAstella	builder	Meyer_Werft
AIDAstella	owner	Costa_Crociere

These examples underscore the need for more in-depth qualitative evaluations. Such evaluations should not only check for factual accuracy or grammaticality but also assess the completeness and informativeness of the information provided.

In addition to the errors mentioned above, Reiter (2020) has highlighted a range of error types that may appear in the outputs of E2E NLG systems. These include the inappropriate use of words, implying incorrect attributes, and suggesting undue importance. Each of these error types can significantly impact the effectiveness and reliability of the generated text. A comprehensive qualitative analysis is therefore needed to provide insights into the extent and types of errors that occur, beyond what is possible through automatic metrics or surface-level human evaluations.

8.3 One-shot REG vs. REG-in-context

In Chapter 3, a clear distinction was made between two REG tasks: REG-in-context and one-shot REG. While REG-in-context incorporates linguistic context in generating referring expressions, one-shot REG operates without this context. This fundamental difference – the presence or absence of context – leads to distinct challenges and mechanisms in each task.

In this section, I delve into the different mechanisms underlying both REG-in-context and one-shot REG, focusing on their central goals of achieving identifiability and naturalness (or humanlikeness). Despite sharing the objective of identifiability, the strategies employed by REG-in-context and one-shot REG may differ significantly due to the role of context.

One of the common issues in one-shot REG is overspecification. This section examines how this problem manifests differently in REG-in-context, where the

additional layer of contextual information plays a role. Another key area of discussion is the use of referential forms. In one-shot REG, the referring expressions are typically descriptions. However, REG-in-context allows for a broader range of forms, including pronouns and proper names.

Finally, given that this book has focused on REG-in-context, an important goal of this section is to assess the implications of the REG-in-context studies for understanding and improving one-shot REG. This involves considering whether the insights and findings from REG-in-context can be extended to enhance our approach to one-shot REG, and to what extent the challenges and solutions identified in REG-in-context are applicable to the one-shot task.

The central goal of both tasks In one-shot REG, a central question revolves around the choice of properties for an RE to uniquely identify a referent. If the sole aim were identifiability, one straightforward approach would be to include *all* properties of a referent. This method would ensure unique identification in most cases. However, as discussed in Chapter 3, achieving identifiability alone is not sufficient for effective REG.

In one-shot REG, algorithms such as the Incremental Algorithm not only aim for identifiability but also aim to produce REs that are natural and humanlike. The criterion of humanlikeness, as defined by van Deemter (2016), assesses how closely the output of an algorithm resembles what a human speaker would naturally produce. Achieving humanlikeness in REs can be approached in various ways, such as generating overspecified expressions or varying the order of attributes (van Deemter et al. 2012).

In contrast, REG-in-context requires an additional layer of consideration. To attain humanlikeness in REG-in-context, it is crucial to take into account the contextual cues and the prominence status of referents within the discourse. Consider examples 5 and 6, generated by two different algorithms discussed in Chapter 7, and observe how the referent *AIDAstella* is mentioned in the bolded portions of these examples:

(5) [FEAT-S] AIDAstella is operated by Rostock based AIDA Cruises. **AIDAstella** was built by Meyer Werft and is owned by Costa Crociere .

(6) [ATT-COPY] AIDAstella is operated by Rostock based AIDA Cruises. **It** was built by Meyer and is owned by Costa Costa .

In the REG-in-context task, the use of a pronominal expression (*It*) to refer to *AIDAstella*, as seen in (6), aligns with how humans typically refer to prominent entities in discourse. Conversely, in the one-shot REG task, the assumption

is that there is no contextual information influencing the status of the referents. All entities – the target and its distractors – are considered to have similar prominence. Therefore, the selection of an RE in one-shot REG does not account for the referent's prominence within a larger discourse, as the task is designed to be context-free. In this scenario, a more descriptive form of RE, like the full name *AIDAstella* as seen in (5), might be preferred for identifiability, even though it may not reflect natural human usage in a contextual setting.

Overspecification Overspecification, a concept widely debated in the context of one-shot REG, refers to instances where an RE includes more properties than necessary for unambiguous identification of the referent (van Deemter 2016). This phenomenon has been observed in both experimental and computational studies, revealing that humans frequently produce overspecified REs (Pechmann 1989, Engelhardt et al. 2006, Koolen et al. 2011, Paraboni et al. 2017, Degen et al. 2020). Interestingly, research indicates that listeners often do not perceive overspecified REs negatively; in some cases, they may even facilitate referent identification (Engelhardt et al. 2006, Arts et al. 2011).

The generation of overspecified REs has been extensively explored and implemented within the framework of one-shot REG. For instance, the Incremental Algorithm (Dale & Reiter 1995) accommodates a certain level of overspecification. This is based on a preference ordering, provided that the overspecified attributes possess discriminatory power. Additionally, non-deterministic algorithms have explored overspecification by varying the order in which a referent's properties are considered, assigning different probabilities to each order (van Deemter et al. 2012, van Deemter 2016, van Gompel et al. 2019).

The phenomenon of overspecification is not exclusive to one-shot REG and often appears in REG-in-context as well. In Chapter 6, it was observed that more than 90% of paragraph-initial REs in REG-in-context are non-pronominal, even though a pronominal form would have sufficed in many instances. This can be considered a form of *form-overspecification*, where a richer form is used even when a simpler reduced one would be adequate for identification. For instance, in (7b), *Mr. Egnuss's* is used instead of a pronoun, despite Edward Egnuss being the only masculine referent in the preceding context.

(7) WSJ-1021

a. [Paragraph 1] Program trading is "a racket," complains **Edward Egnuss**, **a White Plains, N.Y., investor and electronics sales executive**, "and it's not to the benefit of the small investor, that's for sure." But although **he** thinks that it is hurting **him**, **he** doubts it could be stopped.

b. [First sentence of paragraph 2] **Mr. Egnuss's** dislike of program trading is echoed by many small investors interviewed by Wall Street Journal reporters across the country.

This choice might be influenced by a desire to avoid repetition of forms or to mark a transition between paragraphs. While strictly speaking, this RE could be labeled as overspecified based on identification criteria alone, its role in the discourse may justify its usage. This complexity underscores the need for further research to identify the cases of overspecification.

Content-overspecification represents a second, intriguing form of overspecification. While some overspecifications are purely functional, such as including additional location information in spatial and hierarchical domains to aid the addressee in locating referents (Paraboni et al. 2007, Paraboni & van Deemter 2014), other instances serve a more narrative or descriptive purpose. An interesting example of this can be seen in the way *Amanda Gorman* is referred to in (8):[1]

(8) Amanda Gorman became the youngest inaugural poet in U.S. history when she recited her poem "The Hill We Climb" at President Joe Biden's swearing in ceremony Wednesday. **The 22-year-old Los Angeles resident and daughter of a school teacher** began writing at an early age in an attempt to cope with a speech impediment.

In this example, Amanda Gorman is initially introduced with relevant details pertaining to her role as a poet at Joe Biden's inauguration event. The bolded RE in the following sentence provides additional attributes about her age, hometown, and family background. While these details are not strictly necessary for identifying Amanda Gorman, given the context of the first sentence, they enrich the narrative by offering a more comprehensive picture of her.

The inclusion of such content-overspecification in the second sentence extends beyond the mere facilitation of identification. It provides the reader with engaging information that, while not essential to understanding the core narrative, adds depth to the character being described. Furthermore, these additional details can serve as a narrative bridge, smoothly transitioning the reader to subsequent topics or themes within the article.

As highlighted in the previous paragraphs, overspecification is a common phenomenon in both REG tasks. While form-overspecification appears to be unique

[1] `https://www.cnbc.com/2021/01/20/meet-amanda-gorman-the-youngest-inaugural-poet-in-us-history.html`

to REG-in-context, content-overspecification is relevant and observable in both REG-in-context and one-shot REG. Drawing inspiration from one-shot REG research (Paraboni et al. 2007, van Gompel et al. 2019, Degen et al. 2020), REG-in-context research should aim to identify and categorize the different types of overspecification. The goal would be to not only understand these phenomena but also to integrate them into algorithmic solutions for natural language generation. Moreover, the inherent complexity of contextual cues in REG-in-context allows for exploring more varied cases of overspecification.

Proper names In the one-shot REG task, the typical objective is to use one or more properties of a referent to generate a distinguishing description to differentiate the referent from a set of distractors. Traditionally, most REG algorithms focus on generating descriptions rather than considering other categories, such as proper names. However, this conventional approach raises an important question about the naturalness and efficiency of REs in real-world communication.

Figure 8.4: A modified example from the people domain of the TUNA corpus.

Consider a modified example from the people domain of the TUNA corpus, illustrated in Figure 8.4.[2] Suppose participants in a one-shot REG experiment

[2]The image of Barack Obama is not part of the TUNA corpus; it was obtained from the following website: `https://www.nobelprize.org/prizes/peace/2009/obama/facts/`.

are shown this image and asked to produce a distinctive RE for the enclosed image. If restricted to descriptions, participants might generate an RE like *the man with black hair and a tie*. Yet, if allowed to choose freely, many would likely use a proper name, such as *Barack Obama*, which is more direct and intuitively recognizable.

This example raises the question of why a distinguishing RE should always take the form of a description. In natural language usage, people might be more inclined to use a referent's proper name to distinguish it from a set of distractors, especially when given the choice between using a proper name and a description. As van Deemter (2016) suggests, proper names can be incorporated into a knowledge base (KB) as properties of a referent, similar to other properties. By integrating proper names into the REG mechanism, they can be used in conjunction with other properties and relations to create more natural and efficient REs.

In the context of REG-in-context, the use of proper names as properties provides an interesting perspective on the generation of REs. This approach is exemplified in the WSJ corpus, as seen in (9), where the referent *Garry Hoffman* is introduced with multiple attributes.

(9) **Gary Hoffman, a Washington lawyer specializing in intellectual property cases**, said the threat of U.S. retaliation, combined with a growing recognition that protecting intellectual property is in a country's own interest, prompted the improvements made by South Korea, Taiwan and Saudi Arabia.

Table 8.4: An example of a KB for the referent *Gary Hoffman*.

