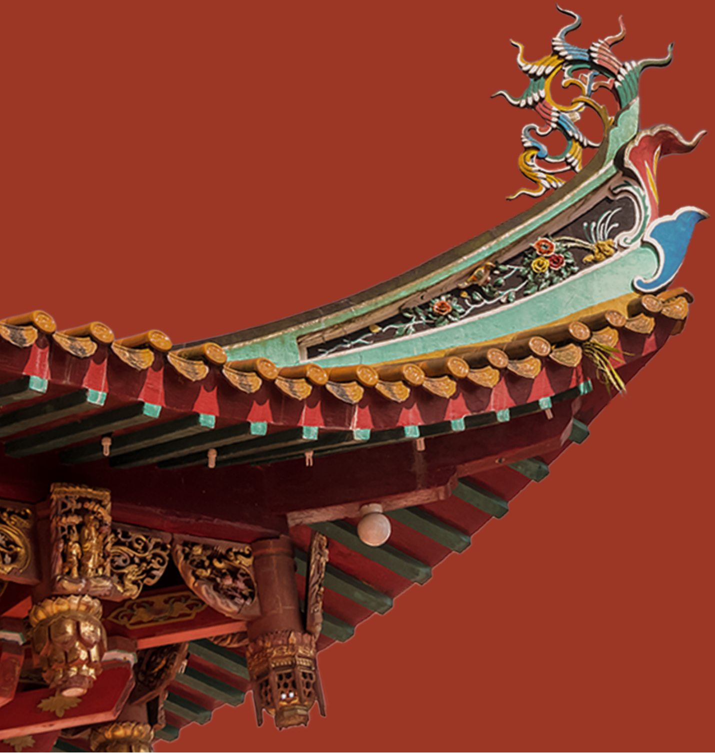
Computational Generation of Chinese Noun Phrases

Guanyi Chen



COMPUTATIONAL GENERATION OF CHINESE NOUN PHRASES

GUANYI CHEN

Natural Language Processing Group,
Department of Information and Computing Sciences,
Faculty of Science,
Utrecht University

SUPERVISED BY
PROF. DR.KEES VAN DEEMTER,
DR. CHENGHUA LIN

Promoter:	Prof.dr. C.J. van Deemter
Copromoter:	Dr. C. Lin
Assessment Committee:	Prof.dr. Y.S. Vinter Seggev
	Prof.dr. E. Reiter
	Prof.dr. T. Zhao
	Prof.dr. E.J. Krahmer
	Dr. B. Basciano

SIKS Dissertation Series No. 2022-12
The research reported in this thesis has been carried out under the auspices of SIKS, the Dutch Research School for Information and Knowledge Systems.

Printed by ProefschriftMaken | www.proefschriftmaken.nl

ISBN 978-90-393-7453-5

Computational Generation of Chinese Noun Phrases

Computationele Generatie van Chinese Noun Phrases
(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor aan de
Universiteit Utrecht
op gezag van de
rector magnificus, prof.dr. H.R.B.M. Kummeling,
ingevolge het besluit van het college voor promoties
in het openbaar te verdedigen op

dinsdag 12 april 2022 des ochtends te 10.15 uur

door

Guanyi Chen

geboren op 27 januari 1993
te Beijing, China

Promoter: Prof.dr. C.J. van Deemter
Copromoter: Dr. C. Lin

Eventueel: Dit proefschrift werd (mede) mogelijk gemaakt met financiële steun van China Scholarship Council

Acknowledgements

This thesis could not have been completed without the guidance and support from people around me. First and foremost, I would like to thank my supervisor, Kees van Deemter for his guidance and encouragement on not only my research but also my life in Aberdeen and Utrecht in the past four years, which enabled me to enjoy my life as a PhD candidate to the full. Working with Kees has taught me the way of doing science, writing science, and respecting science.

I also want to thank my second supervisor Chenghua Lin. Though we could only meet virtually after I moved to Utrecht from Aberdeen, Chenghua continuously provided me with valuable suggestions and enormous support.

Many thanks also go to the members of the assessment committee, for their reviewing my thesis and providing constructive comments. I am deeply grateful to have both computer scientists, Ehud Reiter, Emiel Krahmer and Tiejun Zhao, and linguists Yoad Winter and Bianca Basciano on the committee. This helps me to improve my thesis from both prespective.

A special thank-you goes to Albert Gatt, for his invaluable comments, and Yan Ma, for her proofreading during the writing of this thesis.

My gratitude also goes to my dearest friends in Utrecht: Yupei Du, Qixiang Fang, Lin Li, Danping Wen, Fahime Same, Yang Hu, Tengfeng Li, Wenrui Cao, Zonglin Tian, Yujie Ma, and Man Hu; in Aberdeen: Ruizhe Li, Xiao Li, Chaozheng Wang, Rui Mao, Tianju Wang, and Haoyu Li; in Beijing: Hanqing Zhao, Chen Zhao, Huanyu Ma, Tailun Song, Jinyi Guo, Hao Yuan, Jisi Teng, Mengyu Gao, Jingwen Zhao, Jiyang Zhang, Yinhe Zheng, Xin Liu, Jian Yang, Pinyu Su, and Jinge Yao. Without them, my studies and research work would be way less enjoyable and colourful.

I thank sincerely all my co-authors. Next to the names above, these are Matthew Collinson, Song Liu, Xuan Zhu, Silvia Pagliaro, Louk Smalbil, Minlie Huang, Emiel van Miltenburg, Wei-Ting Lu, Xutan Peng, Mark Stevenson, Jani Jarnfors, Rint Sybesma, Chengkun Zeng, and Jawwad Baig.

I also thank all members in the Utrecht Natural Language Processing group and the Computational Linguistics in AberdeeN (CLAN) research group.

Finally, I take this opportunity to express my deepest gratitude to my parents, Yanping Chen and Ning Wang, and my grandmother, Ying Li. Their love and care have been the inspiration for every turn in my life.

为者常成，行者常至。

— 晏子

Contents

CHAPTER 1

Introduction

This thesis examines some key aspects of the Mandarin Noun Phrase from a computational perspective. More specifically, we will use Natural Language Generation algorithms to shed light on the way in which Noun Phrases in Mandarin are employed to express information. As part of this exploration, we will sometimes compare Mandarin with other languages. To introduce the reader to the main issues in this area, we will first say a few informal words about Natural Language Generation (§1.1), about Noun Phrases (§1.2), about Computational Generation of Noun Phrases (§1.3), and about the hypothesis that Mandarin is a "cool" language (§1.4).

1.1 Natural Language Generation

Natural Language Generation (NLG) systems take non-linguistic data as inputs and produce texts in natural languages automatically by computer programs (see §2.1 for a more detailed review about NLG techniques). For example, given meteorological data, an NLG system can generate weather reports. Classic NLG pipeline roughly divides the generation process into three stages: document-planning (i.e., deciding what to say), micro-planning (i.e., deciding how to say), and surface realisation (i.e., realising plans into their surface form).

It has been pointed out that NLG systems can be broadly grouped into two categories: practical NLG and theoretical NLG. Practical NLG, as the name suggests, focuses on building NLG systems that have practical values, such as those that generate weather reports, news, medical reports, and so on. Systems of this kind often start with researching user requirements (e.g., reporting the weather in ways that can make more users understand the reports). Theoretical NLG aims to mimic the way in which human beings speak, aiming at reaching a better understanding of human language use. NLG work of this kind seeks to understand human patterns of speaking and designing NLG algorithms[1] that imitate those patterns as closely as possible. Practical and theoretical NLG can sometimes go hand

[1] Such algorithms are called "product models" (Sun, 2008; Vicente & Wang, 1998), in which all that matters is the mapping from inputs to outputs. These algorithms are different from "process models", which model *the manner* in which the mapping from inputs to outputs comes about.

in hand – with one and the same algorithm being used in both – yet they can be thought of as slightly different enterprises, which pursue different aims.

1.2 Noun Phrases

Noun Phrases (NPs) in Natural Languages have two important functions: referring and quantifying[2]. When an NP is used for referring, it is called a *Referring Expression* (RE). Consider the following RE:

(1) the student

It refers to Tom, broadly speaking, if and only if two conditions are fulfilled: 1) Tom is a student; and 2) Tom is the only student around. The second condition, in other words, says Tom is the only student in the *context*. The *Production* of REs is widely thought to be non-deterministic. Given the same situation, speakers produce REs differently. For example, to refer to Tom, who is a student wearing glasses, in addition to (1), one could say any of the following REs:

(2) a. Tom
b. the student Tom
c. he
d. the student who wears glasses

In part, the preference of speakers can be influenced by whether there is only one Tom in the context, whether Tom is the only referent that can be referred by pronoun "*he*" in the context, whether Tom is first mentioned in the discourse, whether the RE occurs in subject position, and so on. None of these factors or combinations of factors by itself can decide which RE should be produced, and it is still not fully understood what factors influence a speaker's preference.

Analogous to REs, NPs in their quantification function is called *Quantified Expressions* (QEs), for example:

(3) most students

Variation appears in both understanding and production of QEs. Suppose there are totally 100 students, for some listeners, when connecting (3) with a verb phrase and saying (4), it means approximately 60 out of 100 students wear glasses. Whereas, for some other listener, it might mean 80 out of 100 students wear glasses.

(4) most students wear glasses

For producing QEs, given a situation where 81 out of 100 students wear glasses, one might say any of the following:

(5) a. Most students wear glasses.
b. Almost all students wear glasses.
c. About 80% of the students wear glasses.

[2] NPs can have other functions than reference and quantification. In particular, it can be argued that indefinite NPs and bound anaphors neither refer nor quantify. Nonetheless, reference and quantification are often regarded as the two main functions that NPs can have. See for example Kamp and Reyle (1993) for an integral account.

d. More than three-quarters of the students wear glasses.

Same as REs, studying which factors influence the production of QEs is also valuable.

1.3 Computational Generation of Noun Phrases

Computational generation of NPs (e.g., REs and QEs) uses NLG algorithms to produce NPs. It is about NLG techniques in the micro-planning and the surface realisation stages in the NLG pipeline. These techniques can be categorised into practical ones and theoretical ones. Practical aspects include matters like deciding the syntactic structures of NPs, and deciding which pronoun to use (“*he*” or “*she*”). Theoretical aspects involve issues like investigating factors that influence the choices of expressions in (2) and (5), and modelling such choices computationally.

1.4 Coolness and the Trade-off between Brevity and Clarity

Ever since Grice (1975), linguists have been aware that, when we speak, we trade-off clarity against brevity. This idea has been put forward in particular details in relation to reference (Khan et al., 2006). For instance, for the use of REs, if one intends to be more clear, then s/he has to mention more information about the intended referent, which makes the RE longer, and, therefore, breaches brevity. The reverse is also true.

It has been suggested that East Asian languages (e.g., Mandarin) handle the trade-off between brevity and clarity differently to those of Western Europe (e.g., English; Newnham (1971)). Consequently, in this thesis, we will be interested in learning how this trade-off affects the production of NPs in Mandarin, and the design of Mandarin NLG systems. One major linguistic theory that closely ties to this trade-off is the theory of C.-T. J. Huang (1984), where he suggested Mandarin allegedly leaning more towards brevity, and relying more on communicative context for disambiguation compared to English. Inspired by the “hot-cool” division of media (McLuhan, 1964), C.-T. J. Huang (1984) categorised languages into *cool* languages (i.e., languages that rely more on context) and *hot* languages (i.e., languages that rely less on context). The evidence he provided was focused on the differences between the use of *anaphora* in cool and hot languages. Specifically, Mandarin makes liberal use of zero pronouns (ZP). To exemplify the use of zero pronouns in Mandarin, consider the question:

(6) 你 今天 看见 比尔 了 吗？
nǐ jīntiān kànjiàn bíěr le ma
Did you see Bill today?

Instead of answering “我 看见 他 了 ” (wǒ kànjiàn tā le; *I saw him*), Mandarin speakers often choose a shorter alternative:

(7) ∅ 看见 ∅ 了 。
kànjiàn le
∅ saw ∅.

Here the ∅ symbol indicates the place from where a pronoun appears to have been “dropped" from a full sentence. In this example, the pronouns in both subject and object positions are dropped.

Later on, the coolness theory was extended to cover other major components in Mandarin NPs. Mandarin NPs use determiners and classifiers to express definiteness and plurality. For example, in (8), the determiner “这” (zhè; *these*) and the classifier “些” (xiē) help to form a definite plural NP.

(8) 这 些 书
zhè xiē shū
these books

However, the plurality and definiteness of Mandarin NPs are often not (or not explicitly) specified. For example, a bare noun in Mandarin can be either definite or indefinite and either singular or plural. For example, the NP “书” (shū) can be translated as any of: “*a book*”, “*the book*”, “*books*”, and “*the books*”. Context is needed for deciding which translation is proper (van der Auwera & Baoill, 1998). More discussion about Huang’s coolness theory can be found in §3.1.

1.5 Research Questions

This thesis explores how Mandarin speakers produce NPs and how to build computational models accordingly. Specifically, we came up with the following research question.

> **Main Research Question** To what extent are computational models able to determine what to say and how to say it in Mandarin noun phrases?

To answer this question, we also come up with the following sub-questions.

> **Research Question 1** What noun phrases do Mandarin speakers produce, and how do speakers realise them?

> **Research Question 2** Does the theory of “Coolness” hold in Mandarin and how does it affect the design of computational models of Mandarin noun phrases?

> **Research Question 3** How to build computational models of Mandarin noun phrases and how well do they mimic human behaviours?

In this thesis, we study two types of NPs in Mandarin: referring expressions and quantified expressions. We focus on two stages in the NLG pipeline: micro-planning and surface realisation.

To build computational/NLG models of NPs in Mandarin, we need to first know what are the characteristics of the Mandarin language production and how are they different from other languages, such as English (Research Question 1). These require us to experiment on NPs in both Mandarin and at least one language other than Mandarin, which, in this thesis, is English since English NPs have been widely studied in the past. When we have a clear picture of how Mandarin and English speakers use NPs, we can ask ourselves: do the results validate the Coolness theory of C.-T. J. Huang (Research Question 2)? We are curious about all interpretations of Coolness introduced in §1.4: from the use of ZPs to the clarity-brevity trade-off.

After we have answers to the questions above, we investigate how these matter the way we build computational models/NLG systems. Once the models are built, we need to find a way to evaluate how well do they mimic the production process of Mandarin NPs (Research Question 3).

The contents of the present thesis consist of three parts: the first two parts correspond to the two types of NPs (i.e., REs and QEs) we are interested in respectively, and the last part focuses on the realisation issues. We hereby specify each research question above with respect to the subject matters of each part. In the course of introducing what we will do in this thesis, we mark each work with the notation "[T]" or "[P]" indicating the work emphasises theoretical NLG or practical NLG. "[T, P]" means the work can be seen as a hybrid of theoretical NLG and practical NLG. The distinctions implied by this kind of labelling are not always clear cut, but we believe they are a useful indication of the kind of contribution that the different parts of our work aim to make.

1.5.1 Generating Referring Expressions

In the first part, we focus on the production of REs. Indeed, there are two different kinds of Referring Expression Generation (REG) tasks: one-shot REG (which generates REs individually, in isolation from any linguistic context) and REG in Context (which generates REs from linguistic contexts). More details about the definitions of them and their differences can be found in §2.2.

One-shot Referring Expression Generation

We study one-shot REs in Chapter 4 and the above three research questions can be detailed as follows.

What are the characteristics of Mandarin speakers' use of one-shot referring expressions and how do they differ from English speakers? Fortunately, there exist corpora for both Mandarin REs (van Deemter et al., 2017) and English REs (van Deemter, Gatt, Sluis, et al., 2012), namely MTUNA and ETUNA, and, more importantly, they were built following a very similar experimental setting. In §4.4, we conduct a detailed analysis of the use of REs in MTUNA corpus and an initial comparison study between REs in MTUNA and ETUNA [T][3].

Are Mandarin speakers more likely to produce briefer referring expressions and less clear referring expressions than English speakers? For this question, we are concerned with two linguistic phenomena: one is the use of TYPE in REs. The other is the use of over-specifications and under-specifications. To better understand the use of over- and under-specifications and compare the use of them in different languages, we propose a new perspective of specifications that re-defines and sub-categorises over- and under-specification, and that allows quantitative analysis in §4.4. We then apply this new perspective on both MTUNA and ETUNA corpus [T].

How well can REG algorithms model referring expressions in Mandarin? In §4.5, we annotate the semantics of both MTUNA and ETUNA corpora and examine a number of classic cognitive-inspired REG algorithms on them [T]. We check whether the performance of each algorithm is in line with our expectations: algorithms that favour brevity works

[3] Recall that [T] stands for theoretical NLG and [P] stands for practical NLG

better on MTUNA than other algorithms. For each algorithm, we also conduct a systematic comparison between how close its behaviours are to human behaviours in MTUNA and in ETUNA.

Referring Expression Generation in Context

Subsequently, we study REG in Context in Chapter 5. In a similar vein, building on the three research questions, we are curious about the following issues.

How do Mandarin speakers use zero pronouns? In §5.2, we investigate several factors that influence the use of ZPs in a large scale Mandarin Dataset.

How do zero pronouns affect the design of referring expression generation in context algorithms? Also in §5.2, we show that we can model the use of ZP in a similar way as modelling pronominalisation in other languages [T]. In §5.4, we show that the existence of ZP introduces an extra option when deciding referential forms (which is a sub-task of REG in Context).

To what extent are computational models able to model the use of zero pronouns in Mandarin? In §5.2, we tackle the task using the rational speech act model by assuming that speakers tend to choose a ZP if it is salient enough for successful communication. In §5.3 and 5.4, we attempt to model the task of referential form selection (RFS) in both Mandarin (which includes an option of ZP) and English (which does not consider ZP as an option) using deep learning techniques [T, P]. In each of these two sections, we also conduct interpretability research to understand what linguistic information has been learnt by these black-box deep learning based models [T, P]. In §5.4, we compare the results of modelling RFS in Mandarin and English.

1.5.2 Generating Quantified Expressions

In the second part of the thesis, we study the use of QEs. Instead of QEs, we are concerned with descriptions consisting of multiple QEs so that we can investigate the use of QEs in complex situations. We call descriptions as such quantified descriptions (QDs). Since this subject matter has not been explored before, we start our study with QDs in English. The research questions are elaborated as follows.

How do Mandarin and English speakers use quantified expressions? To understand the quantifier use, we conduct elicitation experiments of QDs in both English (§6.2) and Mandarin (§6.3), which yield the QTUNA and the MQTUNA corpora respectively [T]. We analyse the QDs in each corpus.

How do Mandarin and English speakers use quantified expressions differently? In §6.3.5, we compare the QDs in QTUNA and MQTUNA corpora. Regarding the coolness, we focus on the completeness of QDs, the use of vagueness, and the way how speakers express plurality [T].

How to model quantified descriptions computationally? In §6.4, we propose two algorithms that generate QDs. We evaluate their performance in producing English QDs and discuss how they can be adapted to produce Mandarin QDs [T].

1.5.3 Surface Realisation

The last part of the thesis is about surface realisation. Since almost all sub-tasks in surface realisation are inherently practical, for some research questions we merely focus on one specific sub-task: classifier selection. We adopt research questions as follows.

How do Mandarin speakers use classifiers? In §7.4, we sample a bank of sentences from a corpus for the task of classifier selection. For each sample, we ask human participants to decide the classifiers given its contexts. We then compare the participants' selections with the reference answers to see how well can human beings accomplish such a task [T].

How does surface realisation of Mandarin different from that of English? Before constructing a realiser in Mandarin, we discuss how should a Mandarin realiser should be different from an English realiser in terms of morphology and syntax in §7.2.

How to build a Mandarin realiser? In §7.2, we build a realisation engine for Mandarin based on simpleNLG (Gatt & Reiter, 2009), namely simpleNLG-ZH [P]. We manually evaluate simpleNLG-ZH on generating REs in Mandarin. We also test a number of data-driven approaches on the task of classifier selection in §7.3 and compare its behaviour with that of human beings (P, T).

1.6 Methodology

Experiments in this thesis can be roughly categorised into ones that examine computational models and ones that involve human participants.

1.6.1 Human Experiments

In this thesis, we conduct human experiments aiming to understand how speakers use languages (i.e., elicitation experiments), and to evaluate the qualities of machine-generated texts (i.e., human evaluation).

Elicitation Experiments. To understand the language use, we conduct elicitation experiments where, given a set of situations (which could be text, scene, and so on), each participant is asked to write something in accordance with the instruction. Elicitation experiments can often yield corpora which can later be analysed and used for building and evaluating computational models.

Human Evaluation. Human evaluation is to ask human judges to score aspects like fluency, informativeness, naturalness, and acceptability for each machine-generated text.

Results Analysis. In this thesis, the results of human experiments are always analysed through both hypothesis testing and post-hoc observations. Hypotheses are made after designing each experiment but before conducting it. The hypotheses are then tested building on the results of the experiments. Post-hoc observations are phenomena that are found after looking into the experimental results. Additionally, for analysing data in a more informative way, we often need to annotate the dataset before testing the hypotheses or making observations.

1.6.2 Computational Modelling

Models. In this thesis, we consider a wide range of computational models. This includes rule-based models, statistical models and deep learning based models.

Evaluation and Analysis. Each computational model is validated by either automatic evaluation or human evaluation if the model is aiming at producing texts. The effectiveness of these models is further confirmed by comparing the performance of human beings to state-of-the-art models. Additionally, since deep learning models are, to a large extent, still black-boxes, we also conduct interpretability studies to interpret their behaviours.

1.7 Outcomes of this Thesis

We hereby list datasets, code, software and publications related to this thesis.

1.7.1 Datasets

The annotated MTUNA and TUNA Corpora are annotated versions of the TUNA corpus (van Deemter, Gatt, Sluis, et al., 2012) and the MTUNA corpus (van Deemter et al., 2017). Each RE is annotated with its semantics (G. Chen & van Deemter, 2020) and whether it is an over-specification or an under-specification (G. Chen & van Deemter, 2021). The annotated corpora are available at https://github.com/a-quei/mtuna-annotated.

The QTUNA and the MQTUNA Corpora are corpora of QDs in English and Mandarin respectively. QTUNA contains 1,414 QDs for 30 scenes (G. Chen, van Deemter, Pagliaro, et al., 2019) and MQTUNA contains 465 QDs for 15 scenes. Each QD is annotated with its quantifier use, completeness, and correctness. The datasets are available at https://github.com/a-quei/qtuna.

1.7.2 Code and Software

Neural Referential Form Selector is a referential form selector that involves a number of neural network-based RFS models (G. Chen et al., 2021). It also includes several corresponding probing classifiers aiming at understanding what information each model learns. The code is available at https://github.com/a-quei/probe-neuralreg.

Quantified Description Generators are implementations of the two QD generation algorithms introduced in G. Chen, van Deemter, and Lin (2019). The code is available at: https://github.com/a-quei/quantified-description-generation.

SimpleNLG-ZH is a realisation engine following the tradition of simpleNLG for Mandarin (G. Chen et al., 2018c). The software is available at https://github.com/a-quei/simplenlg-zh.

1.7.3 Publications

Publications that are related to this thesis include:

1. Chen, G., van Deemter, K., & Lin, C. (2018a). Modelling pro-drop with the rational speech acts model. *Proceedings of the 11th International Conference on Natural Language Generation*, 159–164. https://doi.org/10.18653/v1/W18-6519

2. Chen, G., Same, F., & van Deemter, K. (2021). What can neural referential form selectors learn? *Proceedings of the 14th International Conference on Natural Language Generation*, 154–166. https://aclanthology.org/2021.inlg-1.15

3. Chen, G., & van Deemter, K. (2020). Lessons from computational modelling of reference production in Mandarin and English. *Proceedings of the 13th International Conference on Natural Language Generation*, 263–272. https://www.aclweb.org/anthology/2020.inlg-1.33

4. Chen, G., & van Deemter, K. (2021b). Varieties of specification: Redefining over- and under-specification for an enhanced understanding of referring expressions. *Journal Paper in Preparation*

5. Chen, G., van Deemter, K., & Lin, C. (2019). Generating quantified descriptions of abstract visual scenes. *Proceedings of the 12th International Conference on Natural Language Generation*, 529–539. https://doi.org/10.18653/v1/W19-8667

6. Chen, G., van Deemter, K., Pagliaro, S., Smalbil, L., & Lin, C. (2019). QTUNA: A corpus for understanding how speakers use quantification. *Proceedings of the 12th International Conference on Natural Language Generation*, 124–129. https://doi.org/10.18653/v1/W19-8616

7. Chen, G., & van Deemter, K. (2021a). Computational modeling of quantifier use: Elicitation experiments, models, and evaluation. *Journal Paper in Preparation*

8. Chen, G., van Deemter, K., & Lin, C. (2018b). SimpleNLG-ZH: A linguistic realisation engine for Mandarin. *Proceedings of the 11th International Conference on Natural Language Generation*, 57–66. https://doi.org/10.18653/v1/W18-6506

9. Jarnfors, J., Chen, G., van Deemter, K., & Sybesma, R. (2021). Using BERT for choosing classifiers in Mandarin. *Proceedings of the 14th International Conference on Natural Language Generation*, 172–176. https://aclanthology.org/2021.inlg-1.17

A full list of publications can found in my Curriculum Vitae (see Appendix C).

CHAPTER 2

Background

2.1 Natural Language Generation

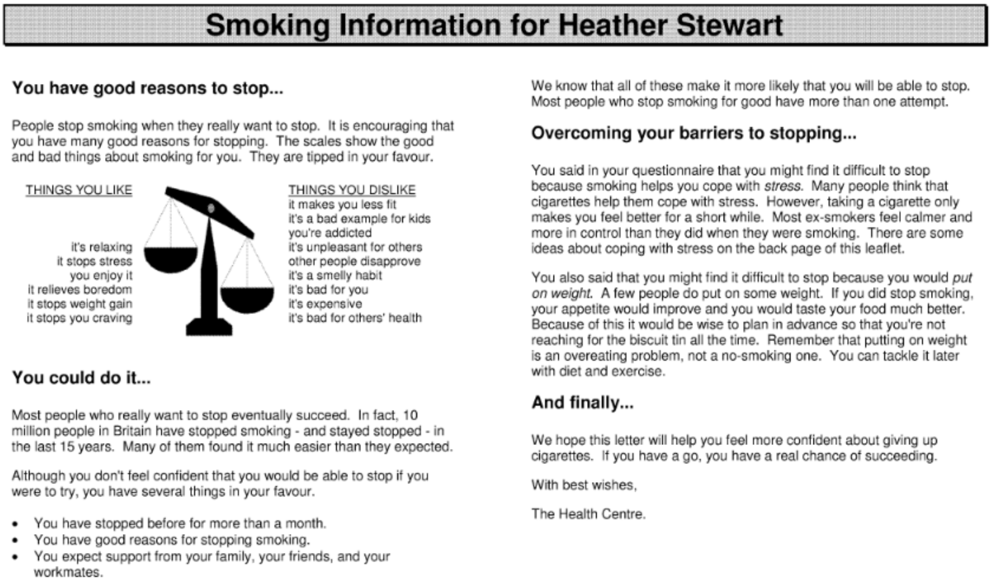
There have been different opinions on defining *Natural Language Generation* (NLG) precisely. For instance, early NLG surveys/books (e.g., Reiter and Dale (2000)) characterised it as:

> the sub-field of artificial intelligence and computational linguistics that is concerned with the construction of computer systems that can produce understandable texts in English or other human languages from some underlying non-linguistic representation of information.

NLG systems following this "narrow" definition take *non-linguistics* (e.g., table, graph, and database) as inputs and output natural language accordingly. Examples of this kind include systems that generate soccer reports (D. L. Chen & Mooney, 2008; Theune et al., 2001), news (Leppänen et al., 2017), weather reports (Reiter et al., 2005), etc.

Nevertheless, this definition is "narrow" because it excludes applications that generate natural language as well but take language as inputs, such as machine translation, summarisation, text simplification. The systems using non-linguistic data as inputs are called *data-to-text generation*, while the ones using linguistic inputs are *text-to-text generation*. The "Broader" definition (Gatt & Krahmer, 2018) demonstrates that NLG consists of either data-to-text generation or text-to-text generation.

In this thesis, we use the term NLG referring to the data-to-text generation and the review in this section will be all about data-to-text generation. As aforementioned in §1.1, NLG can be categorised as practical NLG and theoretical NLG. Recall that practical NLG is about building NLG models and designing algorithms that have practical usages. Commercial NLG systems (Dale, 2020), such as the aforementioned news generation and report generation systems, belongs to this category. In contrast, theoretical NLG is about understanding how human beings produce languages, where algorithms are built for mimicking language production. Much work so far in this category focuses on one specific module of: referring expression generation. This section will be mainly about the practical NLG work and, in the next section, we turn to referring expression generation.



Smoking Information for Heather Stewart

You have good reasons to stop...

People stop smoking when they really want to stop. It is encouraging that you have many good reasons for stopping. The scales show the good and bad things about smoking for you. They are tipped in your favour.

THINGS YOU LIKE
it's relaxing
it stops stress
you enjoy it
it relieves boredom
it stops weight gain
it stops you craving

THINGS YOU DISLIKE
it makes you less fit
it's a bad example for kids
you're addicted
it's unpleasant for others
other people disapprove
it's a smelly habit
it's bad for you
it's expensive
it's bad for others' health

You could do it...

Most people who really want to stop eventually succeed. In fact, 10 million people in Britain have stopped smoking - and stayed stopped - in the last 15 years. Many of them found it much easier than they expected.

Although you don't feel confident that you would be able to stop if you were to try, you have several things in your favour.

- You have stopped before for more than a month.
- You have good reasons for stopping smoking.
- You expect support from your family, your friends, and your workmates.

We know that all of these make it more likely that you will be able to stop. Most people who stop smoking for good have more than one attempt.

Overcoming your barriers to stopping...

You said in your questionnaire that you might find it difficult to stop because smoking helps you cope with *stress*. Many people think that cigarettes help them cope with stress. However, taking a cigarette only makes you feel better for a short while. Most ex-smokers feel calmer and more in control than they did when they were smoking. There are some ideas about coping with stress on the back page of this leaflet.

You also said that you might find it difficult to stop because you would *put on weight*. A few people do put on some weight. If you did stop smoking, your appetite would improve and you would taste your food much better. Because of this it would be wise to plan in advance so that you're not reaching for the biscuit tin all the time. Remember that putting on weight is an overeating problem, not a no-smoking one. You can tackle it later with diet and exercise.

And finally...

We hope this letter will help you feel more confident about giving up cigarettes. If you have a go, you have a real chance of succeeding.

With best wishes,

The Health Centre.

Figure 2.1: A screenshot of an output of the STOP (Reiter et al., 2003) system.

2.1.1 Pipeline NLG

Analogous to many other software systems, practical NLG systems are often decomposed into a number of modules. This (i.e., modularisation) makes an NLG system easier to be maintained and modified. Up to now, most practical NLG systems follow a classic pipeline architecture proposed by Reiter and Dale (2000) since it not only provides a well-defined interface between the components in the architecture/pipeline but also allows each component to be reused independently. Therefore, in this thesis, the term "pipeline NLG" will refer to this specific architecture by Reiter and Dale (2000), details of which will be specified in this subsection.

Overview of the Pipeline

Before detailing each component in the pipeline, we would like to elaborate on what kind of input and output a practical NLG system needs. The exact input and output formats of NLG systems vary with respect to their usage. Seemingly, it is quite certain that the output of an NLG system should be "text". Nonetheless, simply yielding text without concerning display issues in accordance with the users has been proved to be inadequate (Reiter & Dale, 1997). Instead, aspects like format, online display and speech output are worth to be paid attention to in order to render the text in a more useful form. Which form is useful depends on the application. For instance, if the application needs merely text that presents simple information (e.g., weather report (Reiter et al., 2005)), then there is no need for the system to consider structure above sentence level of the generated text. Conversely, if the application is responsible for a complex task (e.g., persuading smokers to stop smoking (Reiter et al., 2003)), then the output needs to not only be well structured but also include graphics (as shown in Figure 2.1). As for the input, it varies in a more

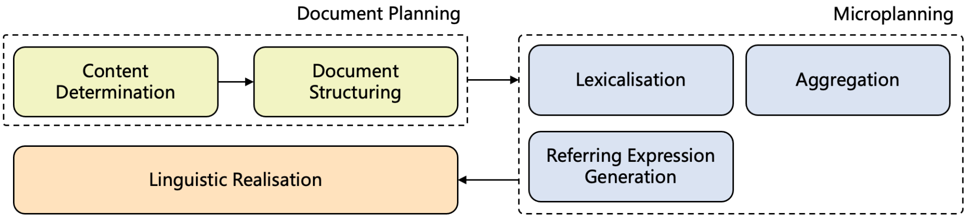


Figure 2.2: The diagram of components in the NLG pipeline.

substantial way. For example, it could be a table, a database, an image and so on. In general terms, Reiter and Dale (1997) characterised the input of an NLG system as a four-tuple, consisting of the *Knowledge Source* to be used, the *Communication Goal* to achieve, a *User Model*, and a *Discourse History*. For example, to generate weather reports, the knowledge source is the database storing necessary weather data for the report. The communication goal is to summarise the weather in a certain period (e.g., a month or a week). Since the weather reports are always not being personalised, there is no explicit user model. Likewise, since the weather report generator is often a single-interaction system, there is also no explicit discourse history.

Roughly speaking, as illustrated in Figure 2.2, Reiter and Dale (2000) proposed a architecture consisting of three major stages: *Document planning*, *Micro-planning*, and *Surface Realisation*. Concretely, document planning is responsible for determining the content and structure of a document. In the micro-planning stage, the system decides in which way (e.g., words and syntactic structure) the determined/planned content will be expressed. At length, a surface realiser maps the abstract representations outputted from the micro-planner into actual text. These three main stages can be further divided into 6 sub-tasks: *content determination*, *document structuring*, *lexicalisation*, *aggregation*, *referring expression generation*, and *surface realisation*.

Document Planning

Document Planning, which is also named *macro-planning* (opposite to micro-planning), has two steps: content determination and document structuring.

Content Determination. As the very first stage of an NLG system, content determination decides "what to say". In other words, at this stage, an NLG system determines which information should be included and which should not. For example, when generating a weather report of a month, it might not be a good idea to enumerate all temperatures in that month. A better strategy is to describe merely the trends of the temperature change as well as the maximum and minimum temperatures. This said, the input data needs to be pre-processed and interpreted before deciding the contents. Therefore, some data-to-text generation architectures separate these pre-processing steps from the content determination. Reiter (2007) introduces two pre-processing steps: *signal analysis* and *data interpretation*.

In earlier days, such a task was accomplished by rule-based or template-based models (Mellish et al., 2006) where the following factors need to be considered (Reiter & Dale, 2000): 1) communicative goal: different communicative goals require different information.

For instance, a monthly weather report would be different from a daily weather report; 2) target audience: contents should be selected with respect to the assumed or known characteristics of the audience; 3) information source: what is worth saying depends on how much and what information is available to the audience.

In recent years, researchers head to data-driven methods, most of which assume content determination as a sequence labelling task. For instance, Barzilay and Lee (2004) leveraged Hidden Markov Models to model topic shift, where each hidden state represents a topic in the document plan. Beyond that, Barzilay and Lapata (2005) proposed to select contents collectively, where all candidates are considered simultaneously for selection. One plight of these data-driven content planners is that they require training sets where the alignments between plan items (e.g., topics) and texts are annotated. However, on the one hand, such annotated datasets are hard to be built. On the other hand, these alignments are not necessary for one-to-one mappings. For example, Koncel-Kedziorski et al. (2014) found that soccer events in data and sentences in associated soccer reports do not one-to-one correspond. To solve this issue, much work focuses on automatically learning alignments between data and text.

Document Structuring. The second step of document planning is document structuring. It is responsible for organising the presentation of the selected contents. The resulting presentation should make the final generated text coherent and fluent. For example, when generating soccer matches or weather reports, it is reasonable to produce one or several sentences for describing the general information in the very beginning. Although some NLG systems are only concerned with the order of the presentation of the information (e.g., Portet et al. (2009)), many researchers have suggested that there are more things than just sequencing should be done. Earlier work attempted to structure the document using hand-crafted domain-dependent rules (McKeown, 1992), which is hard to handle complex discourse relations. This was, to a large extent, solved by using the Rhetorical Structure Theory (RST, E. H. Hovy, 1993; Mann & Thompson, 1988; Scott & de Souza, 1990; Williams & Reiter, 2008).

Making use of data-driven methods is also possible. Machine learning techniques can not only help document structuring (Althaus et al., 2004; Dimitromanolaki & Androutsopoulos, 2003) but also enable simultaneous content selection and structuring (Duboue & McKeown, 2003). There are also approaches for information ordering, such as Barzilay and Lee (2004) and Lapata (2006).

More recently, people in NLG tend to make use of deep neural network for planning content, which will be elaborated in §2.1.3.

Micro-planning

Micro-planning is responsible for deciding which word, syntactic structure, and so forth will be used given the selected and structured contents. As shown in Figure 2.2, it contains three sub-tasks: aggregation, lexicalisation, and referring expression generation.

Aggregation. Micro-planner takes the outputs of the document planner which are usually trees (e.g., RST trees) indicating the structure of the document. However, not all the items in the tree need to be realised in separate sentences, some of which need to be aggregated to make the generated text more fluent and readable (H. Cheng & Mellish, 2000; Dalianis, 1999). The whole process could be seen as an inverse process of sentence splitting in text simplification. For example, for information which could be expressed by two sentences:

(9) a. Yesterday was hot.
b. Yesterday was humid.

A sentence aggregator merges them into a single sentence: "*Yesterday was hot and humid*" to make it more fluent and readable. Reape and Mellish (1999) pointed out that aggregation could be categorised as: (1) syntactic aggregation: such as the aggregation happen in the above example; and (2) semantic aggregation: such as aggregating "*the chair and the table*" to "*the furniture*".

Same with other sub-tasks, research in sentence aggregation starts with using hand-crafted rules (Dalianis, 1999; E. Hovy, 1987; Shaw, 1998). More recent work moves towards data-driven methods. This includes approaches for either semantic aggregation, which sometimes be seen as a part of content selection (Barzilay & Lapata, 2006; H. Cheng & Mellish, 2000; Walker et al., 2001), or syntactic aggregation, which targets at reducing the redundancy of the generated text (Harbusch & Kempen, 2009; Kempen, 2009).

Lexicalisation. Lexicalisation is of choosing which words and syntactic structures should be used to express the selected content (Reiter & Dale, 2000). Lexicalisation matters because given a piece of information, there can be numerous ways to express it in natural language. Gatt and Krahmer (2018) used a scoring event in soccer match news generator as an example. For a scoring event, one could say any of the following:

(10) a. to score a goal
b. to have a goal noted
c. to put the ball in the net

The general target of lexicaliser is not only choosing the proper lexicon and syntax but also generating text with a certain amount of variations, which makes the generated texts interest more readers (Odijk, 1995). Theune et al. (2001) argued that

> Variation is propositional with the length of the generated text, and with the number of similar text to be read.

However, the variation in the generated texts is not the larger the better. For example, Reiter et al. (2005) pointed out that there is less such "idiolect" in weather reports than in soccer reports. Castro Ferreira et al. (2016) found that the variation of *Referential Form* depends on where in a text the variations occur.

The lexicalisation is hard in terms of three aspects: first, the lexical selection is often between semantically similar, near-synonymous (Edmonds & Hirst, 2002), or taxonomically related words (*animal* vs. *dog*, (Stede, 2000)). For example, for the selection of trend verbs (e.g., choosing between "*climb*" and "*soar*" in the sentence "*Microsoft's profit climbed 28%*"), there is a lot of overlaps in the usage of them (G. Chen & Yao, 2019). Given an underlying change (e.g., 20%), there are at least 10 different verbs that are judged appropriately by human readers.

Second, there is not always a crisp concept-to-word mapping for modelling lexicalisation (Power & Williams, 2012; Reiter et al., 2005). This is caused by the fact that most words in natural languages are vague. For instance, it is hard to have a crisp definition when we say an object is "*large*" and when we say an object is "*small*". This makes the choice of word depends on either its intended meaning or its context as the meanings of many vague expressions are context dependent (cf. Barwise and Perry (1981), Kennedy and McNally (2005) and van Deemter (2012)). To address vagueness in language generation, possible solutions include fuzzy logic (Ramos-Soto et al., 2016) and probabilistic logic (Dietz, 2017; Lassiter, 2009).

At length, there are huge variations in word usage (Reiter & Sripada, 2002). A natural way to handle such huge variations is to use machine learning. Up to now, machine learning has been used in the selection of colour words (Zarrieß & Schlangen, 2016), trend verbs (Smiley et al., 2016; D. Zhang et al., 2018), and words in weather reports (X. Li et al., 2016).

Referring Expression Generation. The task of referring expression (REG) is to generate a description of a referent that enables the reader/hearer to identify that referent in a given context (Reiter & Dale, 2000). In this section, our review focus on the REG system embedded in a practical pipeline NLG system whose "context" is natural language. [1]

As illustrated in §1.2, given a referent, there are multiple alternatives referring expressions. For example, when referring to "Joe Biden", we could say any of the following:

(11) a. Joe Biden
b. Mr. President (when this thesis is written)
c. Joe
d. Biden
e. He

The decision depends on the context surrounding the target referent. [2] The REG task looks analogous to the lexicalisation task, but it should have less variation (Castro Ferreira et al., 2016), which is because (in the language of Reiter and Dale (2000)) REG is

> a discrimination task, where the system needs to communicate sufficient information to distinguish one domain entity from other domain entities.

The task of REG contains two stages of choices. The first choice is the choice of referential form, which decides in which form the target referent should be realised. Most of the time, it is about a selection from four alternatives: pronoun, proper name, description, and demonstrative. Such a decision partly depends on whether the target entity is "focused" or "salient". In light of theoretical linguistics, there is a bank of factors that would influence the salient of a referent. For example, Chafe (1976) and Prince (1981) suggested that when a new referent is first introduced to the discourse, it is less likely to be referred to as a pronoun. More details about these factors will be discussed in §2.2.2. By making use of these factors, both rule based (Henschel et al., 2000) and data-driven approaches (Greenbacker & McCoy, 2009; Hendrickx et al., 2008) have been proposed for selecting the referential form. The development of this sub-task has been highly encouraged by a series of shared tasks on the Generation of Referring Expressions in Context (GREC, Belz et al., 2010).

[1] REG has also been used as a tool to understand human language production. "Context" in REG could also be visual scenes or images. Both of these two aspects will be discussed in §2.2. The REG task discussed in this section is also called "REG in Context" or "Discourse REG".

[2] Target referent (entity) is the referent (entity) that the REG module makes decisions on at the moment.

The second choice happens if the selected referential form is *description*, which is about deciding the content of the description. Unfortunately, due to the complexity of the task. The selection of referential content (in the task of REG in context) is often decided by simple one-to-one rules (i.e., each referent is associated with a single description). Most recently, "real" production is enabled by means of deep learning techniques (Cao & Cheung, 2019; Castro Ferreira, Moussallem, Kádár, et al., 2018; Cunha et al., 2020).

More background about REG related tasks, corpora, and algorithms can be found in §2.2.

Surface Realisation

The last stage is to map the plan into its well-formed surface form, i.e., surface realisation. Gatt and Krahmer (2018) categorised the surface realisation (or so-called linguistic realisation) techniques into three lines of approaches: human-crafted templates, human-crafted grammar-based systems, and statistical approaches. The task of surface realisation involves the use of the correct syntax and the generating of the right morphological forms.

Template-based approaches are always used in NLG systems that are small and favour less variation. One advantage of the template method is that the developers could have full control over the outputs at hand, which results in the NLG systems being safer. Thanks to this merit, many large-scale NLG systems and dialogue systems use template methods in their surface realiser. This approves that the template-based systems are able to handle complex situations, making it difficult to distinguish templates from more "real" NLG (van Deemter et al., 2005). More recently, researchers have started to consider marrying templates with data-driven approaches by learning templates automatically from corpora (Angeli et al., 2012; Kondadadi et al., 2013).

Grammar-based methods offer domain-independent alternatives. Most systems fell in this line of work make use of specific types of grammar. For example, KPML (Bateman, 1997) was built upon the systemic-functional grammar (Halliday et al., 2014). This makes these systems difficult to be used since they require developers to be familiar with those grammar formulations. In response to these, simple realisation engines which provide easy-to-use syntax and morphology operation APIs have been developed (a typical example of which is the SimpleNLG system developed by Gatt and Reiter (2009).

Like other modules, recent research on linguistic realisation attempts to use statistical approaches. Most statistical approaches are used in data-driven grammar-based realisers. Using statistical approaches can help the user of grammar-based realiser not care much about the grammar itself so that advanced grammar formulations can be used. For instance, OpenCCG (Espinosa et al., 2008; White & Rajkumar, 2009, 2012; White et al., 2007) uses Combinatory Categorial Grammar (Steedman, 2000) to build a broad coverage English surface realiser. Likewise, there have been realisers that use the Context-free Grammar (Belz, 2008), the Head-driven Phrase Structure Grammar (Carroll & Oepen, 2005; Nakanishi et al., 2005), the Lexical Functional Grammar (Cahill & van Genabith, 2006), and the Tree Adjoining Grammar (Gardent & Narayan, 2015). Additionally, statistical methods also allow the interaction better between the surface realiser and the micro-planner (Gardent & Perez-Beltrachini, 2017). In recent years, deep learning-based approaches have proved to be super effective in linguistic realisation, which will be discussed in §2.1.2 and §2.1.3

2.1.2 End2End NLG

Gatt and Krahmer (2018) distinguishes NLG systems into three architectures: 1) pipeline architecture (see §2.1.1); 2) planning based NLG, which views NLG as planning; and 3) end2end architecture, most of which uses machine learning techniques to output given the input directly.

They also pointed out that there are two major flaws: one is the *generation gap* (Meteer, 1991), indicating the mismatch between the strategic and tactical components. For example, Inui et al. (1992) described a situation where the planned sentence order in the document planning phase might cause ambiguity in linguistic realisation. The other is that the pipeline NLG does make some generation constraints hard to be realised. For instance, pipeline NLG is hard to follow length constraints, especially in early stages (Reiter & Dale, 2000), which is not a problem for End2End systems (Ficler & Goldberg, 2017).

Thanks to the recent development of *Deep Learning* techniques, deep learning-based End2End approaches have dominated the state-of-the-art research in the realm of NLG. It not only helps to produce more fluent and human-like texts but also makes certain tasks that are hard for pipeline NLG (e.g., image captioning, storytelling and so on) become easier. In contrast, these novel techniques also introduce new challenges to NLG systems, which will be discussed in §2.1.3. We will introduce the techniques behind Neural Natural Language Generation (NNLG) and their applications in this section. But before that, we start with the planning based and End2End NLG systems before the age of deep learning.

Beyond the Pipeline Architecture

Planning-based NLG. There has been a long tradition of tackling the Artificial Intelligence problem as planning. It is about identifying a sequence of actions to achieve a particular goal. When adopting this paradigm to NLG, the "particular goal" is a "communicative goal", which is one of the major inputs of an NLG system, and the "sequence of actions" is a series of production operations of language (Clark, 1996).

Building on the fact that all stages (i.e., document planning, micro-planning, and linguistic realisation) can be treated as planning problems, the boundaries between each task in the NLG pipeline get blurred. This is done by unifying the operations for deciding what to say and the operations for deciding how to say into a single set of operations. For instance, Heeman and Hirst (1995) proposed to use KAMP (Appelt & Appelt, 1992) to do REG, where the property selection and surface realisation are done as two sub-goals. Additionally, planning based methods also has their own advantage of being friendly with grammar (e.g., Bateman (1997) and Halliday et al. (2014)) and discourse (e.g., Mann and Thompson (1988)) formalism, but, meanwhile, they also bear the problem of low-speed (Koller & Petrick, 2011).

There also has been work on marrying planning with machine learning tools. Concretely, this line of work views achieving the communicative goal as a stochastic optimisation problem views the generation as a Markov Decision Process and uses the technique of Reinforcement Learning to model such a process (Lemon, 2008; Rieser & Lemon, 2009, 2011). In this way, we could handle the uncertainty caused by the trade-off between informativeness and brevity. Such a strategy has been used for generating restaurant recommendations (Rieser & Lemon, 2009) and referring expressions (Janarthanam & Lemon, 2014).

Early End2End NLG. The fundamental idea behind early End2End NLG is to formulise each sub-task of NLG as a classification task (Duboue & McKeown, 2003; Filippova & Strube, 2007) and optimise all tasks/classifiers jointly/globally. This includes work such as Konstas and Lapata (2013a), which unifies content selection and surface realisation based on the idea of Liang et al. (2009), who learn how to align database records and text segments using a hierarchical hidden semi-Markov generative model. On the basis of the same idea, Konstas and Lapata (2013b) extended Konstas and Lapata (2013a) to include induction rules of context-free grammar. By joint learning all components, the resulting architecture could suffer less from the problem of error propagation (i.e., the error caused by decision in earlier stages will propagate to later stages). In addition to learning alignment using probabilistic models, there are other solutions for joint learning, including Integer Linear Programming (Barzilay & Lapata, 2006; Lampouras & Androutsopoulos, 2013) and Imitation Learning (Lampouras & Vlachos, 2016).

Neural Natural Language Generation

In the past decades, Neural Network based models (or Deep Learning based models; Goldberg (2017) and Goodfellow et al. (2016)) are on their path to dominate our research in NLP. Neural Networks were designed to be good at learning representations through back-propagation (Rumelhart et al., 1986). When applying to NLP tasks, such a function has been proved to be effective for capturing grammatical information and meanings of words (Mikolov et al., 2013; Pennington et al., 2014). Beyond word, recurrent architectures, including the recurrent neural network (RNN, Rumelhart et al., 1986) and its extensions like long and short term memory units (LSTM, Hochreiter & Schmidhuber, 1997), and gated recurrent units (GRU, Cho et al., 2014), are designed for sequential modelling. They are capable of learning representations of sentences or even documents (Tang et al., 2015) and of language modelling (Mikolov et al., 2010). As figured in Gatt and Krahmer (2018), the very first application of deep learning for generating natural language is the work of Sutskever et al. (2011). They assessed the ability of character-based LSTM for generating grammatically correct English sentences.

However, what Sutskever et al. (2011) did was not "real" NLG, but generating random sentences from language models given prompts. For example, given a prompt "*ABC et al. (2008)*" the model is of generating the rest of a potential sentence: "*to be evaluated and motivated by providing optimal estimate*". "Real" Neural NLG (NNLG) that generates natural languages on the basis of certain contexts/inputs was firstly realised by means of conditional RNNs and the RNN Language Model (RNNLM, Mikolov et al., 2010). Building on this idea, Wen et al. (2015) proposed to use RNN to generate delexicalised sentences given dialogue representations. For example, given the following speech act:

(12) inform(name=Seven_Days, food=Chinese)

, where the speech act is "inform", indicating the generated sentence should present certain information and the two key-value pairs decide what the generated sentence need to say, an NNLG generates a delexicalised sentence:

(13) SLOT_NAME serves SLOT_FOOD .

It appears that what a potential NNLG needs to do is micro-planning and surface realisation. We hereby use this task as an example to introduce how RNN is used for NLG.

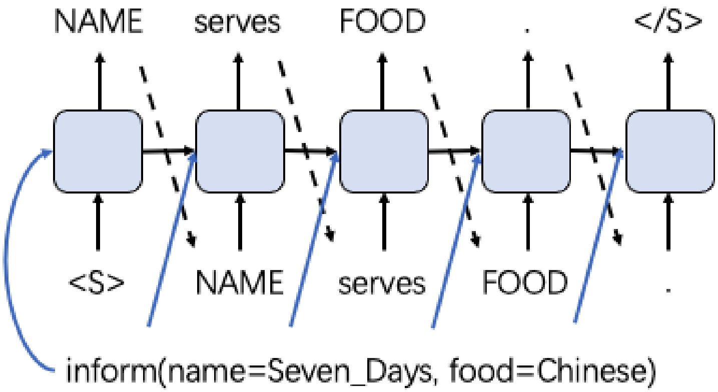


Figure 2.3: Diagram for NNLG from speech act using RNN (Wen et al., 2015).

Suppose we have an input speech act x, and we tend to generate a sentence from the starting symbol "$\langle S \rangle$" referred as w_0. These inputs are firstly mapped into their one-hot representations: $\boldsymbol{x}$ and $\boldsymbol{w}_0$. Figure 2.3 describes the procedure of how to use these inputs to generate outputs using RNN. During generation, at a time step t, the model produces the next token $\boldsymbol{w}_{t+1}$ based on the input $\boldsymbol{x}$ and decoded token at the previous step $\boldsymbol{w}_t$ with the following procedure:

1. Compute the hidden representation $\boldsymbol{h}_t$ with respect to the previous hidden representation $\boldsymbol{h}_{t-1}$, previous token $\boldsymbol{w}_t$, and the input $\boldsymbol{x}$, by:

$$\boldsymbol{h}_t = \sigma\left(W_h \boldsymbol{h}_{t-1} + W_w \boldsymbol{w}_t + W_x \boldsymbol{x}\right), \tag{2.1}$$

   where σ is the Sigmoid function[3]. W_h, W_w and W_x are trainable weights in the model. Note that Wen et al. (2015) proposed to inject input information at every time step is to ease the problem of vanishing gradient of RNNs (Pascanu et al., 2013);

2. Compute the probability distribution of the next token depending on the previous tokens and the input:

$$P(\boldsymbol{w}_{t+1} | \boldsymbol{w}_t, \boldsymbol{w}_{t-1}, ..., \boldsymbol{w}_0, \boldsymbol{x}) = \text{Softmax}\left(W_o \boldsymbol{h_t}\right) \tag{2.2}$$

   where the Softmax function is the normalised multi-class version of Sigmoid.