Property	Value
Name	Gary Hoffman
Job	Lawyer
Place of Work	Washington
Speciality	Intellectual-property

The corresponding knowledge base (KB), shown in Table 8.4, lists *Garry Hoffman* as a property among others, such as his profession and specialization.[3] In

[3]When the REG task is implemented in closed and restricted domains, KBs can be built for all referents and then used to generate REs (Siddharthan & Copestake 2004). In an open domain such as the WSJ corpus used in this work, KBs cannot be easily built for all referents.

this example, the conjunction of multiple attributes – Gary Hoffman's name, profession, place of work, and area of specialization – effectively introduces the referent in a manner that mirrors how humans typically provide context about a person in conversation.

If we view proper names as properties, an interesting question arises: Are the mechanisms for generating first mentions in REG-in-context similar to generation mechanisms in one-shot REG? The answer is a partial yes. While there are similarities, REG-in-context differs from one-shot REG in that the generation of first mentions in REG-in-context is influenced by contextual factors and the prominence of referents. Consider the following example, which could be the opening sentence of a newspaper article:

(10) George Bush, the President of the United States, entered the White House.

In this instance, the RE *George Bush* could refer to either *George H. W. Bush* or *George W. Bush*, with the correct identification depending on whether the article was written in 1992 or 2002. This example illustrates that in REG-in-context, an RE cannot always be identified independently, but rather, its clarity depends on the surrounding context.

In contrast, one-shot REG is traditionally defined as a context-free task. However, when we consider an example like that in Figure 8.4, we might question whether context, in the form of global importance or recognizability, plays a role even in one-shot REG. Does the prominence of a figure like Barack Obama make the use of a proper name more appropriate than a descriptive RE in this context? This observation leads us to reconsider whether one-shot REG is entirely devoid of contextual cues.

Reflecting on these discussions, it becomes apparent that similar mechanisms may be at play in both REG-in-context and one-shot REG, albeit at different levels of complexity. Both tasks deal with the challenge of generating clear and identifiable REs, but REG-in-context introduces additional layers of complexity due to its reliance on contextual information and referent prominence.

8.4 Linguistic vs. computational approaches

This book has endeavored to incorporate linguistic insights into the computational exploration of the REG-in-context task. While this work has drawn upon a range of psycholinguistic concepts, it is important to acknowledge that other relevant topics like alignment, audience design, egocentricity, and non-determinism, though not covered here, also hold significant value for the REG-in-context task

(see van Deemter et al. 2012 for a discussion on the cognitive plausibility of REG models).

In this section, I aim to bridge the gap between linguistic (both theoretical and experimental) and computational perspectives on reference production. This dialogue is particularly necessary in light of the growing separation between these fields in recent years. An analysis by Reiter (2007) of citations in computational linguistics journals from 1995 and 2005 revealed a significant shift. The study found that articles published in 2005 demonstrated markedly less influence from diverse language research communities compared to those from 1995. This trend of isolation has been intensified by the rapid advancement of DL models in recent years, leading to an even more pronounced disconnect between computational linguistics and other linguistic disciplines.

As observed by Rogers & Augenstein (2020), there is a tendency in the mainstream computational linguistics community to prioritize DL-based methods, which might inadvertently marginalize interdisciplinary research efforts. This trend towards intellectual segregation is concerning, as it overlooks the rich insights that linguistic research can offer to computational approaches in natural language processing. In the following discussion, I will explore instances where the paths of linguistics and computational science diverge and identify areas where cross-disciplinary input could be mutually beneficial.

8.4.1 Refining the concepts

In this section, I illustrate with two examples how computational and experimental studies can mutually enhance each other in refining and testing linguistic concepts. Theoretical frameworks of reference, as discussed in Chapter 2, provide a general direction for the choice of REs (Gundel et al. 1993, Ariel 2001, Grosz et al. 1995). However, these theories often lack detailed guidance on practical implementation.

Centering Theory, described in Chapter 2 as a framework for understanding local coherence and salience, illustrates this point. The theory includes various rules, constraints, and parameters, the specifics of which are open to interpretation (Poesio et al. 2004). Poesio et al. demonstrated that while behavioral experiments are valuable for validating centering assumptions, the diverse ways the theory can be implemented pose a challenge. To address this, they adopted a computational approach, applying different parameter settings to an annotated corpus. This method enabled them to compare various interpretations of the theory and formalize some previously unspecified parameters. The insights gained from

this computational analysis can subsequently guide behavioral experiments, providing refined parameter settings for more targeted empirical studies.

Similarly, study C in Chapter 5 followed this approach by computationally testing different implementations of recency, a concept derived from corpus-based (Givón 1983, Ariel 1990) and experimental (Arnold 2010) studies. These studies predominantly used an utterance-based notion of recency, leaving other methods, such as word-based recency, less explored. Our study examined 15 different recency measures and found that higher-level measures like sentences and paragraphs were more influential in the choice of RF than lower-level measures. This finding suggests that the prominence of a referent may decrease not merely due to distance, but rather due to the transitions between sentences or paragraphs. Such a hypothesis, arising from computational analysis, can be further examined through controlled linguistic experiments.

These examples demonstrate the value of using computational methods to test and refine linguistic theories, creating opportunities for new hypotheses that can be validated experimentally. This synergistic approach, combining computational models with linguistic theory and experimentation, underscores the importance of interdisciplinary collaboration in natural language processing, where computational insights can inform experimental designs and vice versa.

8.4.2 Building reliable corpora

In this work, the significance of corpus selection for the REG-in-context task has been a recurring theme. The approaches to corpus development in linguistics and computational science differ markedly, but there are opportunities for these fields to mutually enrich each other in creating more reliable and effective corpora.

Computer scientists typically work with large datasets, often relying on automatic preprocessing libraries and automated rule-based processing. For example, in annotating REs in the WEBNLG corpus, a simple rule is applied: If a determiner precedes a noun phrase, it is classified as a description; otherwise, it is considered a proper name. However, as demonstrated in Chapter 7, adhering strictly to this rule can lead to misclassifications, affecting the performance of classification models. Linguistic corpus-based studies, on the other hand, generally use smaller, meticulously annotated corpora. While this attention to detail ensures high precision, it can sometimes result in datasets that are idiosyncratic and less amenable to reuse for different research purposes. Schmidt et al. (2014) highlight the challenges of linguistic data heterogeneity, pointing out that corpora are often created with specific theories and research questions in mind, which can limit their broader applicability.

The contrast between these approaches suggests a potential for mutual benefit. Linguistic expertise in detailed annotation and exception handling can enhance the precision of large datasets typically used in computational work. Conversely, computational methods for creating standardized, large-scale corpora can contribute to the linguistic need for broader, more versatile datasets. By collaborating, linguists and computer scientists can develop corpora that balance precision with abundance, catering to the needs of both fields. For instance, in the construction of a REG corpus for E2E studies, incorporating delexicalized forms of REs is crucial. While these were automatically generated in study F and may contain errors, a collaborative approach involving linguistic input could improve their accuracy.

In summary, a closer collaboration between linguists and computer scientists could lead to the development of more reliable, interoperable corpora. Such joint efforts can leverage the strengths of both disciplines, resulting in datasets that are not only extensive and standardized but also meticulously annotated and adaptable for various research needs.

8.4.3 Evaluation and experimentation

The fields of computational linguistics and psycholinguistics offer unique methods and perspectives that can be mutually beneficial, particularly in evaluating models and testing linguistic hypotheses. This section explores two ways these fields can interact: (1) using psycholinguistic methods in computational studies, and (2) employing language models as subjects in psycholinguistic experiments.

Applying psycholinguistic methods in computational studies. In Chapter 7, we conducted a Magnitude Estimation task to evaluate the outputs of the WSJ models. Given that most evaluation studies have focused on shorter texts, there was uncertainty about the effectiveness of this method for longer articles. Recent advancements in neuropsychological experiments, particularly in naturalistic language processing involving non-trivial context, suggest new avenues for evaluation methods. Neuroimaging techniques like fMRI and ERP, as highlighted by Alday (2018), are increasingly used in naturalistic language studies and could be adapted to evaluate the outputs of NLG models, especially for longer texts.

Language models as psycholinguistic subjects. In recent years, the NLP community has witnessed the emergence of a new research field that applies psycholinguistic methods to examine the linguistic capabilities of black-box models (Linzen et al. 2016, Futrell et al. 2019, Ettinger 2020). Baroni (2021) coined the term *linguistically-oriented deep net analysis* for this growing area. This field adopts a psycholinguistic viewpoint, conducting sophisticated experiments to probe the

knowledge implicit in a model's behavior, analogous to studying a species' behavior. One interesting aspect of this research, as Baroni (2021) notes, is understanding why models, presumably operating on different principles than the human brain, excel in processing human language.

So far, most studies in this area have focused on syntactic phenomena (Linzen et al. 2016, Futrell et al. 2019, Hewitt & Manning 2019). However, a notable exception is the study by Upadhye et al. (2020), which explored reference processing. They investigated the next-mention biases of two language models, *Transformer-XL* and *GPT-2*, in various contexts: the use of implicit causality verbs, the comparison of motion versus transfer of possession verbs, and aspectual marking in transfer of possession verbs. Their research is particularly compelling as it replicates experiments typically conducted with human subjects, using language models instead. This approach enables a direct comparison between human and model behaviors, offering valuable insights into the similarities and differences in language processing between humans and language models.

These examples illustrate the potential for a collaborative approach between computational linguistics and psycholinguistics. By adopting methods from each other's domains, researchers can develop more sophisticated evaluation techniques and gain a deeper understanding of both artificial and human language processing.

8.4.4 REG in NLG applications

REG is not only a vibrant area of academic research but also an essential component in commercial NLG applications (Reiter 2016b, 2017). The deployment of NLG technologies spans diverse fields such as healthcare, the oil and gas industry, journalism, and financial services, each with its specific requirements and challenges (for a detailed list of use cases, see `https://www.arria.com/industry-expertise/`).

In certain applications, especially those in critical domains like healthcare, accuracy is paramount. As Reiter (2016a) points out, for some applications, "it is essential that the systems produce reasonable texts in *all* cases. For example, a medical decision-support system cannot produce texts that decrease the quality of patient care." In such contexts, it is crucial for the system to recognize potential ambiguities in REs and avoid them to ensure clarity and precision. Conversely, other applications might prioritize more varied and humanlike expressions in both form and content, allowing for a lesser degree of control. This variation in requirements underscores the necessity for context-sensitive approaches in constructing REs. What is considered a high-quality RE in one application might

differ significantly in another, depending on the specific needs and objectives (Reiter 2017).

The development of effective REG systems for these diverse applications can greatly benefit from a synergy between linguistic insights and computational methodologies. Linguistic research provides a deep understanding of the nuances of human language, including how context influences referential clarity and naturalness. Computational approaches, on the other hand, offer the tools and frameworks to implement these insights at scale.

8.5 This book in a nutshell

This book presents seven in-depth studies exploring various aspects of the REG-in-context task, each aimed at developing solutions informed by linguistic theory. The following is an overview of Chapters 4 to 7, summarizing the key studies and their contributions.

8.5.1 Overview of Chapters 4–7

Chapter 4 delves into the critical issue of corpus selection in REG-in-context studies. The central study in this chapter (study A) examined whether the choice of corpus significantly impacts the outcomes of REG-in-context research (question addressed: **QA**). Two hypotheses were tested: (1) The corpora used in the previous REG-in-context studies are not adequate for the task (**HA1**), and (2) the lessons learned in previous REG-in-context studies are not generally valid (**HA2**). The investigation revealed that one of the corpora, GREC-PEOPLE, was not suitable for REG-in-context, highlighting the importance of corpus choice. The study underscored the substantial corpus-dependence of any REG algorithm, raising concerns about the generalizability of past findings. It emphasized the need to critically evaluate whether a chosen language corpus aligns with the specific linguistic phenomena of interest to the researchers, ensuring the relevance and applicability of the research outcomes. The findings of this chapter highlight the significance of corpus selection in NLP and NLG research. It calls for a more refined approach to choosing and evaluating corpora, emphasizing the need to ensure that they are not only technically suitable but also representative of the language use and phenomena under investigation.

Chapter 5 addresses the critical aspect of feature selection for feature-based models in the REG-in-context task. Study B investigated the effectiveness of various feature sets used in previous feature-based REG-in-context studies (question

addressed: **QB**). It tested two hypotheses: (1) a reduced subset of features from each set can perform comparably to the full feature set (**HB1**), and (2) a small set of features drawn from previously published datasets can form a model substantially as accurate as the best-performing existing model (**HB2**). The study's findings indicated that not all features previously used are equally relevant to the task. A linguistically informed subset of each feature set yields results almost identical to those obtained using the full set. Moreover, a consensus set of just six features, drawn from various sets, performed nearly as well as the best-performing model from prior studies, which used 2.5 times as many features.

In addition, study C in this chapter examined the optimal conceptualization of recency for the RFS task (**QC**). It evaluated two hypotheses: (1) recency metrics that encode *higher-level* distances contribute more to RFS than those based on *lower-level* distances (**HC1**), and (2) the effectiveness of recency metrics varies depending on corpus-specific characteristics, such as the genre and structure of texts (**HC2**). The exhaustive evaluation of various recency measures revealed that higher-level metrics like distance in number of sentences and paragraphs are generally more influential than lower-level ones like word count. Furthermore, the study uncovered that the effectiveness of these recency metrics differs across corpora, being more pronounced in the news text corpus (WSJ) compared to the Wikipedia corpus (GREC-2.0).

Chapter 6 delves into the significance of paragraph-related features in REG-in-context models, examining how paragraph structure influences the choice of RF (**QDE**). Study D investigated the impact of paragraph structure on RF choice using data from the WSJ corpus. The chapter evaluated three key hypotheses: (1) paragraph-prominent entities are substantially more likely to be pronominalized (**HD1**); (2) paragraph-new and paragraph-initial REs are substantially more likely to be non-pronominal (**HD2**); and (3) paragraph-new REs are more likely to be pronominal if the referent is prominent in the current and the previous paragraph (**HD3**). The study's findings supported the first hypothesis, revealing that prominent referents are substantially more likely to be pronominalized (almost 4.5 times more likely). It also confirmed that over 90% of paragraph-new and paragraph-initial REs are realized non-pronominally. However, while the data suggested a tendency for cross-boundary pronominalization of prominent referents, it was not sufficient for a conclusive judgment on the third hypothesis.

Additionally, study E in this chapter aimed to assess the utility of incorporating paragraph-related features into REG-in-context models. The hypothesis was that the inclusion of paragraph-related information substantially improves the performance of feature-based REG-in-context models (**HE1**). Results demonstrated that models trained on the WSJ corpus with paragraph-related features did perform

better. However, compared to a strong baseline model, the improvement was modest. This study further hypothesized that the impact of paragraph-related features could be more pronounced in datasets like GREC-2.0 and GREC-PEOPLE, where specific entities play central roles and are repeatedly mentioned.

Chapter 7 addresses the critical issue of selecting appropriate REG-in-context approaches, with a specific focus on evaluating the efficacy of neural REG models in comparison to rule-based and feature-based models. The first study in this chapter, study F, investigated the prevailing assumption that neural REG models are superior (**QF**). The hypothesis (**HF1**) challenged this belief, positing that neural REG models are not always better than rule-based and feature-based models. The systematic evaluation of three different REG-in-context approaches across two distinct corpora revealed that contrary to the prevailing assumption, the linguistically informed feature-based model demonstrated superior performance over other models, including neural ones, particularly when applied to the more complex WSJ corpus. This outcome suggests that a deep understanding of linguistic principles can significantly enhance the effectiveness of REG-in-context models, even in the face of advanced neural architectures.

Additionally, study G conducted eight probing experiments to investigate the types of linguistic features encoded by neural REG-in-context models (**QG**). The experiments revealed that models consistently performed well on features related to referential status and grammatical role. Furthermore, models with an attention mechanism showed a better capacity for encoding linguistic features compared to simpler models. However, the neural classifier designed for pronominalization displayed a somewhat distinct performance pattern. A word of caution is necessary here, as the corpus used for this study may not have been ideally suited for the task, which could affect the interpretability of the results.

8.5.2 Overview of the major lessons

The research presented in this book offered several crucial lessons that contribute to our understanding and implementation of the REG-in-context task:

- The studies in Chapters 4, 5, and 7 (specifically, studies A, C, and F) underscore the critical role of corpus selection in REG-in-context research. This choice influences all subsequent decisions, including feature selection, methodology, and evaluation metrics, highlighting the need for careful consideration in corpus selection to ensure the relevance and effectiveness of REG-in-context models.

- Study B in Chapter 5 demonstrates that a small set of well-chosen, linguistically informed features can achieve high performance in feature-based models. Features such as grammatical role, recency, and animacy are particularly impactful, suggesting that a focused and informed approach to feature selection is key to efficient and effective modeling.
- Chapters 6's studies D and E reveal that incorporating broader contextual information, like paragraph-boundary transitions, enhances the REG-in-context task. However, the degree of this enhancement varies depending on the dataset, underscoring the importance of context-aware modeling.
- The systematic evaluation in study F of Chapter 7 illustrates that rule-based systems with simple rules can perform on par with, or even better than, neural REG systems. Particularly in complex datasets, feature-based models with linguistically informed features excel, emphasizing the value of integrating linguistic knowledge into algorithmic solutions.
- As indicated in study A of Chapter 4 and study F of Chapter 7, the overall accuracy of a model should not be the sole criterion for evaluation. Moving away from overall accuracy performance seems even more important when the dataset contains imbalanced classes. Furthermore, other factors like model transparency and resource usage must be considered for a comprehensive assessment of a model's utility.
- The opacity of neural model architectures, as discussed in study G of Chapter 7, poses a challenge for understanding their internal workings. Methods like probing offer a viable approach to gain insights into the models' latent representations. Furthermore, when feasible, post-hoc explainability methods, including variable importance analysis and SHAP analysis, should be used. These methods aid in clarifying the predictions made by the models.

These lessons collectively advance our knowledge in the field of REG-in-context. They highlight the importance of corpus selection, the efficiency of targeted feature selection, the need for context-aware modeling, and the value of a comprehensive approach to model evaluation and interpretation. This comprehensive understanding paves the way for developing more sophisticated, linguistically informed, and contextually appropriate natural language generation systems.

References

Aissen, Judith. 2003. Differential object marking: Iconicity vs. Economy. *Natural Language & Linguistic Theory* 21(3). 435–483. DOI: 10.1023/a:1024109008573.

Alday, Phillip M. 2018. M/EEG analysis of naturalistic stories: A review from speech to language processing. *Language, Cognition and Neuroscience* 34(4). 457–473. DOI: 10.1080/23273798.2018.1546882.

Almor, Amit. 1999. Noun-phrase anaphora and focus: The informational load hypothesis. *Psychological Review* 106(4). 748–765. DOI: 10.1037/0033-295x.106.4.748.

Alt, Christoph, Aleksandra Gabryszak & Leonhard Hennig. 2020. Probing linguistic features of sentence-level representations in neural relation extraction. In *Proceedings of the 58th annual meeting of the Association for Computational Linguistics*, 1534–1545. Online: Association for Computational Linguistics. DOI: 10.18653/v1/2020.acl-main.140.

Altmann, André, Laura Toloşi, Oliver Sander & Thomas Lengauer. 2010. Permutation importance: A corrected feature importance measure. *Bioinformatics (Oxford, England)* 26(10). 1340–1347.

Ariel, Mira. 1990. *Accessing noun-phrase antecedents*. London: Routledge.

Ariel, Mira. 1991. The function of accessibility in a theory of grammar. *Journal of Pragmatics* 16(5). 443–463. DOI: 10.1016/0378-2166(91)90136-l.

Ariel, Mira. 2001. Accessibility theory: An overview. In Ted Sanders, Joost Schilperoord & Wilbert Spooren (eds.), *Text representation: Linguistic and psycholinguistic aspects*, vol. 8, 29. Amsterdam: John Benjamins. DOI: 10.1075/hcp.8.04ari.

Ariel, Mira. 2004. Accessibility marking: Discourse functions, discourse profiles, and processing cues. *Discourse processes* 37(2). 91–116.

Arnold, Jennifer E. 2001. The effect of thematic roles on pronoun use and frequency of reference continuation. *Discourse Processes* 31(2). 137–162. DOI: 10.1207/s15326950dp3102_02.

Arnold, Jennifer E. 2008. Reference production: Production-internal and addressee-oriented processes. *Language and Cognitive Processes* 23(4). 495–527. DOI: 10.1080/01690960801920099.

Arnold, Jennifer E. 2010. How speakers refer: The role of accessibility. *Language and Linguistics Compass* 4(4). 187–203. DOI: 10.1111/j.1749-818x.2010.00193.x.

Arnold, Jennifer E., Loisa Bennetto & Joshua J. Diehl. 2009. Reference production in young speakers with and without autism: Effects of discourse status and processing constraints. *Cognition* 110(2). 131–146. DOI: 10.1016/j.cognition.2008.10.016.