At length, at each step, Wen et al. (2015) proposed to sample a token w_{t+1} from the distribution $P(\boldsymbol{w}_{t+1} | \boldsymbol{w}_t, \boldsymbol{w}_{t-1}, ..., \boldsymbol{w}_0, \boldsymbol{x})$. The generation would stop if the end symbol $\langle / S \rangle$ is generated. Instead of random sampling, many other work on NNLG produces outputs in ways such as always selecting the word type that has the highest probability among

[3] The Sigmoid function is defined as:

$$\sigma(a) = \frac{1}{1 + e^{-a}}$$

$P(w_{t+1}|w_t, w_{t-1}, ..., w_0, x)$ (which is called greedy decoding. For more discussion about other advanced strategies (e.g., beam search), please check Holtzman et al. (2020)).

Given the results of a series of evaluation experiments, such a model perform well on the given task. Nevertheless, this model takes only simple input (i.e., a single speech act with a limited number of key-value pairs). When it comes to more realistic NLG tasks, whose inputs could be tables, a set of meaning representations, graphs and so on, the one-hot encoding used in Wen et al. (2015) is not competent to encode such inputs. Therefore, an encoder is needed to learn representations of complex inputs. This was done through the encoder-decoder architecture, which has become a standard solution for generating language. The encoder-decoder architecture was firstly introduced in Sutskever et al. (2014) with a form of sequence-to-sequence model (Seq2Seq), which targets text-to-text generation and assumes the inputs of the model follows a sequential structure. Since most NLG tasks do not have sequential inputs, much work applied Seq2Seq model on them by linearising the inputs into the sequential structure in the first place. We will introduce more details of the Seq2Seq Model below with an example of an End2End NLG task introduced in the E2E NLG Challenge (Dusek et al., 2018).

Sequence-to-Sequence Model

The task of E2ENLG[4] is of generating natural language given a set of meaning representations (MRs). For instance, given the following MRs:

(14) name[Juzzman],
eatType[coffee],
food[French],
priceRange[moderate],
rating[3/5],
area[riverside],
kidsFriendly[yes],
near[McDonalds]

the job of a generator is to produce:

(15) The three-star coffee shop, Jazzman, gives families a mid-priced dining experience featuring a variety of wines and cheeses. Find Juzzman near McDonalds.

The inputs are first linearised into sequences. In the case of E2ENLG task, the linearisation could be done following a specific "key order". For instance, Dusek and Jurcicek (2016) followed a order of: `name`, `eatType`, `food`, `priceRange`, `rating`, `area`, `familyFriendly`, `near`. The above example should be linearised into:

(16) name Juzzman eatType coffee food French priceRange moderate rating 3/5 area riverside kidsFriendly yes near McDonalds

Having sequential inputs in hands, we introduce how to produce outputs using the Seq2Seq model. Suppose an input is represented as $\mathcal{X} = \{x_0, x_1, ..., x_N\}$, the goal of

[4] We use the abbreviation E2ENLG for referring to the NLG task used in the E2E NLG Challenge (Dusek et al., 2018) in order to distinguish from the term "End2End NLG" in this thesis.

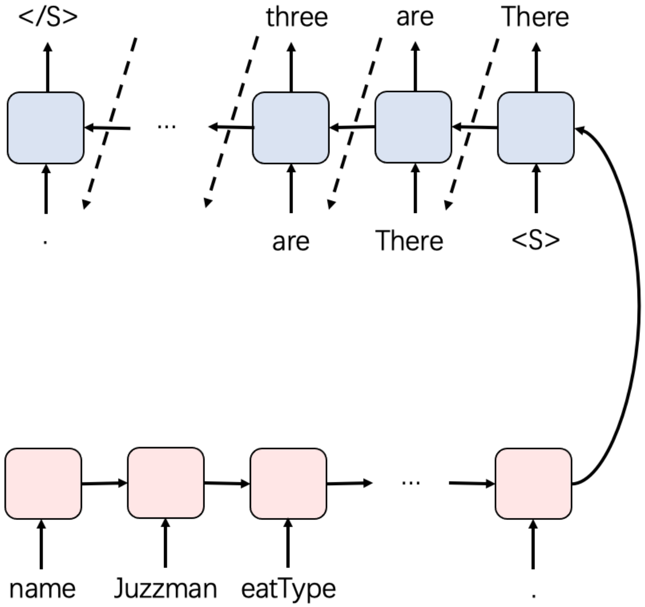


Figure 2.4: Diagram of the Seq2Seq Model, where the red part is the encoder and the blue part is the decoder.

a Seq2Seq model is to learn a function f maps the input to natural language $\mathcal{Y} = \{w_0, w_1, ..., w_M\}$, i.e., $f : \mathcal{X} \mapsto \mathcal{Y}$. A typical Seq2Seq model consists of two components: an encoder and a decoder, as shown in Figure 2.4.

At each encoding step, the representation $\boldsymbol{h}_t^{(e)}$[5] is computed in accordance with the current input $\boldsymbol{x}_t$ and the previous representation $\boldsymbol{h}_{t-1}^{(e)}$:

$$\boldsymbol{h}_t^{(e)} = g\left(W_h^{(e)} \boldsymbol{h}_{t-1}^{(e)} + W_x^{(e)} \boldsymbol{x}_t\right) \tag{2.3}$$

where $g(\cdot)$ is the activation function. The representation at the last time step $\boldsymbol{h}_N^{(e)}$ is then fed into the decoder to be conditioned on. Concretely, at the first decoding step, the decoding is done based on the start symbol w_0 and $\boldsymbol{h}_N^{(e)}$:

$$\boldsymbol{h}_0^{(d)} = g'\left(W_h^{(d)} \boldsymbol{h}_N^{(e)} + W_w^{(d)} \boldsymbol{w}_0\right) \tag{2.4}$$

At each of the rest decoding steps, the hidden representations are computed following a similar manner:

$$\boldsymbol{h}_t^{(d)} = g'\left(W_h^{(d)} \boldsymbol{h}_t^{(d)} + W_w^{(d)} \boldsymbol{w}_t\right) \tag{2.5}$$

Last, the production of each token follows the same way as in Equation 2.2.

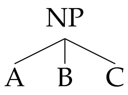
Note that we hereby only review the fundamental backbone of the Seq2Seq model (which will later be used in this thesis). There are plenty of different implementations

[5] We use a superscript (e) indicting the current representation appears in the encoder.

when applying Seq2Seq models to NLG or other NLP tasks. They are not the focus of this thesis, thus will not be further introduced.

In relation to NLG, the format of E2ENLG's inputs is still relatively simple compared to trees, figures, tables, databases and so on. So far, solutions regarding more complex inputs can be aligned into two lines of work. One is to design algorithms responsible for linearisation. For example, in Recurrent Neural Network Grammars (RNNG, Dyer et al., 2016), tree-structured inputs are linearised in a way similar to the "bracket" representations. The following tree is linearised as: NP A B C NP:

(17)



Castro Ferreira, Wubben, et al. (2018) introduced an algorithm to linearise dependency trees into sequences and do surface realisation using the Seq2Seq model. Gong et al. (2020) translated inputs table into a set of simple sentences (e.g., the above example can be translated as: *The name is Jazzman. The type is coffee shop. The food is French ...*).

The other is to make use of structural encoders to replace the sequential encoder in the Seq2Seq model. For example, for encoding tables, T. Liu et al. (2018) designed a hierarchical RNN, and Qasim et al. (2019) and Riba et al. (2019) proposed to use graph neural networks (GNN, Z. Wu et al., 2020). GNNs are also proved to be good at encoding trees (Q. Guo et al., 2021) and graph (Beck et al., 2018; Z. Guo et al., 2019; Koncel-Kedziorski et al., 2019).

Attention Mechanism

One flaw of the vanilla Seq2Seq model is that the representation at the last time step $\boldsymbol{h}_N^{(e)}$ of the encoder cannot memory sufficient amount of information from the input, on the one hand. On the other hand, only feed $\boldsymbol{h}_N^{(e)}$ at the very first step of the decoder also cannot let decoder sufficiently access the information stored in $\boldsymbol{h}_N^{(e)}$. In response to these two issues, a better solution should either have a clever way to make use of the hidden representations at every time step in the encoder or let the decoder access the encoded information more sufficiently.

One better solution is to use the attention mechanism. Generally speaking, as depicted in Figure 2.5, by adding an attention mechanism, the encoder output has become a weighted sum of all hidden representations. Given all encoded hidden states from the encoder $\mathcal{H} = \{\boldsymbol{h}_0^{(e)}, \boldsymbol{h}_1^{(e)}, ..., \boldsymbol{h}_N^{(e)}\}$, at decoding step t, the attention is calculated based on $\boldsymbol{h}_t^{(d)}$. First, the model computes a score (sometimes it is called "similarity") between $\boldsymbol{h}_t^{(d)}$ and each element in $\mathcal{H}$:

$$\text{score}(\boldsymbol{h}_t^{(d)}, \boldsymbol{h}_k^{(e)}), k = 0, 1, ..., N, \tag{2.6}$$

where $\boldsymbol{h}_t^{(d)}$ is also called as a "query". The scoring function score(A, B) could be a additive function (Bahdanau et al., 2015): $V\tanh(W_A A + W_B B)$, a production function (Luong et al., 2015; Vaswani et al., 2017): $A^T B$, and so on. The calculated scores are then normalised to compute weights for elements in $\mathcal{H}$:

$$\alpha_k = \frac{\exp(\text{score}(\boldsymbol{h}_t^{(d)}, \boldsymbol{h}_k^{(e)}))}{\sum_{j=0}^{N} \exp(\text{score}(\boldsymbol{h}_t^{(d)}, \boldsymbol{h}_j^{(e)}))}, k = 0, 1, ..., N. \tag{2.7}$$

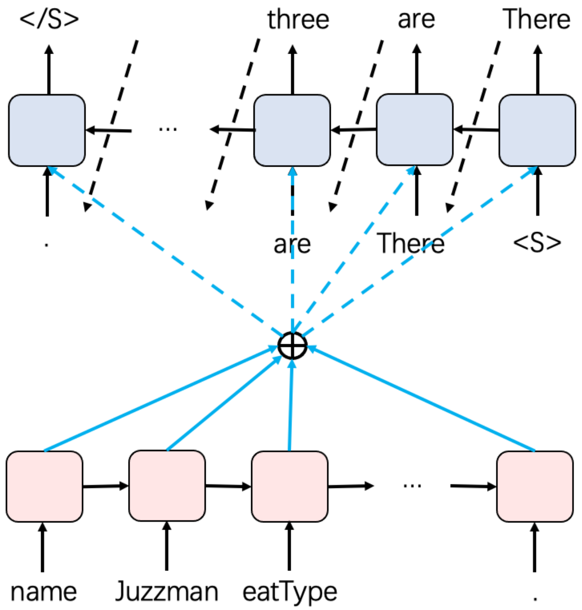


Figure 2.5: Diagram of the Seq2Seq with Attention Model.

At length, the final "context" representation of the input at decoding time step t is the weighted sum of all elements in $\mathcal{H}$:

$$c_t = \sum_{k=0}^{N} \alpha_k \boldsymbol{h}_k^{(e)}. \tag{2.8}$$

A vital extension of the attention mechanism was inspired by the fact that when people read different tokes in a piece of text, they will attend to different contexts. Such an attention mechanism is called intra-attention or self-attention (J. Cheng et al., 2016). Building on this idea, Vaswani et al. (2017) proposed to replace the RNN based encoder and decode in the Seq2Seq model with self-attention and name the resulting model as Transformer. In recent years, Transformer has become one of the mainstream techniques in NLP, but, strictly speaking, it is still a kind of Seq2Seq model.

Recent Advances

Transformer and its successors have hundreds of millions of parameters. This helps it to be able to offer significant improvements on many NLP tasks. For instance, on the task of English-German translation, compared to the Google Neural Machine Translation system (Y. Wu et al., 2016), Transformer improve the BLEU score from 24.6 to 27.3. Additionally, if we increase the number of parameters, then the score can be further increased to 28.4.

Pre-training. To take full advantage of this great amount of parameters, most recently, many researchers propose to *pre-train* Transformer on large-scale unlabelled data as a language model. Subsequently, the pre-trained Transformer is *fine-tuned* on the target tasks. A number of pre-training strategies have been proposed. We hereby list three of them. The first one is the BERT model (Devlin et al., 2019), which has been widely used in natural language understanding (NLU) tasks. [6] BERT is exactly the encoder of Transformer and is pre-training in a way called masked language model (MLM). The idea is first to mask out a certain number of tokens from a sentence, and then ask the model to "recover" the masked tokens. Although BERT was designed for NLU tasks, since it is able to predict the masked token given its context, it can also be used for NLG sub-tasks.

The other two are GPT (Brown et al., 2020; Radford et al., 2019) and BART (Lewis et al., 2020), which were designed for language generation. In contrast to BERT, GPT is the decoder of Transformer. It is trained as a normal language model (i.e., predicting the next word given previous words) on a huge web-crawled corpus. BART also follows the paradigm of MLM, but is trained in a full Seq2Seq architecture.

Creativity. Thanks to the powerful studying and generalising ability, NNLG (especially the Transformer based model) enables us to tackle many tasks that used to be believed hard in the age of pipeline NLG, such as image captioning, visual question answer, question generation and so on. In addition, these, in particular, include tasks that require creativity. For example, T.-H. K. Huang et al. (2016) used NNLG to conduct visual storytelling, i.e., generating a story given a series of images or a video. Other examples include the generation of traditional Chinese poems (X. Zhang & Lapata, 2014), Rap Lyrics (Malmi et al., 2016), product reviews (Zang & Wan, 2017), citation text (Xing et al., 2020), Recipe (Z. Yu et al., 2020b), Metaphor (Z. Yu & Wan, 2019), and Homophonic Pun (Z. Yu et al., 2020a).

Interpretability, Controllability, and Ethics. Different from pipeline NLG or statistical NLG, Deep Learning based End2End NLG is a black-box. For an NLG task, how an NNLG accomplishes the task and what it has done are not lucid. These outputs of NNLG models are not interpretable and controllable compared to pipeline NLG models. In recent years, NLG researchers have started to look at the interpretability and controllability of NNLG models.

Most attempts for interpreting neural language generation models are of text-to-text generation. For example, Ding et al. (2019) and J. Li, Chen, et al. (2016) used gradient-based method to understand how each input token contributes to each decision of Neural Machine Translation (NMT) systems. Kobayashi et al. (2020) tried to do the same thing by means of using the attention weights in the Seq2Seq attention model. There is also work that attempted to understand how well NMT can capture information of morphology (Belinkov, Durrani, et al., 2017), syntax, and semantic (Belinkov, Màrquez, et al., 2017) by using probe classifiers. In the realm of NLG (i.e., data-to-text generation), the interpretation is much harder (than for example MT) since the mappings between its inputs and outputs are of higher complexity. Linzen et al. (2016) conducted behaviour analysis to assess how well an LSTM-based generator capture syntactic dependencies. This was done by checking whether the generated text is correct on subject-verb agreement. For example, in the following example, if the generator generates "are", then it failed to capture the subject-verb agreement.

[6] Interestingly, since there is too much BERT related work, people have started to refer to work of this kind as "BERTology" (Rogers et al., 2020)

(18) a. The key is on the table.
b. * The key are on the table.

Following a similar idea (i.e., behaviour analysis), Petroni et al. (2019) validated whether the pre-trained language model (e.g., BERT or GPT) has learnt knowledge of human beings. Gehrmann et al. (2019) studied factors to distinguish computer-generated languages from human-produced ones.

Regarding controllability, much work has been done to control the content and the style of the generated text. Ficler and Goldberg (2017) did a systematic evaluation of possibilities to control style (e.g., personal, length, or descriptive) and content (e.g., theme and sentiment) in NNLG. They demonstrated that all these variables are controllable to different extents. Recent studies focused on more stable and fine-grained controls. For example, there has been work to build dialogue systems that are emotional (Zhou et al., 2018), personalised (S. Zhang et al., 2018; Zheng et al., 2019), stylised (Zheng et al., 2021) or that are equipped with external knowledge (Madotto et al., 2018). There also has been a long tradition in NLG for research on *Style Transfer* (Jin et al., 2020), which aims at transferring a piece of text from one style to another.

Due to the low interpretability and controllability of NNLG, deploying NNLG systems in real life might cause certain ethic issues. In recent years, one of the spotlighted research subjects in NLP is about the potential biases learnt by neural models, which includes biases regarding gender, race, religion, etc. For language generation models, most studies focused on ethical issues in machine translation and dialogue systems. However, so far, research of this kind is almost blank for data-to-text generation (Sheng et al., 2021).

Non-auto-regressive Decoding. The decoder of the Seq2Seq model always predicts the next word (i.e., w_{t+1}) on the basis of previous decoded results (i.e., $w_t, w_{t-1}, ...$), which is called auto-regressive decoding. One flaw of auto-regressive decoding is that the predictions cannot be parallelised since every decision has to depend on previous decisions. The alternative is the non-auto-regressive decoding (NAD), which, at an early time, was realised as a two-stage procedure (Gu et al., 2018; Ma et al., 2019): 1) predict the length of the output; and 2) predict all words at one time or in a constant number of steps. More recently, some NAD treated the inference/decoding process as a refinement process, which is done by series insertion/deletion operation (Gu et al., 2019; J. Lee et al., 2018; Stern et al., 2019; Susanto et al., 2020). In this way, NAD could achieve on par performance with auto-regressive decoding while accelerating and parallelising the whole process.

2.1.3 Pipeline NLG vs. End2End NLG

Reiter (2018a) and Rohrbach et al. (2018) figured out that compared to pipeline NLG, End2End neural NLG are more likely to produce “Hallucinate”. For instance, in the E2ENLG task, given the input:

(19) name[Cotto], eatType[coffee shop], near[The Bakers]

The TGen system (Dusek & Jurcicek, 2016), which received that highest BLEU score, generates:

(20) Cotto is a coffee shop with a low price range. It is located near The Bakers.

However, none of the information in the input suggests this coffee shop is in a low price range. In other words, it generates content that does not appear in the input, which is called "Hallucination". Possible reasons behind this include: 1) the dataset used for training End2End NLG systems are always noisy, and NNLG models cannot detect these noises but learn from these noises; 2) as what has been introduced, the decoder of a Seq2Seq model is a conditional LM. During inference, it is possible that the decoder relies too much on the previous tokens (as an LM) but too little on the conditions. In response to the first issue, Dušek et al. (2019), Nie et al. (2019), and H. Wang (2019) added an extra pre-processing step on the training data to reduce the amount of noise as much as possible. Regarding the second issue, Balakrishnan et al. (2019) and Kang and Hashimoto (2020) proposed to directly manipulate the output probability while Dziri et al. (2021) tried to solve the problem by refining the generated text. Either way, this makes the advantage of End2End NNLG less significant.

Interestingly, before the age of deep learning, many tasks cannot be done with high quality automatically. For example, it was hard to detect actions in images. Consequently, when building image captioning systems, it was hard to decide the verb in the sentence. To solve this, Mitchell et al. (2012) proposed to hallucinate, i.e., the system chooses the most likely verb using word co-occurrence statistics alone. Therefore, at that moment, in this case, hallucination was believed to be a "good" thing. Deep learning methods can help to come across these difficulties of understanding images, languages, or tables, but, simultaneously, overly produce hallucinations.

Here comes a question: what if we only ask the neural models to accomplish certain sub-tasks instead of doing End2End generation. In other words, if we use deep learning in every sub-task in pipeline NLG (namely, neural pipeline NLG), then how well it can perform compared to a fully End2End NLG system? To answer this question, Castro Ferreira et al. (2019) conducted a systematic comparison on the WEBNLG corpus (Gardent et al., 2017). They concluded that:

> The neural pipeline approaches were superior to the end-to-end ones in most tested circumstances: the former generates more fluent texts that better describe data on all domains of the corpus. The difference is most noticeable for unseen domains, where the performance of end-to-end approaches drops considerably. This shows that end-to-end approaches do not generalise as well as the pipeline ones. In the qualitative analysis, we also found that end-to-end generated texts have the problem of describing non-linguistic representations that are not present in the input, also known as Hallucination.

Faille et al. (2020) further acknowledged that neural pipeline NLG systems are more explainable. In addition, End2End NNLG is also believed to be not good at determining and structuring contents. Therefore, there is a recent trend on separating content planning from the End2End NNLG (Puduppully et al., 2019a, 2019b; Puduppully & Lapata, 2021; Shao et al., 2019; Shen et al., 2020; Z. Wang et al., 2020; Wiseman et al., 2018).

2.1.4 NLG Evaluation

NLG evaluators can be methodologically divided into two types: *intrinsic* and *extrinsic* evaluation methods. An intrinsic evaluator evaluates an NLG system on its own regardless of external aspects, such as the users, the environment, etc. In contrast, the extrinsic evaluation focuses on validating whether a built NLG system is really effective on the offshore platform. NLG systems that are supposed to be practically deployed require extrinsic evaluation to approve their "effectiveness", which depends on the application domain and the communication goal. For instance, the STOP system (Reiter et al., 2003) aims at generating personalised smoking cessation letters in order to persuade smokers to stop smoking. The evaluation was done by an A/B test like experiment on 2,553 smokers and compared the proportion of smokers who received the generated cessation letters and who received nothing. Likewise, since the goal of the SaferDrive is to reduce drivers' dangerous driving behaviours through weekly driving reports, its effectiveness was validated by sending generated reports to real drivers and monitoring their behavioural change. Since the aim of this thesis is not about building a practical NLG system, this sub-section is to be concerned with intrinsic evaluation, which can further be categorised to automatic evaluation and human evaluation.

Automatic Evaluation

Automatic evaluation is to measure the quality of the generated text without any human effort. One of the most famous automatic evaluators is the BLEU score (Papineni et al., 2002). BLEU is a corpus-based evaluation metric that measures token level overlaps between an output with references in the corpus. Concretely, it computes the n-gram precision with a length penalty. [7] Due to the use of n-gram counts, smoothing techniques (C.-Y. Lin & Och, 2004) are also needed against the issue of sparsity. Other word overlap based metrics include metrics like ROUGE (C.-Y. Lin, 2004), METEOR (Banerjee & Lavie, 2005), etc.

Another commonly used type of metric is to compute the edit distance between the outputs and references. There is a long tradition of research on computing edit distance in computational linguistics and computer science, famous ones of which include the Levenshtein distance, and Hamming distance. The fundamental idea behind these methods is to count the number of insertions, deletions and substitutions required to transform the output to the reference. However, when applying such a metric onto natural language, one major flaw is that conducting exact string matching might underestimate the performance of NLG systems because it failed to handle the synonyms. To solve this, Kusner et al. (2015) proposed to leverage word embeddings when calculating edit distance, which was further enhanced by introducing machine learning to make the evaluators trainable (e.g., `BERTScore` (Kusner et al., 2015).

All the above metrics/evaluators evaluate the quality of the generated text by computing how likeness they are compared to the references. Nevertheless, some aspects of text quality, such as human-likeness, coherence, diversity, fatality, do not necessarily rely on the reference. To measure performance on these aspects, reference-free evaluators have been proposed. For example, Mehri and Eskenazi (2020) proposed a trainable evaluator `USR` to evaluate performance of dialogue systems and Hessel et al. (2021) estimated the image captioning quality following a similar paradigm. For diversity, commonly used metrics include DIST (J. Li, Galley, et al., 2016), i.e., computing the proportion of distinct n-grams in the generated texts, and ENT (Y. Zhang et al., 2018), i.e., calculating the n-gram entropy on the generated texts.

[7] Merely computing precision (and overlooking recall) will make a metric prefer shorter outputs.

Human Evaluation

The validity of automatic evaluators is questioned. There has been a bank of studies demonstrating that using only automatic evaluators is insufficient for approving a system's effectiveness (e.g., Belz and Reiter (2006) and Reiter (2018b)), which necessities the human evaluation. [8] Human evaluation is to ask (native) speakers to rate or rank the generated texts from various dimensions. Common dimensions include fluency, naturalness, relevance, grammaticality, readability, clarity, adequacy, diversity, and so on. Most researchers pick 2-4 criteria from the list and which criteria to choose totally depends on the goal of the NLG system. Popular human evaluation designs include (1) Likert scale rating: asking participants to rate on a Liker Scale, which could consist of 2 to 10 points; (2) preference test: asking participants to pick the output s/he prefers from 2 outputs (belonging to two different systems); and (3) magnitude estimation: asking participants to score an output based on a scored output.

In 2019, van der Lee et al. (2019) listed "Best Practices" for human evaluation. We hereby summarise these best practices here. A "good" human evaluation should be conducted 1) on a certain amount readers (the corresponding sample size and the demographics also should be reported); 2) using 7-point Likert scales or continuous ranking (e.g., ranking based magnitude estimation); 3) using test cases with counterbalanced/random order. Additionally, after the human evaluation, statistical testing should be done and reported properly.

2.2 Referring Expression Generation

Recall that, generally speaking, the task of REG is to generate a referring expression (RE) of a referent that enables the reader to identify that referent in a given context. In accordance with different meanings of the term "context", the task of REG can be further divided into two categories[9]:

1. *REG in Context*, where "context" is the *linguistic context*. It asks the algorithm to produce REs for referents appear in a discourse so that the resulting discourse is coherent and contains no referential ambiguity; and

2. *One-shot REG*, where "context" is a set of distractors. It requires the algorithm to produce an RE for the target referent to distinguish it from distractors. Each RE is produced individually, in isolation from any linguistic context.

In this section, we will first introduce the task, the theories, the algorithms, and the evaluation of one-shot REG (§2.2.1) and then introduce those of REG in context (§2.2.2).

2.2.1 One-shot REG

The research of one-shot REG studies RE itself regardless of its linguistic context. This helps us to focus more on some of the core mechanisms of reference. This line of research is mostly about building computational production models of REs to mimic and understand

[8] Interestingly, when BLEU score was firstly invented, it was used as a complementary of the human evaluation (Papineni et al., 2002).

[9] Note that "context" could arguably have other meanings in regard to RE, such as the speaker's and hearer's background knowledge and opinions.

the human's use of REs. This is why NLG researches of this kind are called theoretical NLG. Different from the REG in practical NLG (see §2.1), the aim of one-shot REG is human-likeness. Most one-shot REG studies are merely about determining the content of each RE but not about its surface form. Precisely, van Deemter (2016) defined the one-shot REG as:

> If there exists a set of properties $\{P_1, ..., P_n\}$, where $P_i \in \mathcal{P}$, and where the conjunction of all the P_i in the set singles out the referent r (i.e., $[\![P_1]\!] \cap [\![P_2]\!] \cap ... \cap [\![P_n]\!] = \{r\}$), then find such a set and conjoin its elements. If no such set of properties exists, then say so. Furthermore, make sure that $\{P_1, ..., P_n\}$ are collectively as similar as possible to the set of properties that human speakers would use if they were referring to r in the situation at hand.

In this definition, P is an atomic property, $\mathcal{P}$ is a set of atomic properties in the domain, and $[\![P]\!]$ is the set of elements that share a property P (called the denotation or extension of P).

Gricean Maxim

One of the most vital theoretical bases of REG is the Gricean Maxims (Grice, 1975). It provides an implementation of the idea that communication is cooperative. It contains the following four Maxims: Quality, Quantity, Relation, as well as Manner. In what follows, we would like to explain the details of each maxim and what each maxim means in the context of the production of REs. Most explanations follow those in Dale and Reiter (1995), Dale and Reiter (1996), and van Deemter (2016).

Maxim of Quality. The Maxim of Quality requires:

1. Do not say what you believe to be false.
2. Do not say that for which you lack adequate evidence.

For RE, it requires an RE must be accurate for the intended referent. In other words, for any P_i in $\{P_1, ..., P_n\}$, we always have $r \in [\![P_i]\!]$. In most situations, no NLG system would deliberately say something that is inaccurate (i.e., breaking the Maxim of Quality), nor does REG. One exception is that a system could sometimes produce benign deceit, which, in certain situations, is the most efficient way to achieve a communicative goal (van Deemter & Reiter, 2018). Kutlak et al. (2016) figured out when referring to an object, this can be the case if the hearer's knowledge of the object is incorrect. For example, they showed that the RE "*the man who invented the light bulb*" is the most efficient way to refer to Thomas Edison, but the RE is not true since Edison did not invent the light bulb. In human language production, excepting "lie", there are two possibilities that the maxim of quality is breached. One is when the speaker uses metonymy to ascribe a property to the referent that is not true of the referent but of something associated with it. A classic example (Nunberg, 1978) is that the expression "*the ham sandwich is getting restless*" can be used by waiters to refer to a customer who has ordered a ham sandwich. The other is when the speaker is uncertain about the target referent. For example, one would say "*the man with the Martini*" to refer to a man who is actually with wine since the speaker was unsure whether the drink was wine or Martini.

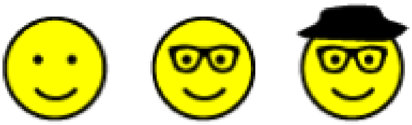


Figure 2.6: A reference game where the target is to produce an RE that can single out the face in the middle (Goodman & Frank, 2016).

Maxim of Quantity. The Maxim of Quantity asks to:

1. Make your contribution as informative as is required.
2. Do not make your contribution more informative than is required.

This requires an RE should provide enough information to enable the hearer/reader to identify the target referent successfully, and, meanwhile, should not include unnecessary information. Let us elaborate on each of these two requirements separately.

REs that violate the first half of the maxim of quantity is called under-specifications. Given the definition of the REG task, the production of under-specifications should be avoided and most classic REG algorithms were designed to do so. Nevertheless, Frank and Goodman (2012) and Goodman and Frank (2016) suggested that producing under-specification is also possible if the used property is salient. For example, in order to single out the face in the middle of the Figure 2.6, one could say:

(21) the one with glasses

Logically, it cannot result in successful communication since there are two faces wearing glasses. During the experiment, the majority of the readers can select the correct face given the expression (21). This is because the readers reasoned that since the property "*hat*" is a salient property in this situation (i.e., Figure 2.6), a speaker must use "*hat*" if s/he intended to refer to the right face. Hence, since s/he did not use "*hat*", s/he should refer to the one in the middle.

The interpretation of the second half of the Maxim of Quantity is disputed. [10] The REs that break the second half can be roughly named as over-specifications. Over-specification is a long-standing theme in linguists' thinking about reference, with many contributions from both psycholinguistics and computational linguistics. A number of possible explanations have been proposed for the frequent occurrence of over-specification. One is the idea that over-specification can be beneficial for hearers (Krahmer & van Deemter, 2012), for example, because it taps into prototypes in the human mind (Levelt, 1993, Chapter 4), which creates so-called conceptual gestalt of the target object by means of highly salient attributes like TYPE and COLOUR. Eikmeyer and Ahlsén (1996), Pechmann (1989), and Schriefers and Pechmann (1988) found that these salience attributes are ubiquitous in referring expressions no matter what the distractors are. This is closely tied to the speakers' belief that salient attributes (which always results in over-specification) help readers to locate the referent quickly, which firstly confirmed on the use of atomic attributes (e.g., TYPE, COLOUR or SIZE) by Arts et al. (2011) where they compared the identification time between over-specifications (e.g., *the round button at the top left* and *the round white bottom*) and minimal descriptions (e.g.,

[10] For more discussion about over- and under-specification, please see §4.

the button). Paraboni et al. (2007) and Paraboni and van Deemter (2014) found something very similar, where the former suggested that over-specifying in hierarchical domain leads to a significant reduction in the amount of search that is needed to identify the referent and the latter indicates that over-specification may not only help the hearer but also is required for a successful communication since minimal descriptions sometimes lead to confusion and misidentification.

Nevertheless, Engelhardt et al. (2006), by conducting a series of eye-tracking and ERP experiments, argued something on the opposite, that is, over-specification may be detrimental to locating the target. They tested on some over-specified prepositional phrases (e.g., *put the apple on the towel in the box* in the context consisting of one apple on a towel and an empty towel) and showed they may lead to "temporary confusion" of the target referent and thus slow down the identification process, which is then re-confirmed by Engelhardt et al. (2011) on more common over-specifications as in *look at red star* in a context in which there is only one star (i.e., *red* is over-specified). More recently, Paraboni et al. (2017) conducted eye-tracking studies on both atomic and relational attributes. They reasoned that the recognisability of superfluous properties matters on whether they speed up or slow down the identification process and concluded that

> Easily recognisable properties may facilitate identification, whereas properties that are more difficult to recognise may have the opposite effect.

Maxim of Relation. The maxim of relation requires that the production of language is relevant. Dale and Reiter (1995) provided an interpretation requiring an RE should not mention properties that have no discriminatory power. The discriminatory power (DP) of proper P is the number of distractors removed by P as a proportion of the total number of distractors. Formally, suppose $m \in \mathcal{M}$ is the set of domain elements not yet ruled out, and, therefore, $\mathcal{M} - \{r\}$ is the set of distractors, DP is computed as:

$$DP(P, \mathcal{M}) = \frac{|[\![P]\!] \cap (\mathcal{M} - \{r\})|}{|\mathcal{M} - \{r\}|}. \tag{2.9}$$

It appears that the interpretation of Dale and Reiter (1995) is a relaxed version of the requirements of the Maxim of Quantity. Concretely, for a situation where there is a choice between properties that do not have equal non-zero DP, the Maxim of Relation provides no preference on which one should be selected since all of them have DP larger than zero, while the maxim of quantity suggests choosing the one with the highest DP. Another interpretation argues that the Maxim of Relation can do the work of all the Maxims combined (Wilson & Sperber, 2002).

Maxim of Manner The last Maxim: Maxim of Manner says:

1. Avoid obscurity of expression.
2. Avoid ambiguity.
3. Be Brief (avoid unnecessary prolixity).
4. Be orderly.

```
Input: A domain of objects containing a target referent $r$, a non-empty set of distractors, a
    set $\mathcal{P}$ of $n$ properties true of $r$.
Output: A distinguishing description $\mathcal{D}$ of $r$ using conjunctions of properties in $\mathcal{P}$ if such
    a distinguishing description exists.
for $m$ in $(0, n]$ do
    Look for a description $\mathcal{D}$ that distinguishes $r$ using $m$ properties
    if a description $\mathcal{D}$ is found then
        return $\mathcal{D}$
    end if
end for
return "No distinguishing description of $r$ exits"
```

Algorithm 2.1: The Full Brevity Algorithm

This maxim advises the language to be clear and against verbosity, (syntactic and lexical) ambiguity as well as any kind of messiness that can make a text difficult to understand. These aspects are mostly studied in lexicalisation, aggregation, or linguistic realisation, but are less focused in the context of REG. In addition to the above clarity requirements, the third rule of this maxim asks for brevity (which is also advocated by the Maxim of Quantity). Some recent efforts have started to understand how clarity should be a trade-off against brevity.

One-shot REG Algorithms

We review REG algorithms, from classic algorithms including the full brevity algorithm, the greedy algorithm, and the incremental algorithm to recent advances such as the Bayesian as well as the probabilistic approaches. For a more detailed review, please check Krahmer and van Deemter (2012) as well as van Deemter (2016).

Full Brevity Algorithm. The first one is the Full Brevity Algorithm (FB, Dale, 1989), which tends to find the shortest possible RE. Algorithm 2.1 describes the FB algorithm. It starts from checking whether there is a single property of the target that rules out all distractors. If it fails, it iteratively checks whether any combination of two properties does this, and so on, until the referent has been singled out or until conjunctions of all possible properties have been attempted.

Apparently, the FB algorithm takes brevity as its priority and provides a restricted interpretation of the Gricean Maxim of Quantity, Relevance and the third rule of Manner. However, finding the shortest RE is an NP-hard problem, and therefore, the FB algorithm is expensive to be implemented. Therefore, later algorithms tend to either approximate such an interpretation of Gricean Maxim or rather violate some Maxims to produce more human-like REs (e.g., over-specifications). One simple alternative is the Local Brevity algorithm (Reiter, 1990). This algorithm starts with an arbitrary distinguishing description[11]

[11] Distinguishing Description is an RE that can successfully single out the target referent. For a more precise definition, please see 4.

```
Input: A domain of objects containing a target referent $r$, a non-empty set of distractors $\mathcal{M}$, a set $\mathcal{P}$ of $n$ properties true of $r$.
Output: A distinguishing description $\mathcal{D}$ of $r$ using conjunctions of properties in $\mathcal{P}$ if such a distinguishing description exists.
1:  $\mathcal{D} := \{\}$
2:  while $\mathcal{M} \neq \emptyset$ and $\mathcal{P} \neq \emptyset$ do
3:      Select a new property $P \in \mathcal{P}$, choosing one whose descriminative power is maximal
4:      if $P$ is false of some distractors then
5:          $\mathcal{D} := \mathcal{D} \cup \{P\}$
6:          $\mathcal{P} := \mathcal{P} - \{P\}$
7:          $\mathcal{M} := \mathcal{M} \cap [\![P]\!]$
8:      end if
9:  end while
10: return $\mathcal{D}$
```

Algorithm 2.2: The Greedy Algorithm

and check whether there is a short possible replacement by replacing two more properties with a single property.

Greedy Algorithm Another approximation is to follow Johnson's greedy heuristic (Garey & Johnson, 1990). Building on this idea, Dale (1989) proposed to select properties one by one and each step choose a property that does best for the referent. The Greedy Algorithm (GR) interprets "best" as removing the maximum number of distractors, i.e., the highest DP. Note that the GR will not always produce the shortest RE.

Algorithm 2.2 sketches the GR algorithm. It starts out with an empty description $\mathcal{D}$. Subsequently, at each step, it selects the property that has the highest discriminative power (line 3), which should not equal zero (line 4). Once a property has been selected, it will be added to the description $\mathcal{D}$ (line 5), be removed from the set of candidate properties $\mathcal{P}$ (line 6), and be used for recording the removed distractors (line 7). The algorithm exits at once either all the distractors have been removed or all the properties have been used. Practically, both FB and GR will add the TYPE of the target referent to fill the position of the noun. In our implementation, we always add TYPE in the very first iteration and remove the distractors from $\mathcal{M}$ accordingly.

Incremental Algorithm Different from the above algorithms, the incremental algorithm (Dale & Reiter, 1995) cares less about the Gricean Maxims but learns from psycholinguistic findings. Pechmann (1989) found that perceptually salient attributes tend to be considered before other attributes when producing REs. Such attributes could include COLOUR and TYPE (i.e., whether the target referent is a dog or a human). In other words, some attributes are more "preferred" than others and such a preference is intrinsic. Building on this idea, IA selects properties one by one (i.e., incrementally) in accordance with a preference order[12].

[12] A preference order is a ordered list of attributes, indicting the preference of attributes. For example, a preference order could be: COLOUR > ORIENTATION > SIZE, suggesting that COLOUR is more preferred than ORIENTATION,

```
Input: A domain of objects containing a target referent $r$, a non-empty set of distractors
    $\mathcal{M}$, a set $\mathcal{A}$ of $n$ attributes at least one of whose values is true of $r$, a linear preference
    order defined on $\mathcal{A}$.
Output: A distinguishing description $\mathcal{D}$ of $r$ using conjunctions of properties in $\mathcal{P}$ if such
    a distinguishing description exists.
 1: $\mathcal{D} := \{\}$
 2: while $\mathcal{M} \neq \emptyset$ and $\mathcal{A} \neq \emptyset$ do
 3:     Select a new property $A_i \in \mathcal{A}$, choosing the most preferred one
 4:     $V_i :=$ FindBestValue$(r, A_i)$
 5:     if $V_i$ is false of some distractors then
 6:         $\mathcal{D} := \mathcal{D} \cup \{V_i\}$
 7:         $\mathcal{A} := \mathcal{A} - \{A_i\}$
 8:         $\mathcal{M} := \mathcal{M} \cap [\![V_i]\!]$
 9:     end if
10: end while
11: return $\mathcal{D}$
```

Algorithm 2.3: The Incremental Algorithm

Algorithm 2.3 presents the IA. In i-th iteration, IA considers the most preferred attribute A_i in $\mathcal{A}$ (line 4). Given the selected A_i and the target the referent r, the FindBestValue function selects the value that removes most distractors. The rest configurations of IA is similar to that of the GR algorithm, both of them select properties incrementally.

In addition to the attribute salience, REG also needs to take the entity/object salience into account. The concept of object salience has been widely discussed in relation to "REG in context". For example, an object is more salient if it has been mentioned in the previous discourse (Passonneau, 1996). More discussion can be found in §2.2.2. In relation to one-shot REG, Krahmer and Theune (2002) proposed to associate each object with a so-called *salience weight*, and interpret an RE like "*the man*" as referring to the man with the highest salience weight. To implement this idea, Krahmer and Theune (2002) added an extra step at the beginning of IA (i.e., line 2 of Algorithm 2.3) that removes all elements that have lower salience weight than r from $\mathcal{M}$.

§2.1 highlighted the importance of variation in NLG, especially lexicalisation. It has been pointed out that, in the human production of REs, variation also plays a vital role. Nevertheless, the REG algorithms introduced so far are all deterministic. In other words, given an input, they will always produce the same output. From now on, we move our attention to non-deterministic REG algorithms.

Probabilistic Models. One straightforward idea is to introduce probabilities into REG algorithms. van Deemter, Gatt, van Gompel, et al. (2012) introduced a non-deterministic version of the IA. The idea was to un-deterministically vary the preference order of IA. For example, the algorithm would check COLOUR before SIZE with a probability P and check SIZE before COLOUR with the rest of the time, i.e., $1 - P$.

which is more preferred than SIZE.

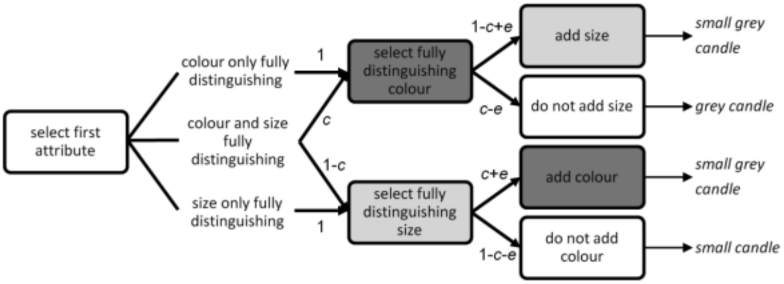

select first attribute
colour only fully distinguishing
colour and size fully distinguishing
size only fully distinguishing
1
c
1-c
1
select fully distinguishing colour
select fully distinguishing size
1-c+e
c-e
c+e
1-c-e
add size
do not add size
add colour
do not add colour
small grey candle
grey candle
small grey candle
small candle

Figure 2.7: An example decision tree of PRO (Figure 4 in van Gompel et al. (2019)).

More recently, van Gompel et al. (2019) noted that the non-deterministic IA produces significant fewer over-specifications than what humans really do and, as a consequence, they proposed a full probabilistic model, namely Probabilistic Referential Over-specification (PRO) model, acknowledging either that the production of REs is non-deterministic or that the production of over-specifications is frequent. Figure 2.7 provides an example decision procedure of the PRO algorithm. As we can see, different from the non-deterministic IA, PRO offers possibilities to continuously consider new attributes even after the target object r has already been singled out.

Bayesian Models. Bayesian methods are believed to be good at modelling uncertainty and, thus, modelling nondeterministic processes. In 2012, Frank and Goodman (2012) proposed to model the production and comprehension of reference in a Bayesian framework. They argued that the way models the production and comprehension as rational speech acts (RSA). The RSA could be seen as a recursive reasoning procedure, which is starting with a so-called literal speaker $S_0(w|r, \mathcal{C})$, where $\mathcal{C}$ is the context. In most cases, S_0 is represented as the likelihood of the word w being chosen to refer to r in the context $\mathcal{C}$: $P(w|r, \mathcal{C})$. Later efforts (e.g., Goodman and Frank (2016) and Monroe and Potts (2015)) suggested taking, the literal speaker should also consider the speech cost of using the word w: $C(w)$, so that the literal speaker can be expressed as a utility function:

$$S_0(w|r, \mathcal{C}) = \exp\left\{\lambda\left(P(w|r, \mathcal{C}) - C(w)\right)\right\} \tag{2.10}$$

With the literal speaker in hand, a pragmatic listener reasons about the intended referent r by maximising:

$$L_1(r|w, \mathcal{C}) = \frac{S_0(w|r, \mathcal{C})P(r, \mathcal{C})}{\sum_{r' \in \mathcal{C}} S_0(w|r', \mathcal{C})P(r', \mathcal{C})}, \tag{2.11}$$

where $P(r, \mathcal{C})$ is the object salience in the context $\mathcal{C}$.

Subsequently, the pragmatic speaker produces REs on the basis of the pragmatic listener. There have been two strategies to model the pragmatic speaker. One is to treat the pragmatic speaker in a similar way as the literal speaker, but replacing the likelihood term with the literal speaker (Langner, 2020; Monroe & Potts, 2015):

$$S_1(w|r, \mathcal{C}) = \exp\left\{\lambda\left(L_1(r|w, \mathcal{C}) - C(w)\right)\right\}. \tag{2.12}$$

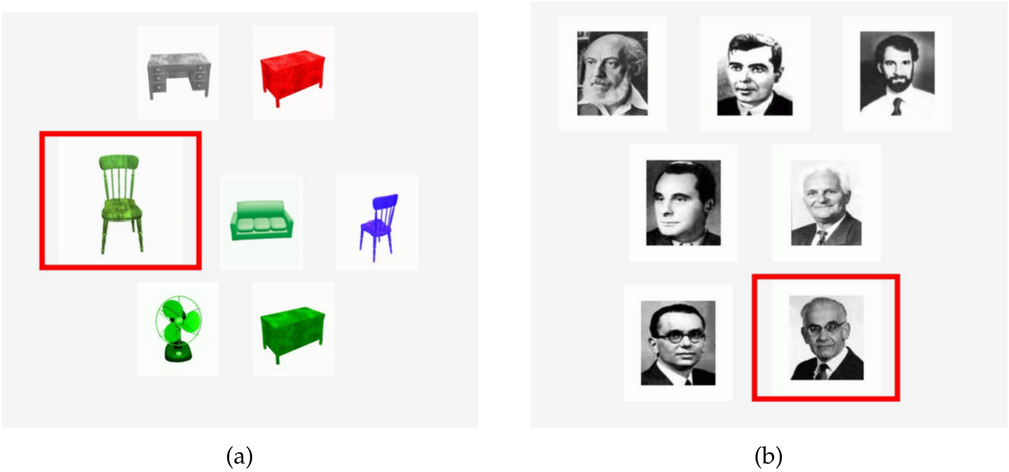


Figure 2.8: Two scenes from the TUNA experiment, in which (a) is a situation from the furniture domain while (b) is from the people domain.

The other is to let the pragmatic speaker also follow a Bayesian reasoning paradigm (G. Chen et al., 2018a; Orita et al., 2015; van Deemter, 2016), in which both the object salience and the attribute salience are considered:

$$S_1(w|r,\mathcal{C}) = \frac{L_*(r|w,\mathcal{C})P(w,\mathcal{C})}{P(r,\mathcal{C})}, \tag{2.13}$$

where $L_*(r|w,\mathcal{C})$ can be either a pragmatic listener (Equation 2.11), or a literal listener (i.e., directing estimating the likelihood $P(r|w,\mathcal{C})$ from a corpus), and $P(w,C)$ is the salience of word w (i.e., attribute salience).

Evaluation

REG, as a content determination task (though in NLG pipeline, it is a sub-task of micro-planning), is usually corpus evaluated. We hereby introduce the evaluation corpus (with a focus on a specific corpus called TUNA) and the evaluation metrics.

Datasets. To assess the performance of the one-shot REG algorithms introduced in this section, there had been a number of datasets being collected for conducting corpus evaluation. Early examples includes the COCONUT corpus (Gupta & Stent, 2005) and the MapTask corpus (Bard et al., 2007). However, these corpora cannot really address the question of how well these REG models handle the human-likeness of the REs produced by human beings. For example, most referents in the COCONET corpus are name entities. Subsequently, Viethen and Dale (2006) built a small scale RE dataset to assess IA, while Viethen and Dale (2008) constructed GRE3D3 to assess REG algorithms that use spatial relations. Here, we focus on a corpus called TUNA, which is able to conduct an exclusive evaluation of REG content determination.

TUNA (Gatt et al., 2007; van der Sluis et al., 2007) is a series of controlled elicitation experiments that were set up to aid computational linguists' understanding of human reference production. In particular, the corpora to which these experiments gave rise were employed to evaluate REG algorithms, by comparing their output with the REs in these corpora. The stimuli in the TUNA experiments were divided into two types of visual scenes: scenes that depict furniture and scenes that depict people. Figure 2.8 shows an example for each of these two types of scenes. In each trial, one or two objects in the scene were chosen as the target referent(s), demarcated by red borders. The subjects were asked to produce REs that identify the target referents from the other objects in the scene (their "distractors"). For example, for the scene in Figure 2.8, one might say *the large chair*. The trials in the people domain were intended to be more challenging than those in the furniture domain.

The resulting corpus, which we will call ETUNA, was subsequently studied for evaluating a set of "classic" REG algorithms (van Deemter, Gatt, Sluis, et al., 2012). Although RE has given rise to a good number of other corpora, with subtly different qualities (e.g., Dale and Viethen (2009)), we focus here on the TUNA corpora for two reasons: firstly the ETUNA corpus was used in a series of Shared Task Evaluation Campaign (Gatt & Belz, 2010), which caused it to be relatively well known. Secondly and more importantly from the perspective of the present paper, ETUNA inspired a number of similarly constructed corpora for Dutch (DTUNA, Koolen & Krahmer, 2010), German (GTUNA, Howcroft et al., 2017), and Mandarin (van Deemter et al., 2017).

Metrics. To evaluate a REG algorithm on the constructed corpus, a metric to measure the similarity between a generated RE and a RE in the corpus is needed. To this end, van Deemter, Gatt, Sluis, et al. (2012) adopted the DICE metric (Dice, 1945), which measures the overlap between two attributes sets:

$$\text{DICE}(\mathcal{D}_H, \mathcal{D}_A) = \frac{2 \times |\mathcal{D}_H \cap \mathcal{D}_A|}{|\mathcal{D}_H| + |\mathcal{D}_A|}$$

where $\mathcal{D}_H$ is the set of attributes expressed in the description produced by a human author and $\mathcal{D}_A$ is the set of attributes expressed in the logical form generated by an algorithm.

Another commonly used metric is the "perfect recall percentage" (PRP), the proportion of times the algorithm achieves a DICE score of 1, which is seen as an indicator of the recall of an algorithm (in contrast to the "precision" advocated by DICE).

In addition, evaluating the performance using only DICE has certain flaws. van Deemter and Gatt (2009) listed some of them:

1. DICE punishes the omission of properties from the oracle more heavily than the addition of properties to it. For example, we have an oracle description: $\{A, B\}$. The description $\{A\}$ would receive a DICE score at $2/3$ while the description $\{A, B, C\}$ receives $4/5$;

2. Descriptions with higher DICE scores are not always distinguishing descriptions. DICE metric targets at comparing sets but is blind towards the goal of a description (i.e., singling out the intended referent);

3. DICE treats all attributes equidistant. However, van Miltenburg et al. (2020) found that similar to the attribute salience, being incorrect on some attributes is more serious than others. For example, when referring to a "*girl with red t-shirt*", saying "*the boy with red t-shirt*" is worse than saying "*the girl with blue t-shirt*".

More discussion about the problems of REG evaluation can be found in §4 when we evaluate REG algorithms on Mandarin TUNA.

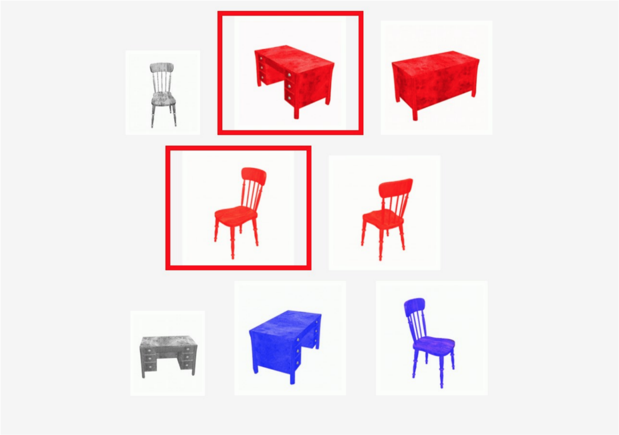


Figure 2.9: The "Referring to Sets" portion of the TUNA corpus.

Referring to Sets

There is also a line of work focusing on generating REs for a set of target objects instead of a single one. As a matter of fact, the TUNA experiment contains trials that ask subjects to produce REs to sets, an example of which is shown in Figure 2.9. One simple but natural idea for designing a REG algorithm for a set of target objects is extending the IA to IA_{plur} by simply replacing the target object r with a set of target objects $\mathcal{R}$. However, as pointed out in van Deemter (2002), the following issues are yet to be addressed using IA_{plur}:

Collective Properties. It does not work for collective properties, such as "*being of the same age*". It can be solved by defining the property set $\mathcal{P}$ as a set with collective properties.

Negation. Negation is sometimes needed and is unavoidable (since the negated relation is not often lexicalised), e.g., "*I bought the two dogs that are not poodles*". One simple solution is adding to $\mathcal{P}$ the properties whose extensions are the complements of those in $\mathcal{P}$.

Disjunction. Classic REG algorithm is also in-able to express disjunction. (i.e., set union) which is needed when referring to sets, e.g., "*the poodles and the white dogs*". One possible solution is making use of the *satellite sets*, i.e., satellite set of an object r is the set of objects from which r cannot be distinguished. However, generating REs with satellite sets is not a kind of incremental generation.

In response to these issues, van Deemter (2002) introduced a two staged algorithm which will run the IA_{plur} on $\mathcal{P}$ and on $\mathcal{P} \cup \mathcal{P}$ again. The algorithm was then optimised by re-structuring the initial content determination results (a set of logical forms). Gatt and van Deemter (2007b) designed a partitioning based algorithm, which partitions the target object set based on the pre-defined preference order. The idea was inspired by the finding that human beings are repeating properties in two disjuncts rather than doing aggregation.

With the increasing size of the searching space, the number of possible alternative distinguishing descriptions also increases dramatically. It is difficult to choose between different distinguishing descriptions that contain combinations of Boolean operators. FitzGerald

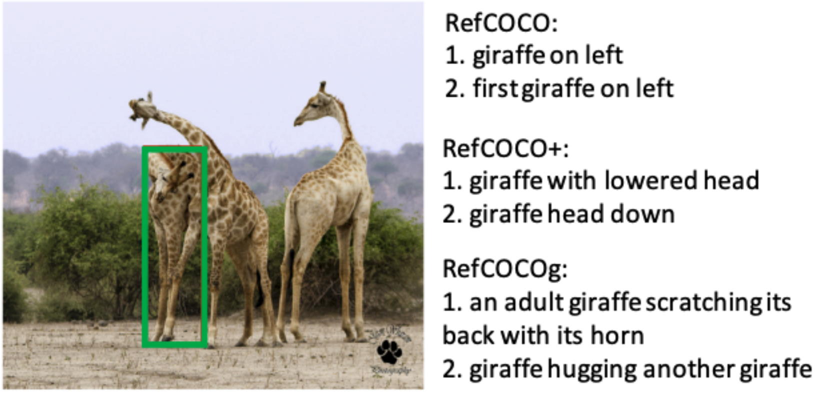

RefCOCO:
1. giraffe on left
2. first giraffe on left
RefCOCO+:
1. giraffe with lowered head
2. giraffe head down
RefCOCOg:
1. an adult giraffe scratching its back with its horn
2. giraffe hugging another giraffe

Figure 2.10: An example input image of RefCOCO corpora and corresponding REs in RefCOCO, RefCOCO+, and RefCOCOg, respectively.

et al. (2013) tried to tackle this problem by learning a distribution over logical forms of REs with a log-linear model. They also constructed a new dataset using the scenes from Matuszek et al. (2012), which contains more situations with multiple targets and each object of which has fewer attributes with more values. Other probabilistic efforts include Monroe and Potts (2015) and X. Li et al. (2018).

When generating REs refer to sets, the problem of *coherence* is unavoidable as we always tend to generate descriptions with multiple clauses. For example, if the RE "*the Kenyan animal and the tiger*" and the RE "*the Kenyan lion and the Chinese tiger*" are both distinguishing REs, the latter one is more probable to be chosen as the former one lacks coherence. Gatt and van Deemter (2007a) attempted to solve this issue based on the intuition that, the word in the RE needs to be as similar to each other as possible. Meanwhile, they also considered that complete coherence may not be compatible with the aim of referring uniquely.

Referring Expression Modelling from Images

In the computer vision community, there is a bank of research on marrying image processing with REG. One of the early attempts is Kazemzadeh et al. (2014), where large-scale data was collected in a way called "ReferItGame". Concretely, it is a two-player game, in which the first player was asked to produce a RE given an image with the annotated target object, while the second player was asked to click on the location where the produced RE describes. At length, 130,525 REs were collected for 96,654 distinct target objects. Following ReferItGame, three datasets RefCOCO, RefCOCO+, RefCOCOg[13] were collected by L. Yu et al. (2016). This time, in each RefCOCO* corpus, each image is associated with multiple REs. Example image-RE pairs are shown in Figure 2.10.

Most work in this line is about RE comprehension. Not much work has been done from the aspect of production. Analogous to text-to-text generation, work targeting generating REs from images used the encoder-decoder architecture. As shown in Figure 2.11, the

[13] The difference between RefCOCO and RefCOCO+ is that RefCOCO+ disallows using location words. RefCOCOg was collected using Mechanical Turk rather than ReferItGame.

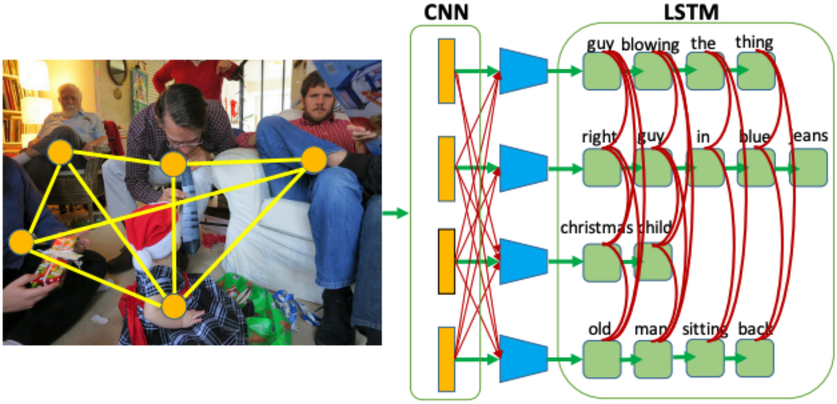


Figure 2.11: An encoder-decoder architecture for REG from images.

model of L. Yu et al. (2016) used a convolution neural network to encode input images and used an LSTM to decode REs. To boost the performance, L. Yu et al. (2016) proposed to encode all objects in an image and decode all REs for them simultaneously. More recently, inspired by the idea of RSA, L. Yu et al. (2017) and Mao et al. (2016) suggested modelling RE generation and comprehension jointly.

2.2.2 REG in Context

Different from the one-shot REG, the task of REG in context is concerned with the generation of REs in discourse context. Belz and Varges (2007) phrase it as follows:

> Given an intended referent and a discourse context, how do we generate appropriate referential expressions to refer to the referent at different points in the discourse?

For example, here is a description from the Wikipedia entry of Joe Biden:

(22) Joseph Robinette Biden (born November 20, 1942) is an American politician who is the 46th and current president of the United States. A member of the Democratic Party, he served as the 47th vice president from 2009 to 2017 under Barack Obama and represented Delaware in the United States Senate from 1973 to 2009.

The input of a REG in context system is the above text where all the underlined text is missing. The goal of the system is to generate all these missing REs given which referent each slot is for.

Pipeline REG

Classic REG in Context was usually understood as a two-step procedure (i.e., a pipeline). At the first step, the referential form (RF, i.e, the syntactic type) is determined. For instance, when referring to Joe Biden at a given point in a discourse, the first step is to decide whether to use a proper name (“*Joe Biden*"), a description (“*the president of the USA*"), a demonstrative (“*this person*") or a pronoun (“*he*"). The second step is to determine the RE content, that is, to choose between all the different ways in which a given form can be realised. For instance, to generate a description of Joe Biden, one needs to decide whether to only mention his job (e.g., *The president* entered the Oval Office.), or to mention the country as well (e.g., *The president of the United states* arrived in Cornwall for the G7 Summit.)

At the second step, the content of the RE is usually determined by a rule-based system in accordance with the decided referential form as well as the referents' meta-information. For example, if the intended referent is a singular referent, is a human, and is a male, then, when the referential form is a pronoun, it is realised as “*he*”.

A common solution for the first step (i.e., referential from selection) is using feature-based machine learning models (see Belz et al. (2010) for an overview). In earlier works, computational linguists linked REG to linguistic theories and built referential form selectors systems on the basis of linguistic features. For example, Henschel et al. (2000) investigated the impact of 3 linguistic features namely recency, subjecthood, and discourse status on pronominalisation, i.e. deciding whether the RE should be realised as a pronoun. Using these features, they used the notion of *local focus* as a criterion for detecting the set of referents that can be pronominalised. This task has attracted many research efforts (e.g., Greenbacker and McCoy (2009) and Hendrickx et al. (2008)) and it has been used in the GREC shared tasks (Belz et al., 2010). Most recently, G. Chen et al. (2021) explored possibilities of using deep learning techniques to directly encode the input context (without doing any feature engineering). Next, we will review the factors that would influence the choice of referential form (which are, therefore, used in feature-based models).

Factors that influence the Selection of Referential Form

Languages display a large inventory of expressions for referring to entities (von Heusinger & Schumacher, 2019). In linguistics, the realisation choice a speaker makes has been associated with accessibility, i.e. activation of mental representations of a referent at a particular point in discourse: attenuated forms such as pronouns are often used to refer to highly accessible or highly activated referents, while richer forms such as descriptions and proper names are employed in referring to less accessible ones (Ariel, 1990; Gundel et al., 1993). Due to the central role of referring in communication, a wealth of research has tried to assess the influence of different features modulating the accessibility of a referent. von Heusinger and Schumacher (2019) refer to these features as *prominence-lending cues*, meaning that they increase the prominence status of their respective referents to some extent.

Referential status or *givenness* has been widely discussed in the literature (see Chafe (1976) and Prince (1981)). When a new character is introduced into the discourse, the chance that this happens by means of a pronoun is slim (unless the referent is situationally given). Pronouns are reserved for referring to previously introduced (or given) referents.

Recency, another well-studied cue, is defined as the distance between the target referent and its antecedent. If a referent is not too far apart from its antecedent, then reduced forms are typically employed to refer to it. When used as a feature in referential form selectors, recency is optimally measured in terms of the number of sentences (Same & van Deemter, 2020).

There are also intra-clausal cues such as *grammatical role* (Brennan, 1995) and *thematic role* (Arnold, 2001) which impact the prominence status of referents. For instance, the subject

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.

Delexicialised Text:
Pre-context: AWH_Engineering_College is in “Kuttikkattoor” , India in the state of Kerala .
Target Entity: AWH_Engineering_College
Pos-context: has 250 employees and Kerala is ruled by Kochi . The Ganges River is also found in India .

Table 2.1: An example data from the webNLG corpus. In the delexicalised text, every entity is underlined.

of a sentence is perceived to be more prominent than the object so that the referent in the subject position has a higher tendency to be pronominalised. Note that the grammatical role of both the antecedent and the current mention does matter.

Discourse-structural features affect the organisational aspects of discourse. Centering-based theories Grosz et al., 1995 often use the notion of local focus to account for pronominalisation. *Local focus* takes the current and previous utterance into account. *Global focus*, on the other hand, situates a referent in a larger space, namely the whole text or a discourse segment (Hinterwimmer, 2019). Concepts such as the importance of a referent or familiarity are associated with the global prominence status of entities (Siddharthan et al., 2011).

Animacy also plays an important role. Fukumura and van Gompel (2011) reported that pronouns were more frequent for referring to animate than inanimate referents.

End2End REG

More recently, this two-step procedure was formulated into a format that goes together well with deep learning: Castro Ferreira, Moussallem, Kádár, et al. (2018) introduced the End2End REG task, built a corresponding dataset based on webNLG (Gardent et al., 2017), and constructed NeuralREG models.

WebNLG Corpus. The webNLG corpus was originally designed to assess the performance of NLG systems (Gardent et al., 2017). Each sample in this corpus contains a knowledge base described by a Resource Description Framework (RDF) triple (Table 2.1). Castro Ferreira, Moussallem, Kádár, et al. (2018) and Castro Ferreira, Moussallem, Krahmer, et al. (2018) enriched and delexicalised the corpus to fit the REG in context task. Table 2.1 shows a text created from an RDF and its corresponding delexicalised version.

Taking the delexicalised text in Table 2.1 as an example, given the entity “*AWH_Engineering _College*”, REG chooses a RE based on that entity and its pre-context (“*AWH_Engineering_College is in “Kuttikkattoor”, India in the state of Kerala . *”) and its pos-context (“*has 250 employees and Kerala is ruled by Kochi . The Ganges River is also found in India .*”).

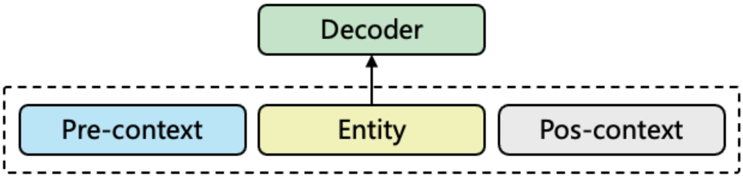


Figure 2.12: The general architecture of the NeuralREG model (Castro Ferreira, Moussallem, Kádár, et al., 2018).

NeuralREG Model. The NeuralREG model is indeed a Seq2Seq model, where the encoder is of encoding the given discourse and the referent while the decoder is of generating the RE. The model proposed by Castro Ferreira, Moussallem, Kádár, et al. (2018) has three encoders: a pre-context encoder, an entity encoder, and a pos-context encoder. Formally, for each $k \in [pre, pos]$, the model encode $x^{(k)}$ to $h^{(k)}$ with a LSTM:

$$h^{(k)} = \text{LSTM}(x^{(k)}). \tag{2.14}$$

These hidden representations are then used for computing the context representation at each decoding step using the attention mechanism (see §2.1.2 for more details), which results in c_{pre} and c_{pos}. At each decoding step, the overall contextual input is the concatenation of the referent representation v_r as well as c_{pre} and c_{pos}.

Unseen Referents. One major issue of the NeuralREG model is that it requires all referents in the test set have to also appear in the training set; consequently, the trained models fail to handle unseen referents. Recently, Castro Ferreira et al. (2019) extended the WEBNLG to include also unseen entities. Cao and Cheung (2019) and Cunha et al. (2020) developed new models to handle them. Concretely, Cao and Cheung (2019) suggested incorporating the Wikipedia data when generating REs. In this way, for unseen referents, the model could acquire corresponding knowledge (mostly text-based) from Wikipedia. One year later, Cunha et al. (2020) proposed to directly make use of the meta-information (e.g., gender and animacy of the referent) extracted from Wikipedia and to use copy mechanism (Gu et al., 2016) to further boost the performance of REG for unseen referents.

2.3 Quantification

In the previous section, we were concerned with one type of NPs, like "*the child*" and "*Joe Biden*", which are responsible for referring. This section focuses on another type of NPs, such as:

(23) a. some children
b. a few children

each of which contains a quantifier. Expressions as such are called Quantified Expressions (QEs). We introduce either the fundamental theories for quantification or computational models for understanding and producing QEs.

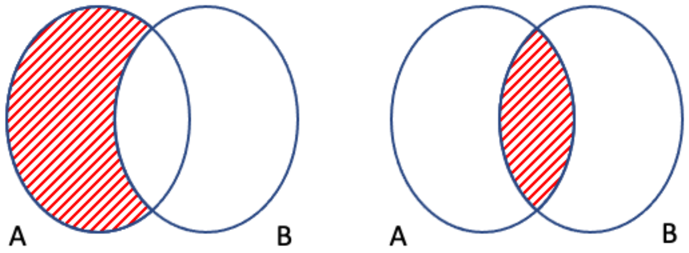


Figure 2.13: The venn diagram of *all* (left) and *no* (right). The shadow areas are empty sets.

2.3.1 Theories of Quantification

Two theories that closely tie to work in this thesis are the Generalised Quantifier theory and the Scalar Implicature theory.

Generalised Quantifier

A long tradition of research in the formal semantics of natural language asks how speakers quantify, as when we say *"Some A are B"*, *"All except two A are B"*, *"Only a few of the A are B"* and so on. This area of work is known as the theory of "Generalised Quantifiers" (GQ) (Peters & Westerståhl, 2006, GQ), because it generalises the idea of quantification beyond the standard logical quantifiers of $\forall$ and $\exists$, even including quantifiers like "*most*" or "*many*", which are not expressible in First-Order Logic (Barwise & Cooper, 1981; Mostowski, 1957; Peters & Westerståhl, 2006; van Benthem et al., 1986).

The concept of GQ has rooted in Frege as well as Aristotle. It was then mathematically formulated by Grice (1975) and was first applied to natural language by Barwise and Cooper (1981). It is based on a general idea that *the semantic values of most QEs in natural languages are relations between sets*. For instance, for a QE:

(24) Every student in this classroom wears glasses.

It mentions two sets: one is the set of students in this classroom, and the other is the set of people who wear glasses. This QE uses the quantifier *every* to express that the former set is a subset of the latter one.

Formally, GQ theory suggests that suppose we have a quantifier Q of the type $\langle 1, 1 \rangle$ [14] as well as two sets A, and B, which are subsets of the universe of discourse $\mathcal{M}$. A QE $Q(A, B)$ describe a relation between A and B. With this methodology in hand, we can define semantics of many quantifiers in natural language. For example,

(25) a. $\text{every}(A, B) \Leftrightarrow A \subseteq B$
b. $\text{some}(A, B) \Leftrightarrow A \cap B \neq \emptyset$
c. $\text{most}(A, B) \Leftrightarrow |A \cap B| > |A - B|$
d. $\text{at least three}(A, B) \Leftrightarrow |A \cap B| \leq 3$

[14] Indeed, in GQ theory, Q can be of any type $\langle n_1, n_2, ..., n_k \rangle$. Here, we use the $\langle 1, 1 \rangle$ type for simplicity.

The GQ theory also allows us to represent the semantics of quantifiers in Venn Diagrams. Figure 2.13 provide examples for the QE all(A, B) as well as the QE no$(A, B))$. More specifically, all(A, B) is saying the shadow area on the left is an empty set, while, no$(A, B))$ expresses the shadow area in the middle is empty.

Restricted Quantifiers. All quantifiers in the above examples are so-called restricted quantifiers since they range over their first inputs. For example, in (24), the quantifier *every* is ranging over the set *students*.

In contrast, there are also unrestricted quantifiers. They appear in two possible forms. One is the type $\langle 1 \rangle$ quantifiers. For example, the QE Everything(A) means A is a subset of the universe $\mathcal{M}$, suggesting that it is ranging over the universe rather than restricted by A. The other is the QEs like more(A, B), which is formally defined as:

(26) more$(A, B) \Leftrightarrow |A| > |B|$

It is not restricted by A, in cases where A and B do not overlap, e.g., "*there are more dogs than cats*".

The concept of restricted quantifiers is closely tied to the concept of *Conservativity*.

Conservativity. For a restricted quantifier, the following two examples are semantically equivalent:

(27) a. Every student wears glasses.
b. Every student is a student who wears glasses.

This phenomenon is called conservativity. Formally, it says,

(28) for each $A, B \subseteq \mathcal{M}$, we have $Q(A, B) \Leftrightarrow Q(A, A \cap B)$.

It expresses restrictness. In other words, it indicates that the truth of (27-a) only depends on elements in A. Conversely, more(A, B), as an unrestricted quantifier, does not follow this pattern, because the corresponding feature $|A| > |B| \Leftrightarrow |A| > |A \cap B|$ is easily falsified.

Universe-restricting and Extensionality. A stronger notion of restrictness is saying that the only thing that can matter to a restrict quantifier is A. In other words, it suggests that $Q_{\mathcal{M}}(A, B)$ [15] is the same as $Q_A(A, B)$. Formally, we have

(29) for each $\mathcal{M}$ and each $A, B \subseteq \mathcal{M}$, we have $Q_{\mathcal{M}}(A, B) \Leftrightarrow Q_A(A, A \cap B)$.

The principle of *extensionality* (van Benthem, 1983) indicates the difference between the conservativity and universe-restricting:

(30) for each $A, B \subseteq \mathcal{M} \subseteq \mathcal{M}'$, we have $Q_{\mathcal{M}}(A, B) \Leftrightarrow Q_{\mathcal{M}'}(A, B)$.

Extensionality also says that the semantics of quantifiers do not change with respect to the change of domains they are in.

[15] We use footnote $\mathcal{M}$ to indicate that Q is of the relation between subsets of $\mathcal{M}$.

Monotonicity. A quantifier $Q(A, B)$ is an upward monotone if and only if the following holds:

(31) if $B \subseteq B'$, the $Q(A, B) \to Q(A, B')$

Quantifiers such as *all* and *most* are upward monotone. For example, in the following example, (32-a) entails (32-b).

(32) a. all students wear black glasses
b. all students wear glasses

Conversely, a quantifier $Q(A, B)$ is a downward monotone if and only if the following holds:

(33) if $B \subseteq B'$, the $Q(A, B') \to Q(A, B)$

Quantifiers including *no* and *few* fall in this category. For example, in the following example, (34-b) entails (34-a).

(34) a. No students wear black glasses
b. No students wear glasses

Scalar Implicature

Scalar Implicature is an implicature of the implicit meaning of a QE. In other words, beyond the semantics of QEs, studies about scalar implicature are interested in the pragmatics QEs. For example, when we say:

(35) Some students wear glasses.

the use of *some* gives rise to an implicature that "*not all students wear glasses*".

Theories have been proposed to interpret such a phenomenon. One is to interpret these implicatures in terms of the Maxim of Quantity (Horn, 1972; Levinson et al., 2000). Concretely, the idea is that, for the above example, if the speaker was in a position to make a stronger statement (i.e., "*all students wear glasses*"), s/he would have. However, since s/he did not, the strong statement must be wrong. Such an account is based on the existence of lexical scales. For example, the above example covers the lexical scale ⟨*some*, *all*⟩. The stronger term (*all*) implies the weaker term (*some*), whereas the weaker term implies the negation of the strong term.

The other denies any role of lexical scalars and views the scalar inference as a contextual process (Carston, 2008; I. Noveck, 2007; Sperber & Wilson, 1986). Specifically, it says the implicature is constructed through a contextual driven ad-hoc concept construction process rather than a lexically based process.

So far, a plethora of empirical work has been done for collecting psychological evidence for each of the above accounts (Bott et al., 2012; Bott & Noveck, 2004; de Carvalho et al., 2016; Dupuy et al., 2016; Feeney et al., 2004; Hartshorne et al., 2015; I. A. Noveck, 2001; Papafragou & Musolino, 2003; Papafragou & Tantalou, 2004; Pouscoulous et al., 2007; Teresa Guasti et al., 2005).

2.3.2 Computational Models

There has been work on computational modelling of both understanding and producing QEs.

Understanding Quantifiers

Computational models were built to answer the research question of "*If a given quantified expression is uttered, what information does it convey*"? In this line of work, some work conducts empirical studies for investigating the meaning of quantifiers. For instance, Yildirim et al. (2013) investigated speakers' use and hearers' interpretation of the quantifiers *some* and *many*. Concretely, each participant was given a set of scenes, and, for each scene, the participant was asked to choose from either using *some* or using *many*. They found that *many* generally represents more quantity than *some*. In a similar vein, Herbelot and Vecchi (2015) looked at *no*, *all*, *most*, *some*, and *few*. Whereas, I. Sorodoc et al. (2016) focused on *no*, *some*, and *all*.

It has been evidenced that hears interpret quantifiers probabilistically (Degen & Tanenhaus, 2011; van Tiel, 2014; Yildirim et al., 2013). Probabilistic models like the Rational Speech Acts model (see §2.2 for more detail) have been used for modelling the meanings of quantifiers. For instance, Carcassi and Szymanik (2021) used RSA to model and compare the meanings of *most* and *more than half*.

Additionally, there is also a bank of work that focuses on disambiguating quantifier scopes. Hobbs and Shieber (1987), Saba and Corriveau (1997), and Srinivasan and Yates (2009) accomplished this task using rule-based systems. Later on, Attali et al. (2021) proposed to use RSA instead.

Producing Quantifiers

As an NLG task, there is work on generating quantifiers in either practical terms or theoretically.

Modelling the Quantifier Use. In practical terms, quantifiers, especially vague quantifiers, play important role in NLG when generating descriptions for time series, locations and so on. Yager (1982) argued that descriptions of information of numerical data are usually based on the concept of fuzzy (vague) quantified statement, such as "*a few researchers are young*" and "*most of the cold days were very humid*". Following this idea, Ramos et al. (2019) propose to use fuzzy logic to produce geographical descriptors that involve quantifiers (four quantifiers were considered: *few*, *some*, *many*, and *most*), e.g., "*many locations in the extreme north are overcast*" or "some locations in the extreme north are partly cloudy". In a similar vein, Kacprzyk et al. (2008) introduced vague quantifiers in the generation of summaries of time series.

Theoretically, works such as Franke (2014) and Qing (2014) built probabilistic speaker models for these two quantifiers, i.e., *some* and *many*, based on the RSA model. More recently, Pezzelle et al. (2018) formalised a cloze test based quantifier selection task, where they asked deep learning-based models to predict which quantifier is used in a given context.

Generating Quantified REs. Referring expressions that contain quantifiers is called Quantified Referring Expression (QRE), such as "*the crate with 10 apples*" and "*the crate with many apples*". Generating expressions as such are called Quantified Referring Expression Generation (QREG) On the human production of QRE, Barr et al. (2013) concluded that if the quantity (of the target items) is subitizable, people prefer numerical expressions (e.g., "10 apples"), because they come naturally to either speakers or listeners (Green & van Deemter, 2011). Additionally, people are more likely to use vague quantifiers (e.g., few and many) for contexts where the gap between the target quantity and the quantities in the distractors is large.

To algorithmically generate QREs, building on the above findings, (Briggs & Harner, 2019) proposed to use a method called *Perceptual Cost Pruning*, assuming that the higher the gap (between the target quantity and the quantities in the distractors) is the more the perceptual cost is. It models human QRE by 1) starting with a complete symbolic knowledge base representing the visual scene; 2) removing facts from the input knowledge base based on a model of the time cost of exact enumeration; 3) using IA on this reduced knowledge base.

CHAPTER 3

Challenges for Mandarin NLG

There is a long tradition of linguists classifying languages in the world, resulting in a research subject, namely, linguistic typology. To explain the specificity of Mandarin Chinese, we introduce three classifications that might be related to building Mandarin NLG systems. In this way, we provide a theoretical basis of potential challenges introduced by Mandarin. Specifically, these classifications include the cool-hot division (§3.1), the analytic-synthetic division (§3.2), as well as the linear-circular division (§3.3). Mandarin Chinese has been proved to be a cool, analytic, and circular language. Since the phenomena related to these characteristics happen in all Chinese languages, we use the term "Chinese" when introducing them (i.e., from §3.1 to §3.3). Meanwhile, we briefly introduce the grammar of Chinese accordingly. For a more detailed introduction of the grammar of Chinese NPs, please check Appendix A. Next, in §3.4, we enumerate possible challenges introduced by Mandarin at each stage of NLG pipeline. Since this thesis is all about challenges in Mandarin NLG, we use the term "Mandarin" in that section.

3.1 Coolness

In 1982, Ross (1982) suggested to classify languages following the "hot-cool" division of the media (McLuhan, 1964). A media is considered as "hot" if its communication process contains limited or no audience participation, while "cool" means that audience participation is required to be active. Likewise, Ross suggested that languages can also be classified on the basis of the explicitness with which they express certain elements. C.-T. J. Huang (1984) elaborated it as:

> A language is "hot" if the information required to understand each sentence is largely obtainable from what is overtly seen and heard in it. A language is "cool" if understanding a sentence requires some work on the reader's or the hearer's part.

The original metaphor of "coolness" was proposed with regards to the use of anaphora. In this sense, English is a "hot" language because English pronouns cannot, in general, be omitted while Chinese is "cool" since such pronouns are usually omittable and are often more naturally omitted. For example, in a dialogue, if a speaker says:

(36) 张三 看见 李四 了 吗？
zhāngsān kànjiàn lǐsì le ma?
Did Zhangsan see Lise yesterday?

In English, except saying "*Yes, he saw him*", none of the followings are acceptable (where *e* represents a omitted pronoun):

(37) a. * Yes, *e* saw him.
b. * Yes, him saw *e*.
c. * Yes, *e* saw *e*.
d. * Yes, I guess *e* saw *e*.
e. * Yes, John said *e* saw *e*.

Conversely, Chinese, as a "cool" language regards all the followings as acceptable:

(38) a. *e* 看见 他 了 。
kànjiàn tā le
[He] saw him.
b. 他 看见 *e* 了 。
tā kànjiàn le
He saw [him].
c. *e* 看见 *e* 了 。
kànjiàn le
[He] saw [him].
d. 我 猜 *e* 看见 *e* 。
wǒ cāi kànjiàn le
I guess [he] saw [him].
e. 张三 说 *e* 看见 *e* 。
zhāngsān shuō kànjiàn le
Zhangsan said [he] saw [him].

The contrast in acceptability suggests that Chinese has larger freedom for the use of *zero pronouns* (ZPs) than English.

There are two different types of ZPs. One is the PRO, which is a pronominal anaphor. PRO is universal across languages, and, therefore, it appears in both Chinese and English. It is always placed in the subject position of a non-finite clause, for example:

(39) He wants PRO to become a millionaire.

The other is *pro*, which is a pure pronominal and is not universal across languages (it does not appear in English), such as the example (38). It can either be recoverable, e.g., referring to the speakers or its antecedents (namely, *Anaphoric Zero Pronoun*), or be irrecoverable when *pro* is in existential sentences, idiomatic expressions, or it refers to a specific entity or event.

Subsequently, van der Auwera and Baoill (1998) pointed out that the concept of "coolness" in language science does not only cover anaphora. It can be interpreted in a way that covers a lot more categories that are not expressed obligatorily if they are obvious enough from the intra-linguistic or the extra-linguistic context. Both of the two fundamental categories in Chinese (i.e., noun and verb) could be related. For nouns, as introduced in Appendix A, a bare noun in Chinese can be either definite or indefinite and either singular or plural. In other words, given a bare noun, its definiteness as well as plurality need to be inferred from its context. Consider the following examples of the noun "狗" (gǒu; *dog*):

(40) a. 狗很聪明。
gǒu hěn cōngmíng
Dogs are intelligent.

b. 我看到狗。
wǒ kàn dào gǒu
I saw a dog/dogs.

c. 狗跑走了。
gǒu pǎo zǒu le
The dog(s) ran away.

The word "狗" in the sentence (40-a) makes a general reference and, thus, is translated as "*dogs*". In the sentence (40-b), the "狗" is an indefinite noun, but whether it refers to a single dog or a set of dogs needs to be decided by wider contexts. Likewise, the plurality of the "狗" in the sentence (40-c) is also hard to be decided without further information, but, certainly, it is definite. For verb, Chinese verb phrases accept zero aspect markers (C.-T. J. Huang, 1987). For example, the Chinese sentence "我去学校" (wǒ qù xuéxiào) could either mean "I am going school" or "I go to school". To understand the exact meaning, wider contexts are needed.

Putting this hypothesis in the context of NLG, it links to the concept of the clarity-brevity trade-off in NLG. Recall that, in Gricean Maxims, both the maxim of quantity and the maxim of manner require the production of language to be clarity and brevity. Nevertheless, in practical terms, there is a trade-off between them since reductions in ambiguity are achieved by increases in length. Building on the existence of such a trade-off, Zipf (2016) hypothesised that individuals maintain an efficient balance between over- and under-specifying in an intended message. In NLP such a phenomenon has been explored/discussed in the context of REG (Khan et al., 2006) and lexical choice (Pimentel et al., 2020). Interestingly, it has been suggested that East Asian languages handle the trade-off between brevity and clarity differently from those of Western Europe (e.g., Newnham (1971)). Pimentel et al. (2020) concluded that, at least for lexical choice, such a trade-off is language-dependent, some languages prefer clarity over brevity whereas some languages prefer brevity over clarity. [1] Based on this idea, we could imagine that the coolness hypothesis might suggest that the Chinese would prefer brevity over clarity in such a trade-off and a Chinese NLG system should be able to capture this.

In addition, since many elements (e.g., definiteness marker, plural marker, pronoun and aspect marker) in Chinese are optional in some circumstances, there could be multiple ways (i.e., omit or not omit these elements) to express the same meaning. Although both omitting them and not omitting them are grammatical and acceptable, as C.-T. J.

[1] Unfortunately, the corpus study in Pimentel et al. (2020) excludes Chinese.

Huang (1984) pointed out, one of them is often more pragmatically natural than the other. Therefore, another general challenge posed by the "coolness" hypothesis is how to choose the pragmatically natural one from multiple alternatives.

3.2 Analyticity

Another classic classification in linguistic typology is the "analytic-synthetic" division. Concretely, analytic languages express concepts using independent words while synthetic languages use inflected words for the same purpose. Simply put, analytic languages have no or very few inflectional morphemes while synthetic language uses a lot. Given this criterion, German, Latin, Russian, as well as ancient English are all synthetic languages. Chinese and Modern English are both analytic languages, but Chinese is more analytical than English.

Chinese is a language that has no inflectional morpheme and, as a trade-off, it needs more syntactic rules than synthetic languages and other less analytical languages (e.g., English). To see in which way Chinese is analytic and why Chinese is more analytical than English, consider the following example:

(41) 张三 的 朋友 都 来 了 。
zhāngsān de dōu lái le
All Zhangsan's friends has come.

Although modern English has much fewer inflectional morphemes than synthetic languages, it is still weakly inflected. In this specific example,

1. Chinese uses a particle "的" to mark possession ("张三 的 朋友"), while English uses a bound morpheme "-'s" in the form of a clitic ("*Zhangsan's friends*");

2. Chinese use a bare noun "朋友" which express plurality in an implicit way, while English uses a bound morpheme "-s" to form a plural noun "*friends*";

3. Chinese use a particle "了" as a perfective marker in this sentence, denoting the friends "has come".

Note that Chinese has a great number of derivational morphemes. For example, the compound of "红" (hóng; *red*) and "花" (huā; *flower*) is "红花" (hónghuā; *saffron crocus*). Please see Packard (2000) for more details.

3.3 Circularity

Languages can also be classified using a so-called "linearity-circularity" division (Kaplan, 1966). This idea considers English a linear language while considers Chinese a circular language because Mandarin speakers prefer using indirect expressions. This is based on the idea that the discourse pattern of English is linear, direct, deductive, and logical while that of Mandarin is inductive, indirect, and non-linear. Real or purported evidence was identified in sales letters. For example, Campbell (1998) found that English sales letters address directly the point in the headline of the letter whereas Chinese sales letters mention something seemingly irrelevant. For instance (in English; from Yunxia (2000)):

(42) a. Headline: Introducing the only credit card to give you $60 to spend on Innovations - FREE!
b. Greetings: How are you? You must be very busy with your work.

Later on, L. Yang and Cahill (2008) compared the essays written by Mandarin and English speakers and checked the place of thesis statements (i.e., the sentences that express the main idea of essays). They found 86% of the English native speakers placed thesis statements in the initial sentence while merely 66% of the Mandarin speakers did the same.

This is because of the, real or purported, Chinese speakers' preference in being "roundabout". This, in part, builds on an old saying from Confucian: *"a word uttered by a gentleman cannot be taken back, even by a team of four horses"*. In other words, Chinese speakers are eager to be "implicit" and "indirect". For example, if a Chinese speaker is asked to do something that s/he cannot do, instead of directly saying "no", s/he prefers saying "It "s not convenient for me today" because it is less negative and more likely to avoid conflicts.

Regarding circularity, there are two possible interpretations. One links circularity to the theory of Hall (1989) where cultures are divided into *high context* cultures (e.g., China, Japan, and Korea) and *low context* cultures (e.g., United States, Australia, and New Zealand). High context cultures are cultures whose rules of communication rely heavily on contextual elements, such as body language, a person's status, and tone of voice. Due to globalisation, communications these years are often cross-cultural. The communication style in a high context culture appears to be indirect to someone from low context cultures (Yunxia, 2000). The other interpretation views circularity as a kind of politeness since indirect utterances are often more polite than direct utterances because the former can offer more options for the addressee (Leech, 2016).

The first interpretation appears to link the idea of circularity to the idea of coolness. On the one hand, circularity, in line with coolness, also leads a language to breach clarity (in order to be indirect). On the other hand, it is in contrast to coolness, which hypothesises Mandarin speakers preferring brevity, since, to be indirect, speakers sometimes need to speak more.

In response to circularity, Chinese NLG systems might need to adapt their communication goal and the way how a generated discourse is structured. Note that since circularity does not affect the Noun Phrase, it's not a topic for this thesis

3.4 Challenges

We hereby list the potential challenges in accordance with the NLG pipeline.

3.4.1 Content Determination

We believe content determination is less language-dependent. [2] For example, when building a system that generates biographies from a knowledge base, the generated texts are mostly in a form like *"Joseph Robinette Biden Jr. is an American politician who is the 46th and current president of the United States ..."* regardless of languages.

One issue that needs to be aware of is ellipses. For instance, the following dialogue is an interaction between the *EasyNav* system (developed by Y. Huang et al. (2001) aiming at introducing the campus of Tsinghua University to their new students) and a real user:

[2] Note that the content determination of REs will be discussed later.

(43) USER: 请问 计算机系 学生 住在 哪里?
qǐngwen4 jìsuànjīxì xuéshēng zhùzài nǎlǐ
Can I ask where is the accommodation for students of the Computer Science Department?
EasyNav: 9 号楼。
9 hàolóu
Building No.9.
USER: 附近 都有 哪些 食堂?
fùjìn dōuyǒu nǎxiē shítáng
Are there any refectories nearby?
EasyNav: 七 八 九 食堂 。
qī bā jiǔ shítáng
Refectory No.7, No.8 and No.9 [are near here]
USER:哪个 最好?
nǎgè zuìhǎo
Which [one] is the best?
EasyNav: ...

In the second turn, although the user simply asked about refectories nearby, pragmatically, s/he actually meant refectories near building number 9. In a similar vein, in the next turn, s/he meant "*which refectory near building number 9 is the best*", but both "*refectory*" and "*near building number 9*" are omitted. Indeed, ellipsis also appears in other languages, including English, but their uses are different. Osborne and Liang (2015) figured out that, in addition to ZP, noun ellipsis, such as those omitted nouns in the above dialogue session, is less restricted and, thus, more frequent in Mandarin than in English. Meanwhile, some other types of ellipsis such as "sluicing" happen only in English. This said, the macro-planner [3] in a Mandarin NLG system needs to handle ellipsis differently from English NLG systems.

3.4.2 Document Structuring

Apparently, the aforesaid issue of ellipsis also matters in the stage of content structuring as most previous work suggested performing ellipsis (for English) when managing the content structure (e.g., using Rhetorical Structure Theory (Theune et al., 2006)).

Besides, there are also other challenges. First, regarding circularity, empirical studies (e.g., L. Yang and Cahill (2008)) are mainly about essay writing. It is interesting to conduct extensive studies on other aspects in human language production to ascertain such a hypothesis (i.e., Chinese is a circular language) and to investigate to what extent it influences the use of language as well as the design of the NLG system. If confirmed, it requires the document structuring of a Mandarin NLG system to be able to manage the information in a circular way. For example, it may sometimes not place the topic sentence at the very beginning. Second, the position and order of Mandarin discourse markers are very fluid. Syntactically, they can be placed in any of the initial positions, the predicate-initial position, and the final position. In contrast, English uses them in a more

[3] We use the term "macro-planner" because some studies (e.g., Theune et al. (2006)) suggested that ellipses are better to be handled during Document structuring.

fixed way, where the discourse makers are mostly in the initial position (Y. Li, 2008). Third, many discourse makers are also omittable. For the sentence

(44) Because he was ill, he did not go to school yesterday.

Literally, its Mandarin translation needs to use the discourse marker “因为 ... 所以 ...” (yīnwéi ... suǒyǐ ..., because ... so ...). However, all of the following expression are acceptable and are expressing the same meaning:

(45) a. 他 病了， 没 来 上课 。
tā bìngle, méi lái shàngkè
b. 他 因为 病 了， 没 来 上课 。
tā yīnwéi bìngle, méi lái shàngkè
c. 他 因为 病 了， 所以 (他) 没 来 上课 。
tā yīnwéi bìngle, suǒyǐ (tā) méi lái shàngkè
d. 他 病 了， 所以 没 来 上课 。
tā bìngle, suǒyǐ méi lái shàngkè

The evidence towards the last two points is coming from a corpus study (Y. Li, 2008) aiming at reminding machine translation researchers to be aware of these issues. There has no empirical study to approve, for example, Mandarin has more freedom with respect to the choice of discourse markers than English. Therefore, they (i.e., the above two points) are both worth further investigation. If they are both true, then they will raise new challenges for document structuring in Mandarin on deciding the position, the order, and even the existence of discourse markers to make the resulting document more pragmatically natural.

3.4.3 Aggregation

On the basis of the coolness hypothesis, one could expect that Mandarin speakers would use more aggregation either semantically or syntactically, in order to shorten the length and reduce the speech cost. On the one hand, this asks the aggregators to be more active in Mandarin NLG systems. On the other hand, since the definiteness and plurality in Mandarin could be expressed implicitly, an aggregation might introduce ambiguities. This said aggregators should be aware that ambiguities as such might result in unsuccessful communications.

3.4.4 Lexicalisation

During lexicalisation, on the one hand, one could expect that Mandarin NLG systems need to make more decisions at this stage compared to English NLG systems. These additional decisions come from:

- Optional elements: definiteness markers, plural markers, and aspect markers are all optional in some circumstances. Omitting them properly would improve the human-likeness/naturalness of the generated texts;

- Synonyms: in addition to the traditionally understood “synonyms” (i.e., words that have similar or the same meaning), 80%-90% of Mandarin words can be expressed by either a short form or a long-form, such as “虎/老虎” (hǔ/lǎohǔ; *tiger*) and “店/商

店" (diàn/shāngdiàn; *shop*). It is crucial for a Mandarin NLG system to choose the correct form, which sometimes highly influences the naturalness of the outputs. For example, in the following example, filling the missing word with either "店" or "商店" always means "*shop*" and is grammatically correct, but using "店" will result in a more natural sentence.

(46) 她 去 日本 旅游 时，必逛 各种 免税 ____ 。
tā qù rìběn lǚyóu shí, bì guàng gèzhǒng miǎnshuì ____
When travels to Japan, she must go duty-free ____.

Initial analysis of such a problem using language modelling tools can be found in L. Li et al. (2020), L. Li et al. (2019). Literature in linguistics (e.g., Arcodia and Basciano (2017)) suggests that, different from the long forms, most short forms are bounded morphemes so that they cannot independently occupy a syntactic slot and have to be combined with other words (e.g., 免税 (duty-free) in the above example).

On the other hand, since predisposing brevity might breach clarity, we expect there are more vague words in Mandarin texts in terms of either frequency or variety. It might also be connected to the aforementioned idea of being roundabout, which can be achieved by being vague and ambiguous. This suggests that, when building Mandarin NLG systems, we need to pay more effort into modelling the meanings of vague terms before producing them.

3.4.5 Referring Expression Generation

We discuss the challenges for one-shot REG and REG in context respectively. Regarding one-shot REG, a TUNA-like experiment was conducted by (van Deemter et al., 2017), in which the MTUNA corpus was introduced. An initial analysis concluded that 1) almost all REs are bare nouns (with properties as pre-modifiers) and number phrases, while demonstratives rarely appear; and 2) syntactic positions of REs matter. Building on MTUNA and findings in the initial analysis, we further consider the following possible challenges:

1. Mandarin allows an NP to have no head noun. Recall that all classic REG algorithms (e.g., the full brevity algorithm, the greedy algorithm, and the incremental algorithm) add the TYPE property regardless of its discriminatory power. This might not always be true for generating REG in Mandarin. For example, "红的" (hóngde; *red*) is an acceptable RE if the target object's TYPE has zero discriminatory power. This requires that, to be more human-like, an REG algorithm needs to treat TYPE non-deterministically;

2. van Deemter et al. (2017) found that only 60%-70% of REs referring to plural referents were marked with numbers. Also considering that Mandarin could express plurality implicitly, the effectiveness of the plural REG algorithms for Mandarin is challenged. For example, if the goal is to single out two chairs from tables, then simply saying "椅子" (yǐzi; *the chair/the chairs*) is ambiguous as it is unclear whether it refers to one of the two chairs or the set of two chairs;

3. Due to the Mandarin's preference for brevity, it is natural to expect brevity-first algorithms (e.g., the full brevity and the local brevity algorithm) would receive better performance than, for example, the greedy algorithm;

4. With a similar reason, although previous studies suggested that, in English, overspecification benefits both speakers and listeners, we expect there are fewer overspecifications and more under-specifications in Mandarin than in English. In aggregate, one-shot REG in Mandarin might ask for an algorithm that is non-deterministic, targeting less at over-specifications and allows under-specifications.

Regarding REG in context, recall that Huang's initial interpretation of the "coolness" hypothesis concerned the use of zero pronouns. This requires a referential form selector to also naturally use non-overt REs. The "coolness" hypothesis also includes that a RE can be definite and plural without explicit markers. When deciding the content of them, a REG model should guarantee that the use of REs as such will not result in referential ambiguity. Additionally, the syntactic structure of NPs in Mandarin could be considerably complex (see Appendix A for more details). The REG algorithms might meet new challenges when generating REs like:

(47) 张三 他们 那 三个 学生
zhāngsān tāmén nà sāngè xuéshēng
the three students that include Zhangsan

3.4.6 Realisation

We consider three possible challenges for linguistic realisation posed by Mandarin. First, since many elements (e.g., aspect marker, particle, plural marker and so on) in Mandarin are sometimes optional, the job of a surface realiser is to provide interfaces that allow as many options as possible. For instance, a good realiser needs to enable its users to either use explicit aspect markers or zero aspect markers. Additionally, to ensure grammaticality, the realiser should also include linguistic constraints on when each of these options can be chosen and when they cannot. Second, as an analytic language, compared to western languages, Mandarin has no inflectional morphology and has more varieties of syntactic structures. This said, a Mandarin surface realiser should have much fewer morphological operators and much more syntactic operators than realisers for western languages. Last, the Mandarin realisers need to handle classifiers, such as the "本" (běn) in the number phrase "三本书" (sānběnshū; *three books*). Dictionaries (i.e., associating each noun in Mandarin with a fixed classifier) can handle most cases but not all. Exceptions include, for example, the choice of the classifier of the noun "房" (fángzi; *house*). In this situation, the choice of classifier relates to the exact meaning of "房", which needs to be reasoned from a wider context. If the classifier "间" (jiān) is chosen, then it (i.e., "一 间 房") will mean "a room", while if the classifier "栋" (dòng) is chosen, then it (i.e., "一 栋 房") will mean "a house".

CHAPTER 4

One-shot Referring Expression Generation

Abstract. *One-shot Referring Expression Generation (REG) is about producing referring expressions (REs) from non-linguistic contexts. One-shot REG algorithms offer computation models for the production of these REs. In earlier work, a corpus of REs in Mandarin was introduced. Building on this corpus, on the one hand, we introduce a new perspective on the various ways in which a RE can refer, or fail to refer, to its target referent. We argue that our perspective enables a more fine-grained understanding of reference phenomena than before, with potential implications for Natural Language Processing. With this new perspective in hand, we offer an in-depth analysis of the corpus, focusing on issues that arise from the grammar of Mandarin, and compare the use of REs between Mandarin and English. Perhaps most strikingly, we found a much higher proportion of under-specified expressions than what previous studies had suggested, not just in Mandarin but in English as well. On the other hand, we annotate the corpus, evaluate REG algorithms on it, and compare the results with earlier results on the evaluation of REG for English.*

—

The publications related to this chapter are:

1. Chen, G., & van Deemter, K. (2020). Lessons from computational modelling of reference production in Mandarin and English. *Proceedings of the 13th International Conference on Natural Language Generation*, 263–272. https://www.aclweb.org/anthology/2020.inlg-1.33
2. Chen, G., & van Deemter, K. (2021). Varieties of specification: Redefining over- and under-specification for an enhanced understanding of referring expressions. *Journal Paper in Preparation*

4.1 Introduction

The primary function of a referring expression (RE) is to help hearers identify what a speaker is thinking about: the intended referent. The task of one-shot Referring Expression

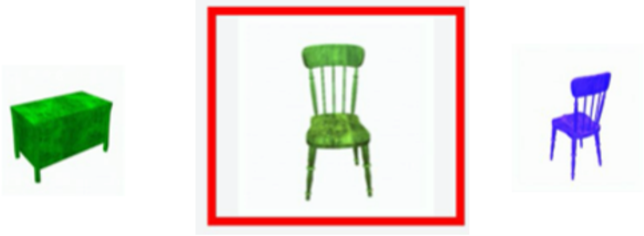


Figure 4.1: A simple scene that requires speakers producing REs to single out the object in the red window from others.

Generation (REG) is to design algorithms to generate such expressions from visual scenes. For example, to generate a RE for the object in the red window of Figure 4.1, an algorithm generates the expression (48-a) to accomplish a successful communication (i.e., singling out the target referent from its distractors). One-shot REG is not a deterministic task because, given a situation, multiple REs can all accomplish the task.

(48) a. the large one
b. the green chair
c. the large chair
d. the large green one

This task, on the one hand, has important practical value in natural language generation (Gatt & Krahmer, 2018), computer vision (Mao et al., 2016), and robotics (Fang et al., 2015), where, most recently, Neural Network based models (e.g., Castro Ferreira, Moussallem, Kádár, et al. (2018) and Mao et al. (2016) and Cao and Cheung (2019)) have started to be used. On the other hand, theoretically, REG algorithms can also be seen as models of human language use (van Deemter, 2016).

In the second line of work, previous studies focus on two aspects (cf., Krahmer and van Deemter (2012)):

(a) Designing and conducting controlled elicitation experiments, yielding corpora which are then used for analysing the use of REs and evaluating REG algorithms to gain insight into linguistic phenomena, e.g., GRE3D3 (Dale & Viethen, 2009), TUNA (Gatt et al., 2007; van Deemter, Gatt, Sluis, et al., 2012), COCONUT (Jordan & Walker, 2005), and MapTask (Gupta & Stent, 2005);

(b) Designing transparent algorithms (in opposite to black-box neural network models) that mimic certain behaviours used by human beings, for example, the maximisation of discriminatory power (Dale, 1989) and/or the preferential use of cognitively "attractive" attributes (Dale & Reiter, 1995); see Gatt et al. (2013) for discussion.

The focus of these studies was mostly on Indo-European languages, such as English, Dutch (Koolen & Krahmer, 2010) and German (Howcroft et al., 2017). Recently, researchers have started to have a look at Mandarin Chinese (van Deemter et al., 2017), collecting a corpus of Mandarin REs, namely MTUNA. More interestingly, they also hypothesise a link between the use of REs and the idea of "coolness" (C.-T. J. Huang, 1984; Newnham, 1971). Concretely, since Mandarin is "cooler" than western languages, Mandarin speakers might favour brevity over clarity when producing REs. So far, only a preliminary analysis has been performed on MTUNA, and this analysis has focused on issues of Linguistic Realisation (van Deemter et al., 2017): (a) the REs in the corpus have not yet been analysed and been compared with those in other languages; and (b) the performance of REG algorithms on the corpus has not been evaluated. Therefore, the proposed hypothesis above has not yet been tested.

To fill this gap, for issue (a), we analyse the REs in MTUNA. Since the idea of coolness suggests that the production of REs in Mandarin might rely more on communicative context for disambiguation than western languages, our analysis focuses on the amount of information that REs use. Specifically, we concentrate on the use of REs that contain more semantic information or less semantic information than is strictly necessary for identifying the intended referent. These two kinds of REs are called over-specification and under-specification, respectively. If Mandarin favours brevity over clarity to a greater extent than languages like English, then one would expect to see less over-specification and more under-specification in Mandarin.

There has been a long tradition of studying the use of REs (Dale & Reiter, 1995; Engelhardt et al., 2006; Engelhardt et al., 2011; Koolen et al., 2011; Paraboni et al., 2017; Pechmann, 1989). They defined over-specification in essentially Gricean terms. Particularly relevant theory is the Maxim of Quantity from the *Gricean Maxims* (Grice, 1975), which is comprised of two rules:

1. the speaker should make the contribution as informative as required;

2. the speaker should not make the contribution more informative than is required.

For these researchers, if an expression violates the second rule of quantity, then it is considered to be an over-specification. Similarly, if an expression breaks the first rule, then it is considered an under-specification. This approach, which has informed many theories and computational models, has a number of important limitations: most obviously, it does not specify what or how much information is "required". This hinders us from conducting in-depth analysis, quantitatively. For example, both expression (48-a) and expression (48-b) provide all the "required" information, allowing hearers to identify the target objects. Likewise, both descriptions contain no redundant information since none of their words could be omitted. Therefore, it is reasonable if we exclude both of them as over-specifications in our analysis. Nevertheless, the description (48-b) mentions two attributes (i.e., `COLOUR` and `TYPE`) whereas (48-a) mentions only one (i.e., `SIZE`), suggesting that, in some sense, the former conveys more information than required. Equally importantly, the above approach lumps together situations that are intuitively very different, designating them all as over-specifications, and nothing else. For instance, it is known from experiments (Levelt, 1993) that an utterance such as (48-c) (which is technically an over-specification because `TYPE` can be removed without causing referential ambiguity, so it's not "required") is much more commonly produced than descriptions like (48-d). This is thought to be because the role of `TYPE` is different from `COLOUR`, for example, because "chair" helps the speaker to construct a grammatical correct noun phrase (NP) in English (Dale & Reiter, 1995). A fine-grained analysis of the over-specification phenomenon should arguably distinguish between those over-specifications that are caused merely by the presence of a logically superfluous TYPE attributes, and those over-specifications that are caused by other attributes.[1]

Therefore, we propose a new perspective on specification (i.e., over-specification as well as under-specification), and developed an annotation scheme for annotating different types of specifications accordingly. With this scheme, we annotated MTUNA, and provide a more detailed analysis of the use of Mandarin REs on the basis of MTUNA.