Arnold, Jennifer E. & Zenzi M. Griffin. 2007. The effect of additional characters on choice of referring expression: Everyone counts. *Journal of Memory and Language* 56(4). 521–536. DOI: 10.1016/j.jml.2006.09.007.

Arnold, Jennifer E., Anthony Losongco, Thomas Wasow & Ryan Ginstrom. 2000. Heaviness vs. Newness: The effects of structural complexity and discourse status on constituent ordering. *Language* 76(1). 28–55. DOI: 10.1353/lan.2000.0045.

Arts, Anja, Alfons Maes, Leo G. M. Noordman & Carel Jansen. 2011. Overspecification in written instruction. *Linguistics* 49(3). 555–574. DOI: 10.1515/ling.2011.017.

Bahdanau, Dzmitry, Kyunghyun Cho & Yoshua Bengio. 2014. *Neural Machine translation by jointly learning to align and translate.* DOI: 10.48550/arXiv.1409.0473.

Bard, Ellen Gurman, Dan Robertson & Antonella Sorace. 1996. Magnitude estimation of linguistic acceptability. *Language* 72(1). 32–68. DOI: 10.2307/416793.

Baroni, Marco. 2021. *On the proper role of linguistically-oriented deep net analysis in linguistic theorizing.* DOI: 10.48550/arXiv.2106.08694.

Barredo Arrieta, Alejandro, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila & Francisco Herrera. 2020. Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion* 58. 82–115. DOI: 10.1016/j.inffus.2019.12.012.

Baumann, Stefan & Arndt Riester. 2012. Referential and lexical givenness: Semantic, prosodic and cognitive aspects: in Gorka Elordieta & Pilar Prieto (eds.), *Prosody and meaning*, 119–162. Berlin: De Gruyter Mouton. DOI: 10.1515/9783110261790.119.

Belinkov, Yonatan, Nadir Durrani, Fahim Dalvi, Hassan Sajjad & James Glass. 2017. What do neural machine translation models learn about morphology? In *Proceedings of the 55th annual meeting of the Association for Computational Linguistics (volume 1: Long papers)*, 861–872. Vancouver: Association for Computational Linguistics. DOI: 10.18653/v1/P17-1080.

Belz, Anja. 2008. Automatic generation of weather forecast texts using comprehensive probabilistic generation-space models. *Natural Language Engineering* 14(4). 431–455.

Belz, Anja & Eric Kow. 2010. The GREC challenges 2010: Overview and evaluation results. In *Proceedings of the 6th International Natural Language Generation Conference*, 219–229. Association for Computational Linguistics.

Belz, Anja, Eric Kow, Jette Viethen & Albert Gatt. 2010. Generating referring expressions in context: The GREC task evaluation challenges. In Emiel Krahmer & Mariët Theune (eds.), *Empirical methods in natural language generation: Data-oriented methods and empirical evaluation*, vol. 5790 (Lecture Notes in Computer Science), 294–327. Berlin: Springer. DOI: 10.1007/978-3-642-15573-4_15.

Belz, Anja & Ehud B. Reiter. 2006. Comparing automatic and human evaluation of NLG systems. In *11th Conference of the European Chapter of the Association for Computational Linguistics*. Trento: Association for Computational Linguistics.

Belz, Anja & Sebastian Varges. 2007. Generation of repeated references to discourse entities. In *Proceedings of the Eleventh European Workshop on Natural Language Generation*, 9–16. Association for Computational Linguistics.

Biau, Gérard. 2012. Analysis of a random forests model. *The Journal of Machine Learning Research* 13(1). 1063–1095.

Biecek, Przemyslaw & Tomasz Burzykowski. 2021. *Explanatory model analysis: Explore, explain, and examine predictive models*. New York: Chapman and Hall/CRC.

Bischl, Bernd, Michel Lang, Lars Kotthoff, Julia Schiffner, Jakob Richter, Erich Studerus, Giuseppe Casalicchio & Zachary M Jones. 2016. Mlr: Machine learning in R. *The Journal of Machine Learning Research* 17(1). 5938–5942.

Bishop, Christopher M. 2006. *Pattern recognition and machine learning* (Information Science and Statistics). New York.

Bohnet, Bernd. 2008. IS-G: The comparison of different learning techniques for the selection of the main subject references. In *Proceedings of the Fifth International Natural Language Generation Conference*, 192–193. Association for Computational Linguistics.

Bolshakov, Igor A. & Alexander F. Gelbukh. 2001. Text segmentation into paragraphs based on local text cohesion. In Václav Matousek, Pavel Mautner, Roman Moucek & Karel Tauser (eds.), *Text, Speech and Dialogue, 4th International Conference, TSD 2001, Zelezna Ruda, Czech Republic, September 11-13, 2001, Proceedings*, vol. 2166 (Lecture Notes in Computer Science), 158–166. Springer.

Branco, Paula, Luís Torgo & Rita P. Ribeiro. 2016. A survey of predictive modeling on imbalanced domains. *ACM Computing Surveys* 49(2). 1–50. DOI: 10.1145/2907070.

Breiman, Leo. 2001. Random forests. *Machine learning* 45(1). 5–32.

Brennan, Susan E. 1995. Centering attention in discourse. *Language and Cognitive Processes* 10(2). 137–167. DOI: 10.1080/01690969508407091.

Brennan, Susan E. 1996. Lexical choice and conceptual pacts in conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 22. 1482–1493.

Brennan, Susan E., Marilyn W. Friedman & Carl J. Pollard. 1987. A centering approach to pronouns. In *25th annual meeting of the Association for Computational Linguistics*, 155–162. Stanford: Association for Computational Linguistics. DOI: 10.3115/981175.981197.

Brilmayer, Ingmar & Petra B. Schumacher. 2021. Referential chains reveal predictive processes and form-to-function mapping: An electroencephalographic study using naturalistic story stimuli. *Frontiers in Psychology* 12. 1–16. DOI: 10.3389/fpsyg.2021.623648.

Cao, Meng & Jackie Chi Kit Cheung. 2019. Referring expression generation using entity profiles. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3163–3172. Hong Kong: Association for Computational Linguistics. DOI: 10.18653/v1/D19-1312.

Carlson, Lynn, Daniel Marcu & Mary Ellen Okurowski. 2002. *RST discourse treebank*. DOI: 10.35111/4W31-M996.

Castelvecchi, Davide. 2016. Can we open the black box of AI? *Nature News* 538(7623). 20.

Castro Ferreira, Thiago. 2018. *Advances in natural language generation: Generating varied outputs from semantic inputs*. Tilburg: Tilburg University. (Doctoral dissertation).

Castro Ferreira, Thiago, Emiel Krahmer & Sander Wubben. 2016a. Individual variation in the choice of referential form. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 423–427. San Diego: Association for Computational Linguistics. DOI: 10.18653/v1/N16-1048.

Castro Ferreira, Thiago, Emiel Krahmer & Sander Wubben. 2016b. Towards more variation in text generation: Developing and evaluating variation models for choice of referential form. In *Proceedings of the 54th annual meeting of the Association for Computational Linguistics (volume 1: Long papers)*, 568–577. Berlin: Association for Computational Linguistics. DOI: 10.18653/v1/P16-1054.

Castro Ferreira, Thiago, Diego Moussallem, Ákos Kádár, Sander Wubben & Emiel Krahmer. 2018a. NeuralREG: An end-to-end approach to referring expression generation. In *Proceedings of the 56th annual meeting of the Association for Computational Linguistics (volume 1: Long papers)*, 1959–1969. Melbourne: Association for Computational Linguistics. DOI: 10.18653/v1/P18-1182.

Castro Ferreira, Thiago, Diego Moussallem, Emiel Krahmer & Sander Wubben. 2018b. Enriching the WebNLG corpus. In *Proceedings of the 11th International Conference on Natural Language Generation*, 171–176. Tilburg University: Association for Computational Linguistics. DOI: 10.18653/v1/W18-6521.

Castro Ferreira, Thiago & Ivandré Paraboni. 2017. Improving the generation of personalised descriptions. In *Proceedings of the 10th International Conference on Natural Language Generation*, 233–237. Santiago de Compostela: Association for Computational Linguistics. DOI: 10.18653/v1/W17-3536.

Castro Ferreira, Thiago, Chris van der Lee, Emiel van Miltenburg & Emiel Krahmer. 2019. Neural data-to-text generation: A comparison between pipeline and end-to-end architectures. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 552–562. Hong Kong: Association for Computational Linguistics. DOI: 10.18653/v1/D19-1052.

Chafe, Wallace L. 1980. The pear stories: Cognitive, cultural, and linguistic aspects of narrative production. http://pearstories.org/english/english.htm.

Chafe, Wallace L. 1976. Givenness, contrastiveness, definiteness, subjects, topics, and point of view. *Subject and topic*. 27–55.

Chen, Guanyi, Fahime Same & Kees van Deemter. 2021. What can neural referential form selectors learn? In *Proceedings of the 14th International Conference on Natural Language Generation*, 154–166. Aberdeen: Association for Computational Linguistics. DOI: 10.18653/v1/2021.inlg-1.15.

Chen, Guanyi, Fahime Same & Kees van Deemter. 2023. Neural referential form selection: Generalisability and interpretability. *Computer Speech & Language* 79. 101466. DOI: 10.1016/j.csl.2022.101466.

Chen, Tianqi & Carlos Guestrin. 2016. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (KDD '16), 785–794. New York: Association for Computing Machinery. DOI: 10.1145/2939672.2939785.

Chiarcos, Christian. 2011. The mental salience framework: Context-adequate generation of referring expressions. In Christian Chiarcos, Berry Claus & Michael Grabski (eds.), *Salience*, 105–140. Berlin: De Gruyter Mouton. DOI: 10.1515/9783110241020.105.

Chiarcos, Christian & Olga Krasavina. 2005. *Annotation guidelines. POCOS-Potsdam coreference scheme draft*. Online. https://www.researchgate.net/publication/228572420_Annotation_Guidelines_PoCoS-Potsdam_Coreference_Scheme_Draft.

Cho, Kyunghyun, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk & Yoshua Bengio. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1724–1734. Doha: Association for Computational Linguistics. DOI: 10.3115/v1/D14-1179.

Cohen, William W. 1996. Learning trees and rules with set-valued features. In *AAAI/IAAI, Vol. 1*, 709–716.