For issue (b), we first annotated the MTUNA corpus in line with the annotation scheme of TUNA (van der Sluis et al., 2006), after which we used this annotation to evaluate the classical REG algorithms and compared the results with those for the English TUNA corpus (abbreviated as ETUNA). We also show that our new perspective on specification can also help with analysing and comparing the results of different REG algorithms.

4.2 The Mandarin TUNA

In Chapter 2, we introduced the background of the TUNA experiment, and the ETUNA corpus. We hereby briefly introduce its Mandarin processor: MTUNA, and highlight some special features of MTUNA together with its initial findings. The different TUNA corpora were set up in highly similar fashion: for instance, they all use a few dozen stimuli, which were offered in isolation (i.e., participants were encouraged to disregard previous scenes and previous utterances), and chosen from the same sets of furniture and people images; furthermore, participants were asked to enter a type-written RE following a question.

Yet there were subtle differences between these corpora as well, reflecting specific research questions that the various sets of authors brought to the task. The stimuli used by MTUNA were inherited from the Dutch TUNA (Koolen & Krahmer, 2010), where there are total of 40 trials. Unlike other TUNAS which always asked subjects essentially the same question, namely *Which object/objects appears/appear in a red window?*, MTUNA distinguished between REs in subject and object position. [2] More precisely, subjects were asked to use REs for filling in blanks in either of the following patterns:

(a) ____ 在红色方块中
'Please complete the sentence: ____ is in the red frame(s)'
(b) 红色方块中的是 ____
'What's in the red frame is ____'

where (a) asked subjects to place the RE in subject position while (b) asked to place it in object position. The initial analysis in van Deemter et al. (2017) focused on how definiteness was expressed in Mandarin REs. They found that most definite REs are bare nouns; and indefinite REs also appear quite often, especially in the subject position.

4.3 Research Questions

We start with analysing the REs in TUNA. The coolness hypothesis stated that Chinese relies more on the communicative context for disambiguation than western languages,

[1] Accordingly, classical REG algorithms, like the incremental algorithm (Dale & Reiter, 1995), give TYPE a special role, i.e., giving every RE a TYPE, regardless of whether this contributes to the ability of the expression to refer uniquely.

[2] This was done because the literature on Mandarin (e.g., Chao (1965)) suggests that Mandarin NPs in pre-verbal position may be interpreted as definite unless there is information to the contrary.

such as English, based on which Chinese is also seen as a discourse-based language while English is a sentence-based language. The existence of primary evidence for this issue in REG was identified in van Deemter et al. (2017), indicating that Mandarin speakers rarely explicitly express number, maximality and giveness in REs, and in G. Chen et al. (2018b), indicting that they sometimes even drop REs. In this study, we were curious about the following research question, which consists of two parts:

RQ1a We were curious about *the use of over-specification and under-specifications in* MTUNA *versus* ETUNA, hypothesising that Mandarin REs use fewer over-specifications and more under-specifications than English;

RQ1b We were curious about *the use of over-specification and under-specifications in* MTUNA *versus* ETUNA*, respectively.* More specifically, in TUNA experiments (for whatever ETUNA and MTUNA), the people domain is designed to be more complex than the furniture domain in the sense of two dimensions: 1) scenes in the people domain use real photographs of people, which allows more alternative attributes for subjects to choose from compared to the artificial pictures in the furniture domain[3]; 2) since all the objects in the people domain are male scientists, the objects in a scene is arguably more perceptually similar to each other. The higher domain complexity makes the targets in the people domain require more efforts to refer to and, consequently, subjects tend to over-specify more frequently. This is consistent with van der Sluis and Krahmer's findings on speakers tending to use more words when referring to targets in difficult tasks. Therefore, it is plausible if one expects that domain complexity has a positive influence on the use of over-specifications in that domain.

As discussed, to analyse the use of over- and under-specifications, we require a new perspective on defining and categorising them in order to allow quantitative analysis. Therefore, in the first study of this chapter, we introduce a new perspective of specification, in which we propose to (re-)define and sub-categorise over- and under-specification. We then use these definitions to annotate both MTUNA and ETUNA corpora and compare the use of specifications.

Analogous to studies of earlier QTUNA corpora, another research question (**RQ2**) is *how classic REG algorithms perform on* MTUNA *and how this is different from the performance on* ETUNA*?* We were curious to see whether the value of each evaluation metric for each algorithm (which will be introduced in §2.2) will change very much, and whether the rank order of the algorithms stays the same. If, as hypothesised, Mandarin prefers brevity over clarity, then the Full Brevity algorithm (which always yields REs with minimal number of properties), is expected to have higher performance on MTUNA than on ETUNA. The expected effect on other classic algorithms is less clear. To answer this question, we annotate the MTUNA corpus (enabling it to be used for evaluation), evaluate REG algorithms on the annotated MTUNA, and compare the results to that on ETUNA. Also, we hope our new perspective can help understand the performance of each algorithm.

It is thought that, since TYPE helps create a "conceptual gestalt" of the target referent (which benefits the hearer (Levelt, 1993, Chapter 4)), speakers tend to include a TYPE in their REs regardless of its discriminatory power. [4] For this reason, algorithms such as

[3] Note that earlier studies by van der Sluis and Krahmer (2007) showed that although the use of attributes in the people domain is less controlled, the subjects prefer to use a certain subset the available attributes than others.

[4] Note that 92.25% of the REs in ETUNA contain a superfluous TYPE (van der Sluis et al., 2007)).

the Incremental Algorithm (Dale & Reiter, 1995) always append a TYPE to the REs they produce. However, Lv (1979) found that the head of a noun phrase in Mandarin is often omitted if this noun is the only possibility given the context. This suggests that, if all objects in a scene share the same TYPE (e.g., all the objects in the people domain of TUNA are male scientists), then it is less likely for Mandarin speakers to express a TYPE. Accordingly, our third research question (**RQ3**) asks *to what extent the role of TYPE differs between English and Mandarin*. This asks, on the one hand, that we need to take the use of TYPE into consideration when analysing the use of over-specifications. On the other hand, we are also curious to what extent this issue affects the performance of the classic REG algorithms.

We have seen that MTUNA asked its participants to produce REs in different syntactic positions. van Deemter et al. (2017) found more indefinite NPs in the subject position, which is inconsistent with linguistic theories (James et al., 2009) that suggests subjects and other pre-verbal positions favour definiteness. Building on these findings, our last research question (**RQ4**) is about *how syntactic position influences the use of over-/under-specification and the performance of REG algorithms*. This would use either our results of evaluating REG algorithms on MTUNA as well as ETUNA or our results of annotating these two corpora using the new perspective.

In a nutshell, to answer the research questions in this section, we need to annotate both MTUNA and ETUNA corpora with both the use of specifications and the used properties.

4.4 Study 1: Modelling Varieties of Specification

In this section, we will first motivate the necessity of a new perspective of specifications using observations from MTUNA and ETUNA. We then offer an explanation of the new perspective we are proposing.

Note that, in this study, we focus on simple situations, in which the wider “context of use" of an NP does not play a role. This will allow us to keep our definitions simple and our annotation scheme easy to use. When the context is taken into account, this can often affect interpretation. Given an appropriate context, for example, we can say “*the dog*” to refer to a particular dog, even though there are many dogs in the world, as long as the intended referent is the contextually most salient dog (Krahmer & Theune, 2002). Viewed in isolation, “the dog" is under-specified but viewed in context, it may be very clear. In the relatively simple corpora on which we focus in this study, i.e., MTUNA and ETUNA, context does not play such a disambiguating role. The role of context is discussed briefly in §4.4.7.

4.4.1 When and How Do Speakers Over-specify?

To understand why over-specification and under-specification occur, we need to know when and where they occur. This requires precise definitions of both over- and under-specification. As discussed, a common practice uses the Gricean Maxim of Quantity. On the basis of the second principle of Gricean Maxim of Quantity, the concept of over-specification is firstly defined as an RE that is more informative than required. In that sense, if we represent an RE as a set of properties, it covers situations in which an RE includes non-required properties while managing to identify the referent. Such a definition works fine in psycho-linguistic studies that investigate over-specifications with only one superfluous property. However, when focusing on general use of RE opinions were divided. Due to some vague terms in the previous definition, we found that:

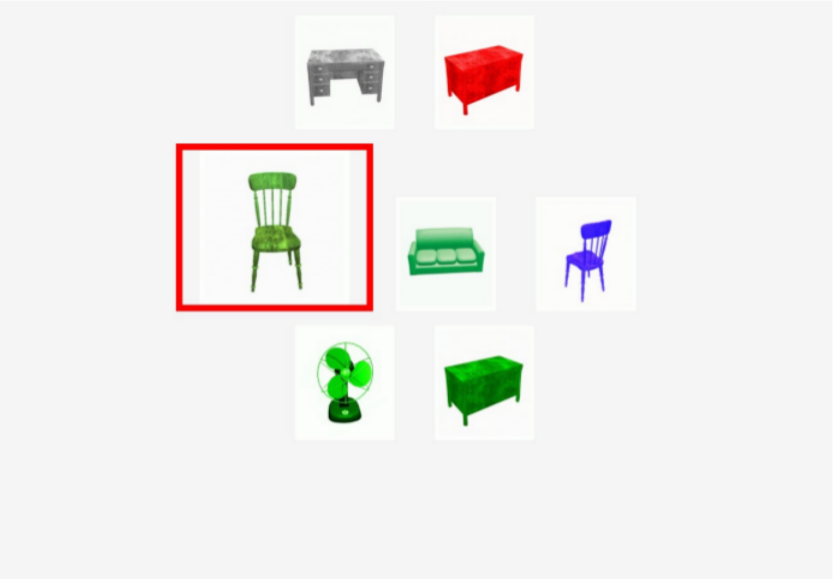


Figure 4.2: A trial from the MTUNA corpus elicits over-specifications that are not covered by the previous definition of over-specification.

- it is unclear what types of over-specification have been included; and
- the definition overlooks vital differences between over-specifications

by providing examples of these problems in the MTUNA and ETUNA[5] corpora.

The first type of problem was found when we look into the REs from the MTUNA experiment, whose scene is shown in Figure 4.2. When subjects referred to the chair in the red window, they could say any of the REs in Example (49), where the "MD" mark indicates the current description uses the minimum number of properties, namely, minimal description:

(49) a. the large one (MD)
b. the large green one
c. the green chair

As there is only one large object in the scene, using only one property: ⟨SIZE, large⟩, is sufficient for successful communication, as in (49-a). However, except over-specifications, such as (49-b) whose property ⟨COLOUR, green⟩ removable without breaking the communication successful, we found substantial amounts of descriptions like (49-c), which uses two properties: ⟨COLOUR, green⟩ and ⟨TYPE, chair⟩. This results in the question of whether the descriptions like (49-c) over-specify? Clearly, description (49-c) uses more properties than the minimal description. But, simultaneously, none of its properties is superfluous. The answer to this question relies on how we interpret the phrase "required" properties in the Gricean Maxim of Quantity. If we interpret it as: properties whose removal would

[5] Since this new perspective is applicable regardless of language, we translate all examples from MTUNA to English in this section.

prevent successful communication, as in Koolen et al. (2011) and van Deemter (2016), then description (49-c) is no over-specification since any removal of its property would result in an under-specification. But if we interpret it as: a set of properties that contains the *minimum* number of properties required for unique identification (as in Dale and Reiter (1995) and Gatt and van Deemter (2007b)), then (49-c) is an over-specification. This suggests that a more precise definition of quantification is needed.

In the corpora, we observed a large amount of RE like:

(50) the large chair

in which only the TYPE attribute is superfluous. As mentioned previously in this study, these cases differ importantly from over-specification like (49-b) because English speakers tend to always include TYPE (Levelt, 1993) (which is reflected in REG algorithms (Dale & Reiter, 1995)). Note that this is bound to be different in languages that allow zero head nouns in noun phrases.

Sometimes, probably because subjects intended to highlight some specific properties of a referent, they describe a single property more than once in a single expression. For example, when referring to a backward table, some subjects said:

(51) the backward table with invisible drawers

in which, semantically, both the phrase *backward* and *with invisible drawers* are talking about the ORIENTATION attribute. Based on Gricean Maxim of Quantity, it is unclear whether description (51) is an over-specification or not if either TYPE or ORIENTATION are superfluous? Moreover, it is also unclear when we are heading to quantify the human use of over-specification if in this example ORIENTATION is a superfluous attribute, should the total number of superfluous attributes be added by 1 or by 2? We hope by introducing an annotation scheme with precise definitions of over-specification with corresponding superfluous attribute counting strategy in this work, we could have good answers to these questions.

Certain issues are particularly related to expressions that refer to sets. Suppose we intend to refer to two red pieces of furniture: one is a chair and the other is a table. Suppose the minimal description contains only one property: ⟨COLOUR, red⟩, we could say:

(52) a. the red chair and the red table
b. the red chair and the table
c. the red furniture

It is hard to tell the difference between these descriptions qualitatively and quantitatively. For example, should the lack of "aggregation" (e.g., merging the two occurrences of *red* in (52-a)) be seen as a type of over-specification? Or should the use of the words *chair* and *table* be seen as a kind of over-specification, because these words offer more details than required?

Not only the Maxim of Quantity but also the Maxim of Quality should play a role in the analysis of an RE corpus, because a substantial number of REs in the corpora are simple "incorrect" because one or more of the properties expressed are not applicable to the target referent. For example, speakers sometimes say *red chair* to a green chair. Earlier analysis of these corpora (Gatt et al., 2007; van der Sluis et al., 2007; van Deemter, Gatt, Sluis, et al., 2012; van Deemter et al., 2017) neglected these incorrect descriptions, which arguably introduced noise into the overall analysis that these papers were able to offer.

4.4.2 Varieties of Referential Specification: a Formal Account

In what follows, we use the insights discussed above to propose a new way of thinking about what it means to talk about an intended referent (to “specify" the referent) that is both linguistically insightful, and formally precise enough that it can be employed to annotate corpora and to quantitatively evaluate NLG algorithms, especially those that produce REs.

It will be useful, in this discussion, to use plenty of examples of concrete utterances. These will frequently include REs from the MTUNA (or ETUNA) corpora. Nonetheless, our definitions and annotation scheme (§4.4.4) are designed to have much wider validity.

Preliminaries

Following the literature on REG (e.g., Krahmer and van Deemter (2012)), we distinguish between attributes and properties. For instance, we call `COLOUR` an *attribute*. Attributes have values. For example, the attribute `COLOUR` may have values such as red, blue, and so on. An attribute-value pair, such as ⟨`COLOUR`, blue⟩ or ⟨`TYPE`, chair⟩, is called a *property* (i.e., something that can be true of an object), for which we will often use the letter P. A RE $\mathcal{D}$ can be represented by a bag pf properties[6] (i.e., multi-set) of n properties: $\mathcal{D} = \{P_1, ..., P_n\}$. We define “$P_i = P_j$” as P_i and P_j express the same value of the same attribute.

Definitions that are well-formed cannot be incorrect. The question is whether a given set of definitions is *useful* because they allow us to make those distinctions that help us think about the phenomenon in question. So although a different set of definitions would have been possible – and we will indicate some points where we are aware that a definition could have been different – we believe that the following set is useful.

The primary goal of an REG algorithm is to enable the hearer to identify the intended referent r in a setting[7] C, which consists of r itself plus a non-empty set of other objects (which are often called distractors, e.g., McDonald (1983)). In other words, producing a *distinguishing description*, which could be formally defined as follows[8]:

Definition 1 (Distinguishing Description)**.** *The description $\mathcal{D} = \{P_1, ..., P_n\}$ is a distinguishing description of the intended referent r if it singles out r from all other elements of C. This is the case if and only if $[\![P_1]\!] \cap ... \cap [\![P_n]\!] = \{r\}$.* [9]

In any distinguishing description, we will call the use of a property (or attribute) *superfluous* if and only if the description would still be distinguishing if that property was removed from the description.

Henceforth, we will suppress the role of the referent r in our definitions. For example, a Distinguishing Description of r will simply be called a Distinguishing Description. If $\mathcal{D}$ singles out something that is not the intended referent, then we will say that it is not a Distinguishing Description but a “Wrong Description".

[6] For instance, $\{A, B, A\}$ is the same as $\{B, A, A\}$ but different from $\{A, B\}$

[7] This “setting" is sometimes called “context", but we avoid this term here to avoid confusion with other types of context.

[8] This definition is implicit in Dale and Reiter (1995).

[9] Note that $[\![P_i]\!]$ is a set of elements that share a property P_i, i.e., the denotation or extension of P_i.

Minimal Description

Dale (1989, 1992) suggested that the best RE for a given referent must always be the shortest possible one: an RE that uses as few properties as possible, also known as a *minimal description*. [10] This idea can be seen as interpreting the Maxim of Quantity as including a requirement that the RE should be as short as possible.

Definition 2 (Minimal Description)**.** *A set of property occurrences* $\mathcal{D} = \{P_1, ..., P_n\}$ *is a minimal description if and only if it is a Distinguishing Description and there is no Distinguishing Description* $\mathcal{D}' = \{P'_1, ..., P'_m\}$ *such that* $m < n$*, that is,* $|\mathcal{D}'| < |\mathcal{D}|$*.*

Here, $|\mathcal{D}|$ is the size of $\mathcal{D}$, that is, the number of property occurrences in $\mathcal{D}$. It is easy to see that, in one and the same situation, a referent may have more than one minimal description..

Over-specification

Previous studies in different areas of research (Engelhardt et al., 2006; Engelhardt et al., 2011; Koolen et al., 2011) motivate their understanding of over-specification on the basis of the second principle of the Gricean Maxim of Quantity: an RE is over-specified if it is more informative than is necessary for successful communication. This clearly covers situations in which an RE includes non-required properties while managing to identify the referent. However, as discussed in the §4.4.1, there are some interesting distinctions that this definition does not make because, as illustrated in example (49-c), a description without superfluous properties may nonetheless not be minimal.

Definition 3 (Over-specified Description)**.** *A set of property occurrences* $\mathcal{D} = \{P_1, ..., P_n\}$ *is an Over-specified Description if and only if it is a Distinguishing Description and it is not a Minimal Description.*

Bearing these issues in mind, we will now sub-categorise the class of Over-specified Descriptions.

Numerical Over-specification. Numerical over-specifications are cases like “the large green one" in (49-c) (in §4.4.1). In this description, no property is superfluous (unlike (49-b), where “green" could be removed) yet it is possible to construct a *shorter* RE by replacing a set of properties in the expression by a smaller set of properties where, crucially, the result is still a distinguishing description:

Definition 4 (Numerical Over-specification)**.** *The description* $\mathcal{D} = \{P_1, ..., P_n\}$ *is a numerical over-specification if and only if* $\mathcal{D}$ *is a Distinguishing Description and there is no* $P \in \mathcal{D}$ *such that* $\bigcap_{P_j \in \mathcal{D} - \{P\}} [\![P_j]\!] = \{r\}$*, but the number of attributes* n *is greater than that of a minimal description of* r*.*

Nominal Over-specification. The special status of the TYPE attribute comes from a long tradition of psycholinguistic work, summarised well in Levelt (1993, Chapter 4), based on the idea that (English) speakers tend to include a head noun in their REs. This idea was combined with the idea that head nouns classify things into broad classes (“types")

[10] An algorithm that achieves this is the *Full Brevity* algorithm (Dale & Reiter, 1995).

of objects , (e.g. Dale and Reiter (1995)), thus differentiating the TYPE attribute from other attributes. Types are also known as *categories* and a large body of theory has arisen about the role of types in language and thought (Rosch et al., 1976). We hereby define the expressions like (50), where there is a superfluous TYPE attribute while none of its other attributes is superfluous attributes, as *nominal over-specifications*. Formally, it can be defined as:

Definition 5 (Nominal Over-specification). *A Nominal Over-specification is a set of property occurrences $\{P_1, .., P_n\}$ in which at least one of $P_1, .., P_n$, say P_i, is a TYPE, and $\{P_1, .., P_n\} - \{P_i\}$ is a Distinguishing Description, but for every $j \neq i$, $\{P_1, .., P_n\} - \{P_j\}$ is not a Distinguishing Description*

Note that we don't need to require explicitly that $\{P_1, .., P_n\}$ is distinguishing because this follows from the requirement that $\{P_1, .., P_n\} - \{P_i\}$ is a Distinguishing Description.

Duplicate-Attribute Over-specification. Similar to the nominal over-specification, expressions like (51) also introduce a new type of over-specification due to its repeated use of the same attribute, which is named as *Duplicate-Attribute Over-specification*.

Definition 6 (Duplicate-Attribute Over-specification). *A description $\mathcal{D} = \{P_1, ..., P_n\}$ is a duplicate-attribute over-specification if and only if there exist two property occurrences $P_i, P_j \in \mathcal{D}$ such that $P_i = P_j$, and $\bigcap_{P_k \in \mathcal{D} - \{P_i\}} [\![P_k]\!] = \{r\}$.*

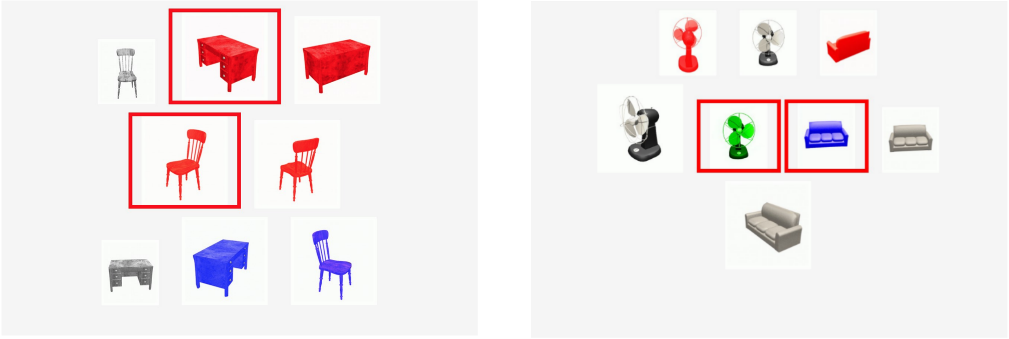
Recall that the clause "$P_i = P_j$" means that P_i and P_j express the same value of the same attribute. Note that this does not preclude considerable variations in surface form. For example, the *left facing chair* and the *chair whose back faces right* express the same property. Other examples in the TUNA corpora include sofa vs. settee, male vs. man, small vs. little, and so on.

Real Over-specification. Now let us turn to the over-specification which was covered by most of the previous studies, called *Real Over-specification*. We also need to note that real over-specification should not overlap with the nominal over-specifications. In other words, if an RE has superfluous properties other than a superfluous TYPE, it should not also be classified as a nominal over-specification anymore. Concretely, a more formal definition can be written as:

Definition 7 (Real Over-specification). *A description $\mathcal{D} = \{P_1, ..., P_n\}$ is defined as a real over-specification if at least one of the $P \in \mathcal{D}$ is $P \neq$ TYPE and such that $\bigcap_{P_j \in \mathcal{D} - \{P\}} [\![P_j]\!] = \{r\}$.*

It could be argued that there exists another special type of over-specification, where the value of an attribute is more specific than necessary. In TUNA, such over-specification occurs frequently in the TYPE attribute in the people sub-corpus. For example when the word *scientist* is used even though the word *person* would have been sufficient. This situation could be modelled by saying that *scientist* is a sub-type of *person*. This phenomenon might be called *Choice-of-Value Over-specification*.

However, we have chosen against this approach, because it would create a systematic ambiguity because an over-specified description could be turned into a minimal description in different ways: by removing a property (e.g. remove a property like "wears glasses"), or by replacing a property by a more general one (e.g. replacing scientist by person); the



(a)

(b)

Figure 4.3: Two scenes from the MTUNA corpus, each of which is a scene asking subjects to produce REs refer to a set of two target referents.

former would make it a real over-specification, but the latter would make it a choice-of-values over-specification.

To acknowledge the fact that, in the situation above, “scientist" is over-specified, we proceed as follows. A sub-type is interpreted as introducing new attributes to its parent type, i.e., dividing a single attribute into multiple attributes. For example, the word *scientist* expresses both TYPE and JOB.

Under-specification

Under-specification is the flip-side of over-specification. It is about expressions that do not successfully single out the target referent from its distractors. As discussed, it break the first principle of the Gricean Maxim of Quantity, i.e., the speaker did not make the contribution as informative as required. For Figure 4.2, the description (53-a) cannot help the reader to successfully identify the intended referent as there are two chairs in the scene.

(53) a. the chair
b. the large one (MD)
c. the green chair
d. the large chair

To be more precise, this kind of specifications is defined as follows:

Definition 8 (Under-specification). *If, for a description $\mathcal{D} = \{P_1, ..., P_n\}$, there exists a real super-set A of r ((i.e., $\{r\} \subsetneq A$)) such that $\bigcap_{P_j \in \mathcal{D}} [\![P_j]\!] = A$, then we call $\mathcal{D}$ an Under-specification.*

Analogous to real over-specification, if a description contains no superfluous property but one of its properties is an attribute whose value is not specific enough, this can also be seen as a special type of under-specification, namely, *Choice-of-Value Under-specification*. In example (54), if there are two chairs in a scene, where one is blue and the other is black, then compare to the minimal description (54-b), the word *dark* in (54-a) is not specific enough to single out the blue chair from the chairs.

(54) a. the dark coloured chair
b. the blue chair (MD)

When referring to multiple target referents, such an under-specification also exists. In mtuna, for the scene 4.3(a), we found the following REs (translated from Mandarin):

(55) a. the objects viewed from the side
b. the red objects viewed from the side
c. the left facing objects (MD)

If we suppose the phrase *viewed from side* means "facing left or right", then (55-a) is choice-of-value under-specified, compared to the minimal description (55-c). [11] To fix such an under-specification, there are two alternatives: 1) making the property more specific to *left facing* and constructing a minimal description; 2) adding a COLOUR property with the value of *red*, which results in a numerical over-specification (55-b). It is hard to decide which repair is better, but such confusion also causes another confusion of which type of under-specification this description should be: it can be either as a real under-specification[12] or as a choice-of-value under-specification. The same problem happens in scene 4.3(b), for which one could say:

(56) a. the green one and the blue one (MD)
b. the front-facing coloured objects (MD)
c. the coloured objects

Similar to (55-a), for (56-c), we can either re-write the word *coloured* to specific colours for the two objects respectively and distribute into two clauses or add the ORIENTATION of these two objects as in (56-b). Similar cases, where only making a property more specific can repair an under-specification, exists, but it does not appear in tuna. A simple example would be the case where all the objects in scene 4.3(a) are red. A simple solution is that we assume that the Choice-of-Value Under-specification does not exist.

Mixed Description. The literature has tended to focus on situations in which a description is either over-specified or under-specified, or minimally specified. Logically, however, there are other possibilities, and these are also encountered in real life.

One example of such cases is what we call a *Mixed Description* for acknowledging that it is an under-specification but has superfluous properties. A Mixed Description is an Under-specified Description from which, nonetheless, one or more property occurrences can be removed without changing the extension of the description. More precisely:

Definition 9 (Mixed Description). *A description $\mathcal{D}$ is a Mixed Description if and only if it is an Under-specified Description and there exists a property occurrence P_i in the description such that* $\bigcap_{P_k \in \mathcal{D}-\{P_i\}} [\![P_k]\!] = \bigcap_{P_k \in \mathcal{D}} [\![P_k]\!]$.

For example, given the scene depicted in Figure 4.4, the under-specification (57-a) describes either the SIZE or the COLOUR of the referent. However, all small objects in the scene are green, which suggest that the use of COLOUR does not add any information if we have already used SIZE. In other words, COLOUR is superfluous in this under-specification. In

[11] The description *facing left or right* is less specific than *facing right*.

[12] If choice-of-value under-specification exists, then we call other under-specifications as real under-specifications.

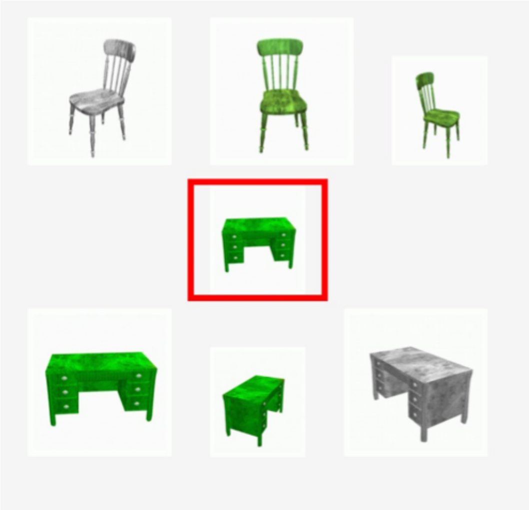


Figure 4.4: A scene from the MTUNA corpus.

contrast, either `ORIENTATION` or `TYPE` in the description (57-b) has certain contribution on singling out distractors. It is, therefore, not a mixed description, but a pure under-specification (which is introduced below).

(57) a. the green small desk
b. the front-facing desk

Pure Under-specification To acknowledge the existence of under-specifications that are not mixed descriptions, we introduce a new category.

Definition 10 (Pure Under-specification)**.** *A set of property occurrences $\mathcal{D} = \{P_1, ..., P_n\}$ is a Pure Under-specified Description if and only if it is an Under-specification and it is not a Mix Description.*

4.4.3 Description Basis

So far, we have introduced a variety of different kinds of over-specifications. In light of their definitions, minimal descriptions and numerical over-specifications have no superfluous property. Therefore, they can serve as description bases of other types of over-specification. Formally, given an over-specified description X has r as its intended referent and this description expresses property occurrences $\{P_1, .., P_n\}$, then this description is considered to be "built around" a minimal description (or a numerical over-specification) if there exists a proper subset X_b of X of $\{P_1, .., P_n\}$ such that X is a minimal description (or a numerical over-specification) of r. X_b is defined as a description basis of X. Theoretically, a description could have multiple description bases.

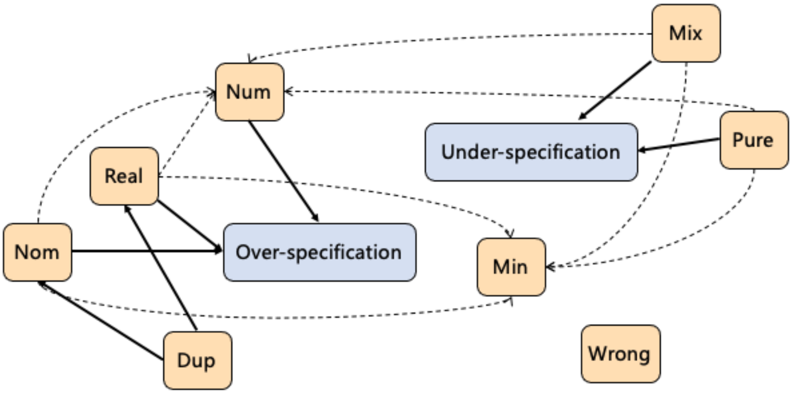


Figure 4.5: Diagram of relationships between each type of specification. In this diagram, "real" is the real over-specification, "num" is numerical over-specification, "nom" is the nominal over-specification, "dup" is the duplicate-attribute over-specification, "min" is the minimal description, "mix" is the mixed description, "pure" is pure under-specification and "wrong" is the wrong description. $A \rightarrow B$ (with solid line) means A is a kind of B, while $A \rightarrow B$ (with dash line) means B can be served as a description basis of A.

Wrong Description

In some cases (in the TUNA corpora, this amounts to approximately 3-4% of all cases) a description is simply *wrong*, i.e., describing the target referent incorrectly. In this study, we do not take a position on what to do with these descriptions[13], but we do offer labels that can flag the issue.

Definition 11 (Wrong Specification). *A Wrong Specification is a set of property occurrences* $\{P_1, .., P_n\}$ *in which at least one of* $P_1, .., P_n$*, say* P_i*, is not true of the intended referent* r*, that is,* $r \notin [[P_i]]$.

It follows that when $\{P_1, .., P_n\}$ is a Wrong Specification, then $r \notin [\![P_1]\!] \cap ... \cap [\![P_n]\!]$.

The logic of reference

Given the definitions above, a number of things follow immediately about the relationships between the various kinds of specifications. In what follows, we state some of the more important of these and visualise the main relationships in a graph. Since most of these consequences of our definitions are fairly immediate, most theorems will be stated without formal proof.

Theorem 1. *All of minimal descriptions, real over-specifications, numerical over-specifications, nominal over-specifications, and duplicate-attribute over-specifications are distinguishing descriptions.*

[13] For example, investigate what kind of role the Gricean Maxim of Quality should play when modelling the production of REs.

```
{
    "LABEL": "Real Over-specification",
    "SUPERFLUOUS": 1
}
```

Figure 4.6: The annotation for the RE (49-b): *the large green one*, in JSON format.

Theorem 2. *A distinguishing description cannot be a mixed description, a pure under-specification, or a wrong description.*

Theorem 3. *Each of the following classes is mutually exclusive: minimal descriptions, real over-specifications, numerical over-specifications, nominal over-specifications, mixed descriptions, pure under-specifications, and wrong descriptions.*

Theorem 4. *Each duplicate-attribution description is either a real over-specification or a nominal over-specification.*

Figure 4.5 describes the relationship between each type of specifications, in which each orange block represents a category we have introduced in this study. Consider theorem 4 for instance. Suppose we have a duplicate-attribute over-specification $\mathcal{D} = \{P_1, ..., P_n\}$ in which there are m ($0 < m < n$) duplicated properties (see Definition 6 for the definition of duplicated properties), represented as $\mathcal{D}_{dup}$ ($\mathcal{D}_{dup} \subseteq \mathcal{D}$). If all properties in $\mathcal{D}_{dup}$ are TYPEs, then $\mathcal{D}$ is a nominal over-specification (because TYPE is the only superfluous attribute in $\mathcal{D}$). Otherwise, it is a real over-specification (because $\mathcal{D}$ contains a superfluous non-TYPE property). The other theorems can be proven using similar reasoning.

4.4.4 Annotating the Use of Over- and Under-specifications

We annotated each expression using a set of key-value pairs stored as a JSON. For example, for the expression (49-b), we annotated it with the annotation shown in Figure 4.6, in which LABEL indicates which types of specification the current description should fall in, and SUPERFLUOUS records the number of superfluous properties. Note that when annotating a numerical over-specification, because none of the properties is superfluous, we set the value of SUPERFLUOUS to zero.

Since both nominal over-specifications, as well as real over-specifications, could have a superfluous TYPE, to take this superfluous TYPE into account and to differentiate it from normal over-specified properties, we employed a new variable, namely, SUPERFLUOUS-TYPE. If a superfluous TYPE is found, we accumulated the variable SUPERFLUOUS-TYPE by one.

When annotating a duplicate-attribute over-specification, since it could either be a real over-specification or a nominal over-specification, we only tracked the number of duplicated properties. To this end, we designed a new variable: DUPLICATE. For example, suppose we have three properties: $P_i, P_j, P_k \in \mathcal{D}$ (each of which could be either a TYPE or other types of property) and $P_i = P_j = P_k$, then we annotate: DUPLICATE : 2.

Here, we provide some relatively complex examples where a real over-specification contains superfluous TYPE property or duplicated properties. For example, if we have the following expressions:

```
{
    "LABEL": ["Real Over-specification",
        "Duplicate-attribute Over-specification"]
    "SUPERFLUOUS": 1,
    "SUPERFLUOUS-TYPE": 1,
    "DUPLICATE": 1,
    "BASIS": "Minimal Description"
}
```

Figure 4.7: The annotation for the RE: *the backward large table with no drawer*, in JSON format.

(58) a. the backward large table with no drawer
b. the large one (MD)

Comparing to the minimal description (58-b), the over-specification (58-a) is a real over-specification, in which there is a superfluous TYPE (*large*), and two superfluous ORIENTATION (*backward* and *with no drawer*). Interestingly, the duplicated properties themselves are superfluous properties. In such a case, we add SUPERFLUOUS with one for acknowledging the superfluous ORIENTATION. It is also a duplicate-attribute over-specification. We add DUPLICATE with one for its duplicated use of ORIENTATION. Therefore, the annotation of description (58-a) is shown in Figure 4.7.

Note that when deciding the number of superfluous properties, we counted the maximum number of properties (including TYPE) that can be removed but the resulting expression is still a distinguishing description. For example, for the scene in Figure 4.2, except the expressions in (49), we could also say:

(59) a. the front-facing green chair
b. the large green chair

For description (59-a), only the phrase *front facing* can be removed, which leads to 1 superfluous property. This implies that removing superfluous properties will sometimes result in a numerical over-specification. As for the description (59-b), either removing the *large* or removing both the *green* and *chair* yields distinguishing descriptions. However, based on the principle above, the latter removal is more favourable since it removes more properties than the former one.

Based on the idea of "description basis", in our annotation, we used a variable BASIS to track which type of specification (minimal description or numerical over-specification) the current description is built around.

As for the under-specifications, we used a variable named UNDERSPECIFIED to record the number of under-specified properties. For instance, for the expression (53-a), we have UNDERSPECIFIED : 1. In contrast with the over-specification, when deciding the amount of under-specified properties, we asked how many properties wound minimally have to be added to make the description distinguishing. To do so, suppose there are two possible fixes that propose to add the same number of properties. We chose the one that generates superfluous properties as little as possible and records that number. For example, we can make (53-a) distinguishing by either adding *large* (i.e., (53-d)) or *green* (i.e., (53-c)). Nevertheless, by adding *large* the fixed description (53-d), it is actually a nominal over-specification with a superfluous `TYPE`. In contrast, by adding *green*, the resulting description (53-c) is a numerical over-specification without any superfluous properties. We, therefore, choose the later fix and mark it as a "Pure Under-specification" that is based on the "Numerical Over-specification" Meanwhile, due to the existence of mixed descriptions, we also need to record the number of superfluous properties.

In addition, we argue that, for under-specifications, it is uninteresting to still be aware of whether a superfluous property is a `TYPE` or not and whether it is a duplicated property or not. Therefore, we use only the variable "`SUPERFLUOUS`" to track the number of superfluous properties in under-specifications.

4.4.5 Referring to Multiple Referents

So far, we introduce a new perspective that works well on singular descriptions (i.e., RE referring to a single target). In fact, this can also be extended and be applicable to plural descriptions (i.e., referring to multiple targets). We hereby list what kind of new risks would be introduced and how we handle them.

First, plural descriptions come in different shapes: they may either separate the set to which they refer into parts (e.g., *the red chair and the blue fan*), or not do this (e.g., *the two grey sofas*). We call the former descriptions *conjunctive*, and the latter *non-conjunctive*. As for non-conjunctive cases, the utterance can be simply tagged as the same as singular cases since it only makes sense to ask how good it is as a description of the set. In contrast, for conjunctive cases, each target can be annotated separately by considering the possibility of aggregation. For the scene 4.3(a), for the former clause of description (60-a) which is talking about the chair is an under-specification since there are more than one *red chairs* in the scene. Meanwhile, for the latter clause, it is a nominal over-specification. However, for a description like (60-b), it can be simplified by aggregating the same properties shared by two objects. We will view this as a type of over-specification due to the possibility of aggregation.

(60) a. the red chair and the left facing table
b. the left-facing chair and the left facing table

Second, if we treat lacking aggregation as a type of over-specification, then two issues need to be tackled: when do we say a non-conjunctive RE is lacking aggregation and how many superfluous attributes are there? To answer these questions, we differentiate two different types of aggregations: syntactic aggregation and semantic aggregation. For example, in the description (61-a), *red* is talking about a property of both *chair* and *table*, then the two *red* can be aggregated to a single property as in description (61-b). Furthermore, *chair* and *table* can be also aggregated to *furniture* to a description like (61-c). In the former aggregation, the two *red* can be aggregated because they are talking about exactly the same property shared by the two target objects, which is so-called *Syntactic Aggregation*. On the contrary, the latter aggregation happened due to the fact that they are both sub-categories of *furniture*. We, therefore, name this specific type of aggregation as *Semantic Aggregation*.

(61) a. the red chair and the red table
b. the red chair and table
c. the red furniture

We argue that the lack of syntactic aggregation is a type of over-specification in order to acknowledge that a description uses more properties than necessary. If a description can be repaired to a minimal description by only syntactically aggregating words[14], then we call such a description a *Lack-of-Aggregation Over-specification*.

As for semantic aggregation, we do not consider it as a type of over-specification by considering the following two reasons. On the one hand, rarely will people produce a so-called *Lack-of-Semantic-Aggregation Over-specification*, which is somehow a result of the fact that descriptions like (61-b) is syntactic ambiguous. On the other hand, sometimes, lacking semantic aggregation could be seen as a particular case of Choice-of-Value Over-specification, which, as discussed, is modelled by viewing a single word conveying more than one property. In order words, we could treat *chair and table* as a more specific version of *furniture*. A more straightforward example would be (62-a), in which is talking about GENDER and TYPE, and, comparing to (62-b) which only talks about TYPE, it has one more superfluous property.

(62) a. the man and the woman
b. the people

Third, we also observed a certain number of descriptions, each of which only refers to one of the multiple target objects. For instance, when using description (63) for scene 4.3(b), then the *the blue sofa* is missing. For this case, we will annotate the fan with what it should be and call *the sofa* as a *Missing Description*.

(63) the green fan

4.4.6 Analysing the Use of REs

We explore the first research question (RQ1) of this Chapter. We start with RQ1b, by analysing the use of over- and under-specifications in MTUNA, which is then extended to the ETUNA corpus. At length, we compare REs in MTUNA and ETUNA (i.e., RQ1a). Before starting the analysis, we introduce the dataset we use.

Dataset

The sources of our dataset are the MTUNA and ETUNA corpora. When evaluating the REG algorithms in Mandarin, we use the whole MTUNA corpora, where there are 10 trials in the furniture domain and 10 trials in the people domain. Sometimes, using TYPE might result in numerical over-specifications. For example, for the scene in Figure 4.2, the description *the green chair* is a numerical over-specification since none of the attributes it used can be removed. This makes the counts of the number of superfluous attributes or the number of each sub-category of over-specifications is under-estimated. Therefore, when analysing the use of over- and under-specifications (i.e., the first research question), we omit the trials which allocate discriminative power to TYPE (i.e., trails whose minimal descriptions or numerical over-specifications have TYPE). At last, we have 7 trials in the furniture domain and 10 trials in the people domain. We name this sub-corpus as MTUNA-NT.

We also apply the scheme to ETUNA in order to compare its REs with those in MTUNA. Unlike MTUNA, in ETUNA, subjects were broken into two groups based on whether

[14] Phrases in different clauses but talking about the same property are seen as the places where the syntactic aggregation should play a role.

				over				under	
	domain	**total**	**mini.**	**real**	**nom.**	**num.**	**(dup.)**	**mix.**	**pure**
MTUNA	furniture	361	46 (12.74%)	118 (32.69%)	132 (36.57%)	2 (0.55%)	7 (1.94%)	14 (3.88%)	49 (13.57%)
	people	361	17 (4.71%)	217 (60.11%)	69 (19.11%)	13 (3.60%))	2 (0.55%)	43 (11.91%)	2 (0.55%)
MTUNA-NT	furniture	252	9 (3.57%)	84 (33.33%)	104 (41.27%)	0 (0%)	3 (1.19%)	11 (4.37%)	44 (17.46%)
	people	361	17 (4.71%)	217 (60.11%)	69 (19.11%)	13 (3.60%))	2 (0.55%)	43 (11.91%)	2 (0.55%)
MTUNA-OL	furniture	252	9 (3.57%)	84 (33.33%)	104 (41.27%)	0 (0%)	3 (1.19%)	11 (4.37%)	44 (17.46%)
	people	218 (%)	15 (6.88%)	145 (66.51%)	37 (16.97%)	2 (0.92%)	2 (0.92%)	18 (8.26%)	1 (0.46%)
ETUNA	furniture	156	1 (0.64%)	59 (37.82%)	62 (39.74%)	0 (0%)	0 (0%)	6 (3.85%)	28 (17.95%)
	people	132	3 (2.27%)	75 (56.82%)	47 (35.61%)	0 (0%)	1 (0.76%)	7 (5.30%)	0 (0%)

Table 4.1: Frequencies with percentages of referring expressions that fall in each type of specifications in MTUNA, MTUNA-NT, MTUNA-OL and ETUNA respectively. Specifically, **total** is the total number of descriptions in each corpus. **mini.** is the minimal over-specification, **real** is the real over-specification, **nom.** is the nominal over-specification, **num.** is the numerical over-specification, **dup.** is the duplicate-attribute over-specification, **wrong** is the duplicate-attribute over-specification, **ordinal** stands for the ordinal description, and **under** is the under-specification.

subjects were encouraged to use so-called locative expressions. Half of the participants were discouraged, although not prevented, from using locative expressions, whereas the other half were not. To conduct a fair comparison, we only use the subjects who were discouraged to use locations. Furthermore, the scenes used in MTUNA and ETUNA also have minor differences. In this study, we only use trails with exactly the same scene from two corpora, which yields 7 trails in the furniture domain and 6 trials in the people domain. We call this set of shared scenes of MTUNA as MTUNA-OL.

The Use of Over-specifications

As discussed, for RQ1b, domain complexity has a positive influence on the use of over-specifications in that domain. Koolen et al. (2011) checked one aspect of this expectation by means of hypothesising that speakers tend to convey more information in a more complex domain (i.e., people domain) than in a simpler domain (i.e., furniture domain). To test this idea, they compared the number of superfluous attributes between two domains and found that, on average, REs in the people domain contain more superfluous attributes. However, this methodology has two shortcomings: on the one hand, a higher number of superfluous attributes does not necessarily result in the more frequent use of over-specified descriptions, for example, because of the existence of numerical over-specifications (which over-specified without containing any superfluous properties), and because of the existence of mixed specifications (which are under-specifications that contain some redundancy at the same time). Furthermore, Koolen and colleagues didn't differentiate TYPE from other attributes when counting the number of superfluous attributes.

We expect that there are a higher proportion of over-specifications in the people domain than in the furniture domain. To confirm this, we counted the number of over-specification in MTUNA-NT rather than the number of superfluous attributes. With the definition of "required/necessary attribute" in §4.4.2, we understand that all of the descriptions that are annotated as any of real over-specification, nominal over-specification, numerical over-specification, and duplicate-attribute over-specification are over-specifications, the number of each of which in MTUNA-NT is shown in Table 4.1, and we then sum up these numbers. This manner of counting resulted in the number of 188 and 299 over-specifications in the furniture sub-corpus and the people sub-corpus respectively.

How should *non*-over-specification be counted? Different approaches are possible. One way is to count all valid descriptions that are not annotated as over-specification, which yields 64 and 62 descriptions, respectively. Using a Chi-square test with Yates correction, we were able to confirm the above hypothesis with moderate significance, $\chi^2(1, N = 613) = 6.1441, p = .0132$. However, this way of counting may overlook mixed descriptions. Therefore, we also tested the above hypothesis (i.e., that there are more over-specifications in the people domain than in the furniture domain) omitting all descriptions that do not result in successful communication (i.e., under-specification) and, this time, we obtained a rejection of the hypothesis, $\chi^2(1, N = 513) = 0.166, p = .6837$. This result is not surprising once one realises that *nearly all* successful descriptions in both the furniture and the people corpus were over-specifications: recall that speakers tend to include a TYPE no matter whether it contributes to distinguishing the target and this biases the analysis (e.g., Levelt (1993)). Therefore, to obtain a more insightful analysis of how domain difficulty influences over-specification, we focused on those over-specifications that are not nominal over-specification, hypothesising that there are more of these over-specifications in the people domain than in the furniture domain. We, therefore, summed up all the real over-specifications, numerical over-specifications and duplicate-attribute over-specifications, resulting in the numbers 84 and 230, for the furniture and the people sub-corpus, respectively. We once again tested the hypothesis and found that, this time, it was confirmed with high significance, $\chi^2(1, N = 513) = 46.4435, p < .0001$.

While a post-hoc analysis of this kind – where different definitions of key phenomena are attempted – has to be treated with some caution, these results at least demonstrate how different insights can be gleaned depending on *what kind* of over-specification one wants to focus on.

Likewise, we analysed the ETUNA using the same strategy and obtained the same results. That is, 1) There are more over-specifications in the people domain than in the furniture domain, $\chi^2(1, N = 470) = 6.80, p < .01$; 2) Within successful communications, there are significant more real over-specifications in the people domain than in the furniture domain, $\chi^2(1, N = 396) = 39.72, p < .0001$.

The Use of Under-specifications

Another type of analysis concerns under-specification. We investigated whether domain difficulty influences the use of under-specification, that is, Are speakers less likely to single out the target object when the domain is more complex? To find out, we counted the descriptions that were annotated as under-specification in MTUNA-NT, and obtained 55 and 45 under-specifications for furniture and people corpus, respectively. Surprisingly, this does not only lead us to reject our hypothesis but, surprisingly, even suggests that the situation is the other way around, $\chi^2(1, N = 613) = 9.5237, p = .0020$. To understand why this happened is a topic for further research. Yet, the current results (15.74% of descriptions in MTUNA-NT are under-specifications) at least show that the role of under-specifications cannot be ignored when analysing REs. The same result is also found in ETUNA ($\chi^2(1, N = 470) = 15.14, p = .0001$).

Comparing ETUNA and MTUNA

Regarding RQ1a, we compare the results of MTUNA-OL and ETUNA. In RQ1a, we expect that there are more over-specifications in ETUNA than in MTUNA and meanwhile there are more under-specifications in MTUNA than in ETUNA. However, both of these two hypotheses are rejected, i.e., there is no difference in the use of over-specification, $\chi^2(1, N = 758) = 3.19, p = .743$, the use of real over-specifications out of successful communications, $\chi^2(1, N = 643) = 1.03, p = .3095$, and on the use of under-specification, $\chi^2(1, N = 758) = 0.31, p = .5742$.

Further Observations

By looking into the annotated dataset, we also have several post-hoc observations:

Duplicate-attribute over-specification. In the whole singular part of the MTUNA-NT corpus, we observed only 5 duplicate-attribute over-specifications. In all these cases, duplication happened when a speaker used one single word to express multiple attributes. Consider the following example in MTUNA-NT:

(64) 年老 的 长者
niánlǎo de zhǎngzhě
'the old old person'

where the word "长者" (zhǎngzhě; *old*) express both AGE and TYPE of the target. Given the small number of cases, this observation needs to be handled with caution, of course.

Numerical Over-specification. We were also curious about the number of specifications that use the numerical over-specification as their description bases. Out of 17 trials of MTUNA, 7 trials allows numerical over-specification, all of which are in the people domain (which explains the fact that no numerical over-specification is found in the furniture domain (see Table 4.1)). 83 out of 260 REs (approximately 31.92%) are built around either numerical over-specifications or numerical over-specified.

Under-specification. As we can see from the Table 4.1, 15.74% and 14.23% descriptions in MTUNA-OL and ETUNA are under-specifications, which is much higher than the previous reported proportions (e.g., Koolen et al. (2011) reported 5%). In other words, approximately 16% of the descriptions used for evaluating the human-likeness of the machine-generated description cannot result in successful communications. In light of earlier discussions of under-specification (e.g. Ferreira et al. (2005), Koolen et al. (2011), and Pechmann (1989)), this finding was surprising, which is why we discuss it further in §4.5.5.

4.4.7 Discussion

The role of context

As was mentioned briefly at the beginning of this chapter, the work described in this chapter has looked at REs in isolation from their context of use. This has allowed us to offer precise definitions of all the key notions involved, such as the notion of under-specification, and various notions of over-specification. When the context of use is taken into account, things can become more challenging. A simple example, where the relevant context is very local, is offered by Stone and Webber (1998), who discuss a situation involving two hats and two rabbits, with one of the two rabbits sitting in one of the two hats. In this situation, one can say "the rabbit in the hat". The expression "the hat" is arguably a minimally distinguishing description. For even though the situation involves more than one hat, the NP as a whole makes it clear which hat is being referred to, and this is what legitimises the use of the definite article.

When the wider context is taken into account, it is extremely common for an NP to be under-specified in isolation (i.e., when only the NP itself is taken into account) whereas, in fact, it is a distinguishing description. For example, when we say "My father bought a dog; the dog eats sausages", the NP "the dog" is under-specified in isolation, but fully specified (i.e., a distinguishing description) when the sentence as a whole is taken into account. These phenomena can also involve aspects of the wider context into account, including even the speaker's and hearer's background knowledge and opinions. These phenomena are widely discussed and described, in both the theoretical (e.g., Pogue et al. (2016)) and the computational literature (e.g. Krahmer and Theune (2002)).

Our point here is that, firstly, all the concepts discussed in the present study can, in principle, be generalised to take context into account. For example, consider our definition of Distinguishing Description: its informal part requires that such a description "singles out r from all other elements of C". This idea remains valid when the context of the use of the description is taken into account. However, the same is not true for the formal part of the definition, which requires that $[\![P_1]\!] \cap ... \cap [\![P_n]\!] = \{r\}$, since this narrow definition would judge "the dog" (in our example above) not to be a distinguishing description, which would go against the aim of capturing whether or not a description manages to identify its intended referent.

Second, although we believe that it would be interesting to generalise (the precise part of) our definitions in such a way that contextual information can help to "disambiguate" an RE (i.e., contextual information makes the RE distinguishing), to apply these new definitions in a new annotation scheme would be far less straightforward, because it would be up to the annotators to decide whether (for example) a given NP, in a given context of use, manages to "single out" the referent. Undoubtedly, different annotators would sometimes resolve such questions differently, in which case one might design protocols for reaching a consensus annotation as is often done when pragmatic information is entered by human annotators (Gatt et al., 2008).

A behavioural perspective on these issues was offered by Arts (2004, Chapter 4), who argued that the second rule of the Gricean Maxim of Quantity is violated if and only if the time that recipients need to identify the intended referent (i.e., identification time) is increased. She showed experimentally that an apparently over-specified RE (like "the round button *on the right*", when the situation contains only one round button) can actually speed up the identification process; she concluded that, therefore, this expression does not break the second rule of Gricean Maxim of Quantity. Arts' perspective suggests that annotators might be asked questions such as "*Tick this box if you believe that the author could have used an alternative RE, which you would have understood substantially faster. If so, then please suggest such an alternative RE.*" How workable such an annotation scheme would be, and to what extent annotators would agree with each other about their annotation decisions, is a matter for further research.

A computational perspective was offered by Dale and Reiter (1995), who proposed an algorithm, called the Incremental Algorithm, which often produces REs that our definitions would classify as Real or Nominal over-specifications. However, if this algorithm is accepted as an implementation of the Gricean Maxims – as the authors proposed – then one would have to call such REs Minimally Specified because they distinguish the intended referent without using any properties that the algorithm considers to be superfluous.

Third, our empirical work in §4.4.6 stays clear of these "contextual" complications, because the TUNA experiments were constructed in such a way that the descriptions in them (i.e., the NPs produced by the participants in the elicitation experiment) had to identify the referent by themselves, instead of relying on the wider context. This was achieved by offering participants simple, artificial situations that bear no resemblance to everyday situations, and by asking them to produce NPs in isolation (answering "Which object/objects appears/appear in a red window?").

Limitations of the Present Analysis

In this study, we have used the English ETUNA corpus and the Chinese MTUNA corpus to illustrate the benefits of the new way in which we propose to look at the ways in which an RE can single out (or fail to single out) its target referent. Although this illustration has been enlightening, we are aware that the above-mentioned corpora have some important limitations. For example, they are not optimal for investigating numerical over-specification. Most trials in ETUNA and MTUNA simply do not allow numerical over-specification. Another example is that since TUNA experiments used abstract scenes, almost all properties have crisp meanings. In TUNA, when saying "large chair", we (and the REG algorithms) know exactly which distractors the property "large" can single out. However, in realistic scenarios, "large" is a gradable property. Therefore, the meaning of "large" is vague (see van Deemter (2012), van Deemter (2016, Chapter 9)) and is context dependent. This poses challenges for applying our new perspective on referring expressions with gradable properties.

If we look into those trials that do allow numerical over-specification[15], we found that 83 out of 260 descriptions actually were built around numerical over-specifications. A similar issue also applies to other types of over-specifications, such as duplicate-attribute and lack-of-syntactic-aggregation over-specifications. For a similar reason, although the annotated corpora also record different types of superfluous attributes, such as the TYPE and duplicated attributes, we cannot make any further conclusions or conduct more fine-grained analysis on how much and what "additional information" is conveyed on biased results on the current two corpora.

In addition, we tried our best to do a fair comparison between Mandarin and English by, for example, using only trails that overlap in two corpora. Nevertheless, we cannot

[15] Four trials in the people domain of MTUNA allow numerical over-specification; these four are associated with 180 descriptions in total.

Attribute	Possible Values	Freq.
TYPE	*chair, sofa, desk, fan*	347
COLOUR	*blue, red, green, grey*	326
ORIENTATION	*front, back, left, right*	185
SIZE	*large, small*	141
X-DIMENSION	1, 2, 3, 4, 5	5
Y-DIMENSION	1, 2, 3	5
OTHER	-	10

Table 4.2: Attributes and their values for REs in furniture domain. Freq. is the frequency of that attribute in MTUNA.

Attribute	Possible Values	Freq.
TYPE	*person*	278
AGE	*young, old*	74
ORIENTATION	*front, left, right*	6
hasBeard	0, 1	169
beardColour	*dark, light*	95
hasHair	0, 1	116
hairColour	*dark, light*	100
hasGlasses	0, 1	165
hasShirt	0, 1	14
hasTie	0, 1	26
hasSuit	0, 1	60
X-DIMENSION	1, 2, 3, 4, 5	6
Y-DIMENSION	1, 2, 3	6
OTHER	-	151

Table 4.3: Attributes and their values for REs in people domain.

say any last word here since this study is not a language comparison work and many experimental settings of MTUNA and ETUNA are different. Therefore, in the future, a careful language comparison study can be conducted whose results could be analysed based on our scheme.

4.5 Study 2: Computational Modelling of Referring Expressions

4.5.1 Annotating the Semantics

In what follows, we introduce how we annotate the MTUNA and the ETUNA corpora. In order to conduct certain analyses as well as evaluate REG algorithms, we need to annotate the semantics (i.e., which properties are used) of each RE in these corpora.

1650 REs were semantically annotated (after omitting some unfinished REs from the

```
{ "sno": "Object",
  "subject_id": "2",
  "object": [
    { "attributes": [
        {"name": "COLOUR",
          "value": "dark"
        },
        {"name": "TYPE",
          "value": "table"
        }]
    }],
  "trial_id": "1",
  "utt": "灰桌子"
}
```

Figure 4.8: An example annotated data sample from mtuna, for the RE 灰桌子 (huīzhuōzi;*grey table*).

corpus) following the scheme of van der Sluis et al. (2006). [16] For simplicity, instead of XML, we use JSON for the annotation. Because the scenes stay the same when different subjects accomplished the experiment, we annotated the scene and the REs in mtuna separately. For the attribute `hairColour`, both van der Sluis et al., 2006 and Gatt et al. (2008) (and all the annotate scheme used by the previous tuna corpora) annotated both hair colour and beard colour as `hairColour`. However, this would cause us to overlook some key phenomena because some participants used the colour of a person's beard to distinguish the target. Therefore, we decided to use `hairColour` and `beardColour` as separate attributes. As pointed out in van Deemter, Gatt, Sluis, et al. (2012), since the attribute `hairColour` depends on `hasHair`, the authors merged these two into a single attribute `Hair` during the evaluation. We did the same thing and obtained two merged attributes: `Hair` and `Beard`.

To avoid compromising the comparison between mtuna and etuna, we did not only annotate mtuna but also re-annotated the etuna corpus, using the same annotators. The attributes and corresponding values we used in our annotation (annotating the mtuna and re-annotating the etuna) are shown in Table 4.2 and 4.3 for furniture and people domain, respectively. An example of the annotated RE sample in mtuna is shown in Figure 4.8.

4.5.2 Evaluating Classic REG Algorithms

This subsection is to answer RQ2 (i.e., how classic REG algorithms perform on mtuna and how this is different from the performance on etuna?). We examine a number of REG algorithms on mtuna and etuna. We then report and analyse the experiment results.

[16] This includes the trials that have one target referent and those that have two targets, but, in this paper, we focus on the former one.

Furniture			People		
Model	DICE (SD)	PRP	Model	DICE (SD)	PRP
IA-COS	**0.875 (0.17)**	**55.7**	IA-GBHOATSS	0.637 (0.26)	16.3
IA-CSO	0.847 (0.21)	55.1	IA-BGHOATSS	0.629 (0.25)	15.5
IA-OCS	0.797 (0.16)	20.5	IA-GHBOATSS	0.617 (0.25)	13.0
IA-SCO	0.754 (0.18)	15.0	IA-BHGOATSS	0.577 (0.24)	7.5
IA-OSC	0.740 (0.20)	18.3	IA-HGBOATSS	0.589 (0.23)	6.1
IA-SOC	0.690 (0.21)	14.7	IA-HBGOATSS	0.559 (0.24)	6.1
-	-	-	IA-SSTAOHBG	0.347 (0.23)	1.9
FB+TYPE	0.830 (0.18)	39.9	FB+TYPE	**0.669 (0.26)**	**23.2**
FB	0.574 (0.25)	3.0	FB	0.446 (0.32)	9.9
GR	0.802 (0.21)	39.3	GR	0.613 (0.29)	19.9

Table 4.4: Experiment results on MTUNA, in which the string after each IA algorithm represents the preference order it uses. For example, "COS" means COLOUR > ORIENTATION > SIZE and "BGHOATSS" stands for hasGlasses > BEARD > HAIR > ORIENTATION > AGE > hasTie > hasShirt > hasSuit.

Experimental Settings

Algorithms. We tested the classic REG algorithms, including 1) the Full Brevity algorithm (FB Dale, 1989): an algorithm that finds the shortest RE; 2) the Greedy algorithm (GR Dale, 1989): an algorithm that iteratively selects properties that rule out a maximum number of distractors (i.e., a property that has the highest "Discriminative Power"); and 3) the Incremental Algorithm: an algorithm that makes use of a fixed "preference order" of attributes (IA Dale & Reiter, 1995). See §2.2.1 for a more detailed review.

Evaluation Metrics. We used DICE (i.e., measuring the overlap between the generated REs and the REs in the corpus) and PRP (i.e., the proportion of times the algorithm achieves a DICE score of 1) for evaluating attribute choice in REG (see §2.2.1 for more details).

Performance of Algorithms on MTUNA

We report the evaluation results on MTUNA and MTUNA-OL in Table 4.4 and Table 4.5. For the FB algorithm, we tested both the version that does not always append a TYPE (named FB in the rest of this chapter) and the version that does always append a TYPE (named FB+TYPE). Moreover, since we did not observe any significant difference in the frequencies of use of each attribute between MTUNA and ETUNA corpora, we let the IA make use of the same set of preference orders as van Deemter, Gatt, Sluis, et al. (2012).

In line with the previous findings in other languages, in the furniture domain, it is IA (with a good preference order) that perform the best in both MTUNA and MTUNA-OL. Interestingly, the people domain yields very different results: this time, FB+TYPE becomes the winner.

An ANOVA test comparing GR, FB+TYPE, and the best IA suggests a significant effect of algorithms on both domains and on both MTUNA and MTUNA-OL (Furniture: $F(2, 1008) = 49.20, p = .002$; People: $F(2, 1065) = 11.97, p < .001$) and MTUNA-OL (Furniture: $F(2, 699) = 14, p < .001$; People: $F(2, 622) = 4.22, p = .015$). As for each algorithm, by

Furniture					People				
	ETUNA		MTUNA-OL			ETUNA		MTUNA-OL	
Model	DICE (SD)	PRP	DICE (SD)	PRP	Model	DICE (SD)	PRP	DICE (SD)	PRP
IA-COS	**0.919 (0.12)**	**62.8**	**0.915 (0.14)**	**65.5**	IA-GBHOATSS	**0.862 (0.17)**	50.0	0.724 (0.22)	22.8
IA-CSO	**0.919 (0.12)**	**62.8**	**0.915 (0.14)**	**65.5**	IA-BGHOATSS	0.861 (0.17)	**50.8**	0.719 (0.21)	21.0
IA-OCS	0.832 (0.14)	26.3	0.823 (0.15)	25.4	IA-GHBOATSS	0.774 (0.20)	27.3	0.674 (0.25)	19.6
IA-SCO	0.817 (0.14)	20.5	0.808 (0.15)	19.4	IA-BHGOATSS	0.761 (0.19)	25.0	0.621 (0.22)	7.8
IA-OSC	0.805 (0.16)	23.7	0.798 (0.17)	23.8	IA-HGBOATSS	0.705 (0.17)	3.8	0.609 (0.22)	4.1
IA-SOC	0.782 (0.16)	19.9	0.767 (0.17)	19.4	IA-HBGOATSS	0.670 (0.19)	4.5	0.570 (0.23)	3.7
-	-	-	-	-	IA-SSTAOHBG	0.339 (0.10)	0.0	0.285 (0.17)	0.0
FB+TYPE	0.849 (0.17)	41.7	0.849 (0.16)	42.5	FB+TYPE	0.847 (0.17)	44.7	**0.734 (0.23)**	**27.4**
FB	0.590 (0.23)	0.6	0.602 (0.24)	3.6	FB	0.556 (0.16)	2.3	0.541 (0.26)	11.0
GR	0.849 (0.17)	41.7	0.849 (0.16)	42.5	GR	0.727 (0.25)	33.3	0.650 (0.28)	21.9

Table 4.5: Experiment results on the MTUNA-OL and ETUNA. Algorithms are listed from top to bottom in order of their performance on ETUNA.

Tukey's Honestly Significant Differences (HSD), we found that IA defeats other algorithms in the furniture domain in both corpora ($p < .001$) and that the victory of FB+TYPE for people domain is significant in MTUNA ($p = .001$) but not in MTUNA-OL ($p = 0.96$).

The scores for algorithms in the people domain are much lower than those in the furniture domain, even lower than the scores for the people domain in ETUNA. This may be because, based on the numbers in Table 4.1, a Chi-Squared Test suggests that, in MTUNA, there are more real over-specifications ($\chi^2(1,722) = 55.95, p < .001$) but fewer nominal over-specifications ($\chi^2(1,722) = 26.57, p < .001$) in the people domain than in the furniture domain. [17] As for the former, real over-specifications are notoriously hard to model accurately by deterministic REG algorithms, which is one of the motivations behind probabilistic modelling (van Gompel et al., 2019) or Bayesian Modelling (Degen et al., 2020); such an approach might have additional benefits for the modelling of reference in Mandarin. The relative lack of nominal over-specifications in Mandarin descriptions of people could be addressed along similar lines, adding TYPE probabilistically. Another evidence is that, in the MTUNA people domain (abbreviated as MTUNA/People), FB outperforms many IAs on PRP, which does not happen in the MTUNA/Furniture.

By comparing the results for MTUNA and MTUNA-OL, we found that the rank order (by performance) of algorithms stays the same, but the absolute scores for the latter corpus are much higher. If we look into the annotations for the trials from MTUNA that are not in MTUNA-OL (see Table 4.2 and Table 4.3), most of these trials have multiple possible minimal descriptions and numerical over-specifications. Every RE in the corpus that results in a successful communication can be seen as either a minimal description or a numerical over-specification, with 0 or more attributes adding to it. When computing the DICE similarity score between a generated RE and human-produced REs, if it is close to a minimal description, it will differ from another minimal description. For example, suppose we have a trial having two minimal descriptions: *the large one* and *the green one*. Our FB produces the second minimal description (as it can only produce one RE at a time). When we compute DICE, we obtain $\frac{2}{3}$ for the RE *the green chair* while 0 for the RE *the large chair*, but, in fact, either of them has only one superfluous attribute. This implies that when a corpus contains multiple minimal REs, this will artificially lower the DICE

[17] This highlights the importance of sub-categorising the different kinds of over-specifications.

scores. [18] For the same reason, the performance of FB increases a lot from MTUNA/People to MTUNA-OL/People because all trials in MTUNA-OL have only one possible minimal description. Another reason lies in the decrease in the number of under-specifications from MTUNA/People to MTUNA-OL/People.

Cross-linguistic Comparison

Table 4.5 reports the results for both MTUNA-OL and ETUNA, from which, except for the fact that FB+TYPE becomes having the best performance, we see no difference in the order of their performance. An interesting observation is that, after correcting a few errors in the annotation of ETUNA (see §4.5.1), the difference between IA and FB+TYPE is no longer significant in the people domain in terms of Tukey's HSD (compare the conclusion in van Deemter, Gatt, Sluis, et al. (2012)). In other words, in both languages, there is no significant difference between the performance of these two algorithms on the people domain. We also checked the influence of language on the performance of FB and FB+TYPE: the influence of the former is significant ($F(1, 349) = 23.63, p < .001$) while that of later is not ($F(1, 349) = 0.36, p = .548$). This suggests that, in fact, it is English speakers who show more brevity, except in terms of use of TYPE. This might also explain the differences in absolute scores for all algorithms in both ETUNA and MTUNA-OL, especially in the people domain. Another possible reason for these differences is the fact that the REs in MTUNA-OL show slightly higher diversity in the choice of content than ETUNA, as the standard deviation for every model is higher.

4.5.3 The Influence of TYPE

Regarding RQ3, we, on the one hand, expect that there are more superfluous TYPEs in English than in Mandarin. To test this, we look at the number of REs that uses TYPE in MTUNA-OL and ETUNA. 98.4% and 95.93% of REs in the furniture and people domains of ETUNA contain TYPE. For MTUNA-OL, those numbers are 91.29% and 74.77%, suggesting that Mandarin speakers are less likely to use superfluous TYPE. On the other hand, as aforementioned, one major difference on the TYPE attribute from other attributes is that TYPE in the people domain has only one possible value (i.e., man) while in the furniture domain there are multiple alternatives (e.g., table, chair or sofa). In other words, the "attribute complexity" of TYPE in the furniture domain is higher than that in the people domain. Additionally, Lv (1979) suggested that the head of a noun phrase in Mandarin Chinese is omissible if the omitted head noun is the only possibility given the context (i.e., in referring expressions, if omitting the head noun results in a distinguishing description, then the head noun is omissible). The Chinese speakers are less likely to "always" produce a superfluous head noun in referring expressions. Therefore, we expect more superfluous TYPE in MTUNA-OL/Furniture than in MTUNA-OL/People. In MTUNA-OL, we observed a smaller proportion of uses of TYPE in the people domain ($\chi^2(1, 485) = 24.16, p < .001$). Moreover, comparing the performance of REG algorithms on the furniture domain of MTUNA and MTUNA-OL, the difference is not as huge as that in the people domain. This implies that the complement of Lv's hypothesis might also hold, namely, if the value of TYPE is *not* the only possibility, then it will not be omitted. We also note that this result is not saying Lv's theory is the only reason why there are more "TYPEless" referring expressions in

[18] An analogous problem has been identified in the task of evaluating image capturing (Yi et al., 2020), where the collision of multiple references for a single image was considered.

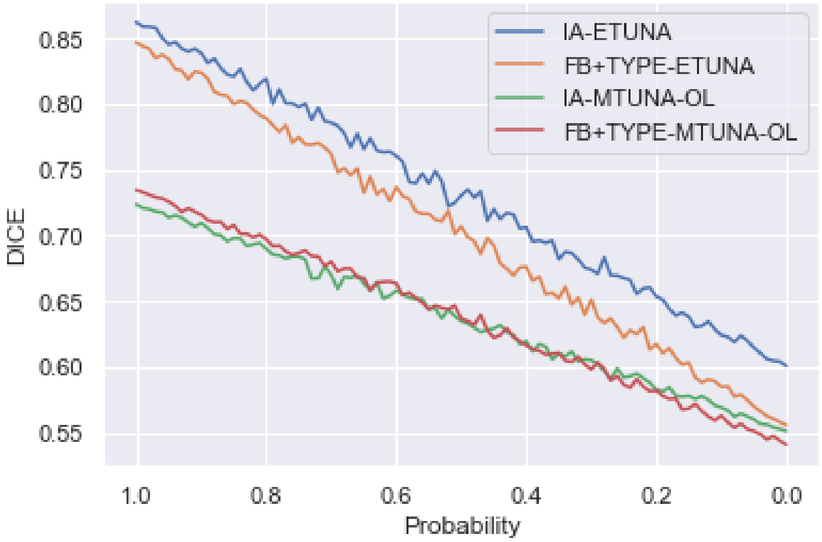


Figure 4.9: Change of the performance with respect to different probabilities of inserting superfluous TYPE for either the FB+TYPE and IA on either the people domain of MTUNA-OL and ETUNA.

the people domain. There are other factors that might have impacts here, such as animacy (i.e., the TYPE is more likely to be dropped for animates than inanimates; Fukumura and van Gompel (2011)).

To further assess the role of TYPE in algorithmically modelling REs, we investigated how introducing uncertainties in whether or not to include a TYPE affects the performance of REG algorithms for the people domain. We tried out different probabilities, and for each probability for inserting the TYPE we ran the algorithm 100 times; we report the average DICE score, drawing the lines indicating the change of performance over different probabilities in Figure 4.9.

We found that: 1) the decrease in performance on MTUNA-OL is smaller than that on ETUNA; 2） IA and FB+TYPE have similar performance for Mandarin while IA performs better for English; 3) The difference between the performance of these algorithms becomes smaller when the influence of TYPE is ignored (i.e., when the probability of inserting TYPE is close to zero), especially for the Full Brevity algorithm. On top of these findings, we observe that although Mandarin speakers are less likely to use superfluous TYPE, always adding TYPE achieves the best performances for all the algorithms. Such a result may be caused by the dependencies between the use of different properties. In other words, introducing uncertainty to only the TYPE cannot sufficiently model the uncertainties in REG: when to drop a TYPE might also depend on the use of other properties.

4.5.4 The effect of Syntactic Position

For RQ4 (i.e., how syntactic position influences the use of over-/under-specification and the performance of REG algorithms), we counted the number of real over-specifications and under-specification in subject and object position. In the MTUNA-OL corpus, there are

Model	Furniture	People
IA (subj.)	0.940 (0.11)†	0.728 (0.23)
IA (obj.)	0.890 (0.16)	0.719 (0.21)
GR (subj.)	0.884 (0.13)†	0.629 (0.30)
GR (obj.)	0.815 (0.18)	0.669 (0.25)
FB+TYPE (subj.)	0.884 (0.13)†	0.736 (0.23)
FB+TYPE (obj.)	0.815 (0.18)	0.733 (0.22)

Table 4.6: The performance of REG algorithms for REs in different syntactic positions, in which IA is the IA with highest performance in the previous experiments, i.e., the IA-COS and IA-GBHOATSS. † indicates that there is significant influence of the syntactic position on that algorithm in that domain.

247 and 239 descriptions in the subject and object positions, respectively. No significant difference on the use of over-specifications was found ($\chi^2(1, 485) = 1.57, p = 0.209$) but a significant difference regarding the use of under-specifications did exist ($\chi^2(1, 485) = 19.27, p < .001$). Considering the fact that there are more indefinite RE in subject position (van Deemter et al., 2017), the present finding might suggest that those indefinite REs are not suitable for identifying a target referent. It appears that further research is required to understand these issues in more details.

As for the computational modelling, Table 4.6 report the performance of each REG algorithms on REs at each position. Generally speaking, all algorithms performed better for REs in subject position than for REs in object position, with one exception, namely the GR algorithm for the people domain; the difference is significant in the Furniture domain, but not in the people domain, possibly because the furniture domain contains more under-specifications.

4.5.5 Discussion

Lessons about RE use

Regarding the “coolness" hypothesis, which focuses on the trade-off between brevity and clarity, we found that the brevity of Mandarin is only reflected in the use of TYPE but not in the other attributes, and, interestingly, no evidence was found that this leads to a loss of clarity; our findings are consistent with the possibility that Mandarin speakers may have found a better optimum than English speakers. They use shorter REs (by omitting superfluous TYPE) and, meanwhile, does not breach clarity.

Although Mandarin speakers are less likely to over-specify TYPE, following Lv (1979), we conclude that TYPE is often omitted *if and only if* it has only one possible value given the domain. This appears to happen “unpredictably” (i.e., in one and the same situation, TYPE is often expressed but often omitted as well). However, we saw that introducing probability for the use of TYPE alone does not work well. This suggests that, to do justice to the data, a REG model may have to embrace non-determinism more wholeheartedly, as in the probabilistic approaches of van Gompel et al. (2019) and Degen et al. (2020).

We found a significant influence of the syntactic position of the RE on the use of under-specifications and on the performance of REG algorithms. This flies in the face of earlier research on REG – which has tended to ignore syntactic position – yet it is in line with the theory of Chao (1965). It gives rise to various questions: *why* are more under-specifications used in subject positions, and why do all REG models perform better for REs in subject positions than for those in object positions? These questions invite further studies including, for example, reader experiments to find out how REs in different positions are comprehended. It would also be interesting to investigate what role syntactic position plays in other languages, where this issue has not yet been investigated.

From the first study (§4.4), we knew that there is a very substantial proportion, of nearly 20%, under-specified REs in both MTUNA and ETUNA. This was surprising, because, at least in Western languages, in situations where Common Ground is unproblematic (Horton & Keysar, 1996), under-specification is widely regarded as a rarity in the language use of adults, to such an extent that existing REG algorithms are typically designed to prevent under-specification completely (see e.g., Krahmer and van Deemter (2012)). Proportions of under-specifications in corpora are often left reported, but (Koolen et al., 2011) reported that only 5% of REs in DTUNA were under-specifications. [19]

These findings give rise to the following questions: 1) Why did previous investigators either find far fewer under-specified REs (e.g., Koolen et al. (2011), see Footnote 8) or ignore under-specification? 2) How does the presence of under-specification influence the performance of the classic REG algorithms (which never produce any under-specified REs, except when no distinguishing RE exists)? and 3) If a REG model aims – as most do – to produce human-like output, then what is the most effective way for them to model under-specification?

Lessons about REG Evaluation

Most REG evaluations so far have made use of the DICE score (Dice, 1945), which measures the overlap between two attribute sets. However, on top of the discussions of van Deemter and Gatt (2007) and of this study, we identify the following three issues for evaluating REG with DICE. First, if a scene has multiple possible minimal descriptions or numerical over-specifications, then this causes DICE scores to be artificially lowered and hence distorted. Second, there is no guarantee that an RE with a high DICE score is a distinguishing description. Third, DICE punishes under-specification more heavily than over-specification. Suppose we have a reference RE d which uses n attributes, a over-specification d_o with one more superfluous comparing to d (so it uses $n+1$ attributes), and a under-specification d_u which can be repaired to d by adding one attribute (using $n-1$ attributes), the DICE score of d_o is $2n/(2n+1)$ while d_u's DICE is $2n-2/(2n-1)$. In other words, d_o has a higher DICE than d_u. Whether this should be considered a shortcoming of DICE or a feature is a matter of debate.

Finally, our analysis suggests that previous TUNA experiments may have been insufficiently controlled. For example, some trials in MTUNA and DTUNA use TYPE for distinguishing the target, causing nominal over-specifications not to be counted as over-specification. Different trials have different numbers of minimal descriptions and different numbers of numerical over-specifications. As shown in §4.4.6, these issues impact evaluation results and this might cause the conclusions from evaluating algorithms with TUNA not to be re-producible.

[19] The difference might be that DTUNA used participants who came into the lab separately, whereas MTUNA participants sat together in a classroom.

Comparisons between corpora need to be approached with caution, and the present situation is no exception. For all the similarities between them, we have seen that there are significant differences in the ways in which the TUNA corpora were set up. [20] Although these differences exist for a reason (i.e., for testing linguistic hypotheses), we believe that it would be worthwhile to design new multilingual datasets, where care is taken to ensure that utterances in the different languages are elicited under circumstances that are truly as similar as they can be.

4.6 Summary

This chapter focused on the use of one-shot REs in Mandarin Chinese and their computational models. To this end, we, on the one hand, analysed the use of REs in Mandarin and compared it with that of English. On the other hand, we built and evaluated REG models on both MTUNA and ETUNA.

In order to better analyse the use of REs in Mandarin. We proposed a new perspective on the different ways in which a description can manage to pick out a referent and the different ways in which it can fail to manage this. This account is more *precise* than its predecessors because it offers precise definitions of key notions such as over-specification; it is also more *fine-grained* because it distinguishes between different kinds of over-specification; finally, it is more *extensive* because it has a place for some varieties of specification (e.g. wrong specification and duplicate attribute specification) that have often been overlooked. Building on this new perspective on description and reference, we introduced a matching annotation scheme. We applied it to the MTUNA corpus (for Mandarin) and the ETUNA corpus (for English).

By analysing the annotated MTUNA and ETUNA, within each language, we found that there are more over-specifications in the more complex domains, while, in contrast, there are fewer under-specifications in the more complex domains. Additionally, we also found that in both MTUNA and ETUNA, there are non-negligible amounts of under-specifications ($> 15\%$) which is not in line with previous researches. When using these annotated corpora, we surprisingly found that there is no significant difference between the use of over-/under-specifications in ETUNA and MTUNA, which is inconsistent with the idea that Mandarin speakers prefer brevity to clarity.

In order to build and evaluate REG algorithms, we annotated the semantics of both ETUNA and MTUNA, and run a number of classic REG algorithms on them. In nutshell, we found two major differences in the computational modelling of REs in English and Mandarin. First, the advantage of using IA no longer exists in the people domain in Mandarin, which is different from that in English. Second, there are different uses of TYPE in Mandarin and in English. That is there are a great number of "TYPEless" ($\approx 25\%$) in MTUNA/People, which is hard to be handled by classic deterministic REG algorithms.

In the next chapter, we move the focus to another category of REG task: REG in Context, where context plays a more vital role than studies in this chapter.

[20] Most TUNA experiments involved type-written REGs, but DTUNA elicited spoken REs. In most TUNA experiments the linguistic context was uniform, but MTUNA elicited REs in different syntactic positions, as we have seen.

CHAPTER 5

Referring Expression Generation in Context

Abstract. *Mandarin speakers use zero pronouns to make their language pragmatically natural. In this chapter, we study the use of zero pronouns in the task of Referring Expression Generation (REG) in context. Specifically, given the context, we included zero pronouns as an option when determining the referential form and built computational models accordingly. To this end, we considered two types of models in the present thesis: models that are based on the rational speech act theory and that are based on deep learning techniques. We conduct three studies. In the first study, we model the use of anapharic zero pronouns in Mandarin with the rational speech act model. We then focus on merging the use of zero pronouns into the "REG in context" task and building neural network based models to tackle a sub-task of REG in context: Referential Form Selection. Since the task of Referential Form Selection (RFS) has not been previously explored using neural models, in the second study, we show how to build neural models to tackle RFS tasked on a well-constructed English REG in context corpus and conduct an interpretability study to understand what had these neural models learnt. In the last study, we extend the second study to RFS in Mandarin, where we construct a Chinese RFS corpus that includes zero pronouns and adopts models used in the second study accordingly.*

—

The publications related to this chapter are:

1. Chen, G., van Deemter, K., & Lin, C. (2018). Modelling pro-drop with the rational speech acts model. *Proceedings of the 11th International Conference on Natural Language Generation*, 159–164. https://doi.org/10.18653/v1/W18-6519

2. Chen, G., Same, F., & van Deemter, K. (2021). What can neural referential form selectors learn? *Proceedings of the 14th International Conference on Natural Language Generation*, 154–166. https://aclanthology.org/2021.inlg-1.15

5.1 Introduction

In light of the coolness hypothesis (see §3.1 and C.-T. J. Huang (1984)), Mandarin is a "cool" language and makes liberal use of zero pronouns (ZP). The analysis of L. Wang et al. (2018) on a large Mandarin-English parallel dialogue corpus shows that 26% of the English pronouns are dropped in Mandarin. Example (38) exemplify zero pronouns in Mandarin, we hereby repeat the example. Considering the question "你今天看见比尔了吗?" (*Did you see Bill today?*). A Chinese speaker can respond in a variety of shorter expressions which are equivalent to "我看见他了" (*Yes, I saw him*), for example, "∅看见他了" (*Yes, ∅ saw him*), "我看见∅了" (*Yes, I saw ∅*), or even "∅看见∅了" (*Yes, ∅ saw ∅*). Here the ∅ symbol indicates the place from where a pronoun appears to have been "dropped" from a full sentence.

Generating zero pronouns (only) where they are appropriate is a difficult challenge for Referring Expression Generation in Context (recall that given a text whose referring expressions (REs) have not yet been generated and given the intended referent for each of these REs, the Referring Expression Generation in Context task is to build an algorithm that generates all these REs; abbreviated as REG in this Chapter), and more specifically for the task of choosing *referential form*, a key step in the classic Natural Language Generation (NLG) architecture (Reiter & Dale, 2000). Traditionally, choosing referential form is framed as modelling speakers' behaviour of deciding whether entities are referred to using a pronoun, a proper name, or a description. However, for "cool" languages, an extra option, namely of choosing a zero pronoun, needs to be added (Yeh & Mellish, 1997) for fully simulating speakers' behaviour.

To this end, we conducted three studies. In the first study, we focused only on the use of ZPs in Mandarin, i.e., choosing between ZPs or overt REs. Concretely, we model the use of ZPs with the Rational Speech Act (RSA) model (Frank & Goodman, 2012) by assuming that Mandarin speakers tend to choose a ZP if it is salient enough for successful communication. The idea of discourse salience has closely related to the concept of prominence status introduced in §2.2.2. Usually, a referent is more salient in the discourse if it is prominent. In this study, we tested our model on the OntoNotes dataset and were concerned with merely the Anaphoric ZPs (AZPs).

Subsequently, we integrated the use ZPs with the use of other referential forms (i.e., pronoun, proper name, and description). In other words, we turned to the tasks of Referential Form Selection (RFS) in Mandarin. In earlier works, computational linguists linked REG to linguistic theories and built English RFS systems on the basis of linguistic features. For example, Henschel et al. (2000) investigated the impact of 3 linguistic features namely recency, subjecthood, and discourse status on pronominalisation, i.e. deciding whether the RE should be realised as a pronoun. Using these features, they used the notion of *local focus* as a criterion for detecting the set of referents that can be pronominalised. The same holds for feature-based models (see Belz et al. (2010) for an overview) where models are trained on linguistically encoded data. More recently, people have started to look at deep learning-based models. For instance, Cao and Cheung (2019), Castro Ferreira, Moussallem, Kádár, et al. (2018), and Cunha et al. (2020) proposed to generate REs in an End2End manner (determine the referential form and the content simultaneously) without any feature engineering (see §2.2.2 for more details). They all used a benchmark dataset called webNLG. The evaluation results suggested that these neural methods perform well in producing fluent REs. However, directly generating REs make us lose the track of how well these neural models simulate human behaviours as most theoretical bases about reference are merely related to the choice of referential forms.

Since there has been no previous work on neural based RFS, and there is only an RFS dataset for English, in the second study, we decided to first focus on neural RFS in English. Specifically, we built neural based RFS models and evaluated them on a well-constructed English REG in Context (in the rest of this chapter, we use REG to represent the task of "REG in Context") corpus (i.e., the webNLG corpus). On the basis of theories of discourse salience/prominence status, through using probing classifiers, we conducted interpretability research in order to understand what information that impact prominence status can a neural RFS learn. In the last study, we extended what we have done to model RFS in Mandarin. We started with building a Mandarin RFS dataset using the OntoNote corpus used in the first study. We then tested and probed neural RFS models on the constructed dataset.

5.2 Study 1: Modelling Pro-drop with the Rational Speech Act Model

In this study, we model the use of zero pronouns in Chinese with the RSA model (Frank & Goodman, 2012) by assuming that speakers tend to choose a ZP if it is salient enough for successful communication (see §5.2.1). For computing discourse salience, we focus on ZPs that are *recoverable*, meaning that they either refer anaphorically to an entity mentioned earlier in the text (i.e., anaphoric ZPs, or AZPs for short), or to the speaker or hearer (i.e., deictic non-anaphoric ZPs or DNZPs for short) (Zhao & Ng, 2007); a ZP is *unrecoverable* if it cannot be linked to any referent, for example:

(65) ∅ 有 二十三 项 高新技术 项目 进区 开发
There are 23 high-tech projects under development in the zone

in which the ∅ cannot be recovered.

5.2.1 Background

Pro-drop raises challenges for a number of NLP tasks including, machine translation (MT), co-reference resolution, and REG. When translating from a pro-drop language, recovering the dropped pronouns of the source language can improve the overall performance of MT L. Wang et al., 2018; L. Wang et al., 2016. Co-reference resolution of ZPs has been widely explored with a variety of techniques including the centring theory (Rao et al., 2015), statistical machine learning (C. Chen & Ng, 2014, 2015; Zhao & Ng, 2007), deep learning (C. Chen & Ng, 2016; Yin, Zhang, et al., 2017; Yin, Zhang, Zhang, et al., 2017) and reinforcement learning (Yin et al., 2018). REG of ZPs for "cool" languages has been addressed through rule-based methods (Yeh & Mellish, 1997) including centring theory (for Japanese; Yamura-Takei et al. (2001)), but we are not aware of any testable computational account. [1] We offer such an account, along probabilistic lines.

Some discourse theories suggest that speakers choose referring expressions (REs) by considering discourse salience (Givón, 1983), i.e., speakers tend to choose pronouns if they believe the referent is highly salient. The intuition behind is that a highly salient referent tends to be highly prominent in the mind of the speaker and/or hearer. Orita et al. (2015) shared a similar view and argued that highly salient REs are highly *predictable*, so they are referred with pronouns (as opposed to full NPs) more often than the less salient ones.

[1] E.g., Yeh and Mellish (1997) did not offer a precise definition of some of the syntactic constraints and the notion of salience that they were using.

A theory that is sometimes used for explaining the relation between discourse salience and human choice of referential forms is Uniform Information Density (UID) (Jaeger & Levy, 2007). UID asserts that speaker tends to optimise information density (quantity of information) of the utterances to achieve optimal communication. In other words, speakers tend to drop a RE when the referent of the RE is predictable (or recoverable), and vise versa.

Apart from salience, production cost (Rohde et al., 2012) and the listener models (Bard et al., 2004), meaning the models that how speakers model listeners' interpretation of the utterance, also have impact on language production. It suggests to us that the salience of the referent may not be enough for modelling speakers' choice. The RSA model (see §5.2.3) used in this study is possible to take all these factors into consideration.

5.2.2 The Rational Speech Acts Model

We briefly introduced the RSA model in the §2.2.2. We hereby explain it again. In realm of NLP, the RSA model (Frank & Goodman, 2012) has been used for a variety of tasks including modelling speakers' referential choice between pronouns and proper names (Orita et al., 2015), the selection of attributes for referring expressions (Monroe & Potts, 2015), and the generation of colour references (Monroe et al., 2017; Monroe et al., 2018). The key idea of RSA is to model human communication by assuming that a rational (pragmatic) listener L_1 uses Bayesian inference to recover a speaker's intended referent r for word w under context C. In this way, RSA claims to offer not only accurate models, but highly explanatory ones as well. Formally, L_1 is defined as

$$L_1(r_s|w, C) = \frac{S_0(w|r, \mathcal{C})P(r, \mathcal{C})}{\sum_{r' \in C} P_S(w|r', \mathcal{C})P(r', \mathcal{C})}, \tag{5.1}$$

where r' denotes a referent in context C, $P(r, \mathcal{C})$ represents the discourse salience of r in the context $\mathcal{C}$, and S_0 is the literal speaker model defined by an exponential utility function:

$$S_0(w|r, C) = e^{\lambda(I(w;r,C) - C(w))}, \tag{5.2}$$

where $I(w; r, C)$ is the informativeness of word w, $C(w)$ represents the speech cost. Note that, here, we replace the $P(w; r, C)$ in Equation 2.10 with a more accurate measure: informativeness $I(w; r, C)$.

Orita et al. (2015) extended the RSA by assuming that speakers estimate listener's interpretation of the (form of) RE w based on discourse information. The speaker chooses w by maximising the listener's belief in the speaker's intended referent r in relation to the speaker's speech cost $C(w)$, where the cost is estimated according to the complexity of the utterance, such as the length of w:

$$\begin{aligned} S_1(w|r) &\propto L_1(r|w) \cdot \frac{1}{C(w)} \\ &= \frac{S_0(w|r, C)P(r)}{\sum_{r'} S_0(w|r', C)P(r')} \cdot \frac{1}{C(w)} \end{aligned} \tag{5.3}$$

Here $L_1(r|w)$ estimates the informativeness of w, and $S_0(w|r, C)$ estimates the likelihood (according to the speaker) that the listener guesses that the speaker used w to refer to r.

5.2.3 Modelling Pro-drop with the RSA Model

We model the decision of whether to use a ZP based on the formulation expressed in Equation 5.3. The speaker model is $S_1(z|r)$, which is the probability that the speaker uses ZP (i.e., drops the RE). We assume that the speaker makes a binary choice (i.e., $z = \{1, 0\}$), with $z = 1$ indicating a ZP and $z = 0$ indicating a non-zero form of RE (NZRE). Note that whether the speaker uses a pronoun or a proper name is not in the scope of this model. To simulate the speaker's choice, we need to estimate the dropping probability $S_0(z|r)$, the discourse salience of the referent $P(r)$, and the cost $C(z)$.

According to the UID theory (see §5.2.1), if a RE is recoverable, then the speaker prefers a ZP over a NZRE to maximise the information density since a ZP is shorter than any other referential form. In that sense, we follow Orita et al. (2015) to estimate the *cost function* $C(z)$ based on the length of the RE, i.e., the total number of words the RE contains. However, the length of the NZRE is not known in advance, thus we use the average length of a set of REs $\mathcal{W}$ instead:

$$C(z = 0) = \text{average\_length}(\mathcal{W}) + 1 \tag{5.4}$$

We experimented with two ways of calculating the average length:

1. *global average length*, meaning that $\mathcal{W}$ is the set of all referring expressions in the corpus; and

2. *local average length*, in which $\mathcal{W}$ is the set of expressions that can refer to referent r. For instance, if r is "*Barack Obama*", then given a corpus for computing local average length in which *he* is referred to, $\mathcal{W}$ might be the set {*Barack Obama*, *Obama*, *he*, *former president*}.

The cost of a ZP is always $C(z = 1) = 1$, which means no discount on $P(z = 1|w)$ and the plus 1 in Equation 5.4 is to make the cost of choosing NZRE different from choosing ZP if $\mathcal{W}$ only contains pronouns (i.e., if length equals to 1).

We assume that the *dropping probability* $S_0(z|r)$ is dependent on whether the referent r is one of the participants in the dialogue (i.e., speaker or listener). For example, in the OntoNotes corpus, 30% of maximally salient entities are dropped, which is much higher than the 10% dropping rate of non-maximally salient entities. If r is one of the participants, we call it *maximally salient entity* (denoted as s). Otherwise, r_s is called *non-maximally salient entity* (denoted as ns). This assumption causes AZP and DNZP to have different proportions in the predicted results. Suppose $P(z = 1|r_s = \text{ns}) = a$ and $P(z = 1|r_s = \text{s}) = b$, then we have $a < b$, which implies that the speaker thinks the listener expects a maximally salient entity (i.e., speaker or listener).

Let $\alpha = \frac{a}{b}$ be the *dropping ratio*, then the probability of dropping a noun phrase that refers to the speaker is:

$$\begin{aligned} S_1(\text{ZP}|\text{Speaker}) &\propto L_1(\text{Speaker}|\text{ZP}) \cdot \frac{1}{C(z=1)} \\ &= \frac{S_0(\text{ZP}|\text{Speaker})P(\text{Speaker})}{\sum_{r'} S_0(\text{ZP}|r')P(r')} \cdot \frac{1}{C(z=1)} \\ &= \frac{N_{\text{Speaker}}}{\alpha \cdot N_{\text{NS}} + N_{\text{S}}} \cdot \frac{1}{C(z=1)} \end{aligned} \tag{5.5}$$

$P(\text{Speaker})$ is the *salience*[2] of the speaker. In general, we take the salience of a referent x to be in proportion to N_x, which is the number of times that x has been referred to in the preceding discourse, hence the use of N_{Speaker}, N_{S}, and N_{NS} in the equation. Note that $N_{\text{S}} + N_{\text{NS}}$ is the total number of REs in the preceding discourse.

Equation 5.5 shows that modelling the dropping probability for maximally salient entities and non-maximally salient entities differently acts as a discount for the number of referents that the ZP can refer to when predicting DNZP. Similarly, using the dropping ratio α, the dropping probability for a noun phrase that refers to a non-maximally silent entity r_ns is estimated as:

$$S_1(\text{ZP}|r_{ns}) = \frac{N_{r_{ns}}}{N_{\text{NS}} + \frac{1}{\alpha}N_{\text{S}}} \tag{5.6}$$

which can be seen as adding a penalty.

The frequencies counted above are all based on the whole preceding discourse of a referent, which might not be reasonable for predicting ZPs. We hypothesise that the informativeness of a ZP depends on only a part of the preceding context. We tested two possible set-ups. One is setting a discourse window to limit the number of sentences that the simulator can look back to. The other uses the idea of recency (Chafe, 1994). Following Orita et al. (2015), we replace each count with:

$$Count(r_i, r_j) = e^{-d(r_i, r_j)/a}, \tag{5.7}$$

where r_j is the same referent as the r_i that has previously been referred to and d is the number of sentences between two REs. Instead of taking the direct raw count 1, $Count(r_i, r_j)$ decays exponentially with respect to how far it is from the predicting RE. The RE that has larger distance contributes less to the overall count of that referent.

For NZREs ($z = 0$), we assume that the number of times that the referent has been referred to is equal to the total number of referents referred to by that NZRE. Thus, the speaker believes that the listener can always resolve the reference by giving them a NZRE. In other words, their informativeness equals 1.

5.2.4 Experiments

Experiment Settings

Dataset. We tested our model on the Chinese portion of OntoNotes Release 5.0 data (E. Hovy et al., 2006)[3]. Documents in the corpus come from six sources, namely Broadcast News, Newswires, Broadcast Conversations, Telephone Conversations, Web Blogs, and Magazines. It has been widely used in (ZP) co-reference resolution tasks. The corpus contains 1,729 documents, including 143620 referring expressions. In Table 5.1, there is the basic statistics about the recoverable zero pronouns in OntoNotes corpus.

Baseline. In this work, we used the modified rule 1 in Yeh and Mellish (1997), i.e., the RE in the subject position will be a ZP if it was referred to a referent in the immediately preceding sentence, as the baseline. The modification is inspired by the fact that 99.2% ZPs in OntoNotes corpus are in the subject position.

[2] Our use of the term salience is similar to E. Hovy et al. (2006)'s use of "recoverability".

[3] The OntoNotes dataset is available at: https://catalog.ldc.upenn.edu/ldc2013t19.

# of Recoverable Zero Pronouns	17,129
# of Anaphoric ZPs	14,675
# of Deictic Non-anaphoric ZPs	2,454

Table 5.1: Basic statistics of different types of recoverable ZPs in OntoNotes.

Discourse	Model	Cost	Total Acc.	ZP Acc.	AZP Acc.	DNZP Acc.	NZRE Acc.
-	baseline	-	78.57	40.88	42.90	28.81	83.67
Discourse Window	full	global	77.10	46.16	38.34	92.95	81.29
		local	81.79	22.53	25.50	4.81	89.81
	-dropping ratio	global	77.05	43.77	41.88	55.09	81.56
		local	81.44	23.67	27.09	3.19	89.26
Recency	full	global	75.64	**50.56**	43.08	**95.35**	79.03
		local	80.08	25.36	28.81	4.77	87.49
	-dropping ratio	global	74.04	50.26	**48.29**	62.02	78.04
		local	79.26	27.47	31.63	2.6	86.28
Whole	full	global	86.24	8.35	5.18	27.30	96.79
		local	**86.67**	3.67	4.27	0.08	**97.91**
	-dropping ratio	global	86.13	6.23	6.38	5.33	96.95
		local	86.61	3.84	4.47	0.04	97.81

Table 5.2: Accuracy of each model, recall that AZP and DNZP are two sub-categories of ZP.

Experiment Results

Table 5.2 shows the results (reported in accuracy) of various models on the OntoNotes dataset. The dropping ratio α was empirically set to 0.1 and the decay parameter a of recency was set to 0.8. The window size was 1, so the simulator only looks at the current sentence and preceding sentence.

As expected, the models that look back to the whole preceding discourse perform badly on predicting ZPs (i.e., 8.35% of accuracy), especially DNZPs. They tend to predict all REs as NZREs, which even performs worse than the model using simple rule (i.e., the baseline). In contrast, limiting the discourse history by applying discourse windows or replacing frequency with recency have a negative impact on predicting NZREs, more specifically pronouns. Such an impact is caused by the idea that every NZRE can always be resolved by the listener, which is not correct for pronouns. However, so far, we cannot calculate the informativeness of pronouns properly since we do not know which referent (speaker or listener) a deictic pronoun in the corpus refers to. For example, in the corpus, both the speaker and listener will use "I" to refer to themselves, so we don't know whether "I" refers to the speaker or the listener. This setting will lead to over-estimation of the informativeness of pronouns. Additionally, computing cost by average length (as we do) over-estimates the costs of pronouns, whose lengths are generally shorter than proper names.

The baseline model's performance is not bad, especially for predicting AZPs. This is partly because the rule predicts that all REs in object position are NZREs[4] and this is nearly always correct. At the same time, if the referent was referred to in the immediately preceding sentence (as the baseline model requires), then it is clearly more salient than if it

[4] Recall that 99.84% REs in object position are NZREs.

wasn't. The baseline model is therefore quite similar to the model with discourse window, but its decisions are made in a simpler way (i.e., based on a simple "if-then" rule).

With respect to overall accuracy for predicting ZPs and NZREs, models with recency perform similarly to those that use a discourse window. However, recency offers better prediction on AZPs. Adding a dropping ratio could significantly improve the performance on predicting DNZPs without decreasing the accuracy of AZPs and NZREs very much (i.e., accuracy increase from 62.02% to 95.35%). For the choice of cost function, we found that using global average length is the best.

5.2.5 Discussion

This study has explored the possibilities of using the RSA model for probabilistic simulation of speakers' use of ZPs (i.e., pro-drop), and investigated factors that influence speakers' choice. Our model performs respectably yet, as mentioned in §5.2.4, it under-estimates the probability of choosing a pronoun. Solving this problem will require a more fine-grained annotation of the corpus, indicating which person each occurrence of the deictic pronouns "I" and "you" refers to.

5.3 Study 2: Neural Referential Selection in English

Following on from the first study, we turn to take other referential forms into considerations, including pronoun, proper name, description and demonstrative, and to make use of advanced learning from data techniques, i.e., deep learning. In other words, we tackle the task of RFS using neural network based models. However, as aforesaid, there has been no previous work on Neural RFS and there has been no existing REG (in Context) dataset in Mandarin. Therefore, in this study, we introduce the task of RFS on the basis of an English REG corpus: the webNLG corpus and propose a number of neural models to tackle the task by adopting the state-the-of-the-art neural REG model of (Castro Ferreira, Moussallem, Kádár, et al., 2018).

In addition, neural models are always considered black-boxes. It was unclear to what extent these neural models can encode linguistic features, and, thus, it was hard to link the behaviours of these neural models to linguistic theories of reference (see §2.2.2 for more details). To conduct a model inspection, we make use of the probing classifiers. Using probing tasks is a well-established method to analyse whether a model's latent representation encodes specific information. This approach has been widely used for analysing models in machine translation (Belinkov, Durrani, et al., 2017), language modelling (Giulianelli et al., 2018), relation extraction (Alt et al., 2020), and so on. There had also been various works on co-reference resolution and bridging anaphora (Pandit & Hou, 2021; I.-T. Sorodoc et al., 2020) which, similar to this study, target the understanding of reference. More precisely, for a probing task, a diagnostic classifier is trained on representations from the model. Its performance embodies how well those representations encode the information associated with the probing task. To understand what linguistic features neural models encode when modelling REs, we introduce 8 probing tasks, each of which is associated with a linguistic feature influencing the choice of RF and examine our RFS models on these probing tasks in order to interpret and explain their behaviour.

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.

Delexicialised Text:
Pre-context: AWH_Engineering_College is in "Kuttikkattoor" , India in the state of Kerala .
Target Entity: AWH_Engineering_College
Pos-context: has 250 employees and Kerala is ruled by Kochi . The Ganges River is also found in India .

Table 5.3: An example data from the webNLG corpus. In the delexicalised text, every entity is underlined.

Type	Classes
4-Way	Demonstrative, Description, Proper Name, Pronoun
3-Way	Description, Proper Name, Pronoun
2-Way	Non-pronominal, Pronominal

Table 5.4: 3 different types of RF classification.

5.3.1 The RFS Task

Before formally defining the RFS task, we first recall the End2End REG task that has been detailed in §2.2.2. Based on webNLG, Castro Ferreira, Moussallem, Kádár, et al. (2018) first introduced the task of End2End REG. Taking the delexicalised text from webNLG in Table 5.3 as an example, given the entity "*AWH_Engineering_College*", REG chooses a RE based on that entity and its pre-context ("*AWH_Engineering_College is in "Kuttikkattoor" , India in the state of Kerala .* ") and its pos-context ("*has 250 employees and Kerala is ruled by Kochi . The Ganges River is also found in India .*").

Akin to End2End REG, given the previous context $x^{(pre)} = \{w_1, w_2, ..., w_{i-1}\}$ (where w is either a word or a delexicalised entity label), the target referent $x^{(r)} = \{w_i\}$, and the post context $x^{(pos)} = \{w_i, w_{i+1}, ..., w_n\}$, a RFS algorithm aims at finding the proper RF $\hat{f}$ from a set of K candidate RFs $\mathcal{F} = \{f_k\}_{k=1}^{K}$.

Regarding the possible RFs for the RFS task, we test 3 different classifications, depicted in Table 5.4. Due to the small number of demonstrative noun phrases in the dataset, we decide to also conduct a 3-way classification in which descriptions and demonstratives are merged. Also, most emphasis in the linguistic literature is on the pronominalisation issue. Therefore, we also include a 2-way classification task in the study.

As stated, the main goal of the study is to understand which linguistic features are encoded by RFS neural models. Additionally, we are curious whether models trained solely for pronominalisation capture different contextual features in comparison with the other two classifications.

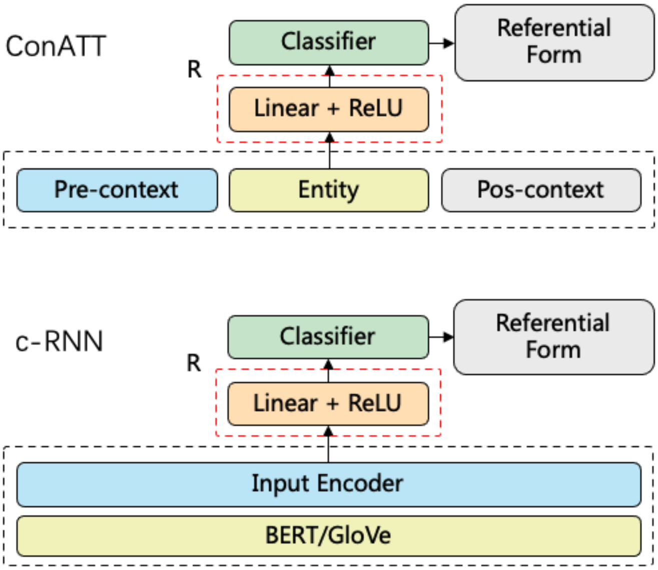


Figure 5.1: Figure of the ConATT model (above) and the c-RNN model (below).