Comrie, Bernard. 1989. *Language universals and linguistic typology: Syntax and morphology*. Chicago: University of Chicago Press.

Cunha, Rossana, Thiago Castro Ferreira, Adriana Pagano & Fabio Alves. 2020. Referring to what you know and do not know: Making referring expression generation models generalize to unseen entities. In *Proceedings of the 28th International Conference on Computational Linguistics*, 2261–2272. Barcelona: International Committee on Computational Linguistics. DOI: 10.18653/v1/2020.coling-main.205.

Daelemans, Walter, Jakub Zavrel, Ko van der Sloot & Antal van den Bosch. 2007. *TiMBL: Tilburg memory-based learner*. https://languagemachines.github.io/timbl/.

Dahl, Östen & Kari Fraurud. 1996. Animacy in grammar and discourse. In Thorstein Fretheim & Jeanette K. Gundel (eds.), *Reference and referent accessibility*, 47. Amsterdam: John Benjamins. DOI: 10.1075/pbns.38.04dah.

Dale, Robert. 1989. Cooking up referring expressions. In *27th Annual Meeting of the Association for Computational Linguistics*, 68–75.

Dale, Robert. 1992. *Generating referring expressions - constructing descriptions in a domain of objects and processes* (ACL-MIT Press Series in Natural Language Processing). Cambridge: MIT Press.

Dale, Robert & Ehud B. Reiter. 1995. Computational interpretations of the Gricean maxims in the generation of referring expressions. *Cognitive science* 19(2). 233–263.

De la Fuente, Israel. 2015. *Putting pronoun resolution in context: The role of syntax, semantics, and pragmatics in pronoun interpretation*. Université paris Diderot. (Doctoral dissertation).

Degen, Judith, Robert D. Hawkins, Caroline Graf, Elisa Kreiss & Noah D. Goodman. 2020. When redundancy is useful: A Bayesian approach to "overinformative" referring expressions. *Psychological Review* 127(4). 591–621. DOI: 10.1037/rev0000186.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee & Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. Minneapolis: Association for Computational Linguistics. DOI: 10.18653/v1/N19-1423.

Di Eugenio, Barbara, Johanna D Moore & Massimo Paolucci. 1997. Learning features that predict cue usage. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, 80–87. Association for Computational Linguistics.

Donnellan, Keith S. 1966. Reference and definite descriptions. *The Philosophical Review* 75(3). 281–304.

Donnellan, Keith S. 1972. Proper Names and Identifying Descriptions. In Donald Davidson (ed.), *Semantics of natural language*, 356–379. Dordrecht: Springer Netherlands.

Dosilovic, Filip Karlo, Mario Brcic & Nikica Hlupic. 2018. Explainable artificial intelligence: A survey. In *2018 41st international convention on information and communication technology, electronics and microelectronics (MIPRO)*. IEEE. DOI: 10.23919/mipro.2018.8400040.

Dušek, Ondřej, Jekaterina Novikova & Verena Rieser. 2018. Findings of the E2E NLG challenge. In *Proceedings of the 11th International Conference on Natural Language Generation*, 322–328. Tilburg University: Association for Computational Linguistics. DOI: 10.18653/v1/W18-6539.

Engelhardt, Paul E., Karl G. D. Bailey & Fernanda Ferreira. 2006. Do speakers and listeners observe the Gricean maxim of quantity? *Journal of Memory and Language* 54(4). 554–573. DOI: 10.1016/j.jml.2005.12.009.

Ettinger, Allyson. 2020. What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics* 8. 34–48. DOI: 10.1162/tacl_a_00298.

Favre, Benoit & Bernd Bohnet. 2009. ICSI-CRF: The generation of references to the main subject and named entities using conditional random fields. In *Proceedings of the 2009 Workshop on Language Generation and Summarisation*, 99–100. Association for Computational Linguistics.

Filippova, Katja & Michael Strube. 2006. Using linguistically motivated features for paragraph boundary identification. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, 267–274. Sydney: Association for Computational Linguistics.

Fox, Barbara A. 1987a. Anaphora in popular written English narratives. In Russell S. Tomlin (ed.), *Coherence and grounding in discourse* (Typological Studies in Language 11), 157–174. Amsterdam: John Benjamins.

Fox, Barbara A. 1987b. *Discourse Structure and Anaphora: Written and Conversational English* (Cambridge Studies in Linguistics). Cambridge: Cambridge University Press. DOI: 10.1017/CBO9780511627767.

Frank, Michael C. & Noah D. Goodman. 2012. Predicting pragmatic reasoning in language games. *Science (New York, N.Y.)* 336(6084). 998–998. DOI: 10.1126/science.1218633.

Frege, Gottlob. 1960. *Translations from the philosophical writings of Gottlob Frege*. Peter Geach & Max Black (eds.). London: Blackwell. 244.

Fukumura, Kumiko, Jukka Hyönä & Merete Scholfield. 2013. Gender affects semantic competition: The effect of gender in a non-gender-marking language. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 39(4). 1012–1021. DOI: 10.1037/a0031215.

Fukumura, Kumiko & Roger P. G. van Gompel. 2010. Choosing anaphoric expressions: Do people take into account likelihood of reference? *Journal of Memory and Language* 62(1). 52–66. DOI: 10.1016/j.jml.2009.09.001.

Fukumura, Kumiko & Roger P. G. van Gompel. 2011. The effect of animacy on the choice of referring expression. *Language and Cognitive Processes* 26(10). 1472–1504. DOI: 10.1080/01690965.2010.506444.

Futrell, Richard, Ethan Wilcox, Takashi Morita, Peng Qian, Miguel Ballesteros & Roger Levy. 2019. Neural language models as psycholinguistic subjects: Representations of syntactic state. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 32–42. Minneapolis: Association for Computational Linguistics. DOI: 10.18653/v1/N19-1004.

Gardent, Claire, Anastasia Shimorina, Shashi Narayan & Laura Perez-Beltrachini. 2017. Creating training corpora for NLG micro-planners. In *Proceedings of the 55th annual meeting of the Association for Computational Linguistics (volume 1: Long papers)*, 179–188. Vancouver: Association for Computational Linguistics. DOI: 10.18653/v1/P17-1017.

Gatt, Albert & Anja Belz. 2009. Introducing shared tasks to NLG: The TUNA shared task evaluation challenges. In Emiel Krahmer & Mariët Theune (eds.),

Empirical methods in natural language generation, 264–293. Berlin: Springer Berlin Heidelberg.

Gatt, Albert & Emiel Krahmer. 2018. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research* 61. 65–170.

Gatt, Albert, Emiel Krahmer, Kees van Deemter & Roger P. G. van Gompel. 2014. Models and empirical data for the production of referring expressions. *Language, Cognition and Neuroscience* 29(8). 899–911. DOI: 10.1080/23273798.2014.933242.

Gatt, Albert, Francois Portet, Ehud B. Reiter, Jim Hunter, Saad Mahamood, Wendy Moncur & Somayajulu Sripada. 2009. From data to text in the neonatal intensive care unit: Using NLG technology for decision support and information management. *AI Communications* 22(3). 153–186. DOI: 10.3233/AIC-2009-0453.

Gatt, Albert & Kees van Deemter. 2007a. Incremental generation of plural descriptions: Similarity and partitioning. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, 102–111. Prague: Association for Computational Linguistics.

Gatt, Albert & Kees van Deemter. 2007b. Lexical choice and conceptual perspective in the generation of plural referring expressions. *Journal of Logic, Language and Information* 16(4). 423–443.

Gatt, Albert & Kees van Deemter. 2009. Generating plural NPs in discourse: Evidence from the GNOME corpus. In *Proceedings of production of referring expressions*.

Gatt, Albert, Ielka van Der Sluis & Kees van Deemter. 2007. Evaluating algorithms for the generation of referring expressions using a balanced corpus. In *Proceedings of the Eleventh European Workshop on Natural Language Generation*, 49–56. Association for Computational Linguistics.

Gatt, Albert, Roger P. G. van Gompel, Kees van Deemter & Emiel Krahmer. 2013. Are We Bayesian referring expression generators. In *Proceedings of the CogSci Workshop on the Production of Referring Expressions*, 1–6. Cognitive Science Society.

Gernsbacher, Morton Ann. 1989. Mechanisms that improve referential access. *Cognition* 32(2). 99–156. DOI: 10.1016/0010-0277(89)90001-2.

Giulianelli, Mario, Jack Harding, Florian Mohnert, Dieuwke Hupkes & Willem Zuidema. 2018. Under the hood: Using diagnostic classifiers to investigate and improve how language models track agreement information. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural*

Networks for NLP, 240–248. Brussels: Association for Computational Linguistics. DOI: 10.18653/v1/W18-5426.

Givón, Talmy. 1983. *Topic continuity in discourse: A quantitative cross-language study* (Typological Studies in Language v.3). Amsterdam/Philadelphia: John Benjamins. DOI: 10.1075/tsl.3.

Givón, Talmy. 1992. The grammar of referential coherence as mental processing instructions. *Linguistics* 30(1). 5–56. DOI: 10.1515/ling.1992.30.1.5.

Gkatzia, Dimitra & Saad Mahamood. 2015. A snapshot of NLG evaluation practices 2005 - 2014. In *Proceedings of the 15th European Workshop on Natural Language Generation (ENLG)*, 57–60. Brighton: Association for Computational Linguistics. DOI: 10.18653/v1/W15-4708.

Goldberg, Yoav. 2017. Neural network methods for natural language processing. *Synthesis Lectures on Human Language Technologies* 10(1). 1–309. DOI: 10.2200/s00762ed1v01y201703hlt037.

Goodfellow, Ian, Joshua Bengio & Aaron Courville. 2016. *Deep learning*. Cambridge: MIT Press.

Greenbacker, Charles & Kathleen E. McCoy. 2009a. UDel: Generating referring expressions guided by psycholinguistic findings. In *Proceedings of the 2009 Workshop on Language Generation and Summarisation*, 101–102. Association for Computational Linguistics.

Greenbacker, Charles F. & Kathleen E. McCoy. 2009b. Feature selection for reference generation as informed by psycholinguistic research. In *Proceedings of the CogSci 2009 Workshop on Production of Referring Expressions (PRE-Cogsci 2009)*.

Grice, Herbert P. 1975. Logic and conversation. In Peter Cole & Jerry L. Morgan (eds.), *Speech acts*, 41–58. Leiden: Brill.

Grosz, Barbara J., Aravind K. Joshi & Scott Weinstein. 1983. Providing a unified account of definite noun phrases in discourse. In *21st Annual Meeting of the Association for Computational Linguistics*, 44–50.

Grosz, Barbara J., Aravind K. Joshi & Scott Weinstein. 1995. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics* 21(2). 203–225.

Gundel, Jeanette K. 2003. Information structure and referential givenness/newness. How much belongs in the grammar? *Journal of Cognitive Science* 4. 177–199.

Gundel, Jeanette K. 2008. *Reference: Interdisciplinary perspectives*. Oxford New York: Oxford University Press.

Gundel, Jeanette K., Nancy Hedberg & Ron Zacharski. 1993. Cognitive status and the form of referring expressions in discourse. *Language* 69(2). 274–307. (8 April, 2022).

Gupta, Samir & Sivaji Bandopadhyay. 2009. JUNLG-MSR: A machine learning approach of main subject reference selection with rule based improvement. In *Proceedings of the 2009 Workshop on Language Generation and Summarisation*, 103–104. Association for Computational Linguistics.

Hajičová, Eva. 1993. *Issues of sentence structure and discourse patterns* (Theoretical and Computational Linguistics 2). Prague: Charles University.

Hendrickx, Iris, Walter Daelemans, Kim Luyckx, Roser Morante & Vincent Van Asch. 2008. CNTS: Memory-based learning of generating repeated references. In *Proceedings of the Fifth International Natural Language Generation Conference*, 194–195. Association for Computational Linguistics.

Hendrickx, Iris & Veronique Hoste. 2009. Coreference resolution on blogs and commented news. In *Discourse anaphora and anaphor resolution colloquium*, 43–53. Springer.

Hendriks, Petra. 2016. Cognitive modeling of individual variation in reference production and comprehension. *Frontiers in Psychology* 7. 1–17. DOI: 10.3389/fpsyg.2016.00506.

Henschel, Renate, Hua Cheng & Massimo Poesio. 2000. Pronominalization revisited. In *Proceedings of the 18th Conference on Computational Linguistics-Volume 1*, 306–312. Association for Computational Linguistics.

Hewitt, John & Percy Liang. 2019. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2733–2743. Hong Kong: Association for Computational Linguistics. DOI: 10.18653/v1/D19-1275.

Hewitt, John & Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4129–4138. Minneapolis: Association for Computational Linguistics. DOI: 10.18653/v1/N19-1419.

Himmelmann, Nikolaus P. & Beatrice Primus. 2015. Prominence beyond prosody - a first approximation. In Amedeo De Dominicis (ed.), *Ps-prominences: Prominences in linguistics. Proceedings of the international conference*, 38–58. Viterbo. https://kups.ub.uni-koeln.de/24935/.

Hinds, John. 1977. Paragraph structure and pronominalization. *Paper in Linguistics* 10(1-2). 77–99.

Hinterwimmer, Stefan. 2019. Prominent protagonists. *Journal of Pragmatics* 154. 79–91.

Hitzeman, Janet & Massimo Poesio. 1998. Long distance pronominalisation and global focus. In *Proceedings of the 17th International Conference on Computational Linguistics*. Association for Computational Linguistics. DOI: 10.3115/980451.980937.

Hobbs, Jerry R. 1978. Resolving pronoun references. *Lingua. International review of general linguistics. Revue internationale de linguistique générale* 44(4). 311–338.

Hobbs, Jerry R. 1979. Coherence and coreference. *Cognitive Science* 3(1). 67–90. DOI: 10.1207/s15516709cog0301_4.

Hochreiter, Sepp & Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9(8). 1735–1780.

Hofmann, Thomas R. 1989. Paragraphs, & anaphora. *Journal of Pragmatics* 13(2). 239–250. DOI: 10.1016/0378-2166(89)90093-3.

Hovy, Eduard, Mitchell Marcus, Martha Palmer, Lance Ramshaw & Ralph Weischedel. 2006. OntoNotes: The 90% solution. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, 57–60. New York City: Association for Computational Linguistics.

Howcroft, David M., Anya Belz, Miruna-Adriana Clinciu, Dimitra Gkatzia, Sadid A. Hasan, Saad Mahamood, Simon Mille, Emiel van Miltenburg, Sashank Santhanam & Verena Rieser. 2020. Twenty Years of confusion in human evaluation: NLG needs evaluation sheets and standardised definitions. In *Proceedings of the 13th International Conference on Natural Language Generation*, 169–182. Dublin: Association for Computational Linguistics.

Huang, Yan. 2000. Discourse anaphora: Four theoretical models. *Journal of Pragmatics* 32(2). 151–176. DOI: 10.1016/s0378-2166(99)00041-7.

Jamison, Emily. 2008. Using discourse features for referring expression generation. In *Proceedings of the 5th meeting of the midwest computational linguistics colloquium (MCLC)*.

Jamison, Emily & Dennis Mehay. 2008. OSU-2: Generating referring expressions with a maximum entropy classifier. In *Proceedings of the Fifth International Natural Language Generation Conference*, 196–197. Association for Computational Linguistics.

Jordan, Pamela & Marilyn A. Walker. 2000. Learning attribute selections for non-pronominal expressions. In *Proceedings of the 38th annual meeting on Association for Computational Linguistics*, 181–190. Association for Computational Linguistics.

Jordan, Pamela W. & Marilyn A. Walker. 2005. Learning content selection Rules for generating object descriptions in dialogue. *Journal of Artificial Intelligence Research* 24. 157–194.

Jurafsky, Daniel & James H Martin. 2021. *Speech and language processing*. 3rd edn. https://web.stanford.edu/~jurafsky/slp3/.

Kaiser, Elsi. 2003. Word order, grammatical function, and referential form: On the patterns of anaphoric reference in Finnish. *Nordlyd* 31(1). DOI: 10.7557/12.28.

Kaiser, Elsi. 2010. Effects of contrast on referential form: Investigating the distinction between strong and weak pronouns. *Discourse Processes* 47(6). 480–509.

Kaiser, Elsi & John C. Trueswell. 2011. Investigating the interpretation of pronouns and demonstratives in Finnish. In Edward A. Gibson & Neal J. Pearlmutter (eds.), *The processing and acquisition of reference*, 323–354. Cambridge: The MIT Press. DOI: 10.7551/mitpress/9780262015127.003.0013.

Karmiloff-Smith, Annette. 1985. Language and cognitive processes from a developmental perspective. *Language and Cognitive Processes* 1(1). 61–85. DOI: 10.1080/01690968508402071.

Kass, Robert E & Adrian E Raftery. 1995. Bayes factors. *Journal of the american statistical association* 90(430). 773–795.

Kehler, Andrew. 2002. *Coherence, reference, and the theory of grammar* (Center for the Study of Language and Information Publication Lecture Notes). Cambridge: Cambridge University Press.

Kehler, Andrew, Laura Kertz, Hannah Rohde & Jeffrey L Elman. 2008. Coherence and coreference revisited. *Journal of semantics* 25(1). 1–44.

Kibble, Rodger & Richard Power. 1999. Using centering theory to plan coherent texts. In *Proceedings of the 12th Amsterdam colloquium*, 187–192. AC2012.

Kibrik, Andrej A, Mariya V Khudyakova, Grigory B Dobrov, Anastasia Linnik & Dmitrij A Zalmanov. 2016. Referential choice: Predictability and its limits. *Frontiers in Psychology* 7. 1429.

Koolen, Ruud, Albert Gatt, Martijn Goudbeek & Emiel Krahmer. 2011. Factors causing overspecification in definite descriptions. *Journal of Pragmatics* 43(13). 3231–3250. DOI: 10.1016/j.pragma.2011.06.008.

Krahmer, Emiel & Mariët Theune. 2002. Efficient context-sensitive generation of referring expressions. *Information sharing: Reference and presupposition in language generation and interpretation* 143. 223–263.

Krahmer, Emiel & Kees van Deemter. 2012. Computational generation of referring expressions: A survey. *Computational Linguistics* 38(1). 173–218. DOI: 10.1162/COLI_a_00088.

Krahmer, Emiel & Kees van Deemter. 2019. Computational Generation of Referring Expressions: An Updated Survey. In Jeanette K. Gundel & Barbara Abbott (eds.), *The Oxford Handbook of Reference*, 410–456. Oxford University Press. DOI: 10.1093/oxfordhb/9780199687305.013.19.

Kuhn, Max, Steve Weston, Mark Culp, Nathan Coulter & Ross Quinlan. 2018. *C50: C5.0 Decision Trees and Rule-Based Models*. `https://CRAN.R-project.org/package=C50`.

Kunz, Jenny & Marco Kuhlmann. 2020. Classifier probes may just learn from linear context features. In *Proceedings of the 28th International Conference on Computational Linguistics*, 5136–5146. Barcelona: International Committee on Computational Linguistics. DOI: 10.18653/v1/2020.coling-main.450.

Lambrecht, Knud. 1994. *Information structure and sentence form*. Cambridge: Cambridge University Press. DOI: 10.1017/cbo9780511620607.

Levenshtein, Vladimir Iosifovich. 1966. Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady* 10(8). 707–710.

Linzen, Tal, Emmanuel Dupoux & Yoav Goldberg. 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics* 4. 521–535.

McCoy, Kathleen E. & Michael Strube. 1999. Generating anaphoric expressions: Pronoun or definite description? In *The relation of Discourse/Dialogue structure and reference*.

Mei, Hongyuan, Mohit Bansal & Matthew R. Walter. 2016. What to talk about and how? Selective generation using LSTMs with coarse-to-fine alignment. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 720–730. San Diego: Association for Computational Linguistics. DOI: 10.18653/v1/N16-1086.

Mellish, Chris, Donia Scott, Lynne Cahill, Daniel Paiva, Roger Evans & Mike Reape. 2006. A reference architecture for natural language generation systems. *Natural Language Engineering* 12(1). 1–34. DOI: 10.1017/S1351324906004104.

Modi, Ashutosh, Ivan Titov, Vera Demberg, Asad Sayeed & Manfred Pinkal. 2017. Modeling semantic expectation: Using script knowledge for referent prediction. *Transactions of the Association for Computational Linguistics* 5. 31–44.

Molnar, Christoph. 2019. *Interpretable Machine learning*. Morrisville: Lulu. com.

Mountassir, Asmaa, Houda Benbrahim & Ilham Berrada. 2012. An empirical study to address the problem of Unbalanced Data Sets in sentiment classification. In *2012 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 3298–3303. DOI: 10.1109/ICSMC.2012.6378300.