5.3.2 Neural Referential Form Selection Models

We build NeuralRFS models by 1) adopting the best NeuralREG model from Castro Ferreira, Moussallem, Kádár, et al. (2018) (see §2.2.2 for more details); and 2) proposing a new alternative that is simpler, and can easier incorporate pre-trained representations.

ConATT

We adopt the CATT model from Castro Ferreira, Moussallem, Kádár, et al. (2018), which achieves the best performance on REG among the models they tested in their study. The above diagram in Figure 5.1 depicts our ConATT model. Given the inputs, we first use Bidirectional GRU (BiGRU, Cho et al., 2014) to encode $x^{(pre)}$ as well as $x^{(pos)}$. Formally, for each $k \in [pre, pos]$, we encode $x^{(k)}$ to $h^{(k)}$ with a BiGRU:

$$h^{(k)} = \text{BiGRU}(x^{(k)}). \tag{5.8}$$

Subsequently, different from Castro Ferreira, Moussallem, Kádár, et al. (2018), we encode $h^{(k)}$ into the context representation $c^{(k)}$ using self-attention (Z. Yang et al., 2016). Concretely, given the total N steps in $h^{(k)}$, we first calculate the attention weight $\alpha_j^{(k)}$ at each step j by:

$$e_j^{(k)} = v_a^{(k)T} \tanh(W_a^{(k)} h_j^{(k)}), \tag{5.9}$$

$$\alpha_j^{(k)} = \frac{\exp(e_j^{(k)})}{\sum_{n=1}^{N} \exp(e_n^{(k)})}, \tag{5.10}$$

where v_a is the attention vector and W_a is the weight in the attention layer. The context representation of $x^{(k)}$ is then the weighted sum of $h^{(k)}$:

$$c^{(k)} = \sum_{j=1}^{N} \alpha_j^{(k)} h^{(k)}. \tag{5.11}$$

After obtaining $c^{(pre)}$ and $c^{(pos)}$, we concatenate them with the target entity embedding $x^{(r)}$, and pass it through a feed-forward network to obtain the final representation:

$$R = \text{ReLU}(W_f[c^{(pre)}, x^{(r)}, c^{(pos)}]), \tag{5.12}$$

where W_f is the weight in the feed-forward layer. R is also used as the input of the probing classifiers. R is then fed for making the final prediction:

$$P(f|x^{(pre)}, x^{(r)}, x^{(pos)}) = \text{Softmax}(W_c R), \tag{5.13}$$

where W_c is the weight in the output layer.

c-RNN

In addition to `ConATT`, we also try a simpler yet effective structure, which uses only a single BiGRU. We name the framework it follows as the centred recurrent neural networks (henceforth `c-RNN`), which is sketched in the bottom diagram of Figure 5.1. Specifically, instead of using two separate BiGRUs to encode pre- and pos-contexts, we first concatenate $x^{(pre)}$, $x^{(r)}$, and $x^{(pos)}$, and then encode them together:

$$h = \text{BiGRU}([x^{(pre)}, x^{(r)}, x^{(pos)}]). \tag{5.14}$$

Suppose that the target entity is in position i of the concatenated sequence, we extract the i-th representation from h_i for obtaining $R = \text{ReLU}(W_f h_i)$. After obtaining R, the rest of the procedure is the same as `ConATT`.

Pre-training

As a secondary objective of this study, we want to see whether RFS can benefit from pre-trained word embeddings and language models, whose effectiveness has not yet been explored in REG. [5] For both `c-RNN` and `ConATT`, we try the GloVe embeddings (Pennington et al., 2014) to see how pre-trained word embeddings contribute to the choice of RF. [6] For `c-RNN`, we try to stake it on the `BERT` (Devlin et al., 2019) model. In order to let `BERT` better encode the delexicalised entity labels, we first re-train `BERT` as a masked language model on the training data of WEBNLG. We then freeze the parameters of `BERT` and use the model to encode the input, which is then fed into `c-RNN`.

Machine Learning (ML) based Model.

We use `XGBoost` (T. Chen & Guestrin, 2016) from the family of Gradient Boosting Decision Trees to train RFS classifiers. 5-fold-cross-validation was used to train the models. The

[5] Previously, only Cao and Cheung (2019) used pre-trained embeddings, but no ablation study was done.

[6] We also explored other ways of using `BERT`, such as using only `BERT` plus a feed-forward layer to obtain h, or not freezing parameters of `BERT` while training. The resulting models had low performance in all cases.

Feature	Definition	2-way	3-way	4-way
Syn	Description is provided in the main text.	✓	✓	✓
Entity	Values: Person, Organisation, Location, Number, Other	✓	✓	✓
Gender	Values: male/female/other	✓	✓	✓
DisStat	Description is provided in the main text.	✓	✓	✓
SenStat	Description is provided in the main text.	-	✓	✓
DistAnt_S	Description is provided in the main text (DistAnt).	✓	✓	✓
DistAnt_W	Distance in number of words (5 quantiles)	✓	-	✓
Sent_1	Does RE appear in the first sentence?	✓	✓	✓
MetaPro	Description is provided in the main text.	✓	✓	✓
GloPro	Description is provided in the main text.	✓	✓	✓

Table 5.5: Features used in the XGBoost models.

	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

Table 5.6: Evaluation results of our RFS systems on WEBNLG. Best results are **boldfaced**, whereas the second best results are underlined.

classifiers were first trained on a wide range of features obtained from the webNLG corpus (16 features). After running a variable importance analysis, we selected a subset of features for the final models. The detailed list of features is presented in Table 5.5.

5.3.3 Evaluating RFS Models

Implementation Details

We tuned hyper-parameters of each of our models on the development set and chose the setting with the best macro F1 score. For the BERT model, we used the cased BERT-BASE and added all entity labels into the vocabulary to avoid tokenisation. [7] When re-training BERT on webNLG, we set the masking probability to 0.15 and trained it for 25 epochs.

For the XGBoost models, we set the learning rate to 0.05, the minimum split loss to 0.01, the maximum depth of a tree to 5, and the sub-sample ratio of the training instances to 0.5.

We report the macro averaged precision, recall, and F1 on the test set. We ran each model 5 times, and report the average performance. As for the dataset, we used v1.5 of webNLG (Castro Ferreira et al., 2019) and used only seen entities.

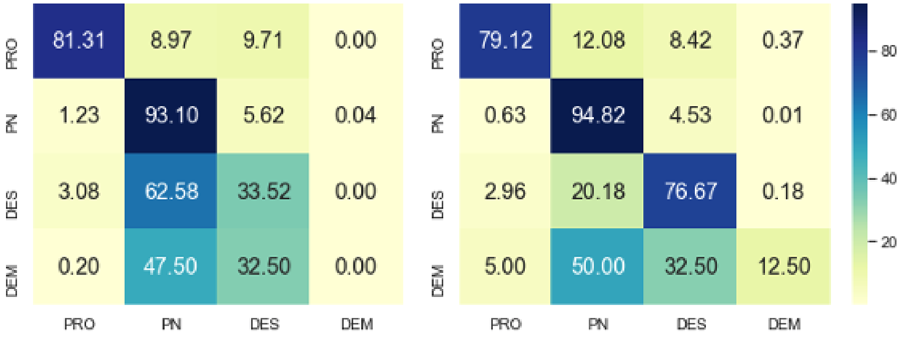


Figure 5.2: Confusion Matrices for 4-way classification results of XGBoost (left) and c-RNN+GloVe (right), where PRO, PN, DES, and DEM are pronoun, proper name, description and demonstrative respectively.

Results

Table 5.6 shows the results of different classification tasks. Generally, all neural variants outperform the machine learning baseline. The performance difference is small in the case of binary classification, while it is much bigger for 3- and 4-way classifications. This is because the 2-way classification (i.e., pronominalisation) is clearly less complex than the other two alternatives, and, thus, the feature set used by the baseline results in almost similar outcomes to neural models.

Comparing neural variants to each other, the results show that the simpler c-RNN wins over ConATT in 4-way classification, and has on par performance with ConATT for 3- and 2-way classifications. One possible explanation is that ConATT first breaks down the input into three pieces (i.e., the target entity as well as pre- and pos-context), encodes them separately, and merges the encoded representations back before being sent to make predictions. This "divide and merge" procedure might hinder the model from learning some useful information.

Regarding the effectiveness of incorporating pre-trained models, GloVe embeddings have a positive impact on c-RNN only in case of 4-way predictions and have no contribution to 2- and 3-way classifications. Moreover, it has a negative effect on ConATT: the performance diminishes when GloVe is used. It is surprising to see that in the case of c-RNN, BERT has a negative effect on 4- and 3-way predictions (the F1 score was reduced from 64.86 and 83.63 to 62.15 and 82.15 respectively). For pronominalisation, BERT slightly boosts the performance (from 89.09 to 89.42), but this boost is not as much as BERT's boosting effect on other NLP tasks. This is probably because although BERT was re-trained on webNLG delexicalised sentences, the entity labels still function as noise for BERT.

To obtain insights into the behaviours of the deep learning and classic ML-based models for RFS, we depict the confusion matrices of XGBoost and the best performing neural model c-RNN+GloVe in Figure 5.2 for the 4-way classification. The confusion matrices suggest that both models do a good job in selecting pronouns and proper names (that is why the performance difference in the 2-way classification is small), and both perform poorly in choosing demonstratives (probably due to the fact that demonstratives are extremely

[7] The code for cased BERT-BASE can be found at: huggingface.co/bert-base-cased

infrequent in webNLG). The main difference between the two models is in distinguishing proper names from descriptions. The XGBoost model wrongly predicted the descriptions as proper names in 62.58% of the cases, while the neural c-RNN+GloVe model did this wrong prediction in 20.18% of the times. This difference in the performance of the two models might be because the neural models learnt some useful features from the discourse which are not covered in our feature engineering procedure. Furthermore, after looking into the webNLG dataset, we noticed that various RE cases are annotated incorrectly. For example, webNLG annotates "*United States*" as a proper name, and "*the United States*" as a description. The incorrect annotations might increase the confusion between choosing description and proper name and, as a consequence, reduce the overall performance.

5.3.4 Probing RFS Models

We use a logistic regression classifier as our probing classifier. Concretely, for each input, we first use a model discussed in §5.3.2 to obtain its representation R. As mentioned in §5.3.2, we ran each model five times and reported their averaged scores. For the probing tasks, we used the representations of the models with the best RFS performance on the development set.

Probing Tasks

Following the discussion about factors that influence the choice of referential forms in §2.2.2, we formulate the following probing tasks.

Referential Status. The referential status of the target entity influences the choice of RF in both linguistic (Chafe, 1976; Gundel et al., 1993) and computational studies (Castro Ferreira et al., 2016). In this study, we define referential status on two levels: discourse-level and sentence-level. The former (**DisStat**) has two possible values: (a) discourse-old (i.e., the entity has appeared in the previous discourse); and (b) discourse-new (i.e., the entity has not appeared in the previous discourse). Sentence-level referential status (**SenStat**) also consists of three values: (a) sentence-old (i.e., the RE is not the first mention in the current sentence); (b) sentence-new (i.e., the RE is the first mention of the entity in the sentence); and (c) discourse-new (i.e., the RE is the first mention of the entity in the discourse).

Syntactic Position. Entities in subject position are more likely to be pronominalised than in object position (Arnold, 2010; Brennan, 1995). Therefore, in the syntax probing task (henceforth **Syn**), we do binary classification: subject or object.

Recency. Recency has been used as a vital feature in many of the previous REG or RFS systems (Greenbacker & McCoy, 2009; Kibrik et al., 2016). It measures the distance between the target entity and its closest antecedent. There are various ways of estimating the recency of a target entity given its context. We hereby use two measures:

1. The number of sentences between the target entity and its antecedent (**DistAnt**), which consists of four possible values: the entity and its antecedent are (a) in the same sentence; (b) one sentence away, (c) more than one sentence away; and (d) the entity is a first mention (to distinguish first mentions from subsequent mentions); and

2. Whether there is an intervening referent between the target and its nearest antecedent (**IntRef**) (Greenbacker & McCoy, 2009). In other words, it checks whether the target and the preceding RE are coreferential. This feature has three possible values: (a) the target entity is the first mention; (b) the previous RE refers to the same entity; and (c) the previous RE refers to a different entity. Note that the existence of intervening markable might signal the existence of a competition (if the intervening referent has the same animacy and gender values as the target RE).

Discourse Structure Prominence. As mentioned in §2.2.2, the "organisational" properties of discourse may influence the prominence status of the entities. We introduce three probing tasks capturing different properties of the discourse.

1. *Local prominence* (**LocPro**): The idea of local prominence is coming from Centering Theory (Grosz et al., 1995). It is a hybrid feature of DisStat and Syn. Concretely, we use the implementation of Henschel et al. (2000): an entity is *locally prominent* if it is "discourse-old" and "realised as subject". It is a binary feature with two possible values: (a) locally prominent; and (b) not locally prominent;

2. *Global prominence* (**GloPro**): This feature is based on the notion of global salience in Siddharthan et al. (2011), asking whether the entity is a minor or major referent in the text. According to them, "the frequency features are likely to give a good indication of the global salience of a referent in the document" (p. 820). We define a binary feature in which the most frequent entity in a text is marked as globally prominent.

3. *Meta-prominence* (**MetaPro**): In line with global prominence, we also want to explore to what extent prominence beyond a single text (e.g. on a text collection level) may impact the way people refer. In the context of the current circumstances, the sentence "I received *my vaccine* today" is unambiguous, and the RE *my vaccine* needs no extra modification (e.g. my COVID-19 vaccine); however, a couple of years from now, a richer RE may be needed to refer to the vaccine. The idea behind this exploratory feature is that people might use less semantic content to refer to the referents which are well known outside of the text. Based on the number of mentions of a target entity in the whole WEBNLG, four possible values, each of which representing an interval, are assigned to each RE: (a) $[0, 50)$; (b) $[50, 150)$; (c) $[150, 290)$; and (d) $[290, \infty)$. For example, the category $[0, 50)$ contains those entities that occur fewer than 50 times in the corpus.

Importance Analysis

We conducted a feature importance analysis to find out which features that are used in the probing tasks had the highest contributions to the feature-based ML models. This analysis functions as a sanity check to find out whether the representations have learnt the features contributing the most to the RFS task.

To assess the importance of the features used in the probing tasks, we trained `XGBoost` models, only using the features above, and calculated the model-agnostic permutation-based variable importance of each model (Biecek & Burzykowski, 2021). Concretely, we measured the extent to which the performance changes if we removed one of the features. Figure 5.3 depicts the performance change for each feature. According to the figure, DisStat and Syn contribute the most. LocPro is the least important feature because it is a hybrid

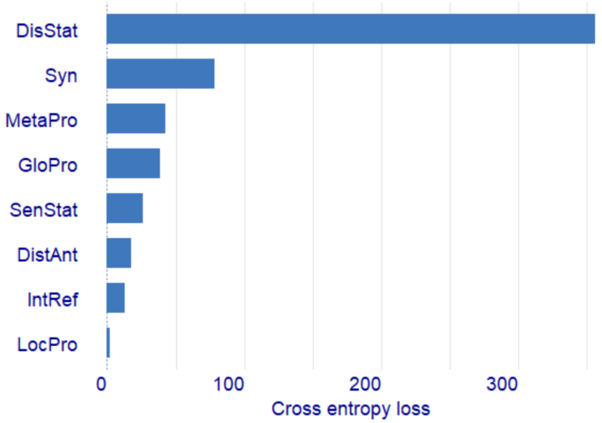


Figure 5.3: Feature importance of XGBoost classifiers for 4-way predictions. Higher loss shows greater importance of a feature.

combination of DisStat and Syn. Removing it while keeping DisStat and Syn will not hurt the performance of the model a lot. Considering that DisStat and Syn are both highly vital features, LocPro is much more important than what the experiment suggests. In addition to DisStat and Syn probing tasks, we also expect high performance for the LocPro task.

Probing Results

We mentioned earlier in this study that we conducted probing tasks to find out whether the RFS models' latent representations encode the linguistic features. High performance in probing tasks would indicate that the features are encoded in the latent representations of the models.

We evaluate probing tasks using the accuracy and macro-averaged F1 scores. Each probing classifier was trained 5 times. Here, we report the averaged value. Additionally, we used 2 baselines:

1. `random`: it randomly assigns a label to each input; and
2. `majority`: it assigns the most frequent label in the given probing task to the inputs.

Results of Each Probing Task. Compared to the `random` baseline, all neural models have achieved higher performance on all tasks.

1. Referential status and syntactic position: all models exhibit consistently high performance on DisStat, SenStat, and Syn. This shows that, at least for the webNLG corpus, all neural models can learn information of referential status and syntactic position;
2. Recency (i.e., DistAnt and IntRef): all models perform worse compared to the referential status and syntax probes. Although they do not have bad accuracy scores, their F1 scores are lower than that of DisStat, SenStat, and Syn, and are closer to the baselines. This finding is consistent with the results of the importance analysis, where DistAnt and IntRef were found to be less important (compared to DisStat and

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)

Table 5.7: Results of each probing task. Results are reported in the format of A(B), where A is the accuracy and B is the macro F1.

Syn). One possible explanation is that, in the webNLG corpus, 67% of the documents contain only one sentence, making recency-related features play a smaller role. As another possible explanation, in line with the previous probing works on co-reference and bridging anaphora (Pandit & Hou, 2021; I.-T. Sorodoc et al., 2020), models have more difficulty capturing long-distance properties;

3. Discourse structure prominence: since LocPro is a hybrid of DisStat and Syn, all models handled it to a large degree. Meanwhile, neural models appear to handle GloPro and MetaPro worse than other features since the performance of their corresponding probing tasks is closer to the baselines. [8] These results are in contrast with

[8] Note that, for MetaPro, the `Majority` has a low F1 score because the distribution of the values of MatePro is balanced.

the importance analysis results, which suggested that both GloPro and MetaPro are important features (ranking 3 and 4 in Figure 5.3). Learning GloPro and MetaPro requires a model to have an overall understanding of the whole input document or the whole corpus, which the neural models might not be able to acquire.

Comparing `c-RNN` and `ConATT`. When evaluating our RFS models, we concluded that the `c-RNN` model works better than `ConATT` on 4-way RF classification. Nevertheless, when probing, we observed that `ConATT` does a better job in many tasks, including DisStat, LocPro, GloPro, and MetaPro. To understand why, we look into the webNLG dataset and found that REs in webNLG are not representative of the realistic use of REs. Specifically, it has three shortcomings: 1) it consists of rather formal texts that may not reflect the everyday use of REs, and in which very simple syntactic structures dominate; 2) the texts are extremely short, with an average length of only 1.4 sentences. Consequently, 86.91% of the REs in webNLG are first mentions; 3) 21% of the documents talk about the entity "*United_States*". Therefore, although `ConATT` learns more contextual features, it still has a lower performance. `ConATT`'s better learning of referential status (i.e., DisStat) is probably a benefit of using self-attention, which helps the model capture longer dependencies than RNNs.

The Effect of Pre-training. As mentioned earlier, the secondary objective of this study is to find out whether RFS can benefit from pre-trained word embeddings and language models. The effect of incorporating the `GloVe` embeddings is not significant to `c-RNN` and `ConATT`. The major contribution of `BERT` is helping with learning DisStat (which is, again, probably a result of using self-attention). Akin to the above discussion, since the majority of the entities in webNLG are first mentions, the increased accuracy boost in the DisStat task is not enough to boost the overall performance of RFS.

Comparing Different RF Classifications. It also appears that models learn different information using different label sets (classes). For example, 2-way classification (i.e., pronominalisation) helps `c-RNN` learn more about referential status. But in case of models with attention mechanism (i.e., `ConATT`, `ConATT+GloVe` and `c-RNN+BERT` models), referential status is learnt better in 4-way classification models. Also, in case of `ConATT(+GloVe)`, we observed that more fine-grained classifications help the model learn more about meta prominence (i.e., MetaPro).

5.3.5 Discussion

Our aim is to understand whether neural models capture the features associated with the task of RFS. To this end, we defined 8 probing tasks in which we focused on referential status, syntactic position, recency, and discourse structure. The probing results suggest that the probe classifiers always performed better than the `random` and the `majority` baselines. The performance was consistently good in the tasks associated with referential status, syntax and local prominence.

It is worth noting that probing has its own shortcomings. For instance, on the one hand, low probing performance does not always mean the feature is not encoded, but could also mean that such a feature does not matter to RFS. To mitigate this issue, we conducted a complementary ML-based variable importance analysis; in this analysis, discourse status and syntactic position came out as the factors with the highest contributions. These features were also predicted very well in the probing tasks. However, these results should still be taken with a pinch of salt: the variable importance has been conducted on the ML model and not on the neural models. We cannot be certain that the same features contribute to all the models similarly: a feature might be quite important in the machine learning model, but not as important in the neural models. On the other hand, some researches have questioned the validity of probing methods. They found out that it is difficult to distinguish between "learning the probing task" and "extracting the encoded linguistic information" (Hewitt & Liang, 2019; Kunz & Kuhlmann, 2020) for a probing classifier. This suggests that higher performance of a probing classifier does not necessarily mean more linguistic information has been encoded. This prevents us from directly quantifying *how well* the linguistic information has been learnt using the performance of probing classifiers and requires us to make conclusions more carefully.

From our probing efforts, we concluded that:

1. All neural models have learnt some information about the features associated with the probing tasks, but how well they have learnt this information is yet to be assessed;

2. The webNLG corpus, which has often been used for the study of discourse REG, is not ideally suitable for studying discourse-related aspects of RFS, because the texts are too short and the majority of REs are first mentions. This leads to bias in the evaluation of RFS and REG algorithms;

3. When it comes to the question of how well an RFS feature can be learnt, it matters what neural architecture and label set are used, and whether the model is pre-trained or not. Using an attention mechanism and more fine-grained label sets help a model learn more information;

4. All models perform poorly in terms of learning those features, such as GloPro and MetaPro, that do not derive from the text itself but from the wider context in which it is written and read.

5.4 Study 3: Neural Referential Selection in Mandarin

In the previous study, we built a number of NeuralRFS models and examined them on the webNLG dataset, an English REG dataset. This study turns to the RFS task in Mandarin. As aforesaid, one challenge of RFS in Mandarin is that the RFS, as a classification task, has an extra option: zero pronouns. However, there is no suitable dataset available. We, therefore, need to construct an RFS dataset in Mandarin (§5.4.1). As concluded from Study 2, the webNLG is not ideal for studying human referential behaviour because its texts are too formal and short and its REs are predominantly "first mention" noun phrases. Therefore, we are aiming for a dataset whose texts are natural and long, and whose REs are less often first mentions than in webNLG. To this end, we build our dataset on the basis of the OntoNotes dataset (which was used in the first study), whose contents come from six sources, namely Broadcast News, Newswires, Broadcast Conversations, Telephone Conversations, Web Blogs and Magazines.

Subsequently, in §5.4.1, we extend the models proposed in §5.3.2 to handle Mandarin texts. In most Mandarin NLP tasks, the input text can be encoded either in a word-based way or in a character-based way, and most pre-trained models (e.g., BERT) have only a

Pre-context: 风 或许 是 百步蛇 成为 排湾族 祖灵 的 原因 。 百步蛇 属于 蛇类 中 演化 较 晚 的 蝮蛇科 ， 与 原始 的 蟒蛇 、 盲蛇 不同 。 蝮蛇科 蛇类 具有 分泌 毒素 的 毒腺 ， 杜铭章 解释 ， 相较于 灵活 的 四 足 动物 ， 蛇类 既 乏 四肢 ， 眼力 、 听力 又 都 不 佳 ， 毒牙 几乎 是 蛇类 猎食 与 御敌 的 唯一 工具 。 杜铭章 近 几 年 带领 学生 陆续 研究 了 台湾 特有种 菊池氏 龟壳花 与 赤尾_青竹丝 等 蛇类 生态 ， 但
Target Entity: 百步蛇
Pos-context: 仍 乏人问津 。 针对 百步蛇 ， 缺乏 生殖 周期 、 摄食 偏好 、 活动 范围 、 生活 习性 等等 的 深入 调查 ， 保育 也 只 是 空谈 。 百步蛇 不 是 冷血 动物 直 到 前年 动物园 开始 纪录 百步蛇 的 繁衍 过程 ， 人们 才 有 机会 了解 百步蛇 的 真 面貌 。

Table 5.8: An example data sample from the OntoNotes corpus.

Type	Classes
4-Way	Description, Proper Name, Pronoun, Zero Pronoun
3-Way	Proper Name, Pronoun, Zero Pronoun
2-Way	Overt Referring Expression, Zero Pronoun

Table 5.9: 3 different types of Mandarin RF classification.

character-based version available. Thus, in this study, we try both strategies: word-based modelling and character-based modelling.

Lastly, we evaluate and probe the models following the same paradigm as Study 2 (see §5.3.2 and §5.3.4).

5.4.1 Dataset Construction

Akin to the first study, we construct the dataset based on the OntoNotes dataset and process the data to the same format as that of the webNLG dataset. We follow the following construction process.

First, for each RE in OntoNotes, we used 3 previous sentences as the pre-context and 3 proceeding sentences as the pos-context. With the help of the syntax tree of the sentence that the target referent is in and the surface form of the target referent, we automatically annotated each RE with its category. For example, if the surface form is "*pro*", then the category is ZP. If the RE is an NP in the syntax tree, then it was annotated as a description. In this study, we consider three different classifications. The binary classification asks classifiers to conduct pro-drop (i.e., whether the target referent should be realised as a ZP or an overt RE). The 3-way classification is to do pronominalisation and pro-drop simultaneously. Table 5.9 lists the details of all three classifications.

Second, we extracted all REs for every referent. Different from study 1, this time, we only focus on the referents that are referred by at least one proper name or one description. If all REs that refer to a given referent are ZPs or pronouns, then we disregarded this referent from our corpus. In other words, we do not investigate deictic ZPs and deictic pronouns in this study.

Third, to obtain the entity label for each referent, following Castro Ferreira, Moussallem, Kádár, et al. (2018), we used its proper name and replaced the blanks with "_". If a referent has no proper name, we used its shortest description instead. Subsequently, both the pre-contexts and pos-contexts were delexicalised using entity labels. Table 5.8 shows a sample from the constructed dataset.

Fourth, we split the whole dataset into training set and test set in accordance with the CoNLL 2012 Shared Task (Pradhan et al., 2012). We then sampled 10% of documents from the training set as the development data.

As a result, we obtained a dataset in which the training set contains 73607 samples, the development set contains 10008 samples, and the test set contains 12096 samples. From now on, we refer to this newly constructed dataset as OntoNotes.

5.4.2 Mandarin RFS Models

We planned to test exactly the same set of neural models in §5.3.2 on the OntoNotes dataset. Nevertheless, the way in which `BERT` encodes Mandarin texts is different from how it encodes English texts. For English, before encoding an input, `BERT` calls a word segmentation algorithm (e.g., BPE (Sennrich et al., 2016) and WordPiece (Y. Wu et al., 2016)) to break each token in the input into subwords (which is often morphemes). For Mandarin, since morphemes in Mandarin are always characters (see §A for more details), Mandarin `BERT` is fully character-based. To conduct a fair comparison, we grouped the models into two categories: word-based models and character-based models.

Word-based Models

For Mandarin, both `ConATT` and `c-RNN` can be used in the same way as in §5.3.2. Regarding the use of pre-trained word embeddings, since there is no commonly used pre-trained Mandarin `GloVe` embedding available, we use the one trained by `Word2Vec` (Mikolov et al., 2013) instead. [9] More specifically, it is trained through the Skip-Gram with Negative Sampling (SGNS) technique on the Chinese Wikipedia corpus using all word, character, and N-gram features (S. Li et al., 2018).

Character-based Models

We adapt all neural models in §5.3.2 to be character-based. To this end, we need to re-process the OntoNotes corpus. We broke all inputs into characters, including the pre-contexts, the pos-contexts, and the entity labels. Therefore, for a target referent r, the input of RFS models defined in §5.3.1 is re-formalised as: the previous context $x^{(pre)} = \{c_1, c_2, ..., c_{i-1}\}$ (where c is a character), the target referent $x^{(r)} = \{c_i, c_{i+1}, ..., c_j\}$, and the post context $x^{(pos)} = \{c_{j+1}, c_{j+2}, ..., c_n\}$. Since entity labels were also broken into characters, the underlines "_" in them are no longer meaningful. Following Cunha et al. (2020), we removed all "_" in entity labels. Additionally, due to the limitation of the computing resources, for the `BERT` model, we use "bert-base-chinese", which only accepts inputs shorter than 512 characters. [10] Thus, we removed all inputs whose total lengths (including all the pre-contexts, the pos-contexts, and the target referents) are longer than 512 characters. We call the resulting corpus OntoNotes-c. Its training set has 70428 samples, its development set has 9217 samples, and its test set has 11607 samples.

In what follows, we describe how we adapt the `ConATT` and the `c-RNN` to become character-based.

[9] The pre-trained embeddings are available at: https://github.com/Embedding/Chinese-Word-Vectors.

[10] The pre-trained bert-base-chinese model is available at: https://huggingface.co/bert-base-chinese.

	4-way			3-way			2-way		
Model	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
`c-RNN`	53.55	51.51	52.18	56.11	53.33	54.42	64.91	63.64	64.22
`+SGNS`	**59.14**	**57.65**	**57.63**	**59.15**	55.46	56.78	66.76	**68.57**	67.58
`ConATT`	53.69	52.93	53.05	55.25	54.04	54.55	64.60	65.85	65.01
`+SGNS`	57.75	55.98	56.42	**59.34**	55.40	**56.87**	**67.04**	68.30	**67.59**
`c-RNN-char`	54.60	54.65	54.19	56.78	53.50	54.68	67.66	62.89	64.59
`+SGNS`	57.78	56.75	57.16	59.57	56.19	57.46	67.74	65.33	66.37
`+BERT`	**68.22**	**69.48**	**68.17**	**70.36**	**68.60**	**69.13**	**78.35**	**73.51**	**75.59**
`ConATT-char`	54.27	53.08	52.98	53.67	49.47	50.79	63.25	56.92	58.28
`+SGNS`	55.88	54.94	54.18	55.01	53.06	53.87	64.98	61.38	62.69

Table 5.10: Evaluation results of our word based RFS systems on the OntoNotes dataset as well as our character based RFS systems on the OntoNotes-c. Best results are **boldfaced**, whereas the second best results are underlined.

ConATT-char. In Study 2, we used self-attention (see Equation 5.8-5.11) to encode the pre-context and pos-context. Here, in `ConATT-char`, we also encode the entity label using the same method, and obtain its representation $c^{(r)}$. Subsequently, the final representation R is computed by:

$$R = \text{ReLU}(W_f[c^{(pre)}, c^{(r)}, c^{(pos)}]), \tag{5.15}$$

The rest of the procedure is the same as the one that was used by `ConATT`.

c-RNN-char. As in Equation 5.14, the `c-RNN-char` model first concatenates $x^{(pre)}$, $x^{(r)}$, and $x^{(pos)}$, and encodes the concatenated input using a single BiGRU to obtain the representation of the whole input h. Subsequently, we extract the i-th representation (i.e., the position of the first character in $x^{(r)}$) and the j-th representation (i.e., the position of the last character in $x^{(r)}$) from h and add them together for obtaining:

$$R = \text{ReLU}\left(W_f(h_i + h_j)\right) \tag{5.16}$$

After obtaining R, the rest of the procedure is the same as the `c-RNN`.

5.4.3 Evaluating Mandarin RFS Models

Implementation Details

We tuned hyper-parameters of each of our models on the development set and chose the setting with the best macro F1 score. For the `BERT` model, we used the cased BERT-BASE-CHINESE. We report the macro averaged precision, recall, and F1 on the test set. We ran each model 5 times, and report the average performance.

Results

Table 5.10 lists the results of our models on the OntoNotes and OntoNotes-c, respectively. On the OntoNotes dataset, probably because of the use of the self-attention, `ConATT` performs better than `c-RNN`. Specifically, `ConATT` significantly improves the recall score in

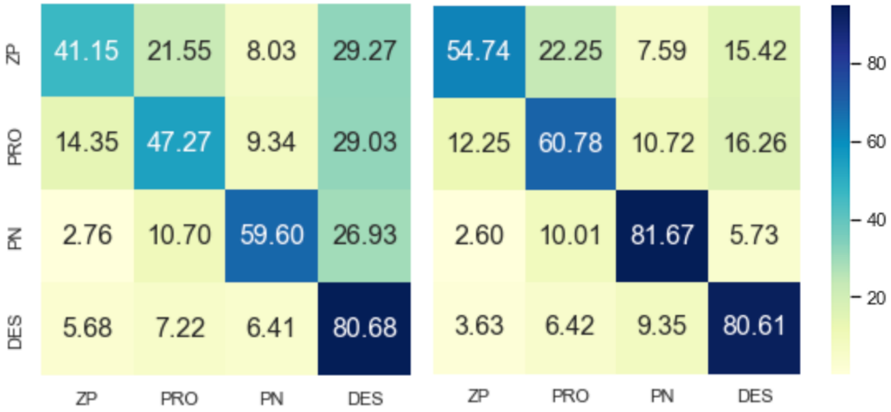


Figure 5.4: Confusion Matrices for 4-way classification results of c-RNN-char+SGNS (left) and c-RNN-char+BERT (right), where ZP, PRO, PN, and DES are zero pronouns, pronoun, proper name, and description respectively.

all three classifications compared to c-RNN. The use of the pre-trained word embeddings SGNS appears to make a considerable contribution. It helps to improve the performance of both c-RNN and ConATT.

On the OntoNotes-c, the results are very different: c-RNN outperforms ConATT, especially when doing 3-way and 2-way classification. One possible explanation is that, in character-based modelling, the inputs are too long to be handled by the self-attention module of ConATT. It is also worth noting that the self-attention mechanism of ConATT is different from that of BERT. To be precise, ConATT uses a random initialised vector as the attention query (Z. Yang et al., 2016) while BERT uses the input as the attention query (Vaswani et al., 2017).

Additionally, BERT dramatically improves the performance of each of the three classifications. For example, in 4-way classification, using BERT improves F1 by 13.98 points compared to c-RNN-char and by 11.42 points compared to c-RNN-char+SGNS. To see in which way BERT can improve the performance, we print the confusion matrix for the 4-way classification using c-RNN-char+SGNS and c-RNN-char+BERT in Figure 5.4. As we can see from the confusion matrix, c-RNN-char+SGNS is finding it hard to distinguish ZPs, pronouns and proper names from descriptions. More than 25% of them are misclassified as descriptions. BERT can solve the problem to a large extent. For instance, only 5.73% of the proper name are misclassified as descriptions after using BERT.

Focusing on the use of ZPs, we print the confusion matrix for the 2-way classification in Figure 5.5. Akin to the 4-way classification, c-RNN-char+SGNS also does not work well on distinguishing ZPs from overt REs and BERT reduce the misclassification rate from 61.95% to 44.56%. Moreover, we also observed that only 4.92% of the overt REs are misclassified as ZPs. This suggests that BERT, by learning from contexts and pre-trained on large scale datasets, can work considerably well on using ZP in a natural way.

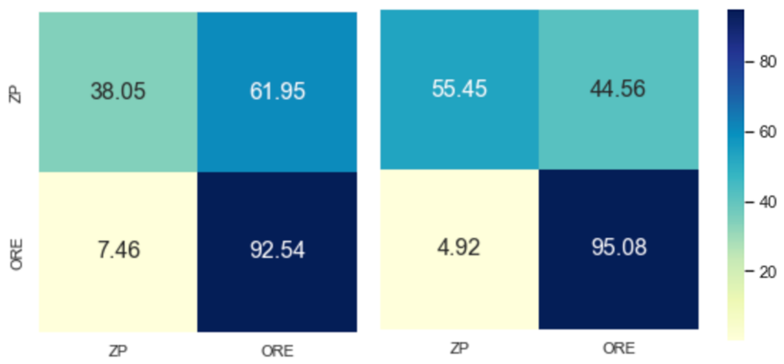


Figure 5.5: Confusion Matrices for 2-way classification results of `c-RNN-char+SGNS` (left) and `c-RNN-char+BERT` (right), where ZP and ORE are zero pronoun and overt RE, respectively.

5.4.4 Probing Mandarin RFS Models

In this study, we use the same probing classifier as in Study 2: a logistic regression to analyse the latent representations learnt by models in §5.4.2. We use the representations of the models with the best RFS performance on the developments as the inputs to the probing classifier.

Probing Tasks

We use all probing tasks defined in §5.3.4 excepting the MetaPro task. The MetaPro asks the classifier to identify referents that appear way more frequently than other referents. However, compared to webNLG, OntoNotes is more natural and was collected from a wider range of resources. There is no referent appears more than 50 times in the corpus. We therefore decide not to include it in this study.

Probing Results

We use the same two baselines as in Study 2: `random` and `majority`. We evaluate each probing task using the accuracy and macro-averaged F1 scores. Each probing classifier was trained 5 times and we hereby report the averaged value.

Table 5.11 and Table 5.12 report the probing results of our word-based models and character-based models, respectively.

Results of Each Probing Task. Focusing on the results of each probing task, we made the following observations. First, by comparing the performance between baselines and neural models, we found that all models can learn a certain amount of information about nearly all features, except the GloPro. Every model (including the baselines) obtains a similar performance on the GloPro task. There are two possible explanations. One is that the neural models are not good at counting how many times a referent appears in a discourse, and, thus, they failed to figure out the dominant referent. The other is because we constructed

Model	Type	DisStat	SenStat	Syn	DistAnt	IntRef	LocPro	GloPro
Random	-	50.20	33.18	50.11	25.02	33.56	50.12	50.00
		(49.93)	(32.70)	(49.79)	(23.81)	(33.01)	(46.44)	(44.27)
Majority	-	57.30	42.70	57.79	42.70	42.70	76.27	81.13
		(36.43)	(19.95)	(36.62)	(14.96)	(19.95)	(43.27)	(45.09)
c-RNN	4-way	64.81	47.24	75.23	45.63	47.34	79.02	82.02
		(62.05)	(41.68)	(73.79)	(26.16)	(43.46)	(64.50)	(46.68)
	3-way	64.42	47.74	75.79	45.95	46.75	78.83	82.00
		(61.18)	(43.44)	(74.64)	(27.32)	(42.40)	(65.34)	(45.20)
	2-way	62.23	46.15	77.29	45.62	45.35	77.93	82.13
		(58.11)	(39.44)	(76.25)	(25.81)	(39.07)	(61.09)	(45.26)
c-RNN +SGNS	4-way	66.60	51.03	79.43	48.24	50.83	80.30	82.14
		(63.90)	(47.74)	(78.34)	(29.86)	(47.91)	(66.50)	(48.23)
	3-way	66.09	49.03	79.69	47.14	49.74	80.11	82.13
		(63.20)	(45.44)	(78.89)	(27.48)	(46.71)	(67.78)	(45.23)
	2-way	63.21	46.91	78.64	45.47	45.41	78.53	82.12
		(60.35)	(41.54)	(77.54)	(25.85)	(39.42)	(62.08)	(45.14)
ConATT	4-way	65.37	47.84	75.26	45.99	47.65	78.85	82.19
		(62.50)	(44.97)	(73.43)	(27.95)	(43.54)	(62.81)	(46.70)
	3-way	64.25	47.09	76.01	45.81	46.82	78.60	82.12
		(61.63)	(42.35)	(74.87)	(26.40)	(42.42)	(62.62)	(45.13)
	2-way	62.65	44.51	75.05	44.25	45.19	78.44	82.12
		(56.79)	(38.98)	(73.37)	(23.46)	(38.38)	(62.24)	(45.12)
ConATT +SGNS	4-way	66.85	49.12	79.11	47.12	49.62	80.19	82.18
		(63.43)	(45.94)	(77.93)	(28.19)	(46.33)	(66.93)	(45.75)
	3-way	65.63	47.30	77.18	45.71	48.28	78.89	82.21
		(62.69)	(43.01)	(75.86)	(26.16)	(46.09)	(64.06)	(45.77)
	2-way	63.32	46.73	77.49	46.22	47.05	78.90	82.20
		(57.15)	(41.14)	(76.26)	(25.71)	(44.13)	(60.89)	(45.97)

Table 5.11: Results of our baselines as well as word based models on each probing task on the OntoNotes dataset.

each input using only 3 sentences preceding the target referent and 3 sentences following the target referent. This sometimes makes the dominant referent appear only once in the given discourse. In other words, there is no or a very small difference between the frequency of the prominent referent and other referents, which hinders the classifier from distinguishing them.

Second, all models work remarkably well on the task of DisStat, Syn, and LocPro. This suggests that all models have learnt information about the referential status and the grammatical role of the target referents. Since LocPro is a hybrid of DisStat and Syn, it is no surprise that our models can handle it well.

The performance of SenStat and IntRef is slightly lower than that of the above three tasks. Such a decrease in performance is understandable because learning these two specific

Model	Type	DisStat	SenStat	Syn	DistAnt	IntRef	LocPro	GloPro
c-RNN-char	4-way	64.60 (61.80)	48.76 (43.39)	76.30 (74.74)	45.75 (27.73)	47.84 (44.65)	79.11 (63.44)	81.97 (46.64)
	3-way	63.55 (61.19)	47.52 (41.52)	77.13 (76.11)	45.69 (26.43)	46.60 (41.13)	78.11 (61.70)	82.02 (45.76)
	2-way	61.32 (58.06)	46.09 (36.30)	77.95 (76.96)	45.23 (24.11)	45.71 (36.49)	77.86 (58.82)	82.11 (45.54)
c-RNN-char +SGNS	4-way	66.07 (62.90)	50.93 (46.96)	78.41 (77.18)	47.64 (30.78)	50.57 (47.81)	80.11 (66.16)	82.24 (48.20)
	3-way	64.70 (62.87)	48.24 (42.54)	79.02 (77.81)	46.27 (27.51)	47.48 (43.59)	79.35 (64.17)	82.01 (46.11)
	2-way	62.48 (60.45)	46.30 (38.24)	78.50 (77.12)	45.38 (24.27)	44.82 (37.61)	77.72 (64.09)	81.93 (46.12)
c-RNN-char +BERT	4-way	75.32 (73.96)	59.69 (57.66)	78.86 (78.15)	56.66 (37.12)	60.27 (56.90)	81.95 (69.68)	81.96 (46.60)
	3-way	74.46 (73.77)	58.41 (56.29)	80.48 (79.67)	55.91 (35.77)	59.39 (55.96)	82.71 (73.24)	81.91 (45.59)
	2-way	69.20 (68.10)	55.16 (52.08)	80.68 (79.84)	51.74 (29.71)	51.73 (52.36)	81.43 (71.30)	82.05 (45.07)
ConATT-char	4-way	65.07 (61.91)	48.40 (43.15)	70.38 (67.48)	45.95 (26.41)	48.16 (44.15)	77.89 (57.31)	82.22 (47.27)
	3-way	62.93 (59.54)	45.14 (39.55)	70.38 (68.78)	43.85 (24.47)	45.28 (39.13)	77.34 (55.27)	82.06 (45.73)
	2-way	60.55 (52.10)	44.21 (32.85)	68.33 (65.67)	43.75 (21.78)	44.36 (32.66)	76.37 (49.38)	82.07 (45.35)
ConATT-char +SGNS	4-way	66.09 (61.97)	49.43 (44.63)	75.87 (74.65)	46.04 (28.19)	49.20 (46.61)	79.50 (64.49)	82.22 (47.27)
	3-way	62.84 (58.79)	46.51 (38.78)	75.15 (74.09)	44.99 (24.66)	45.76 (38.51)	78.12 (60.19)	82.06 (45.73)
	2-way	62.65 (60.09)	46.76 (39.53)	74.17 (72.90)	44.31 (22.13)	44.84 (34.88)	77.53 (61.43)	82.07 (45.35)

Table 5.12: Results of our character based models on each probing task on the OntoNotes-c.

features requires a model to not only check whether the target referent has appeared in the pre-context, but also roughly locate them. They are clearly more challenging tasks compared to DisStat and Syn.

Besides GloPro, all models receive the worst performance on the DistAnt task. Compared to SenStat and IntRef, this task asks each model to locate the previous mention of the target referent in a more fine-grained way: checking whether the previous mention appears in the current sentence or the previous sentence. In other words, models' lower performance on DistAnt is, in part, because the task is harder than other tasks.

Comparing Word-based Models and Character-based Models. When comparing the probing results of word-based and character-based models, for most cases, we found no significant difference. The only exception is that ConATT-char and ConATT-char+SGNS learn significant less information about the syntactic positions of the target referents than ConATT and ConATT+SGNS. This partly explains why they perform worse in RFS classification.

Comparing ConATT and c-RNN. We found no significant difference between the information learnt by c-RNN and ConATT. This is not in line with the fact that, in RFS classification, ConATT slightly outperforms c-RNN. Further study is needed to explain why.

Among character-based models, c-RNN-char learns significantly more information about the syntactic position as well as slight more information about the referential status (i.e., SenStat) and the recency (i.e., IntRef) than ConATT-char, which is consistent with the winning of c-RNN-char in RFS classification.

The Effect of Pre-training. In line with the performance of RFS classification, SGNS helps every model (i.e., c-RNN, ConATT, ConATT-char, and ConATT-char+SGNS) to learn significantly more information about nearly every feature except GloPro. Moreover, using BERT can further improve the abilities of these models to acquire information about features except GloPro. We also observed that the benefit of using BERT is slightly less on learning syntactic position information than learning other features. This is probably because deciding the syntactic position of a RE relies more on tokens around it, but less on dependencies between it and other REs in the discourse, which is what BERT is good at.

Comparing Different RF Classifications. We found no significant difference between the amount of information of each feature learnt by models trained on 4-way classification and those trained on 3-way classification. However, if we train a model on merely 2-way classification (i.e., whether the target referent is realised as an over RE or a ZP), the model will learn less information about every feature except GloPro. This suggests that fine-grained classifications provide more supervision signals to make a model learn more linguistic information than coarse-grained classifications since, at least in the RFS task, fine-grained classifications are closer to human behaviours.

5.4.5 Discussion

WebNLG vs. OntoNotes

The webNLG and OntoNotes datasets are about different languages and were constructed using different methodologies. Additionally, all referents in the webNLG test set appear in its training set while only a few referents in the OntoNotes test set appear in its training set. Therefore, it is hard to use the results on the two datasets to conduct controlled comparisons between languages or between different dataset construction methodologies. Nonetheless, intuitively, compared to webNLG, the texts in OntoNotes appear to be more natural and the REs in OntoNotes are closer to the human behaviours. Regarding this intuition, we have the following observations.

First, as discussed, the difficulty of each probing task follows the following order: GloPro $\succ$ DistAnt $\succ$ {SenStat, IntRef} $\succ$ {DisStat, Syn, LocPro}, where A $\succ$ B means A is harder than B. Theoretically, if a probing task is hard, then it is hard for an RFS model to learn the corresponding task and, thus, the probing classifier has a lower performance. This happens when the OntoNotes dataset is used. For example, since DistAnt is harder than SenStat and IntRef, every model performs better on either SenStat or IntRef than on DistAnt. However, unfortunately, when using the webNLG dataset, we found no clear correlation between the difficulties of probing tasks and the performance of a probing classifier.

Second, the aim of the probing study is to understand what and how much linguistic information each model can learn and use the results to interpret the model's behaviours. Intuitively, if a model learns more linguistic information than other models, it will achieve better RFS classification performance. However, in the second study, we found that the models that learnt more information (i.e., `ConATT` and `c-RNN+BERT`) performed worse than those acquired less linguistic information. The situation is different when testing models on OntoNotes, whose texts and uses of REs are more realistic than webNLG. As discussed in the §5.4.4, in most cases, the model that performs poorly on probing tasks does not work well on RFS classification.

Third, pre-trained word embeddings and language models have been proved effective in many NLP tasks. However, in Study 2, we found that neither word embeddings (i.e., `GloVe`) nor pre-trained language models (i.e., `BERT`) help RFS classification. Such an abnormal phenomenon no longer exists in the present study because, when using the OntoNotes dataset, models that incorporate pre-trained word embeddings and language models always achieve better results compared to those that do not incorporate them.

In aggregate, the above three observations suggest that OntoNotes is more suitable for studying human behaviours on reference.

The Use of ZPs

In §5.4.2, we concluded that among all models we have tested, `c-RNN-char+BERT` performed the best. It works remarkably well on using ZPs in a pragmatically natural way. Now, we look at how well it models the use of ZPs more closely. We found no significant difference in its performance of selecting ZPs when doing 4-way classification (Figure 5.4) and 2-way classification (Figure 5.5). Let's focus on the 4-way classification in order to find out which referential form is always confused with ZPs by the model. We observed that the use of ZP was quite often confused with the use of pronouns. According to linguistic theory, both pronominalisation and pro-drop happen when the target referent is salient enough in the given discourse. Therefore, it is understandable that ZPs and pronouns are easily confused since it is hard for a model to make such a fine-grained decision of when the target referent is salient enough for pronominalisation but not salient enough for pro-drop. Additionally, the use of ZPs is also easily confused with the use of description. One possible explanation is that the OntoNotes dataset is not a balanced dataset, 50% of which are descriptions while only 13.6% of which are ZPs. Such an unbalance distribution makes the trained model the trained model to be biased towards non-descriptions (i.e., ZPs, pronouns, and proper names).

Regarding the learnt linguistic information, we found, when doing 2-way classification (i.e., deciding whether to use ZP or not), models were good at acquiring information about the syntactic position and about referential status. This is in line with the use of ZPs in OntoNotes. Specifically, we found 9827 REs out of 9897 REs that are ZPs in OntoNotes training set and 8944 REs out of 9897 REs are discourse-old. This suggests that `c-RNN-char+BERT` did well on modelling human use of ZPs rather than simply learnt artefacts in the corpus.

5.5 Summary

In this chapter, we focused on the use of REs in linguistic contexts (i.e., contexts are texts). Particularly, we were interested in the use of Zero Pronouns (ZPs) in Mandarin.

In the first place, we attempted to use the RSA framework as a tool to understand how Mandarin speakers choose between ZP and overt RE. The model we built took various factors, including the salience of referents, recency, as well as speech cost, into consideration. Benefiting from a Bayesian decision-making process, this simple statistical model worked respectably.

Building on the findings of the first study: factors like recency and referent salience do affect the use of ZPs in Mandarin.

We then decided to broaden our focus from merely the use of ZPs to all types of referential forms, such as pronoun, proper name, description, and demonstrative. To this end, we defined the task of RFS based on the webNLG corpus following a similar paradigm of the End2End REG task (Castro Ferreira, Moussallem, Kádár, et al., 2018) and tackled the task by means of neural methods. Considering that these tasks have seldom been explored using neural methods and that the webNLG corpus is in English, in the second study, we focused on RFS in English and left RFS in Mandarin to the third study. By evaluating a number of neural-based RFS models on the webNLG corpus, we surprisingly found that the simpler c-RNN model outperformed the ConATT model as well as BERT. To interpret the results and the behaviours of neural models, we conducted probing studies and introduced several probing tasks. The probing studies' results suggested that, on the one hand, these neural models learnt information that has been proved effective for the choice of RFs by theoretical linguists. On the other hand, we also found that more complex models, e.g., BERT, learnt more useful information than simpler c-RNN The reason why these complex models cannot achieve better results on webNLG than c-RNN is probably that the texts in webNLG are too short and too formal to study discourse-related aspects of RFS.

In the third study, we extended the work in Study 2 to Mandarin Chinese. We started with building an RFS/REG corpus based on the Chinese OntoNotes dataset, whose texts are more natural. This time, c-RNN still performed remarkably well. It received scores that are similar to the ConATT model. Moreover, models that incorporate pre-training language models (i.e., BERT) or word embeddings significantly defeated models that do not use them. The results of the probing study are consistent with the winning of BERT: it learnt significantly more linguistic information than other models.

So far, we have talked about one-shot REG and REG in Context. These two types of REG tasks are both about one specific function of NP: Referring. Next, we will look at another function of NP: Quantifying.

CHAPTER 6

Quantified Description Generation

Abstract. *A prominent strand of work in formal semantics investigates the ways in which human languages quantify over the elements of a set, as when we say* "All A are B", "Few A are B"*, and so on. Building on a growing body of empirical studies on this subject matter, which sheds light on the meaning and the use of quantifiers, we extend this line of work by computationally modelling how human speakers textually describe complex scenes in which quantitative relations play an important role. We first describe a series of elicitation experiments in English in which human speakers were asked to perform a linguistic task that invites the use of quantified expressions. We explain how we analysed the resulting corpus. We then extend such experiments into Mandarin Chinese. We provide an initial analysis of the use of quantified descriptions in Mandarin and compare that in English. At length, we explain how these experiments inspire to build computational models of human quantifier use that was subsequently evaluated.*

—

The publications related to this chapter are:

1. Chen, G., van Deemter, K., & Lin, C. (2019). Generating quantified descriptions of abstract visual scenes. *Proceedings of the 12th International Conference on Natural Language Generation*, 529–539. https://doi.org/10.18653/v1/W19-8667

2. Chen, G., van Deemter, K., Pagliaro, S., Smalbil, L., & Lin, C. (2019). QTUNA: A corpus for understanding how speakers use quantification. *Proceedings of the 12th International Conference on Natural Language Generation*, 124–129. https://doi.org/10.18653/v1/W19-8616

3. Chen, G., & van Deemter, K. (2021). Computational modeling of quantifier use: Elicitation experiments, models, and evaluation. *Journal Paper in Preparation*

6.1 Introduction

The aim of this chapter is to report on our work on a computational model of human speakers' use of quantified noun phrases (NPs) in descriptions of simple scenes in both English and Mandarin. Let us clarify our aim by putting our work in its historical context.