Mulkern, Ann E. 1996. The game of the name. In Thorstein Fretheim & Jeanette K. Gundel (eds.), *Reference and referent accessibility*, 235. Amsterdam: John Benjamins. DOI: 10.1075/pbns.38.14mul.

Nayak, Anmol & D Natarajan. 2016. Comparative study of naive bayes, support vector Machine and random forest classifiers in sentiment analysis of twitter feeds. *International Journal of Advance Studies in Computer Science and Engineering (IJASCSE)* 5(1). 14–17.

Nie, Feng, Jin-Ge Yao, Jinpeng Wang, Rong Pan & Chin-Yew Lin. 2019. A simple recipe towards reducing hallucination in neural Surface realisation. In *Proceedings of the 57th annual meeting of the Association for Computational Linguistics*, 2673–2679. Florence: Association for Computational Linguistics. DOI: 10.18653/v1/P19-1256.

Novikova, Jekaterina, Ondřej Dušek, Amanda Cercas Curry & Verena Rieser. 2017. Why We need new evaluation metrics for NLG. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2241–2252. Copenhagen: Association for Computational Linguistics. DOI: 10.18653/v1/D17-1238.

Orăsan, Constantin & Iustin Dornescu. 2009. WLV: A confidence-based Machine learning method for the GREC-NEG'09 task. In *Proceedings of the 2009 Workshop on Language Generation and Summarisation (UCNLG+Sum 2009)*, 107–108. Suntec: Association for Computational Linguistics.

Orita, Naho, Eliana Vornov, Naomi Feldman & Hal Daum'e III. 2015. Why discourse affects speakers' choice of referring expressions. In *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)*, 1639–1649.

Padurariu, Cristian & Mihaela Elena Breaban. 2019. Dealing with data imbalance in text classification. *Procedia Computer Science* 159. 736–745. DOI: 10.1016/j.procs.2019.09.229.

Pandit, Onkar & Yufang Hou. 2021. Probing for bridging inference in transformer language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4153–4163. Online: Association for Computational Linguistics. DOI: 10.18653/v1/2021.naacl-main.327.

Papineni, Kishore, Salim Roukos, Todd Ward & Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of Machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, 311–318. Association for Computational Linguistics.

Paraboni, Ivandré, Alex Gwo Jen Lan, Matheus Mendes de SantAna & Flávio Luiz Coutinho. 2017. Effects of cognitive effort on the resolution of overspecified descriptions. *Computational Linguistics* 43(2). 451–459. DOI: 10.1162/coli_a_00288.

Paraboni, Ivandré, Kees Van Deemter & Judith Masthoff. 2007. Generating referring expressions: Making referents easy to identify. *Computational linguistics* 33(2). 229–254.

Paraboni, Ivandré & Kees van Deemter. 2014. Reference and the facilitation of search in spatial domains. *Language, Cognition and Neuroscience* 29(8). 1002–1017.

Passonneau, Rebecca J. 1996. Using centering to relax Gricean informational constraints on discourse anaphoric noun phrases. *Language and Speech* 39(2-3). 229–264.

Pechmann, Thomas. 1989. Incremental speech production and referential overspecification. *Linguistics* 27(1). 89–110.

Pennington, Jeffrey, Richard Socher & Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543. Doha: Association for Computational Linguistics. DOI: 10.3115/v1/D14-1162.

Poesio, Massimo. 2000. *The GNOME annotation scheme manual, Version 4*. Online. https://cswww.essex.ac.uk/Research/nle/corpora/GNOME/anno_manual_4.htm.

Poesio, Massimo. 2004. Discourse annotation and semantic annotation in the GNOME corpus. In *Proceedings of the Workshop on Discourse Annotation*, 72–79.

Poesio, Massimo, Rosemary Stevenson, Barbara Di Eugenio & Janet Hitzeman. 2004. Centering: A parametric theory and its instantiations. *Computational linguistics* 30(3). 309–363.

Portet, François, Ehud B. Reiter, Albert Gatt, Jim Hunter, Somayajulu Sripada, Yvonne Freer & Cindy Sykes. 2009. Automatic generation of textual summaries from neonatal intensive care data. *Artificial Intelligence* 173(7). 789–816. DOI: 10.1016/j.artint.2008.12.002.

Prat-Sala, Mercè & Holly P Branigan. 2000. Discourse constraints on syntactic processing in language production: A cross-linguistic study in English and Spanish. *Journal of Memory and Language* 42(2). 168–182. DOI: 10.1006/jmla.1999.2668.

Prince, Ellen F. 1992. The ZPG letter: Subjects, definiteness, and information-status. In William C. Mann & Sandra A. Thompson (eds.), *Pragmatics & Beyond New Series*, vol. 16, 295. Amsterdam. DOI: 10.1075/pbns.16.12pri.

Prokhorenkova, Liudmila Ostroumova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush & Andrey Gulin. 2018. CatBoost: Unbiased boosting with categorical features. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi & Roman Garnett (eds.), *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, 6639–6649.

Pu, Ming-Ming. 2019. Zero anaphora and topic chain in Chinese discourse. In Chris Shei (ed.), *The Routledge handbook of Chinese discourse analysis*, 188–200. London: Routledge. DOI: 10.4324/9781315213705-13.

Récanati, François. 1993. *Direct reference : From language to thought*. Oxford: Blackwell.

Reiter, Ehud B. 2007. Last words: The shrinking horizons of computational linguistics. *Computational Linguistics* 33(2). 283–287. DOI: 10.1162/coli.2007.33.2.283.

Reiter, Ehud B. 2016a. *Commercial and academic perspectives on NLG (and AI?)* Online. https://ehudreiter.com/2016/12/21/commercial-and-academic-nlg/.

Reiter, Ehud B. 2016b. Method and Apparatus for Referring Expression Generation. US9355093B2.

Reiter, Ehud B. 2017. A commercial perspective on reference. In *Proceedings of the 10th International Conference on Natural Language Generation*, 134–138. Santiago de Compostela: Association for Computational Linguistics. DOI: 10.18653/v1/W17-3519.

Reiter, Ehud B. 2018. *Hallucination in Neural NLG*. Online. https://ehudreiter.com/2018/11/12/hallucination-in-neural-nlg/.

Reiter, Ehud B. 2020. *Accuracy errors go beyond getting facts wrong*. Online. https://ehudreiter.com/2020/04/27/accuracy-errors-go-beyond-getting-facts-wrong.

Reiter, Ehud B. & Anja Belz. 2009. An investigation into the validity of some metrics for automatically evaluating natural language generation systems. *Computational Linguistics* 35(4). 529–558.

Reiter, Ehud B. & Robert Dale. 2000. *Building natural language generation systems*. Cambridge: Cambridge University Press.

Reiter, Ehud B., Somayajulu Sripada, Jim Hunter, Jin Yu & Ian Davy. 2005. Choosing words in computer-generated weather forecasts. *Artificial Intelligence* 167(1-2). 137–169. DOI: 10.1016/j.artint.2005.06.006.

Rogers, Anna & Isabelle Augenstein. 2020. What can We do to improve peer review in NLP? In *Findings of the Association for Computational Linguistics:*

EMNLP 2020, 1256–1262. Association for Computational Linguistics. DOI: 10.18653/v1/2020.findings-emnlp.112.

Rohrbach, Anna, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell & Kate Saenko. 0010. Object hallucination in image captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 4035–4045. Brussels: Association for Computational Linguistics. DOI: 10.18653/v1/D18-1437.

Rosa, Elise C. 2015. *Semantic role predictability affects referential form.* The University of North Carolina at Chapel Hill. (Doctoral dissertation).

Rösiger, Ina. 2018. BASHI: A corpus of wall street journal articles annotated with bridging links. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Rudin, Cynthia. 2019. Stop explaining black box Machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* 1(5). 206–215. DOI: 10.1038/s42256-019-0048-x.

Russell, Bertrand. 1905. On denoting. *Mind; a quarterly review of psychology and philosophy* 14(56). 479–493.

Saha, Sriparna, Asif Ekbal, Olga Uryupina & Massimo Poesio. 2011. Single and multi-objective optimization for feature selection in anaphora resolution. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, 93–101.

Same, Fahime, Guanyi Chen & Kees Van Deemter. 2022. Non-neural models matter: A re-evaluation of neural referring expression generation systems. In Smaranda Muresan, Preslav Nakov & Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 5554–5567. Dublin: Association for Computational Linguistics. DOI: 10.18653/v1/2022.acl-long.380.

Same, Fahime, Guanyi Chen & Kees van Deemter. 2023. Models of reference production: How do they withstand the test of time? In C. Maria Keet, Hung-Yi Lee & Sina Zarrieß (eds.), *Proceedings of the 16th International Natural Language Generation Conference*, 93–105. Prague: Association for Computational Linguistics. DOI: 10.18653/v1/2023.inlg-main.7.

Same, Fahime & Kees van Deemter. 2020a. A linguistic perspective on reference: Choosing a feature set for generating referring expressions in context. In *Proceedings of the 28th International Conference on Computational Linguistics*, 4575–4586. Barcelona: International Committee on Computational Linguistics. DOI: 10.18653/v1/2020.coling-main.403.

Same, Fahime & Kees van Deemter. 2020b. Computational interpretations of recency for the choice of referring expressions in discourse. In *Proceedings of the First Workshop on Computational Approaches to Discourse*, 113–123. Association for Computational Linguistics. DOI: 10.18653/v1/2020.codi-1.12.

Schmidt, Thomas, Christian Chiarcos, Timm Lehmberg, Georg Rehm, Andreas Witt & Erhard Hinrichs. 2014. Avoiding data graveyards: From heterogeneous data collected in multiple research projects to sustainable linguistic resources. In Thomas Schmidt, Christian Chiarcos, Timm Lehmberg, Georg Rehm, Andreas Witt & Erhard Hinrichs (eds.). Institut für Deutsche Sprache, Bibliothek.

Scott, Kate. 2019. Null referring expressions. In Kate Scott (ed.), *Referring expressions, pragmatics, and style: Reference and beyond*, 114–129. Cambridge: Cambridge University Press. DOI: 10.1017/9781316822845.006.

Searle, John R. 1969. *Speech acts: An essay in the philosophy of language*. Cambridge: Cambridge University Press.

Siddharthan, Advaith & Ann Copestake. 2004. Generating referring expressions in open domains. In *Proceedings of the 42nd annual meeting of the Association for Computational Linguistics (ACL-04)*, 407–414. Barcelona. DOI: 10.3115/1218955.1219007.