Quantified NPs are studied in different research traditions. For example, much work has been done by formal semanticists, often building on the idea that the prime function of an NP is to express quantitative relations between sets of individuals. The study of Generalised Quantifiers, as it is often called, can be understood as an attempt to understand the huge variation in quantifier patterns: not only we can say things of the form *All A are B* and *All except 2 A are B*, but also *Most A are B* and *Few A are B*, which are not expressible in First Order Predicate Logic (FOPL). Quantifiers can also play other logical roles, for instance when we say "*There are (some/few/etc.) A*", where the quantifier has only one set argument. Clearly, a speaker who describes a situation by using quantified NPs faces a large range of options, many of which express different propositions.

Building on earlier logical work (Mostowski, 1957), these issues were studied in Barwise and Cooper (1981) and further elaborated in works such as Keenan and Moss (1985) and van Benthem et al. (1986). Key questions include "What is the subset of all the theoretically possible quantifiers that natural languages can actually express, and what do these quantifiers have in common?" Connected with this is a long tradition of work on interactions between quantifiers, focusing on issues such as quantifier scope ambiguity (e.g., Kurtzman and MacDonald (1993) and Montague (1973)) and intensionality (e.g., Montague (1973)). An overview of work in these combined "logical" traditions can be found in Peters and Westerståhl (2006).

A more empirical strand of work asks how human speakers produce and comprehend quantified NPs, focusing not so much on the range of variation that fascinates formal semanticists, but more on properties of one particular quantifier (Kotek et al., 2015; Lidz et al., 2011), or differences between small sets (e.g., pairs) of quantifiers (Geurts & Nouwen, 2007; Lappin, 2000; Moxey & Sanford, 1993; Solt, 2016; Zajenkowski & Szymanik, 2013), often focusing on vague quantifiers, and focusing on quantifiers in a fixed sentence position (e.g., the position *Q* in the sentence "*Q* of the circles are round"). A smaller body of work links the two traditions of research on natural language quantifiers by investigating the relation between quantifiers' logical types and human comprehension of quantified expressions (QEs, Szymanik et al., 2016).

In recent years, many areas of human behaviour have been "simulated" using computer programs, including human memory, logical reasoning, and so on (see e.g., Sun (2008)), resulting in a methodological paradigm sometimes referred to as computational modelling. This paradigm has been extended to human language production as well (van Deemter, 2016, Section 16.1). In the spirit of this work, we want to construct a computational model of human quantifier use. Unlike *process models* which characterise the *manner* in which humans perform a given task, our models merely characterise the input-output behaviour between scenes perceived and descriptions uttered. Models of this kind are known as *product models* (see Sun (2008) and §1.1). Product models often focus on predicting how a human speaker would verbally describe a given visual scene (without claiming that the steps that our algorithms take resemble processing steps undertaken in the human mind); in other cases, they focus on producing outputs that are optimal for hearers or readers. [1]

[1] For further discussion of these perspectives, please see van Deemter, 2016, particularly Chapter 16.1.

The models presented in this work will be evaluated both in terms of the extent to which the descriptions they produce are perceived to resemble human-produced descriptions, and, especially, in terms of their utility for human readers.

We consider our models to be a valuable addition to the more ubiquitous computational models that focus on *interpreting* natural language because the former embodies an insight into *what utterance is most appropriate in a given situation*: thus, the model embodies an understanding of expressive *choice*. In a nutshell, "Why do we say what we say?", addressing both the strategic aspect of this question (i.e., What do we say?) and the tactical aspect (i.e., How do we say it?). The expressive choice is the defining challenge of the research field of NLG (see §2.1).

Given that modelling the full range of speakers' use of quantifiers is an extremely ambitious goal, we focus on simple situations, where there is only a limited range of objects to talk about, and a limited range of things to say about them, embedded in a simple communicative setting that minimises the role of such "complicating" factors as background knowledge and expectations that the speakers or hearers may have about the domain. To build a good model, one needs to know:

1. What utterances, including what quantified expressions, are likely to be uttered by a speaker in a given situation?

2. If a given quantified expression is uttered, what information does it convey?

Aspects of these questions have been addressed before. For instance, Yildirim et al. (2013) investigated speakers' use and hearers' interpretation of the quantifiers *some* or *many*. Herbelot and Vecchi (2015) looked at *no*, *all*, *most*, *some*, and *few*; I. Sorodoc et al. (2016) focused on *no*, *some*, and *all*. However, these studies only focused on a small set of quantifiers.

Building on evidence that hearers interpret quantifiers probabilistically (Degen & Tanenhaus, 2011; van Tiel, 2014; Yildirim et al., 2013), works such as Franke (2014) and Qing (2014) built probabilistic speaker models for these two quantifiers, i.e., *some* and *many*, based on Bayesian pragmatics (Frank & Goodman, 2012). [2] To the best of our knowledge, there have been no attempts to model computationally how a wider range of quantifiers are used by human speakers, let alone in a setting that allows unlimited choice of sentence patterns.

To get a first glimpse of the challenge, consider a table with four coffee cups, three of which are red while the remaining one is white. Each of the following expressions describes this scene truthfully:

(66) a. There are some red cups on the table.
b. At least three cups are red.
c. Fewer than four cups are red.
d. All the red objects are coffee cups.
e. Three of the four cups are red.

Each of these sentences could be uttered felicitously in some contexts. For example, (66-a) might make a fine answer to the question, *Is the table empty now?*. However, as a description

[2] Barr et al. (2013) elicited noun phrase patterns of the form *the square with Q dots/dashes/etc*; though this gave the authors a range of different quantifiers, the sentence pattern was once again fixed; moreover, the paper does not attempt a computational model. More recently, Pezzelle et al. (2018) formalised a cloze test based quantifier selection task, where they asked models to predict which quantifier is used in a given context.

of the scene as a whole (e.g., answering the question, *Can you tell me what's on the table?*), (66-e) would probably be more effective. An early computational investigation of the question of what quantifiers are called for in a given situation (Creaney, 1996) was based on the principle of informativity. This principle asserted that the speaker should always choose the logically *strongest* expression that holds true in a given situation. Although the idea of looking at the logical strength of an expression makes sense, Creaney's idea runs into obvious difficulties over pairs of expressions that are logically independent of each other, such as the pair of (66-b) and (66-c), where either of the two expressions conveys some information that the other one does not. Examining the evidence, we suspect that no single "principle" can tell us what makes the best description of a visual scene, and that a radically different, more empirically guided approach is called for, to inform the generation algorithm. The present work offers such an approach.

A tricky problem that our enterprise runs into is that even simple quantified expressions harbour a considerable amount of ambiguity and vagueness. The ambiguity of *most* and *many* is well-attested (Kotek et al., 2015; Lappin, 2000; Lidz et al., 2011; Solt, 2016; Zajenkowski & Szymanik, 2013), but even apparently simple quantifiers such as *all* and *some* can be far from clear: if I say *some A are B*, can I be taken to convey that there is more than one A? Do I imply that some A are *not* B? These issues are widely acknowledged (e.g., Peters and Westerståhl (2006)), but far from resolved.

Additionally, almost all the work on quantification so far are merely considered quantifiers in English. Very few of them have their eyes on other languages, such as Mandarin. The uses of quantifiers in Mandarin are very different from those in English. For example, a corpus study of vague quantifiers in English and Mandarin (A. Y. Wang & Piao, 2007) suggested that there are more vague quantifiers in Mandarin than in English and vague quantifiers are more fine-grained. For example, the quantifier *many* can be translated to "许多" (xǔduō), "多" (zhòngduō), "很多" (hěnduō), and "众多" (zhòngduō), but they are interpreted as different quantities by Mandarin speakers. In line with the topic of this thesis, in addition to quantification in English, we are also curious about how Mandarin quantifiers are used and how they are different from the use of English quantifiers.

To obtain more insight into these issues, we decided to study situations in which the sentence patterns are not given in advance, and where speakers are free to describe a visual scene in whatever way they want, using as many sentences as the speaker chooses, and using any sentence pattern that they choose. This setup has the advantage not only of allowing participants to use language in a slightly more natural way than in earlier experiments (i.e., uttering full sentences); it also had the advantage of leaving the decision of whether or not to use a quantifier to the speaker herself. Last but not least, it permits the use of quantifiers of all possible logical types.

For this purpose, we conduct a series of *elicitation* experiments, in which each participant is asked to produce descriptions of visual scenes. For example, for the scene presented in Figure 6.1, a participant in our experiment may say "*Half of the objects are blue squares, the other half are circles in both colours.*" For want of a better name, we call such a stretch of text a *Quantified Description* (QD). Such an elicitation experiment is first conducted in English and yield an English corpus that we call QTUNA. We believe that this corpus will be a source of inspiration for researchers in various research areas, including students of Generalised Quantifiers (in the intersection of Linguistics and Logic) and psychologists interested in human language production. Subsequently, we extend the QTUNA experiment to Mandarin quantifiers, and, in a similar way, we produce a Mandarin version of QTUNA, namely MQTUNA.

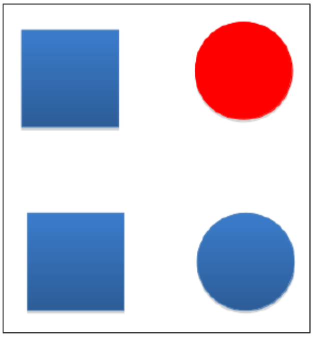


Figure 6.1: An example scene in QTUNA with 4 objects. Other scenes contain 9 and 20 objects.

Based on an analysis of QTUNA, we design two rule-based "Quantified Description Generation' (QDG)' algorithms, which mimic the types of quantified descriptions that human speakers use in any given situation; a rule-based approach is chosen because it allows us to link with the theoretical literature on quantification, and with computational models of other linguistic phenomena. We then evaluate our algorithms on QTUNA and find that these work rather well, both in terms of describing scenes in the QTUNA corpus and in terms of describing scenes of different sizes (i.e., domain sizes not occurring in the corpus). At length, we sketch possible ways to apply such algorithms on the MQTUNA corpus.

6.2 Study 1: Understanding the Use of Quantifiers in English

We introduced the TUNA experiment in §2.2.1 and used TUNA corpus to understand the use of referring expressions in Chapter 4. In order to understand how people use quantification in English, we decide to use a similar methodology while taking on-board the lessons that were learned from TUNA and adapting the method to the study of quantification. This new experiment is called the QTUNA experiment, which led to the QTUNA corpus.

6.2.1 Eliciting Quantified Descriptions

As discussed, we want to find out how a broad range of quantified NPs is used as part of a wider communicative task. Instead of showing our subjects a scene and asking them how they would explain to a hearer how many so-and-so's (e.g., circles) are red (e.g., *Many circles are red.*), we asked them to describe *the scene as a whole*. We made the scenes complex enough so that one simple QE would never suffice. Scenes came in different sizes; we use the variable N to represent the size, i.e., the number of objects in a given scene.

Each participant was presented with a series of abstract visual scenes in certain domain size. Instead of using realistic photographs, we decided to use "synthetic visual scenes" because this makes it easy to construct and modify the scenes in whatever way necessary (cf., Pezzelle and Fernández (2019) and Testoni et al. (2019)). Each scene contains N objects, each of which is either a circle or a square in either blue or red. Our instructions to participants (see Figure 6.3) asked participants to try to produce a QD that would allow a reader to *reconstruct* the scene modulo location (i.e., to reconstruct the scene except for the

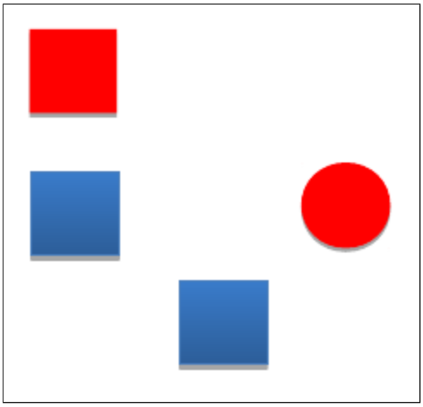


Figure 6.2: An example scene with the size of 4.

location of each object), thus ensuring a focus on quantity. Pilot experiments had taught us that without the "modulo" clause, many participants focused on location to such an extent that led to a large reduction in the number of different quantifiers used.

Materials

As shown in Figure 6.2, in each scene, there are N objects. In this study, we tested three different values of N: 4, 9, and 20. Each object has two attributes: shape and colour. Both of these two attributes have two different values, so there were 4 possible combinations of attributes (i.e., blue square, blue circle, red square and red circle). Since there were at least 4 objects in each scene, the number of attribute combinations can vary from one (i.e., all the objects are the same) to four. In our experiment, we ensured that all these variations are presented (i.e., there were scenes with 4, 3, 2, and 1 number of attribute combinations). In addition, we took care to balance shape and colour. For example, in the $N = 4$ experiment, from the set of scenes where there are 2 combinations, we selected one in which the two combinations differ in terms of colour (2 red squares and 2 blue squares), and one in which they differ in shape (2 red circles and 2 red squares).

Furthermore, instead of placing the objects in a grid (as was done in our earliest pilots), we placed objects in a more random layout as in Figure 6.2. The changes that we made on the basis of our pilots proved to be very effective in letting speakers produce descriptions that meet the requirements spelt out above, leading to a richly varied set of QDs.

Designing a workable set of instructions for participants proved to be a challenging task, so we decided to start with a series of pilot experiments before conducting the real experiment. Apart from the requirement of avoiding participants mentioning the location of each object, we also needed to discourage them from performing what we called *enumeration*, as when different kinds of objects in the scene are listed one by one. This had happened frequently in some of our pilot experiments, causing only a small range of (mostly existential) quantifiers to be used. For example, a scene like Figure 6.2 was described as follows:

(67) There are two blue squares, one red square, and one red circle.

We'd like you to describe each situation in one or more grammatically correct English sentences. (...)

1. *Based on your description, a reader will try to "reconstruct" the situation. We use the word "reconstruct" loosely here, because the only thing that matters is the different types of objects that the sheet contains. Therefore, please do not say *where* in the grid a particular object is located (e.g., "top left", "in the middle", "on the diagonal").*
2. *Each object is a circle or a square, and either red or blue. Your reader knows this.*
3. *Please do not "enumerate" the different types of objects. For example, do not say "There is a red circle, two blue circles, and ...".*
4. *Every situation contain four objects. Your reader knows this in advance, and he/she will take this information into account when interpreting your description.*

Figure 6.3: The sketch of how an instruction looks like, taking $N = 4$ as an example.

While these descriptions are as legitimate as others, they do not show us a wide range of quantifier patterns. To ensure that descriptions fulfill a concrete purpose, we also wanted to encourage descriptions that are logically "complete", by which we mean that participants should do their best to produce a description that allows readers to *reconstruct* the situation in all respects except the location of the objects; we made this exception because pilot experiments had shown that if describing the location of objects is part of what we ask speakers to do, then expressions of location (e.g., "*far away in the upper-left corner, right next to ...*") will tend to dominate their descriptive task, distracting from the quantification phenomena we wish to focus on.

In an early pilot experiment, we tried to encode the above requirements explicitly in the instructions, saying things like, "*do not use numerals when describing the situation*" and "*do not describe the location of objects*". However, this did not work well, because many subjects still used enumerations and locations. After a number of pilots, we decided to omit these explicit rules. Instead, we asked subjects to avoid enumeration as much as possible and added two examples in the instructions, explaining how one of them would allow a reader to reconstruct the situation whereas the other did not. For instance, in the $N = 4$ experiment, the two examples are:

(68) a. There are equally many circles as squares. All squares are blue. Half the circles are blue.
b. Half of the objects are blue squares.

Figure 6.3 depicts what the instruction for the $N = 4$ experiment looks like. The avoidance of enumeration may have diminished the ecological validity (Schmuckler, 2001) of our findings, but we believe that this is more than out-weighted by the increased richness of the resulting descriptions.

Design

We considered one variable in this study: the domain size N to find out how domain size impacts the use of quantifiers. We conducted three different experiments, with domain sizes (N) of 4, 9, and 20 respectively, each containing 10 different scenes. Figure 6.2 and Figure 6.4 show two examples from the $N = 4$ and the $N = 9$ experiment respectively.

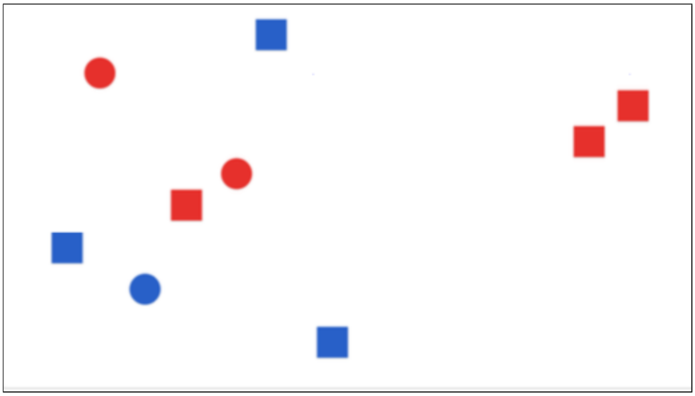


Figure 6.4: An example scene with size of 9.

Participants and Procedure

Participants were asked for a self-rating of their fluency in English (*native speaker*, *fluent*, *not fluent*). Participants who rated themselves as *not fluent* were not included in the corpus. At length, 66, 63, and 58 participants (excluding those who did not finish their experiment) completed $N = 4$, $N = 9$, and $N = 20$ experiments, respectively. Participants in each experiment were asked to read the instruction first and complete the experiment (10 scenes) in one sitting.

6.2.2 The QTUNA Corpus

Our experiments yield a corpus with 3 sub-corpora corresponding to the 3 scene sizes. Here, we introduce the corpus and its annotation.

Data Cleaning

We manually filtered out all descriptions from subjects who showed a misunderstanding of the task: (1) writing gibberish; (2) describing the scene by enumerating the objects in it; or (3) describing the scene by expressing locations (e.g., ".. at the bottom right corner of the screen"). The resulting corpus QTUNA contains 656, 380, and 378 valid descriptions for the three domain sizes, which contain 1401, 638, and 543 QEs.

Annotation

Since we want to design algorithms that mimic how people use quantifiers, we needed to annotate the descriptions in the corpus with their semantic representations. QEs in QTUNA are quite different from the referential descriptions in TUNA, where the core of the annotation for a given utterance is always simply a set of properties (e.g., $\{\texttt{shape} = \textit{square}, \texttt{colour} = \textit{red}\}$ when the expression said "The square that is red"). A new annotation scheme needs to be designed, which records quantifier patterns as well as the ways in which they were filled.

Recall that QEs express relations between sets. Following Barwise and Cooper (1981), we annotated the QEs in a form in which each n-ary quantifier is a function Q that takes a number of set terms as arguments. For example, a QE with a binary quantifier can be written as: $Q(A, B)$.

To keep the annotation task – and the later construction of the generation algorithm – manageable, we made a few simplifications. For example, we took the view that *all* and *every* in *all/every objects are red* express the same quantifier. Table 6.1 lists the top-10 most frequently used quantifiers and their frequencies in our corpus. In our annotations, $A, B, ...$ are sets. BS, BC, RS, RC, R, B, C and S stand for blue square, blue circle, red square, red circle, red object, blue object, circle and square set, respectively. O refers to the set of all objects in a situation. [3] For example, for the QE:

(69) All objects are red squares.

our annotation says $\text{All}(O, RS)$. More annotation examples can be found in Table 6.2.

In the current version of the corpus, anaphors were replaced by their corresponding antecedents. For example, the description:

(70) Most of the objects are blue. Half of them are squares.

was labeled as $\text{Most}(O, B) \wedge \text{Half}(B, S)$. Note that the pronoun *them* can refer to all the objects or only the blue objects, causing an unwanted ambiguity. When annotating such cases in the corpus, we chose a “charitable” approach: if one interpretation causes a given description to be logically complete and another causes it to be logically incomplete, then annotation sides with the former. This rule applies to all kinds of ambiguities that we encountered in our annotation work. Whenever a description contains more than one QE, our annotation records their left-to-right order.

6.2.3 Hypotheses

First of all, we were curious to see how much variation in linguistic descriptions the scenes of the QTUNA experiments would permit, and how much variation we would see between speakers. We were curious what quantifiers and quantifier patterns would be used and how these would be expressed linguistically; knowing this is also essential for the computational models that we were to develop later.

Following our pilot experiments, we were also curious to know how much information would be conveyed. How often did speakers under-specify (i.e., when they did not say enough to allow a hearer to reconstruct the scene), over-specify (i.e., saying more than necessary), and how often did they use vagueness (e.g. saying things like “a few ..")? What information would be expressed explicitly and what information would be left implicit (i.e., left to be inferred by the reader). Furthermore, we were interested in knowing whether the fact that one attribute (e.g., shape) is more easily expressed as a noun than the other (e.g., colour) has implications for its position in a quantified pattern. Given that most diagrams require several QEs for their logically complete descriptions, we were also interested in what order quantifiers tended to appear in a description. We, therefore, set out to address the following questions.

[3] There are also notations for second-order sets, which will be discussed later.

Notation	Surface Form(s)	Example Quantified Expression(s)	Frequency			
			N=4	N=9	N=20	Total
`all`	all; every; each	*All A are B. / All of the A are B.*	436	147	91	674
`most`	most	*Most A are B. / Most of the A are B.*	27	63	56	146
`more`	more	*There are more A than B.*	67	23	37	127
`half`	50%; half	*Half of A are B.*	76	12	15	103
`equal`	equivalent; equal/same number	*There are/is the same number of A and B.*	72	8	23	103
`some`	some	*There are some A. / Some A are B.*	2	30	66	98
`majority`	majority	*A majority of A are B. / The majority of A are B.*	24	23	14	61
`only`	only	*There is only A. / Only A are B.*	38	13	4	55
`half-rest`	half ..., the other half ...; half ..., the rest/remaining ...	*Half of A are B, and the other half are C.*	38	0	5	43
`more-half`	more than half	*More than half of the A are B.*	28	1	3	32

Table 6.1: Top-10 most frequently occurring quantifiers with English examples and frequencies in the three QTUNA sub-corpora.

N	Description	Meaning
4	*There are 4 squares. All objects are blue.*	$\exists_{=4}(S) \wedge \text{All}(O, B)$
9	*Most of the items are red circles, but there are a couple of blue squares.*	$\text{Most}(O, RC) \wedge \exists_{\geq 2}(BS)$
20	*All the objects in the picture are circles and majority of them is blue.*	$\text{All}(O, C) \wedge \text{Majority}(O, B)$

Table 6.2: List of example descriptions from the QTUNA corpus, with their annotations. N indicates scene size (i.e., the total number of objects in the scene).

How often do speakers manage to describe a scene completely and correctly?

We say a description is complete if the scene described is the only one (modulo location) from all possible scenes of the same size that fits the description, given the background assumptions conveyed in the instructions to participants (i.e., that there are only circles and squares, and that they can only be red or blue). Since producing a complete description requires much more work (or, sometimes, is impossible) in a larger domain, we hypothesised ($\mathcal{H}_1$) that *larger domains give rise to a smaller proportion of complete descriptions than smaller ones*.

This hypothesis is not easy to test because speakers frequently rely on inference when describing a scene, and because the meaning of quantifiers like *most*, *some*, or *few* is not cast in stone. Consider the following two examples from one of our pilots:

(71) a. Half of the objects are blue.
b. Everything is blue. Most things are square.

For the description (71-a), given that there are only two colours (i.e., blue and red), we infer that the other half are red. Or, in the description (71-b), if *most* means not just "*more than half*" but also "*not all*", then the description completely describes a scene with 3 blue squares and 1 blue circle, despite not saying this explicitly.

For similar reasons, since describing larger domains requires more work, so the task itself becomes harder, and mistakes (e.g., counting mistakes) become more likely. We, therefore, expected ($\mathcal{H}_2$) that, *in larger domains, there are more descriptions that convey incorrect information*. Information is incorrect if it is not true with respect to the scene. For example, the description "*all objects are blue*" is incorrect if it describes a situation where all objects are red.

When do people use vague quantifiers?

People frequently use vague quantifiers, such as *many*, *some*, and *most* (see e.g. Moxey and Sanford, 1993). We wanted to see how the proportion of vague quantifiers in our corpus changes with scene size. The larger a domain, the harder it is to see at a glance how many objects there are in each of its set-theoretic regions (e.g., A, B, $A \cup B$, $A \cap B$, $A - B$, $B - A$, and the domain O of objects as a whole). We, therefore, hypothesised ($\mathcal{H}_3$) that, *as the domain size (N) increases, more vague quantifiers appear*.

Are larger scenes described more elaborately?

Since there is more to describe in a large domain than in a small one, we expected ($\mathcal{H}_4$) that *participants produce longer descriptions in larger scenes*.

Left-to-right order of QEs.

Recall that most descriptions in the QTUNA corpus consist of multiple QEs. In pilot studies, speakers tended to employ two discourse structures. The first start by describing the whole scene, e.g., *"all objects are blue"*, followed by a more detailed statement, e.g., *"half of them are squares"*. A second, more frequent, discourse structure cuts the set of objects into two parts, each of which is described separately. Focusing on the second discourse structure, we hypothesised ($\mathcal{H}_5$), *The most important information tends to be stated first*. More precisely, there are two types of situations. In the most common type, a scene is described using a succession of two QEs, each of which has two set-arguments (i.e., each has the form $Q(A, B)$, which is by far the most common form). Such quantifiers can be understood as being "about" the intersection of the two arguments (i.e., about $A \cap B$). Hypothesis ($\mathcal{H}_5$) says that the first of the two QEs is usually "about" a larger set than the second. For instance, *"3/4 of A are B, 1/4 are C"* is much more frequent than *"1/4 A are C, 3/4 are B"*). The second type of situation covered by $\mathcal{H}_5$ is similar, except the second QE is left implicit. For instance, *"3/4 of A are B,"* is much more frequent than *"1/4 A are C"*.

Are there any differences between the use of colour and shape?

Given the well-documented primacy of colour over shape in referring expression (Pechmann, 1989; van Deemter, Gatt, Sluis, et al., 2012), it seemed plausible to us that colour and shape play different roles in QEs as well. [4] Based on our pilot experiments, in which colour was often realised as an adjective, we hypothesised that ($\mathcal{H}_6$), in k-ary ($k > 1$) QEs, *shape occurs more often in the former argument places (i.e., the A position in the QE:* Q of A are *B) and colour in later positions*. For example, we expected to see more expressions like *"all circles are red"* than ones like *"all blue objects are circular"*.

6.2.4 Hypothesis testing

We tested the hypotheses introduced in §6.2.3. In order to test the first hypothesis $\mathcal{H}_1$, we annotated each description in QTUNA for being complete or not. [5] Considering the above-mentioned challenges, when annotating the corpus, instead of relying on the formalisation of the meaning of quantifiers, it was annotated, for each description, whether it is logically complete or not. Annotating for completeness was about whether the situation can be fully reconstructed based on a description. In accordance with our rule of charitable interpretation, if a description was ambiguous, the final category (complete or incomplete) was decided using the "best" interpretation of the description. Completeness annotation was performed by two annotators. Where disagreements occurred, the annotators discussed their initial judgement and made a final decision together. In this way, we found 46, 205 and 355 incomplete descriptions from 656, 380 and 378 descriptions of the three sub-corpus respectively (Table 6.3). As one can see, incompleteness frequencies appear to grow with

[4] As is usual, we take k-ary QEs to be ones whose quantifier relates k sets

[5] See, for example, Coventry et al. (2010) for problems assessing the meaning of *most*.

	$N = 4$	$N = 9$	$N = 20$
Quantified Description	656	380	378
Quantified Expression	1401	638	543
Complete Description	610	175	23
Incomplete Description	46	205	355
Vague Quantifier	57	201	234
Wrong Description	7	12	47
Larger Part First	123	145	99
Smaller Part First	72	54	10

Table 6.3: The number of quantified descriptions, quantified expressions, incomplete descriptions, vague quantifiers, and wrong descriptions in each sub-corpus of QTUNA.

scene size. Fewer than 1/10 descriptions in $N = 20$ sub-corpus are complete, most of which come from scenes with only one property combination (i.e., all the objects in a scene look-alike) or two property combinations. We conduct a binary logistic regression analysis (setting completeness as the outcome variable and domain size as the predictor) on the annotated data. The result confirms our hypothesis $\mathcal{H}_1$ that there are less complete descriptions in larger domain ($p < .001$, adjusted $p < .001$). [6]

For the second hypothesis $\mathcal{H}_2$, we annotated, for each description, whether it is correct or incorrect (a "wrong description"). If the property was debatable, it was considered to be correct. Such cases often occur with colour terms, for example, the colour of a red circle was sometimes described as *orange*; since only red and blue were permitted, there was no confusion possible so we considered such descriptions to be correct. We found 7, 12 and 47 wrong descriptions for the three scene sizes. The high proportion of correctness (minimally 87.57% for $N = 20$) indicates that most of our participants understood the instructions, yet it suggests an overall association between the domain size and the error frequency, which is confirmed by a binary logistic regression analysis (setting correctness as the outcome variable and domain size as the predictor; $p < .001$, adjusted $p < .001$).

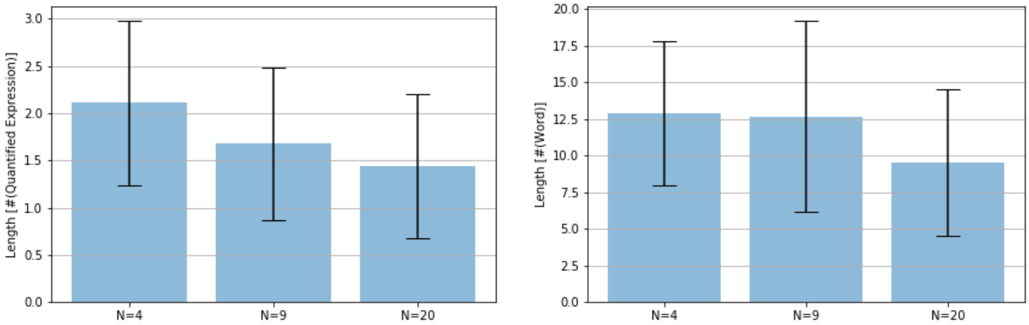
$\mathcal{H}_3$ asserts that vague quantifiers appear more frequently in larger scenes. In accordance with common practice (e.g., Kenney and Smith (1996)), we understand a quantifier to be vague if it permits so-called borderline cases (i.e., cases in which it is unclear whether the QE is true or false). We counted the number of QEs that use vague quantifiers (e.g., *many* and *few*). [7] The number of QEs was compared with the total number of QEs, as listed in Table 6.3. The trend of more vague quantifiers in larger domains (i.e., $\mathcal{H}_3$) was confirmed ($p < .001$, adjusted $p < .001$) as by a binary logistic regression analysis. [8]

To test $\mathcal{H}_4$, we also calculated the length of each description, as defined by both the number of QEs (Figure 6.5(a)) and the number of words (Figure 6.5(b)) in the description. The results show the opposite of what we expected, that is, the length of descriptions

[6] Adjusted p is the p-value obtained by applying Bonferroni correction, where the p-value is multiplied by 6 as there are 6 hypotheses.

[7] A quantifier like *most* was always counted as vague, despite the fact that it might acquire a precise meaning when N=4 (because when we say that *most* of a set of four 4 A are B, we can arguably only mean that *three* of the four A are B.

[8] If we had decided to count *most* as a precise (i.e., non-vague) quantifier when used in the N=4 domain, then this would have further strengthened the support for $\mathcal{H}_3$.



(a) (b)

Figure 6.5: The length of descriptions with respect to the domain size by means of (a) the number of QEs; (b) the number of words.

decreased. We believe that a plausible explanation lies in the fact that speakers produced fewer complete descriptions in larger domains, as in $\mathcal{H}_1$: after all, when a task is made more complicated (in this case, because we move from smaller to larger scenes), the effect can be that participants try less hard to perform the task perfectly (i.e., they lower their standards).

Regarding the last two hypotheses, we first counted the number of descriptions that describe the larger part of a scene first (i.e., descriptions like "*3/4 of A are B, 1/4 are C*" or "*3/4 of A are B*"), and those that describe the smaller part first (i.e., descriptions like "*1/4 of A are B, 3/4 are C*" or "*1/4 of A are B*"), the numbers for each N being shown in Table 6.3. This confirmed ($\chi^2(2, N = 503) = 27.29, p < .001$, adjusted $p < .001$) the hypothesis that the most important information tends to be stated first ($\mathcal{H}_5$) by a Chi-squared test. In a similar way, we then counted the number of descriptions that place shape in the first argument while placing colour in the later argument (i.e., descriptions like "*all circles are blue*"), and the number of descriptions that order the two attributes the other way around (e.g., *all blue objects are circular*). As for shape, 489 descriptions used it in the first argument place and 121 in the second; for colour, those two numbers are 112 and 514 respectively. Consequently, a Chi-square test confirms this hypothesis $\mathcal{H}_6$ ($\chi^2(1, N = 1236) = 479.59, p < .001$, adjusted $p < .001$).

6.2.5 Post-hoc Observations regarding the QTUNA corpus

We also made a number of post-hoc observations. These should be distinguished from the earlier-listed hypotheses, which were formulated before we saw the data of the experiment.

3-ary Quantifiers

Besides binary quantifiers, we found a substantial number of 3-ary quantifiers. One class of examples is *half ..., the other half ..., one ..., the rest ..., half ..., the rest ...* and so on. Note that an expression such as (72-a) should not be confused with (72-b).

(72) a. Half of A are B, the other half are C
b. Half of A are B and half of A are C

In (72-b), the sets A and B can have a non-empty intersection, but (72-a) means that $\frac{1}{2}$ of As are B, and $(A - B) \subseteq C$.

Higher-Order Quantifiers

We found a remarkable number of "higher-order" quantifiers, where quantification is not over objects but over sets of objects. For example, the word "both" in the following example quantifies over the set of colours:

(73) Half of the objects are in both colours.

Frequent examples of higher-order quantification can be found in descriptions of a situation in $N = 4$ sub-corpus where all the objects are different. Many subjects used the descriptions equivalent to (74).

(74) All possible objects are shown.

This description quantifies over elements of the Cartesian product of the colour set and the shape set (i.e., $\text{Some}(O, BS) \wedge \text{Some}(O, BC) \wedge \text{Some}(O, RS) \wedge \text{Some}(O, RC)$).

Descriptions that Rely on Implicit Information.

This section describes a set of experiments, each of which assumes a small and precisely defined domain of possibilities (e.g. scenes of N objects with only two attributes (shape and colour), each of which has only two possible values). In these cases, one can frequently infer more than say explicitly by considering the complementary relationship of two values of one attribute. For example, if a subject says:

(75) Half of the objects are blue.

The reader is entitled to infer that the other half of the objects are red. Descriptions of this kind were marked as logically complete descriptions despite the appearance of incompleteness.

6.2.6 Discussion

In this study, we investigated the use of quantifiers in English by conducting an elicitation experiment and analysed the resulting corpus, namely QTUNA. We analysed the completeness, the correctness, as well as the vagueness of the human-produced QDs and found that the domain size is a major factor that influences the use of quantifiers.

With the QTUNA experiment, two questions are yet to be answered:

1. What descriptions will be produced if the domain size is further increased? One might expect that, similar to the findings of this study, the participants would produce even more vague quantifiers, more incompleteness, etc. More details of this open question will be discussed in Chapter 8;

2. What types of QEs are produced in other languages? The use of quantifiers in Mandarin is studied in the §6.3.

6.3 Study 2: Understanding the Use of Quantifiers in Mandarin

In this study, we investigate the use of quantifiers in Mandarin following a similar paradigm in study 1. Regarding this subject matter, we are interested in how Mandarin speakers use quantifiers. On the one hand, one previous corpus study on a machine translation corpus suggested that there are much more variations in QEs in Chinese than in English (A. Y. Wang & Piao, 2007). For example, they found that in this parallel corpus, the quantifier *many* was translated to "许多" (xǔduō), "多" (duō), "很多" (hěnduō), and "众多" (zhòngduō). Although all of them were translated as *many*, for Mandarin speakers, they represent different quantities. The quantity of "很多" is interpreted to be more than that of "多", and "多" more than "许多", and so on. According to A. Y. Wang and Piao (2007), they literally mean *some many*, *many*, *very many*, and *many many*, respectively. Building on this finding, we are curious how much variation will result if we do a QTUNA like experiment on Mandarin speakers.

On the other hand, as discussed in the §3.1, we expected Mandarin Chinese to prefer brevity but scarifie clarity. Therefore, in this study, we are curious whether the findings in the first study still stand in hold true for Mandarin QEs as well. Additionally, we are also curious how would Mandarin speakers use quantifiers differently from English speakers. For example, the coolness hypothesis would make it plausible for us to expect Mandarin speakers will use QEs in a less complete way and use more vague quantifiers than English speakers. We will elaborate these hypotheses in §6.3.5.

6.3.1 Eliciting Quantified Descriptions in Mandarin

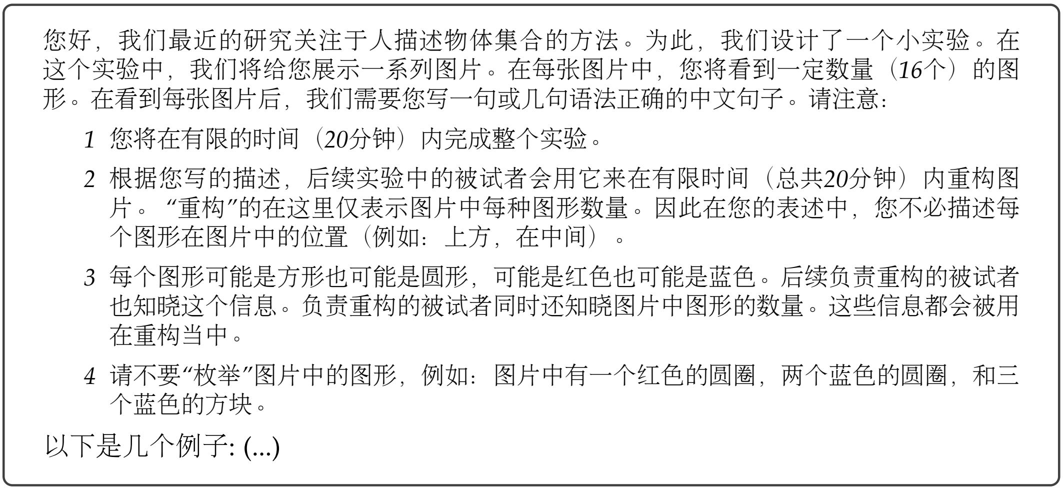
In MQTUNA, we followed the same methodology as the QTUNA experiment. Roughly speaking, we re-used some of the scenes of the QTUNA experiment, inherited the same experimental design as QTUNA, and adopted the instructions to Mandarin from QTUNA.

Materials

To prepare materials for the MQTUNA experiment, we sampled scenes from QTUNA following two steps. In the first step, we eliminated all scenes all of whose objects have the same property. In other words, we removed all the scenes that can be described by simply using one QE in the form of "*all objects are ...*". Subsequently, for each domain size (i.e., 4, 9, or 20), we randomly sampled 5 scenes from QTUNA. Unlike QTUNA, where experiments with different domains were conducted separately, we planned to deliver a single MQTUNA experiment which includes all three domain sizes. In other to minimise the impact of presentation order of scenes with different domain sizes, we randomised the order of the sampled scenes. [9] In the second step, in order to make each subject familiar with the experiment, we added a practice situation that uses a $N = 4$ scene whose objects are the same. We put this practice situation at the very beginning of the experiment. When analysing the results, we did not take it into account.

For the instruction, we simply translated the instruction of QTUNA and adapted some to make the instruction in fluent Mandarin.

[9] That is to get rid of the case where, for example, a subject see all $N = 4$ scenes before s/he see $N = 20$ scenes.



您好，我们最近的研究关注于人描述物体集合的方法。为此，我们设计了一个小实验。在这个实验中，我们将给您展示一系列图片。在每张图片中，您将看到一定数量（16个）的图形。在看到每张图片后，我们需要您写一句或几句语法正确的中文句子。请注意：

1 您将在有限的时间（20分钟）内完成整个实验。

2 根据您写的描述，后续实验中的被试者会用它来在有限时间（总共20分钟）内重构图片。“重构”的在这里仅表示图片中每种图形数量。因此在您的表述中，您不必描述每个图形在图片中的位置（例如：上方，在中间）。

3 每个图形可能是方形也可能是圆形，可能是红色也可能是蓝色。后续负责重构的被试者也知晓这个信息。负责重构的被试者同时还知晓图片中图形的数量。这些信息都会被用在重构当中。

4 请不要“枚举”图片中的图形，例如：图片中有一个红色的圆圈，两个蓝色的圆圈，和三个蓝色的方块。

以下是几个例子: (...)

Figure 6.6: The sketch of the instruction of MQTUNA.

Design, Participants, and Procedure

Same as QTUNA, we only considered one variable: the domain size, and tested $N = 4$, $N = 9$, and $N = 20$. Data from 32 participants were collected. All of our participants are Mandarin native speakers. Participants were asked to read the instruction first and complete the experiment (16 situations) in one sitting.

6.3.2 The MQTUNA Corpus

We cleaned the dataset in the same way as QTUNA. This resulted in 465 valid QDs, in which there are 155 QDs for each domain size, and 1175 QEs, in which there are 383, 386, and 406 QEs for $N = 4$, $N = 9$, and $N = 20$ sub-corpus, respectively. In table 6.4, there are some example QDs from the QTUNA dataset.

We annotated the MQTUNA in the same way as QTUNA (see §6.2.2 for more details). In line with the annotation of MQTUNA, for simplification, we viewed quantifiers that have the same meaning (e.g., “所有” (all) and “全部” (all)) as the same quantifier and used a single notation to represent them. Table 6.5 enumerate the top-10 quantifiers and their usage in MQTUNA. Note that both “大多数” (dàduōshù) and “多数” (duōshù) are literally translated as *most*, but, “大多数” is often interpreted to be more than “多数”. Additionally, although “少数” (shǎoshù) is translated as *minority*, it is always viewed as a vague quantifier and, thus, it is, in some contexts, translated as *a few*.

As for the quantifier use, same with QTUNA (Table 6.1), quantifier “所有” (suǒyǒu; *all*) and “一半” (yíbàn; *half*) are two of the most frequent quantifiers. We also notice that, in the top-10 frequent quantifiers of MQTUNA, 4 of them are vague quantifiers, including “绝大多数” (juédàduōshù; *overwhelming majority*), “大多数” (dàduōshù; *most*), “多数” (duōshù; *most*), “少数” (shǎoshù; *minority*).

We also observed that most QEs are realised by means of the following three forms.

(76) a. Q A 是 B 。

N	Description
4	所有都是蓝色，方块是圆形三倍。 *All objects are blue. The number of squares is triple that of circles.*
4	所有图形都是蓝色的。但是只有一个圆。 *All objects are blue but there is only one circle.*
4	圆圈与方块数量相同。一半圆圈是红色。 *There are equal numbers of circles and squares. Half of the circles are red.*
9	所有的圆圈是红色的。方块都是蓝色的。方块的数量少于圆圈的数量。 *All circles are red. All squares are blue. There are fewer squares than circles.*
9	方块是圆圈数量的三倍。全部为红色。 *The number of squares is triple that of circles. All of them are red.*
9	所有图形都是蓝色。大多数是方块。只有少数是圆形。 *All objects are blue. Most of them are squares. Only a minority of them are circles.*
20	图中红色蓝色方块圆球数量相差不大。 *There is no big difference between the numbers of all combinations.*
20	一半红色，一半蓝色。红色方块比蓝色方块多。蓝色圆圈多于红色圆圈。 *Half of the objects are red, the other half of them are blue. There are more red squares than blue squares and more blue circles than red circles.*
20	方块和圆圈各占一半。红色多于蓝色。 *Half of the objects are squares, the other half of them are circles. There are more red objects than blue objects.*

Table 6.4: List of example descriptions from the MQTUNA corpus, with their annotations. *N* indicates domain size.

Q A shì B
Q A are B.
b. A 中 Q 是 B 。
Q zhōng Q shì B
(lit.) In A, Q are B.
c. B 在 A 中 占 Q 。
B zài A zhōng zhàn Q
B takes up Q of A.

For example, suppose the quantifier is “大多数” and the target QE is 大多数(S, R)，we could say any of the following:

(77) a. 大部分 方块 是 红色的 。
dàbùfèn fāngkuài shì hóngsède
Most squares are red.
b. 方块 中 大部分 是 红色的 。
fāngkuài zhōng dàbùfèn shì hóngsède
Among squares, most are red.
c. 红色的 在 方块 中 占 大部分 。

Notation	English	Surface Form(s)	Example Quantified Expression(s)	Frequency			
				N=4	N=9	N=20	Total
所有	all	(所有)...都..., (全部)...都...	(全部)A都是B / A中(全部)都是B *All A are B*	100	127	53	280
一半	half	一半，百分之五十	一半A是B / A中的一半是B / B在A中占一半 *Half A are B*	101	19	28	148
相同	equal	数量相同, 一样多, 个数一样	A与B数量相同 *There is an equally number of A and B*	59	11	29	99
绝大多数	overwhelming majority	绝大部分, 绝大多数	A中绝大多数是B / 绝大多数A是B / B在A中占绝大多数 *Almost A are B*	7	50	37	94
各半	half ... rest ...	各半, 一半一半, 一半...另一半...	BC在A中各半 / A中BC各半 / 一半的A是B, 另一半是C *Half of A are B, the other half of A are C*	60	6	24	90
比-多	more	比...多	A比B多	10	28	48	96
大多数	most	大多数, 大部分	A中大多数是B / 大多数A是B / B在A中占大多数 *Most A are B*	7	35	33	75
少数	minority	少数, 少部分	A中少数是B / 少数A是B / B在A中占少数 *Minority of A are B*	5	31	24	60
有	exist	有, 存在	图片中有A (*There are A in the scene*)	4	12	18	34
多数	most	多数	A中多数是B / 多数A是B / B在A中占多数 *Most A are B*	5	4	20	29

Table 6.5: Top-10 most frequently occurring quantifiers with their English translation and Mandarin examples as well as frequencies in the three MQTUNA sub-corpora.

	$N = 4$	$N = 9$	$N = 20$
Quantified Description	155	155	155
Quantified Expression	383	386	406
Complete Description	122	19	5
Incomplete Description	33	136	150
Vague Quantifier	25	143	184
Wrong Description	7	14	30
Larger Part First	30	129	116
Smaller Part First	6	9	11

Table 6.6: The number of quantified descriptions, quantified expressions, incomplete descriptions, vague quantifiers, and wrong descriptions in each sub-corpus of MQTUNA.

hóngsède zài fāngkuài zhōng zhàn dàbùfèn
Red squares take up most of squares.

We also note that "所有" cannot be realised using the pattern (76-c). This is because, in Mandarin, "所有" can only be used as a modifier and, thus, it cannot stay alone after "占".

6.3.3 Hypotheses and Hypothesis Testing

We believed that our findings from analysing QTUNA about quantification are universal across languages. We, therefore, expected that all findings in 6.2.4 still stand in MQTUNA. To confirm this, we did all the counting work in 6.2.4 again on MQTUNA following the same counting principles. Table 6.6 reports the counts of different types of descriptions, expressions, and quantifiers. In what follows, we summarise the results:

1. $\mathcal{H}_1$ hypothesises that there are more incomplete QDs in larger domains. We identified 33, 136, and 150 incomplete descriptions from the three sub-corpora, respectively. A binary logistic regression test confirmed $\mathcal{H}_1$ ($p < .0001$, adjusted $p < .0001$);

2. $\mathcal{H}_2$ assumes that there are more incorrect QDs in larger domains. We counted the number of wrong descriptions and observed more wrong descriptions in larger domains. Wrong descriptions are likely to appear when the domain size is large and two groups of objects in a scene have a similar amount. For example, when there are 11 red objects and 9 blue objects in the scene, the most frequent wrong description is:

   (78) 一半 的 图形 是 红色的 。
   yíbàn de túxíng shì hóngsè de
   Half of the objects are red.

   We found 7, 14, and 30 wrong descriptions in the three sub-corpora, respectively. A binary logistic regression test validated this hypothesis ($p < .0001$, adjusted $p < .0001$);

3. $\mathcal{H}_3$ is about the use of vague quantifiers in MQTUNA. It hypothesises that there are more vague quantifiers in larger domains. We counted the number of QEs that contains vague quantifiers and we found 57, 201, and 234 vague quantifiers from 383, 386, and 406 QEs of the three sub-corpora, respectively. This shows a clearly increasing trend in the use of vague quantifiers with respect to the rise of domain size and such a trend was confirmed significant also by a binary logistic regression test ($p < .0001$, adjusted $p < .0001$);

4. $\mathcal{H}_4$ hypothesises that QDs in larger domains are also longer. We tested whether Mandarin speakers tend to speak more QEs for scenes with larger domain sizes by looking at the number of QEs in each QDs. Slightly different from QTUNA, QDs in larger domains in MQTUNA on average consist of more QEs than those in smaller domains. We computed the Pearson correlation between the domain size and the QD length. Unfortunately, the difference was not significant ($p = 0.1025$, adjusted $p = 0.615$) and, therefore, the $\mathcal{H}_4$ was also rejected in MQTUNA;

5. Regarding $\mathcal{H}_5$, same as QTUNA, a Chi-squared test validated that when describing a scene, a participant is more likely to describe the larger part first ($\chi^2(2, N = 502) = 315.55, p < .0001$, adjusted $p < .0001$);

6. $\mathcal{H}_6$ asserts that if a QE is in the form of Q(A, B), where one of A and B is describing the shape and the other is describing the colour, then A is more likely to be the one about shape. Among the whole MQTUNA corpus, there are 320 QDs of that form, 295 of which describe shape in A position while only 25 of which describe colour in A position. A Chi-square test confirm this hypothesis ($\chi^2(2, N = 640) = 455.63, p < .0001$, adjusted $p < .0001$);

In a nutshell, all except the fourth hypothesis were confirmed. For this hypothesis, though Mandarin speakers used slightly longer descriptions in larger domains, the difference was not significant.

6.3.4 Post-hoc Observations

In addition to the post-hoc observations of QTUNA, such as the 3-ary quantifiers, higher-order quantifiers, and complementary which are still standing in QTUNA, we also made the following additional post-hoc observations.

A-drop

Let us coin a new phrase, analogous to the well-known term “Pro-drop”. For a QE Q(A,B), we call the phenomenon of not explicitly mentioning the phrase in the position A as A-drop. For example, the QE (79) drops the phrase “图形” (túxīng; *object*) which is understood to be in position A.

(79) 大部分 是 红色的 。
dàbùfèn shì hóngsède
(lit.) Most are red.

Similar to pro-drop (see §3.1 for more details), A-drop is frequent in Mandarin QEs. In MQTUNA, we found that 304 out of 1175 QEs (approximately 25.87%) omit A. This happens in two situations:

1. Parallelism: if a phrase referring to the same set appears in the previous sentence and is also in position A, then it is omissible. It is worth noting that, as introduced in §6.3.2, A does not necessarily serve as the subjects of QEs and phrases in the A position are omissible regardless of which syntactic positions they are in. For example, in addition to (79), one could also say "红色的 占 大部分" (hóngsède zhàn dàbùfèn; *red objects take up most*);

2. The phrase in the position A refers to "object", for example, the QE (79).

Plurality

Building on the fact that Mandarin can express plurality implicitly (see §3.1), a QE is sometimes less informative than an English QE, which probably makes the whole QD incomplete. For example, consider the following Mandarin QE for $N = 4$:

(80) 图片 里 有 红色方块 和 蓝色圆圈 。
túpiàn lǐ yǒu hóngsèfāngkuài hé lánsèyuánquān
There are red squares and blue circles in this picture.

There could be either one red square or multiple red squares given this QE. However, in English, since both phrases are in their plural forms and given the domain size is 4, we can easily know that there are 2 red squares and 2 blue circles.

6.3.5 Initial Comparison between the Quantifier Use in English and Mandarin

In Chapter 1 and §3.1, we linked the idea of coolness (C.-T. J. Huang, 1984) to the trade-off between brevity and clarity in NLG and argued that coolness might suggest that Mandarin prefers brevity over clarity. This said, one can expect that QDs in Mandarin are shorter but are less clear than QDs in English. Based on this idea, we came up with the following 4 research questions. In the analysis below, we use the situations in QTUNA that use the same scenes as those in MQTUNA. Also, note that we call the current analysis an "initial" analysis because it is not a carefully designed language comparison study.

First, the major expected characteristic of being a brevity-favouring language is that the length of Mandarin QDs should generally be shorter than that of English. Concretely, we expected the QDs in MQTUNA to be shorter than QDs in QTUNA in terms of the number of QEs per QD. We computed the average length of QDs in the two corpora and printed them in Figure 6.7. As we can see, the trends do not go in the predicted direction. The QDs in MQTUNA are generally longer than those in QTUNA. Additionally, as mentioned, in QTUNA, the length is decreasing with respect to the increase of the domain size, but, in MQTUNA, there is no such trend. It needs more controlled experiments to say the last words on this phenomenon and to explain why.