Siddharthan, Advaith, Ani Nenkova & Kathleen McKeown. 2011. Information status distinctions and referring expressions: An empirical study of references to people in news summaries. *Computational Linguistics* 37(4). 811–842.

Smith, Carlota S. 2003. *Modes of discourse*. Cambridge: Cambridge University Press. DOI: 10.1017/cbo9780511615108.

Sorodoc, Ionut-Teodor, Kristina Gulordava & Gemma Boleda. 2020. Probing for referential information in language models. In *Proceedings of the 58th annual meeting of the Association for Computational Linguistics*, 4177–4189. Online: Association for Computational Linguistics. DOI: 10.18653/v1/2020.acl-main.384.

Sporleder, Caroline & Mirella Lapata. 2004. Automatic paragraph identification: A study across languages and domains. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, 72–79. Barcelona: Association for Computational Linguistics.

Sporleder, Caroline & Mirella Lapata. 2006. Broad coverage paragraph segmentation across languages and domains. *ACM Transactions on Speech and Language Processing* 3(2). 1–35. DOI: 10.1145/1149290.1151098.

Stark, Heather A. 1988. What do paragraph markings do? *Discourse Processes* 11(3). 275–303.

Stent, Amanda J. 2011. Computational approaches to the production of referring expressions: Dialog changes (almost) everything. In *PRE-CogSci Workshop*.

Stevenson, Rosemary J., Rosalind A. Crawley & David Kleinman. 1994. Thematic roles, focus and the representation of events. *Language and Cognitive Processes* 9(4). 519–548. DOI: 10.1080/01690969408402130.

Stoia, Laura, Darla Magdalene Shockley, Donna K Byron & Eric Fosler-Lussier. 2006. Noun phrase generation for situated dialogs. In *Proceedings of the Fourth International Natural Language Generation Conference*, 81–88. Association for Computational Linguistics.

Strawson, Peter F. 1950. On referring. *Mind; a quarterly review of psychology and philosophy* 59(235). 320–344.

Strobl, Carolin, Anne-Laure Boulesteix, Thomas Kneib, Thomas Augustin & Achim Zeileis. 2008. Conditional variable importance for random forests. *BMC bioinformatics* 9(1). 307.

Tomlin, Russell S. 1987a. *Coherence and grounding in discourse. Symposium.* Russel S. Tomlin (ed.) (Typological Studies in Language 11). Amsterdam: John Benjamins.

Tomlin, Russell S. 1987b. Linguistic reflections of cognitive events. In Russell S. Tomlin (ed.), *Coherence and grounding in discourse* (Typological Studies in Language 11), 455–479. Amsterdam: John Benjamins.

Torroba Hennigen, Lucas, Adina Williams & Ryan Cotterell. 2020. Intrinsic probing through dimension selection. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 197–216. Online: Association for Computational Linguistics. DOI: 10.18653/v1/2020.emnlp-main.15.

Upadhye, Shiva, Leon Bergen & Andrew Kehler. 2020. Predicting reference: What do language models learn about discourse models? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 977–982. Online: Association for Computational Linguistics. DOI: 10.18653/v1/2020.emnlp-main.70.

van der Lee, Chris, Albert Gatt, Emiel van Miltenburg, Sander Wubben & Emiel Krahmer. 2019. Best practices for the human evaluation of automatically generated text. In *Proceedings of the 12th International Conference on Natural Language Generation*, 355–368.

van der Lee, Chris, Emiel Krahmer & Sander Wubben. 2017. PASS: A Dutch data-to-text system for soccer, targeted towards specific audiences. In *Proceedings of the 10th International Conference on Natural Language Generation*, 95–104. Santiago de Compostela: Association for Computational Linguistics. DOI: 10.18653/v1/W17-3513.

van Deemter, Kees. 2002. Generating referring expressions: Boolean extensions of the incremental algorithm. *Computational Linguistics* 28(1). 37–52.

van Deemter, Kees. 2016. *Computational models of referring: A study in cognitive science*. Cambridge: MIT Press.

van Deemter, Kees. 2019. Computational models of referring: Complications of information sharing. In Jeanette K. Gundel & Barbara Abbott (eds.), *The Oxford handbook of reference*, 475–495. Oxford: Oxford University Press. DOI: 10.1093/oxfordhb/9780199687305.013.20.

van Deemter, Kees, Albert Gatt, Roger P. G. van Gompel & Emiel Krahmer. 2012. Toward a computational psycholinguistics of reference production. *Topics in Cognitive Science* 4(2). 166–183. DOI: 10.1111/j.1756-8765.2012.01187.x.

van Deemter, Kees, Ielka van der Sluis & Albert Gatt. 2006. Building a semantically transparent corpus for the generation of referring expressions. In *Proceedings of the Fourth International Natural Language Generation Conference*, 130–132. Association for Computational Linguistics.

van Gompel, Roger P. G., Kees van Deemter, Albert Gatt, Rick Snoeren & Emiel J. Krahmer. 2019. Conceptualization in reference production: Probabilistic modeling and experimental testing. *Psychological Review* 126(3). 345–373. DOI: 10.1037/rev0000138.

van Miltenburg, Emiel, Miruna Clinciu, Ondřej Dušek, Dimitra Gkatzia, Stephanie Inglis, Leo Leppänen, Saad Mahamood, Emma Manning, Stephanie Schoch, Craig Thomson & Luou Wen. 2021. Underreporting of errors in NLG output, and what to do about it. In *Proceedings of the 14th International Conference on Natural Language Generation*, 140–153. Aberdeen: Association for Computational Linguistics.

Viethen, Jette. 2011. *The generation of natural descriptions: Corpus-based investigations of referring expressions in visual domains*. Australia: Macquarie University. (Doctoral dissertation).

Viethen, Jette, Thomas van Vessem, Martijn Goudbeek & Emiel Krahmer. 2017. Color in reference production: The role of color similarity and color codability. *Cognitive science* 41. 1493–1514. DOI: 10.1111/cogs.12387.

Vogels, Jorrig. 2014. *Referential choices in language production: The role of accessibility*. Tilburg: Tilburg center for Cognition & Communication (TiCC). (Doctoral dissertation).

Vogels, Jorrig. 2019. Both thematic role and next-mention biases affect pronoun use in Dutch. In *CogSci*, 3029–3035.

von Heusinger, Klaus & Petra B. Schumacher. 2019. Discourse prominence: Definition and application. *Journal of Pragmatics* 154. 117–127. DOI: 10.1016/j.pragma.2019.07.025.

Vonk, Wietske, Lettica G. M.M. Hustinx & Wim H. G. Simons. 1992. The use of referential expressions in structuring discourse. *Language and Cognitive Processes* 7(3-4). 301–333. DOI: 10.1080/01690969208409389.

Walker, Marilyn A. & Ellen F. Prince. 1996. A bilateral approach to giveness: A hearer-status algorithm and a centering algorithm. In Thorstein Fretheim & Jeanette K. Gundel (eds.), *Pragmatics & Beyond New Series*, vol. 38. Amsterdam: John Benjamins. DOI: 10.1075/pbns.38.17wal.

Weischedel, Ralph, Martha Palmer, Mitchell Marcus, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, Michelle Franchini, et al. 2012. *OntoNotes release 5.0 with OntoNotes DB Tool v0. 999 beta*. https://catalog.ldc.upenn.edu/docs/LDC2013T19/OntoNotes-Release-5.0.pdf.

Weischedel, Ralph, Palmer, Martha, Marcus, Mitchell, Hovy, Eduard, Pradhan, Sameer, Ramshaw, Lance, Xue, Nianwen, Taylor, Ann, Kaufman, Jeff, Franchini, Michelle, El-Bachouti, Mohammed, Belvin, Robert & Houston, Ann. 2013. *OntoNotes release 5.0 LDC2013t19*. DOI: 10.35111/XMHB-2B84.

Weiss, Gary M & Foster Provost. 2003. Learning when training data are costly: The effect of class distribution on tree induction. *Journal of artificial intelligence research* 19. 315–354.

West, Darrell. 2018. *The future of work : Robots, AI, and automation*. Washington, D.C: Brookings Institution Press.

Wright, Marvin N & Andreas Ziegler. 2017. Ranger: A fast implementation of Random Forests for High Dimensional Data in C++ and R. *Journal of Statistical Software* 77. 1–17. DOI: 10.18637/jss.v077.i01.

Yang, Zichao, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola & Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1480–1489. San Diego: Association for Computational Linguistics. DOI: 10.18653/v1/N16-1174.

Zarrieß, Sina & Jonas Kuhn. 2013. Combining referring expression generation and Surface realization: A corpus-based investigation of architectures. In *Proceedings of the 51st annual meeting of the Association for Computational Linguistics (volume 1: Long papers)*, 1547–1557.

Name index

Referring expression generation in context

Reference production, often termed Referring Expression Generation (REG) in computational linguistics, encompasses two distinct tasks: (1) one-shot REG, and (2) REG-in-context. One-shot REG explores which properties of a referent offer a unique description of it. In contrast, REG-in-context asks which (anaphoric) referring expressions are optimal at various points in discourse.

This book offers a series of in-depth studies of the REG-in-context task. It thoroughly explores various aspects of the task such as corpus selection, computational methods, feature analysis, and evaluation techniques. The comparative study of different corpora highlights the pivotal role of corpus choice in REG-in-context research, emphasizing its influence on all subsequent model development steps. An experimental analysis of various feature-based machine learning models reveals that those with a concise set of linguistically-informed features can rival models with more features. Furthermore, this work highlights the importance of paragraph-related concepts, an area underexplored in Natural Language Generation (NLG). The book offers a thorough evaluation of different approaches to the REG-in-context task (rule-based, feature-based, and neural end-to-end), and demonstrates that well-crafted, non-neural models are capable of matching or surpassing the performance of neural REG-in-context models. In addition, the book delves into post-hoc experiments, aimed at improving the explainability of both neural and classical REG-in-context models. It also addresses other critical topics, such as the limitations of accuracy-based evaluation metrics and the essential role of human evaluation in NLG research.

These studies collectively advance our understanding of REG-in-context. They highlight the importance of selecting appropriate corpora and targeted features. They show the need for context-aware modeling and the value of a comprehensive approach to model evaluation and interpretation. This detailed analysis of REG-in-context paves the way for developing more sophisticated, linguistically-informed, and contextually appropriate NLG systems.