Second, linking clarity to the concept of completeness in QDs, we expected there are more incomplete QDs in MQTUNA than in QTUNA. We counted the complete and incomplete QDs in the two corpora, respectively, and reported the results in Table 6.7. In total, 379 out of 710 QDs are complete in QTUNA while merely 146 out of 465 QDs in

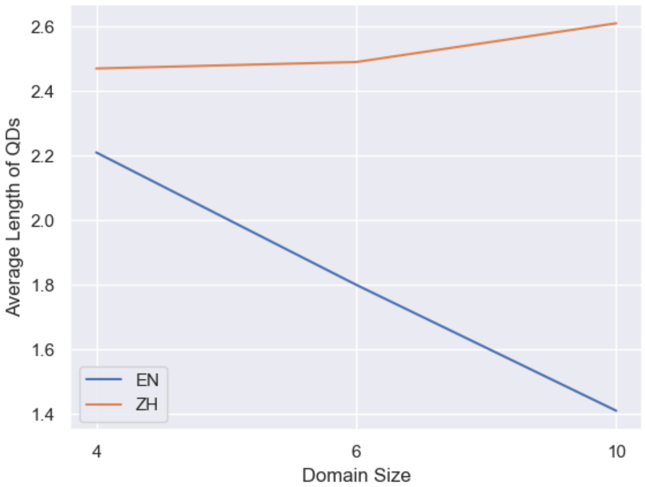


Figure 6.7: Average QD length in QTUNA (blue) and MQTUNA (orange).

	QTUNA		MQTUNA		
N	Complete	Incomplete	Complete	Incomplete	p-value
4	298	32	122	33	$p < .001$
9	77	113	19	136	$p < .0001$
20	4	186	5	155	$p = .5$
all	379	331	146	319	$p < .0001$

Table 6.7: The number of complete and incomplete QEs in QTUNA and MQTUNA, respectively. N means domain size.

MQTUNA are complete. This confirms the hypothesis with a Chi-squared test ($\chi^2(2, N = 1175) = 54.93, p < .0001$, adjusted $p < .0001$). Additionally, considering that QDs in MQTUNA are generally longer than those in QTUNA, Mandarin speakers tend to produce longer descriptions but are less complete. Interestingly, if we look into more detail, we found that such a difference exists only in situations whose domain size is 4 or 9. When the domain size is large enough (i.e., 20), then the difference does no longer exist. In other words, both English and Mandarin speakers find it hard to come up with a logically complete description if the domain size is large.

Third, we are curious about the differences between the use of vague quantifiers in English and in Chinese. For a similar reason (i.e., Mandarin favours brevity and breaches clarity), we expect Mandarin speakers use more vague quantifier English speakers. In English, we identified 222 vague quantifiers from 1342 QEs (approximately 16.54%). In Mandarin, we found 352 vague quantifiers from 1175 QEs (approximately, 29.96%). A Chi-squared test confirms the hypothesis ($\chi^2(2, N = 2517) = 64.04, p < .0001$, adjusted $p < .0001$).

Last but not least, since QTUNA and MQTUNA were collected following similar paradigms, we believed there should not exist a significant difference in terms of speakers' production of incorrect descriptions. In other words, we expected that there is no difference between the proportion of incorrect QDs in QTUNA and MQTUNA. However, this hypothesis was rejected because we identified 39 incorrect QDs out of 710 QDs in QTUNA and identified 51 incorrect QDs out of 465 QDs in MQTUNA ($\chi^2(2, N = 1175) = 11.91, p < .001$, adjusted $p < .01$). One possible explanation is that the participants of MQTUNA tended to write longer descriptions than those of QTUNA (which has been approved above) and it is more likely to write something wrong if one writes longer texts.

6.3.6 Discussion

In this study, we conducted an elicitation experiment about quantification in Mandarin, namely MQTUNA, and analysed the resulting corpus.

By comparing the findings of MQTUNA with the corresponding findings of QTUNA in the first study, we assessed the coolness hypothesis in the context of quantification. As for brevity, on the one hand, we found no evidence suggesting that Mandarin speakers produce shorter QDs than English speakers. Moreover, what we found went in the opposite direction, i.e., English speakers produced shorter QDs than Mandarin speakers. Here, we have to note that we have not said the last word on this research question as the current study does not make a direct comparison between English and Mandarin. On the other hand, we identified several phenomena in relation to "coolness" and in line with the coolness hypothesis. For example, A-drop frequently happens in MQTUNA, and plurality is always expressed implicitly. As for clarity, it indeed was breached and we observed significantly more incomplete QDs as well as significantly more vague quantifiers in MQTUNA than in QTUNA.

6.4 Study 3: Computational Modelling of Quantification

In this study, we are aiming to construct a generation algorithm that is able to perform the same task as was given to the participants in the QTUNA and the MQTUNA experiments, which came in three scene sizes ($N = 4$, $N = 9$, and $N = 20$). However, we do not want the algorithm to be limited to these scene sizes: we want them to perform well on scenes of any reasonable size. We do not target scenes sized lower than 4 because we suspected that these involve quantification to a much smaller extent (see the literature on "subitizable" sets, from Kaufman et al. (1949) on-wards); scenes in which there are more objects than can be counted in a few seconds were similarly beyond the scope of this study because they are likely to involve guesswork and estimation on the part of the hearer, which is not the focus of the present work (although future work may extend in that direction). Consequently, neither very small nor very large scenes were tested in our evaluation experiments.

Even though our main aim is to model human behaviour since our speakers (in previous studies) were asked to produce QDs that are correct (i.e., truthful) and complete (i.e., giving as much information as can reasonably be expected), it seemed reasonable to design our algorithms with these objectives in mind. In this study, we introduce algorithms that endeavour to meet the above requirements as well as they can; later on, we will evaluate our algorithms based on these two criteria and compare their performance with that of our participants.

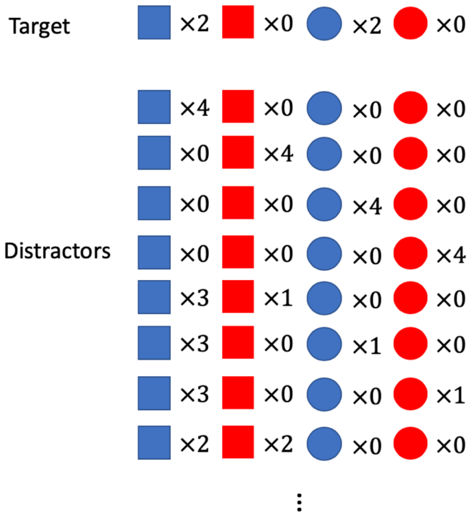


Figure 6.8: The target scene as one among many possible scenes ($N = 4$).

You may recall that participants were told that, “*based on your description, a reader will try to “reconstruct” the situation*”. We know that this was not always easy, particularly for larger and more varied domain sizes; consequently, our speakers did not always produce correct and complete descriptions. Therefore, we decided to also evaluate our algorithms using an additional criterion that asks explicitly how “human-like” the descriptions generated by it are. Below, we introduce the fundamental idea behind our algorithm, and we sketch a pipeline architecture for producing QDs, all the way from a scene to an actual sentence. We then propose two QDG algorithms. So far, these algorithms have only been evaluated in English. We believe they are easy to be generalised to Mandarin QD.

6.4.1 The idea behind our algorithms

The basic idea is to regard the production of a QD as an attempt to identify, within the set of all possible scenes, what specific scene we are looking at. In other words, the idea is to view the task of our participants as – very broadly – analogous to the task of referring to an object.

Let us unpack this idea a little further, deliberately opting to use terminology familiar from work on REs, in order to emphasise what the two problems have in common (despite the differences between the two, which will be discussed below). Let us call the scene that the algorithm aims to describe the *target scene*. Given a certain scene size and domain assumptions provided to our participants (i.e., what colours and shapes are permitted), the algorithm can compute how many possible scenes of this size there are. For example (as shown in Figure 6.8), if the target scene ($N = 4$) has two blue squares and two blue circles, then possible “distractor” scenes include a scene with 4 blue squares, a scene with 4 red squares, and so on. Generation algorithms operate by accumulating QEs that are true of the target scene but false of at least one distractor. For instance, if one says “*all objects are blue*”, then this is true of the target scene but it will “remove" many other scenes, including the scene consisting of 4 red squares, for instance. The algorithm repeats this step until a

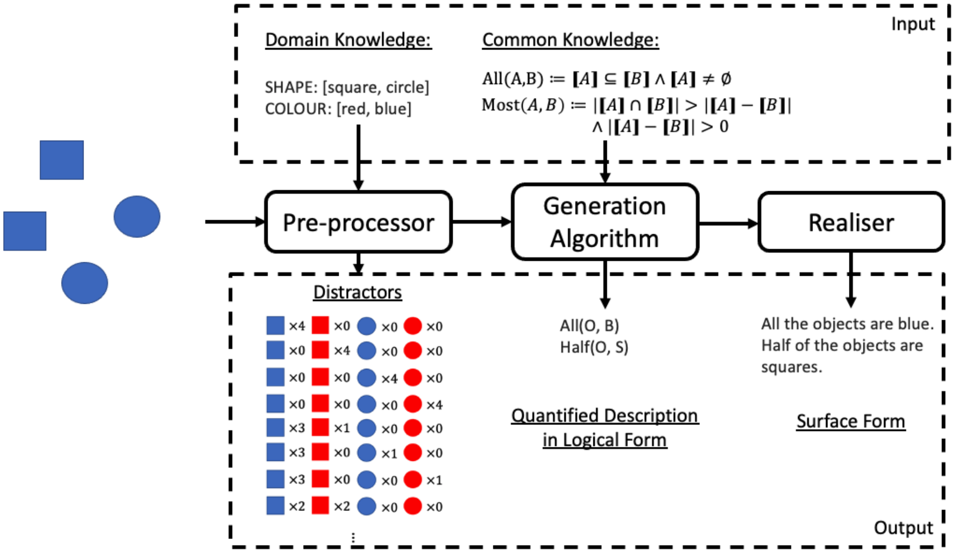


Figure 6.9: The pipeline of how we generate a QD based on a given scene, consisting of three steps: pre-processing, generating a QD in logic form, surface realisation.

stopping criterion is met. In simple situations, a reasonable stopping criterion is that *all* distractor scenes have been removed, though as we shall see, this idea cannot always be upheld. Let us see how these ideas can be made precise.

6.4.2 Generation Pipeline

As introduced in §2.1, NLG systems often use a pipeline architecture in which the content of the generated text is determined before its linguistic form. We construct our Quantified Description Generation (QDG) pipeline in line with this setup: the QDG algorithms introduced in this section are responsible for determining the content of the description (i.e., essentially a logical form), which is then turned into its linguistic form, which is a process known as *Linguistic Realisation* or *Surface Realisation* in NLG. In addition, in order to extract the required information from the given scenes, an extra pre-processing module was inserted at the beginning of the pipeline.

Concretely, the generation pipeline consists of 3 components: a pre-processor, a QD generator (which runs one of the algorithms below), and a surface realiser. As shown in Figure 6.9, given a target scene s, with its domain knowledge $\mathcal{K}_d$ (which records, among other things, how many objects and how many possible properties there are, as will be detailed later), the pre-processor calculates what kinds of distractors there are and, constructs a set $\mathcal{S}$ of all possible scenes. The system then calls a generation algorithm to construct a description $\mathcal{D}$ containing a set of L QEs, that is, $\mathcal{D} = \{q_l(v)\}_{l=1}^{L}$, where $q(\cdot)$ is a quantified pattern with quantifier q (e.g., the *all* quantifier with two arguments can be written $\text{All}(\cdot, \cdot)$) and v is a property tuple. If v is capable of filling the slots of a quantified pattern $q(\cdot)$, we say that the pattern $q(\cdot)$ accepts v, and we write $q(v)$. The generation algorithm makes a selection from a set of quantified patterns $\mathcal{Q}$, based on the common

Quantifier	Semantics	Pragmatics
All(A, B)	$[\![A]\!] \subseteq [\![B]\!]$	$[\![A]\!] \neq \emptyset$
Only(A)	$[\![A]\!] = [\![O]\!]$	-
Half(A, B)	$\lvert [\![A]\!] \cap [\![B]\!] \rvert = \lvert [\![A]\!] - [\![B]\!] \rvert$	$[\![A]\!] \neq \emptyset$
Some(A, B)	$\lvert [\![A]\!] \cap [\![B]\!] \rvert \geq 2$	$\lvert [\![A]\!] \rvert > \lvert [\![B]\!] \rvert$
Some(A)	$\lvert [\![A]\!] \rvert \geq 2$	$\lvert [\![O]\!] \rvert > \lvert [\![A]\!] \rvert$
Only-1(A)	$\lvert [\![A]\!] \rvert = 1$	-
More(A, B)	$\lvert [\![A]\!] \rvert > \lvert [\![B]\!] \rvert$	$[\![B]\!] \neq \emptyset$
Fewer(A, B)	$\lvert [\![A]\!] \rvert < \lvert [\![B]\!] \rvert$	$[\![A]\!] \neq \emptyset$
Equal(A, B)	$\lvert [\![A]\!] \rvert = \lvert [\![B]\!] \rvert$	$[\![A]\!] \neq \emptyset$
Most(A, B)	$\lvert [\![A]\!] \cap [\![B]\!] \rvert > \lvert [\![A]\!] - [\![B]\!] \rvert$	-
Half-rest(A, B, B′)	$\lvert [\![A]\!] \rvert = 2 \lvert [\![B]\!] \rvert = 2 \lvert [\![B']\!] \rvert$	-
Minority(A, B)	$\lvert [\![A]\!] \rvert > 2 \lvert [\![B]\!] \rvert$	-
All-Comb(O)	All property combinations appear.	-

Table 6.8: List of quantifiers in English used in our quantified description generation system and their meanings.

knowledge $\mathcal{K}_c$ (i.e., meanings of all quantifiers) defined on $\mathcal{Q}$. Finally, with a set of logical forms $\mathcal{D}$, a simple template-based surface realiser is employed to map the logical form $\mathcal{D}$ into actual natural language text.

This generation system requires two types of knowledge:

Domain Knowledge. This is the list of all possible attributes and their possible values, with which the pre-processor could compute what distractors there are, and thus construct the set S. This knowledge is stored as a set of key-value pairs. For example, matching the current experimental setting of QTUNA, its domain knowledge is $\{\text{SHAPE} : [square, circle], \text{COLOUR} : [red, blue]\}$.

Common Knowledge This is a body of knowledge that corresponds to the quantified patterns in $\mathcal{Q}$. For a quantified pattern $q(\cdot)$, this knowledge base includes the meaning of the quantified pattern and a set of possible property tuples that could be assigned to v. The meaning of a quantified pattern has two parts: its semantics and its pragmatics. For example, the semantics of $\text{All}(A, B)$ asserts that $[\![A]\!] \subseteq [\![B]\!]$. The pragmatics says that $[\![A]\!]$ is not empty. Determining the semantics and pragmatics of each English quantifier term is difficult in general, but the QTUNA corpus allowed us to choose definitions that match majority usage in that corpus. The reason why we distinguish between semantics and pragmatics will become clear in the following section. Table 6.8 lists the quantifiers we considered in the current version of the QDG algorithm. We decided to use only the most frequent quantifiers. Note that, since we assign each quantifier a precise (i.e., non-fuzzy) meaning, which causes quantifiers like *some* and *a few* to have exactly the same meaning, we chose the most frequent one among the quantifiers with the same meaning. Quantifiers like *few* and *many*, which have attracted a lot of attention from researchers, are not included in our system since they have extremely low frequency in our corpus (that is, *few* appears 2 times and *many* 13 times).

```
Input: A target scene $s$, a set $\mathcal{S}$ of all possible scenes, a set of quantified patterns $\mathcal{Q}$, the
    common knowledge $\mathcal{K}_c$ defined on $\mathcal{Q}$.
Output: A quantified description $\mathcal{D}$ of $s$ that uses conjunctions of single or multiple $q(v)$.
 1: $\mathcal{D} := \{\}$
 2: while $\mathcal{S} \neq \{s\}$, and $|\mathcal{D}| < \delta$ do
 3:     $q(v) :=$ Pluralise$(q(v), s)$
 4:     $q(v) :=$ FindBestQuantifiedExpression$(s, \mathcal{S}, \mathcal{Q}, \mathcal{K}_c)$
 5:     if $q(v) = \emptyset$ then
 6:         break
 7:     end if
 8:     $\mathcal{D} := \mathcal{D} \cup \{q(v)\}$
 9:     $\mathcal{S} := \{s' \in \mathcal{S} : q(v)$ is true for $s'\}$
10: end while
```

Algorithm 6.1: The Greedy Algorithm for Generating Quantified Descriptions.

6.4.3 A Greedy Algorithm

As said, we view the QDG task as a task of ruling out distractor scenes. One can view this as a search problem, namely, the problem of finding a set of QEs that removes all (or as many as possible) distractors. This search can be performed by means of a *greedy* algorithm: working iteratively, this algorithm keeps selecting (and including into the QD) QEs that jointly rule out the largest possible number of distractor scenes.

We sketch the greedy algorithm for QDG (abbreviated as QDG-GREEDY) in Algorithm 6.1. The algorithm takes a target scene s, a set $\mathcal{S}$ of all possible scenes with the same domain as s (calculated by the pre-processor), a set of quantified patterns $\mathcal{Q}$ with their corresponding meanings (stored in $\mathcal{K}_c$) as inputs, and outputs a set $\mathcal{D}$ of QEs in logical form.

The algorithm initialises the description $\mathcal{D}$ as an empty set, then inserts QEs $q(v)$s iteratively into $\mathcal{D}$. During each iteration, QDG-GREEDY pluralises the $q(v)$; by this we mean adding a plural marker where necessary – namely whenever a property appears multiple times in the target scene s (e.g., Some(S, R) acquires a plural marker if there is more than one red square in the scene.

For example, suppose the QE is All(S, R) (meaning that all the squares are red) and the target scene contains two red squares; the expression is pluralised as All$(\langle S, pl\rangle, \langle R, pl\rangle)$ indicating that multiple squares in the target scene are red, in which, from now on, each argument is represented as a tuple and pl stands for plural while sg stands for singular. Pluralisation serves two purposes. The first is to determine the pragmatics of $q(v)$, which is then used for deciding how many distractors are left after selecting a certain QE. For instance, the plurality of All$(\langle S, pl\rangle, \langle R, pl\rangle)$ could rule out distractors that contain only one red square. The second purpose is to decide the surface form of the QE in English. The algorithm then calls the function `FindBestQuantifiedExpression` (line 4) to choose the QE that rules out the most distractors from all possible QEs. Specifically, `FindBestQuantifiedExpression` checks, for each possible QE $q(v)$, whether this expression fits the target scene based on the meaning (including both semantics and pragmatics) of $q(v)$ defined in $\mathcal{K}_c$. If yes, it calculates the number of distractors that can be ruled out by only using $q(v)$'s semantics. We call the number of distractors that a QE $q(v)$ rules out in a given situation the *Discriminatory Power* of $q(v)$. The `FindBestQuantifiedExpression` function will return the QE with the highest discriminatory power. If none of the candidate QE has discriminatory power, then the function returns an empty set.

To see why only the semantics, and not the pragmatics, of a QE is used for computing discriminatory power (i.e., for deciding whether to include a given QE into the QD) consider, by way of an example, the expression $\text{All}(C, B)$ (i.e., *All circles are blue*). Its semantics says (see Table 6.8) that the set of circles is a subset of the set of blue objects, and its pragmatics says, among other things, that there exists at least one circle. If the pragmatics of the expression contributes to its discriminatory power, then the algorithm would end up adding this QE to a description even when the QE's sole contribution is the (pragmatic) requirement that at least one object is a circle – as will happen when other QEs, previously added to $\mathcal{D}$ (e.g., $\text{All}(O, B)$), already ensure that the set of circles is a subset of the set of blue objects. [10] Additionally, as listed in Table 6.8, a number of quantifiers have the same pragmatics. So, if the pragmatics is taken into account when the algorithm determines the discriminatory power of a QE, then some very different QEs would end up having the same discriminatory power.

Line 5 of the algorithm checks whether the $q(v)$ is empty or not. If yes, then the algorithm concludes the *while* loop (line 6). If $q(v)$ is not empty, it is added to $\mathcal{D}$ (line 7) and the distractor scenes are removed from $\mathcal{S}$ based on both semantics and pragmatics of $q(v)$. Line 2 of the Algorithm 6.1 talks about the *Stop Criteria*. Generation terminates when all distractors are removed from $\mathcal{S}$ or the length of the generated description $\mathcal{D}$ reaches an upper bound δ. The idea of setting an upper bound comes from the observation that, in qtuna, descriptions were remarkably constant across domain sizes (see $\mathcal{H}_4$ in §6.2.3).

Note that in line 4 of this algorithm, the `FindBestQuantifiedExpression` is likely to find multiple QEs that have the same discriminatory power (i.e., several "best" expressions). Instead of trying to choose intelligently (and in order to increase the variation in generated QDs), the `FindBestQuantifiedExpression` randomly return one of these "best" expressions.

6.4.4 An Incremental Algorithm

We have seen that the Greedy algorithm iteratively selects that QEs that have the higher discriminatory power. From a cognitive viewpoint, however, there could be thought to be something slightly suspeccious about an algorithm that needs to perform such a complicated calculation: alter all, `FindBestQuantifiedExpression` has to check, for each quantifier pattern and all its possible values, how many scenes would be ruled if these were selected. Moreover, when we examined the qtuna dataset more closely, we found that some quantifiers patterns are far more frequent than others, and some choices of properties to fill a given pattern are far more frequent than others. For example, akin to what $\mathcal{H}_5$ (in §6.2.3) indicates, we found that if *all* fits in any of the properties in a scene, subjects tend to use *all* to construct a QE. Building on these observations, a natural idea would be to compose an ordered sequence of quantifiers, and an ordered sequence of fillers (i.e., property tuples), reflecting the different degrees of "popularity" of different quantifiers and different fillers. The algorithm can then make use of this ordered sequence

[10] If plurality is also treated as a part of pragmatics, then the pragmatics of the QE $\text{All}(\langle C, pl\rangle, \langle B, pl\rangle)$ says that there are at least two circles. This would exacerbate the above effect.

```
Input: A target scene $s$, a set $\mathcal{S}$ of all possible scenes, a set of quantified patterns $\mathcal{Q}$, the
    common knowledge $\mathcal{K}_c$ defined on $\mathcal{Q}$, a Quantifier Preference Order defined on $\mathcal{Q}$, a
    set of all possible property tuples in the domain $\mathcal{V}$, and a property preference order
    defined on $\mathcal{V}$.
Output: A quantified description $\mathcal{D}$ of $s$ that uses conjunctions of single or multiple $q(v)$.
$\mathcal{D} := \{\}$
for each $q$ in $\mathcal{Q}$ (in order of the Quantifier Preference Order) do
    for each $v$ in $\mathcal{V}$ such that $q$ accepts $v$ (in order of the Property Preference Order) do
        $q(v) :=$ Pluralise$(q(v), s)$
        if $q(v)$ is true for $s$, and $\mathcal{D} \not\models q(v)$ then
            $\mathcal{D} := \mathcal{D} \cup \{q(v)\}$
            $\mathcal{S} := \{s' \in \mathcal{S} : q(v)$ is true for $s'\}$
        end if
    end for
    Until $\mathcal{S} = \{s\}$ or $|\mathcal{D}| \geq \delta$
end for
```

Algorithm 6.2: The Incremental Algorithm for Generating Quantified Descriptions

to determine in what order to consider the different types of expressions for inclusion in the generated description. Analogous to the "preference orders" of attributes (like colour, size, etc.) that are employed in the generation of REs (see 2.2.1), one would ultimately like to understand the reasons behind these preference orders, for instance in terms of codability (cf., van Deemter (2016, chapter 3) for discussion). Lacking such a deep understanding for the moment, we considered the following two types of sequences:

Quantifier Sequence. Inspired by the fact that some quantifiers occur more frequently than other quantifiers (as shown in Figure 6.1), QEs that use frequent quantifiers like *all*, *half* or *most* should have high priority (i.e., they should occur early in the Preference Order).

Property Sequence. Analysis of QTUNA (see $\mathcal{H}_6$ in Study 2) suggested that for patterns of the form All(A, B), the first argument, A, is more often a SHAPE property, whereas B is more often a COLOUR. For example, the algorithm should prefer the property tuple (S, R) than (R, S).

Further details of both the Quantifier Sequence and the Property Sequence are given below. The algorithm incrementally generates the description by considering possible quantifiers and fillers one by one, starting at the top of the sequence, working its way from the top of the preference order downwards. Given the analogy with the incremental algorithm for REG (Dale & Reiter, 1995), we call the algorithm the incremental algorithm (abbreviated as QDG-IA). Likewise, we will speak of the *Quantifier Preference Order* (instead of quantifier sequence) and the *Property Preference Order* (instead of Property Sequence).

Note that in addition to the inputs of the QDG-GREEDY algorithm, as shown in Algorithm 6.2, QDG-IA requires two pre-defined preference orders defined above. Given these inputs, the QGD-IA algorithm will go through all the quantified patterns $\mathcal{Q}$ in order of the quantifier preference order. In each iteration, for the selected quantified pattern $q(\cdot)$, QDG-IA will test all possible property tuples accepted by that pattern in the order of property preference order. Recall that the information which $q(\cdot)$ accepts which property tuple can be found in $\mathcal{K}_c$. The algorithm then calls the Pluralise function on the QE, which is the same manipulation done by QDG-GREEDY.

Line 5 of Algorithm 6.2 involves some important deviations from Dale and Reiter's algorithms. Here, our algorithm first tests whether $q(v)$ is correct as a QE for s; the test is performed by using both its semantics and pragmatics. Subsequently, the algorithm tests whether $q(v)$ does not follow from the description $\mathcal{D}$ (i.e., $\mathcal{D} \not\models q(v)$) [11], ensuring that $q(v)$ rules out one or more further scenes (i.e., it is not logically superfluous). Crucially, the latter test uses only the semantics of $q(v)$, not the pragmatics. In the case of the present algorithm, the different roles of semantic and pragmatic information (in this case: the information provided by the plural form) is possibly even more striking than in the case of the Greedy algorithm. For example, suppose we want to generate a QD for a scene that consists of 2 blue squares and 2 blue circles, and the quantifier *all* has the highest priority in the quantifier preference order. In its first iteration, the algorithm produces a QE like "*all objects are blue*". In the second iteration, if the pragmatics is used for validation, the algorithm could add "*all circles are blue*", whose semantics contribute no new information at all, but whose pragmatics (i.e., the claim that there are at least two circles) rules out all those distractor scenes that contain less than two circles (which would cause it to pass the second test of line 5). The resulting description, "*All objects are blue and all circles are blue and ..*" (which can be made logically complete by adding "*... and there are squares*") would sound strange because, intuitively, the second clause is logically redundant given the first. [12]

Once the above two conditions have been validated, $q(v)$ is appended at the end of the description and the scenes for which $q(v)$ is not true are removed from $\mathcal{S}$. Both semantics and pragmatics are used for removing such distractors. The generation terminates according to the same criteria as the QDG-GREEDY algorithm.

As for the design of preference orders, we started with testing the following settings, once again based on analysis of the corpus. The quantifier preference order is a linear preference order, namely:

$$\begin{aligned}&\text{All}(\cdot,\cdot) \succ \text{Everything}(\cdot) \succ \text{Only}(\cdot) \succ \text{Half}(\cdot,\cdot) \succ \text{Half-rest}(\cdot,\cdot,\cdot) \succ \text{Equal}(\cdot,\cdot)\\ &\succ \text{Most}(\cdot,\cdot) \succ \text{More}(\cdot,\cdot) \succ \text{Minority}(\cdot,\cdot) \succ \text{Fewer}(\cdot,\cdot) \succ \text{Some}(\cdot,\cdot) \succ \text{Some}(\cdot)\\ &\succ \text{Only-1}(\cdot).\end{aligned}$$

The second-order quantifier All-comb is only applicable to a small number of scenes but is used very frequently for those scenes. Therefore, although it has a relatively low overall frequency across the whole corpus, we still assign it a high priority. [13] The property preference order was designed by following some constraints, for example, SHAPE

[11] Logical consequence is implemented by calculating the set of scenes that are removed by a given expression (or set of expressions). Thus, $\mathcal{D} \models q(v)$ means that the set of distractor scenes removed by $q(v)$ is a subset of the set of distractor scenes removed by $\mathcal{D}$.

[12] These observations might have applications in other areas of language use as well, for instance, Gricean conversational implicatures (Grice, 1975). Imagine the Gricean scenario in which an academic referent "praises" one of his students for having nice handwriting (implying that the student is academically inept and should not be hired). Our observations suggest that it would be odd for this academic to make the same utterance as part of a conversation in which the student's handwriting had already been favourably commented upon.

[13] A $\succ$ B means that A follows B in the preference order.

properties have higher priorities in the first argument places and compounded properties (e.g., RS and BC) are more preferred than singular properties (e.g., R, C, and B).

However, when we ran the algorithm, we found that some quantified patterns that have low preference are never chosen by the algorithm, causing the generated descriptions to only use a very limited set of patterns. For example, the pattern All$(\cdot,\cdot)$ has a higher preference than the pattern Only$(\cdot)$, and consequently the latter is never chosen, because its meaning is covered by the former. For example, the meaning of "*there are only squares*" is covered by that of "*all objects are squares*". To increase variety, we introduced a probability θ, with which the QDG-IA performs a one-off re-ordering move; for the work reported in this study, we set θ to 0.1. Re-ordering was not performed across the entire preference order, but only within certain groups of quantifiers that have high meaning overlap with each other. To be precise, we used the following partitioning of the Preference Order of quantifiers (each $\{\cdot\}$ represents a partition):

$$\begin{aligned}&\text{All-Comb} \succ\\ &\{\text{All}(\cdot,\cdot) \succ \text{Everything}(\cdot) \succ \text{Only}(\cdot)\} \succ\\ &\{\text{Half}(\cdot,\cdot) \succ \text{Half-rest}(\cdot,\cdot,\cdot) \succ \text{Equal}(\cdot,\cdot)\} \succ\\ &\{\text{Most}(\cdot,\cdot) \succ \text{More}(\cdot,\cdot) \succ \text{Minority}(\cdot,\cdot) \succ \text{Fewer}(\cdot,\cdot)\} \succ\\ &\{\text{Some}(\cdot,\cdot) \succ \text{Some}(\cdot)\} \succ \text{Only-1}(\cdot).\end{aligned}$$

Once the algorithm has decided to conduct a one-off move, the order of quantifiers within that part are re-ordered at random.

6.4.5 Surface Realisation

Surface Realisation is typically the last stage in an NLG pipeline, where abstract structures are turned into concrete sentences. In the present case, Surface Realisation turns the logical forms produced by the Greedy and Incremental algorithms into actual stretches of English text. Though this is not the stage of the pipeline on which our computational model focuses, it cannot be omitted because, without Surface Realisation, it would be much more difficult for human judges to evaluate the output of the algorithm: people are used to interpreting and judging text, not abstract representations.

Our system uses a simple template-based surface realiser (see e.g. (van Deemter et al., 2005) for comparison with other types of Linguistic Realisation). For each quantified pattern, there is a specific template. For example, for All$(\cdot,\cdot)$, we have a template:

(81) All of ⟨ARGUMENT-1⟩ ⟨COPULA⟩ ⟨ARGUMENT-2⟩

where ⟨COPULA⟩ will be realised into *is* or *are* depending on the plurality of the first argument of the generated QE that uses this pattern. When filling these slots with chosen properties, some simple syntactic and morphological operations are employed. For example, if a COLOUR property takes the first place of a quantified pattern, a noun is appended to package it into a noun phrase (i.e., *red* → *red object*). If a property has a plural suffix, the surface form of the property is mapped into its plural form. A number of further constraints, specific to particular quantified patterns, were also encoded in the realiser.

The present work has focused on the way in which speakers use a variety of quantifiers, which is why Linguistic Realisation of sentences and texts was kept simple and could be improved in many ways. One significant limitation of the way in which the abstract patterns generated by the algorithms of the previous sections are put into words is that our wordings do not use *anaphora* yet. This is despite the occurrence of many different types of anaphoric expressions in our corpus, for example as when a QE is followed by "*Half of them are red*". Anaphoric patterns were particularly prevalent in QEs with 3-ary quantifiers, for example as in "*Half of the objects are red, the other half are blue*". Using anaphora judiciously without creating unwanted ambiguities is quite doable in general, but the topic is not without its problems (e.g., Kamp and Reyle, 1993, Chapter 4). We expect that, by addressing these issues, future Linguistic Realisation modules will be able to produce even more human-like descriptions of the scenes on which we are focusing.

6.4.6 Evaluating the Generated Quantified Descriptions

Although our algorithms were informed by extensive elicitation experiments, we wanted to gain additional insights into the quality of generated descriptions. We were curious how "human-like" these descriptions are, and how correct and informative.

Previous studies on evaluating the human-likeness of a computational language production model tend to use corpus-based evaluation: the model generates outputs (e.g., sentences or logical forms) and these outputs are compared with a corpus using a similarity measure (e.g., van Deemter, Gatt, Sluis, et al. (2012)), such as DICE (Dice, 1945) or BLEU (Papineni et al., 2002). However, there are two insurmountable problems with using such a methodology in the present situation.

First, the quality of a QD cannot easily be measured automatically. Consider the QE $\text{Few}(O, S)$ once again. Suppose the target scene is a situation in which 5 out of 20 objects are squares; then is it correct to say that $\text{Few}(O, S)$, or does this underestimate the number of squares? And if $\text{Few}(O, S)$ is all that is said about the proportion of objects are squares, is this sufficiently informative or not? We are not aware of any existing metric or algorithm that would give us reliable answers to these questions. Therefore, we decided not to use corpus-based evaluation, but to conduct two evaluation studies: a human judgement study (i.e., asking expert human judges to rate the generated QDs) and a scene reconstruction study (i.e., asking human subjects to reconstruct the input scenes given the generated QDs).

Second, since we designed our algorithms based on the QTUNA corpus, it would be insufficient to evaluate them on the same corpus again, since this would fail to distinguish between training and test data. (Borrowing terminology from the machine learning community, it would risk letting the model over-fit the corpus.) To avoid this problem, we selected our experimental materials not only from our QTUNA corpus but also from scenes that do not appear in QTUNA.

Concretely, we divided the evaluation experiments into experiment A and experiment B. For experiment A, we randomly selected 3 or 4 scenes from each of the 3 sub-corpora of QTUNA to construct a set of, in total, 10 scenes, each of which was paired with 3 descriptions: one by QDG-IA, one by QDG-GREEDY, and one selected at random from our corpus. A number of example scenes, paired with their descriptions, are listed in Table 6.9. For experiment B, we focused on three new domain sizes namely $N = 6$, $N = 10$, and $N = 16$. For each of these, we sampled 6 scenes, each of which was paired with 2 descriptions: one by QDG-GREEDY and one by QDG-IA. Finally, we have 66 scene-description pairs ready to be evaluated.

To assess the quality of each description, we used two different methods: a method based on quality judgements by human experts and a task-based method in which readers were asked to reconstruct the scenes that are described.

Scene	Model	Description
BS:2 RS:2 BC:0 RC:0	Human	All the objects are squares and half of them is blue.
	QDG-IA	Every object is square. There are equally many blue squares and red squares.
	QDG-GREEDY	Half of the objects are blue squares, the rest are red squares.
BS:2 RS:2 BC:5 RC:0	Human	Two objects are red squares. Two objects are blue squares and the remainder is blue.
	QDG-IA	Every circle is blue. Half of the squares are blue. More than half of the objects are blue circles.
	QDG-GREEDY	Half of the squares are red, the rest are blue. Most of the objects are blue circles.
BS:9 RS:2 BC:8 RC:1	Human	There is a mixture of squares and circles. Most of them are blue. Some of them are red.
	QDG-IA	All possible objects are shown. A minority of the objects are red squares. Less than half of the objects are blue circles. Less than half of the objects are blue squares. Less than half of the objects are circles.
	QDG-GREEDY	All possible objects are shown. A minority of the objects are red squares. Less than half of the objects are blue circles. Less than half of the objects are blue squares. Less than half of the objects are circles.

Table 6.9: Examples of quantified descriptions produced by humans, by QDG-IA, and by QDG-GREEDY. The numbers in the Scene column represent the number of objects of each type (e.g., the first scene consists of two blue squares and two red squares).

Human Judgement

Settings. We recruited 4 annotators, who were academics from Utrecht University, and none of whom had been involved in our research. Two were young lecturing staff in computational linguistics and two were senior lecturing staff in computational logic and formal argumentation. All the 66 scene-description pairs (from both experiments A and B) were put together and randomly allocated to our four judges. Each of them judged 33 scene-description pairs. Thus, each scene-description pair was judged by two judges and was judged from three aspects: correctness, completeness, and naturalness.

However, the correctness and the completeness of a description is not an "all or nothing" affair, especially when larger domains are involved, which frequently give rise to descriptions that contain vague quantifiers. The same is true for the perceived naturalness of the description. As is often done in NLG (Gatt & Krahmer, 2018), we used a gradable scale. Judges were asked three Likert-scaled questions in each case:

1. Naturalness: *On a scale of 1-5, how likely do you think it might be that this description was uttered by a human?* [1=very unlikely, 5=very likely];

2. Informativity: *On a scale of 1-5, do you believe the description is as informative as it can be expected to be?* [1=description is not even nearly informative enough, 5=description gives as much information as is possible];

	Model	Naturalness	Informativity	Correctness
E.A	Human	3.45	4.05	4.6
	QDG-IA	2.85	3.95	4.55
	QDG-GREEDY	3.45	3.8	4.8
E.B	QDG-IA	3.7	3.8	4.83
	QDG-GREEDY	3.46	4.2	4.83

Table 6.10: Average scores for each algorithm and for human-produced descriptions, by naturalness, informativity, and correctness as annotated by our four human judges. "E.A" stands for "Experiment A".

3. Correctness: *On a scale of 1-5, how correct do you consider this description to be?* [1=the description is not ad all correct, 5=everything the description says is correct].

Note that when judges make judgements the words naturalness, informativity and correctness were invisible. In addition, our instructions said "*Please note that we are mainly interested in the logic of how people describe the scene, and less in the details of the wording, so please disregard minor syntax errors and typos*". Because in experiment A, the first question was asked about a human-produced description as well as two algorithm-generated descriptions. This setup allowed us to perform what is essentially a *Turing Test*. The other two questions offered invaluable formative evaluation.

On the basis of the nature of the task and the algorithms of QD production, we formulated a number of evaluation hypotheses:

1. Humans perform better at naturalness than QDG-IA and QDG-GREEDY ($\mathcal{EH}_1$);

2. Both algorithms perform better at informativity and correctness than humans, because both of them were explicitly designed to optimise informativity and correctness ($\mathcal{EH}_2$);

3. QDG-IA performs better at naturalness than QDG-GREEDY ($\mathcal{EH}_3$). We reasoned that, in REG, the incremental algorithm offered greater human-likeness than the greedy algorithm (Dale & Reiter, 1995; van Deemter, Gatt, Sluis, et al., 2012), so why should things be different this time?

Results. Table 6.10 shows the scores from the judges. Both algorithms scored well over 3 in all except one cell, confirming our impression that the descriptions tended to be of very respectable quality.

As for our evaluation hypotheses, our first evaluation hypothesis, $\mathcal{EH}_1$, was rejected: in terms of naturalness, QDG-GREEDY performed well above expectation, gaining the exact same score as the human speakers. QDG-IA had a slightly lower score, but this did not amount to a significant difference (as tested by a paired t-test).

Though "no difference" results always need to be approached with caution, the rejection of $\mathcal{EH}_1$ might be interpreted as an NLG algorithm passing a kind of Turing test (focusing on a limited type of language use, of course), since it suggests that the perceived quality of the algorithm was indistinguishable from human speakers. In an effort to understand the low naturalness performance of QDG-IA, we had a closer look at the cases where QDG-IA had particularly low scores. We found that these almost always contained vague quantifiers (e.g., *few*, *most*), where the semantic and pragmatic definitions of which our algorithms made use were especially tentative. Moreover, vague quantifiers were used disproportionally often in the scenes of Experiment A, and far less in the scenes of Experiment B; accordingly, the QDG-IA scored much better on naturalness in Experiment B. We surmise that a possible reason for the low naturalness performance of QDG-IA is that the semantics of the vague quantifiers in $\mathcal{K}_c$ is not as accurate as it could have been. For instance, the currently-used semantics of *most* is the same as that of *more than half*, which is a precise quantifier. The effect of using more accurate, empirically based, definitions of vague quantifiers, which requires further comprehension experiments, will be investigated in our further work.

Our analysis of the second evaluation hypothesis, $\mathcal{EH}_2$, shows some of the hidden difficulties of the description task that our algorithms solve. Human speakers, and both of our algorithms, all performed similarly well in terms of informativity and correctness. To understand why, we decided to separately calculate the average informativity score for those descriptions in experiment B that were *logically complete* (i.e., the algorithm stopped when $S = \{s\}$). For this reduced set of descriptions, the average scores for QDG-IA and QDG-GREEDY were a mere 3.88 and 4.1, instead of the score that one might expect, namely 5. One possible explanation is that our algorithms judged the logical correctness and completeness of these descriptions by taking both their semantics and their pragmatics into account (as discussed in §6.4.3 and §6.4.4), which is something our judges may have disagreed with. Alternatively, judges may sometimes have had a lapse of concentration.

The last evaluation hypothesis, $\mathcal{EH}_3$, was also rejected, as there was no significant difference between the naturalness performance of QDG-IA and QDG-GREEDY. This may be because the preference order that we proposed for quantified patterns has much higher complexity than that of properties (or attributes) in the task of REG. In particular, the number of quantifiers is considerable, and, because of our "one-off" re-ordering move, our preference order of quantifiers was not linear. It is possible that a different preference order would have led to better results for QDG-IA, but it seems equally possible that the idea of using a preference order to determine the choice of quantifier patterns – on which the Incremental Algorithm is based – is simply not on the right track, and that a simpler "greedy" approach leads to results that are equally good.

Scene Reconstruction

Settings. We recruited 20 undergraduate students from Utrecht University 13 of whom are majored in Artificial Intelligence; the other 7 study a variety of other subjects. The descriptions in experiment A were randomly allocated to all participants. Each description was used for reconstructing the paired scene four times (i.e., by four participants). The descriptions in experiment B were allocated in the same way, except that each pair was assigned twice instead of four times. Each participant reconstructed a total of 8 or 9 scenes.

Given a description and the domain size of the paired scene, we demanded each participant to write down the number of objects (i.e., the number of BS, RS, BC, and RC) in the scene by asking "*please tell us about a scene that could be described by the description*". We chose to ask participants to write numbers instead of drawing scenes, in order to encourage them to disregard the location of each object in the scene. Participants had not seen any of the QTUNA scenes before, which makes the reconstruction task become tough. Therefore, before starting, we provided each participant with two examples to show them how the

	Model	N=4	N=9	N=20	All
Experiment A	Human	8.33	6.25	15	9.86
	QDG-IA	0	11.81	18.33	10.05
	QDG-GREEDY	2.08	8.33	15	8.47
	Model	N=6	N=10	N=16	All
Experiment B	QDG-IA	1.39	10	10.42	7.27
	QDG-GREEDY	1.39	8.33	7.81	5.84

Table 6.11: Average SWAP(%) for each algorithm and for human-produced descriptions. N represents the domain size.

reconstructed scenes might look like. In addition, considering that some descriptions have multiple possible corresponding scenes, we told participants: *"In those cases, please choose an answer (number) that you consider to be consistent with the description"*.

Given the above settings and the hypotheses of the human judgement study, we formulated two evaluation hypotheses. We hypothesised that

1. Reconstructions based on descriptions generated by QDG-IA and QDG-GREEDY are more similar to the input scenes than are those produced by humans ($\mathcal{EH}_4$). We expected this because these algorithms, especially the QDG-GREEDY are designed to be as logically complete as possible. Since the more complete the generated descriptions are, the easier for them to be reconstructed.

2. Secondly, we hypothesized that (2) descriptions produced by QDG-IA let readers reconstruct scenes more accurately than QDG-GREEDY ($\mathcal{EH}_5$).

Similarity between reconstructions. A key part of our analysis is the metric that we used to measure the similarity between a reference scene and a reconstructed scene. Given a reference scene and a reconstructed scene, we care about how many "swaps" are needed to convert one into the other. Concretely, we propose the SWAP metric, which takes the absolute differences between the cardinalities of each of the four types of object in the reference scene and in the reconstructed scene, takes the sum of these, then divides that sum by 2 times the domain size.

For instance, suppose the reference scene is: $\{\mathrm{BS}:2, \mathrm{RS}:1, \mathrm{BC}:0, \mathrm{RC}:1\}$, where each number represents the cardinality of the relevant type of object. Suppose one of the reconstructed scenes is: $\{\mathrm{BS}:2, \mathrm{RS}:1, \mathrm{BC}:1, \mathrm{RC}:0\}$. Then

$$\mathrm{SWAP} = \frac{|2-2|+|1-1|+|0-1|+|1-0|}{2\times 4} = 1/4. \tag{6.1}$$

The lower the SWAP score, the better the reconstruction, with 0 as a minimum and 1 as a maximum.

Results. Table 6.11 reports the SWAP score for both experiments A and B. We analyse these results, focusing on our hypotheses first. The SWAP scores in experiment A show that hypothesis $\mathcal{EH}_4$ is only confirmed for the smallest domain size ($N = 4$) while for larger domain sizes, QDG-IA generates less reconstructable descriptions than our human speakers. When domain size is small, precise (i..e, non-vague) quantifiers tend to be used (c.f., §6.2). Consequently, our algorithms always generate logically complete descriptions, so they enjoy low SWAP scores (i.e., high re-constructability). Conversely, when domain size becomes larger, then logical completeness becomes less and less achievable (cf., §6.2), and a larger number of vague quantifiers are used. Consequently, the SWAP scores go up, so the re-constructability of descriptions produced by human speakers and algorithms goes down.

As for our hypothesis $\mathcal{EH}_5$, by looking at results from both experiments A and B, the hypothesis is rejected; in fact, the data go in the opposite direction, since descriptions generated by the QDG-GREEDY have better re-constructability than QDG-IA. One possible explanation is that QDG-IA may have generated less complete quantified descriptions than QDG-GREEDY (since QDG-GREEDY always looks at QEs that have the highest discriminatory power). In other words, when readers were simply reading these descriptions together with paired scenes (i.e., in the human judgment study), this difference may not have been noticed (note that QDG-IA had a similar level of informativity as QDG-GREEDY), the difference may have been "enlarged" in the reconstruction experiments.

Besides, from the results in Table 6.11, we also found that when focusing on hard cases (i.e., Experiment A) re-contructability decreases with the increase of domain size. In contrast, when we use randomly selected scenes (i.e., Experiment B), although differences between small and large domains exist, it appears that if the domain size is large enough, then no significant difference exists (i.e., there is no significant difference between the SWAP score when $N = 9$ and when $N = 20$).

6.4.7 Initial Comparison Between two Evaluation Protocols

Considering the results of these two evaluation studies, we made two post-hoc observations. On the one hand, although algorithms were designed to produce QDs that are logically complete, in the human judgment study, we observed that the algorithm-produced QDs did not receive a higher informativity score than human-produced ones. A similar phenomenon occurs in the large domain reconstruction study. Machine-generated descriptions had better re-constructability in merely the small domain (i.e., $N = 4$) than human-generated ones, where vague quantifiers are rarely used and received worse or equal re-constructability scores in larger domains (i.e., $N = 9$ and $N = 20$) than human-generated ones. This said, logically complete descriptions not only, from the speakers' perspective, are less often produced than logically incomplete ones but also, from the readers' perspective, contribute nothing to help readers to comprehend its meaning.

On the other hand, the greedy algorithm can generate descriptions that have slightly better re-constructability scores than the incremental algorithm, but they were not judged to be significantly more informative. This inconsistency might result from the differences between these two kinds of evaluation protocols. Explaining the reason behind this requires larger-scale experiments that assess such differences.

6.4.8 Discussion

In this study, we proposed two generation algorithms that aim to mimic the language production behaviour recorded in the corpus, understood as what is known in the computational modelling community as a *product model*, that is, a model that focuses on the relation between inputs and outputs without any claims about the manner in which this is done. We then evaluated these algorithms, looking at scenes of a variety of sizes, including scenes that contained numbers of objects not seen in the corpus. Our evaluations suggest that our algorithms produce descriptions that are both natural (i.e., human-like) and useful.

Computational models of language use can offer a wealth of insight into the choices that human speakers and writers make when they use language. Let us take stock to see what lessons may be drawn from our computational modelling exercise. Additionally, since we designed our algorithms in accordance with the collected English QDs and have merely evaluated our algorithms in English, we hereby also discuss potential issues of applying our algorithms to model Mandarin QDs.

Quantified Descriptions and Referring Expressions

Computational models of the production of REs have been studied widely (Dale & Reiter, 1995; Dale & Viethen, 2009; Krahmer & van Deemter, 2012; van Deemter, Gatt, Sluis, et al., 2012; van Gompel et al., 2019). They aim to mimic how human speakers use RE to single out a referent for a hearer. For example, given a scene such as Figure 2.8(a), a participant could say *"the large chair*", *"the large front facing sofa*", *"the front facing sofa*" or *"the green chair*". Each of these expressions lets readers identify the target reference from the context.

Quantification is not reference, of course. Nonetheless, it is illuminating to compare the two phenomena and, in fact, the algorithmic approach we have chosen to model quantification resembles some algorithms originally discussed in Dale and Reiter (1995), where an RE is constructed by accumulating properties (e.g., `COLOR`, `SIZE`) one by one, each of which is thought to "remove" from consideration a set of "distractor objects", that is, potential referents that differ from the target referent in one or more respects. We have emphasised this similarity by using terms familiar from REG (e.g., "target", "removing distractors", "preference order" and so on). In a nutshell,

- In the generation of both REs and QDs, the task can be viewed as a step-wise addition of descriptive information that narrows down an initial set of possibilities (a set of possible referents in one case, and a set of scenes in the other case) to a small set – typically a singleton set.
- In both situations, the "narrowing down" metaphor gives rise to a range of possible algorithms. In each case, for instance, a "greedy" algorithm might proceed by always adding the information that most effectively reduces the size of the current set of possibilities. In other words, the notion of *discriminatory power*, which is crucial for models of reference, looms large in the modelling of quantification as well.
- In both cases, the effect of adding information must be understood in the context of the Common Ground of the speaker and hearer. When the speaker is unsure as to what the hearer knows (e.g., what the initial set of possibilities is), for example, the question can arise of whether it is practically feasible – in a reasonable time, and using a description that is not too lengthy or complicated – to reduce the initial set of possibilities to a singleton set. In the realm of reference, for example, Kutlak et al. (2016) model a situation in which the aim of a RE is not to uniquely pick out one single referent. Below we will discuss similar situations in the realm of quantification.

These similarities should not close our eyes to the important differences that exist between the two tasks. Firstly, in the most often studied versions of the reference task, distractors are concrete objects, which are observed by the speaker and the hearer; in our quantification task, the distractors are a set of *possible* scenes, only one of which is observable, namely the target scene. This makes our quantification task much more abstract than most versions of the reference task. In our generation system, this is reflected by the stage in which the pre-processor computes the set $\mathcal{S}$ of all possible scenes from the properties that are given.

Secondly, in the reference task, properties (such as *red*) take the place that QEs have in the quantification task. QEs are much more complicated than properties, hence the distinction between choosing a pattern $p(\cdot)$ (line 2 of Algorithm 6.2) and choosing a value v to fill the pattern (line 3).

Thirdly, the algorithms proposed in the present study have had to find a way to take both the semantics and pragmatics of quantifier patterns into account. In a nutshell, semantics is about literal meaning whereas pragmatics is about other ways in which language use can convey information. That said, the distinction between semantics and pragmatics is much debated within Theoretical Linguistics, and the precise boundary between the two is notoriously difficult to draw (Levinson, 1983). The way in which this distinction works in relation to reference is relatively well understood, but the distinction has proved to be much harder to deal with in connection with quantification, because if semantic information is lumped together with pragmatic information, our algorithms tend to generate descriptions that are unnecessarily unwieldy (see our explanation in §6.4.3). Whether the solution embodied in our algorithms generalises to other types of pragmatic information is a question for further research.

Representing the meanings of quantifiers

Our generation algorithms embody specific assumptions concerning the meaning of each quantifier. For example, when an algorithm adds the QEs "*All circles are blue*" to a description, we assume that "All A are B" means $[\![A]\!] \subseteq [\![B]\!] \wedge [\![A]\!] \neq \emptyset$; consequently, our algorithms remove from the set S all those scenes for which this logical conjunction does not hold true. Although we have done our best to choose representations of quantifier meaning that are consistent with both the Linguistics literature and the way in which quantifiers are used in our corpus, we cannot claim yet to have found the optimal representation in each case. For example, various authors (Moxey & Sanford, 1993; Nouwen, 2010) have pointed out that human quantifier use is guided not only by raw numbers of objects alone but by (speakers' and) hearers' expectation about the number as well; for example, a child in The Netherlands who has seen 10 animals on a given day may say she has seen *many elephants* (if that's what they were) but *a few cows* (if that's what they were). Although the geometrical scenes on which we have focused in this study has sought to minimise these issues, there is surely a lot of progress to be made; in fact, it is perhaps remarkable that our algorithms work as well as our evaluation suggests they do.

A class of quantifiers where this disclaimer is particularly opportune are "vague" quantifiers, that is, quantifiers where there can exist borderline cases in which it is debatable whether or not the quantifier applies; cases in point are quantifiers like *many*, *few*, *all except a few*, and so on. In all these cases, the set-up of our generation algorithms forces us to use a crisp cut-off point – deciding, for example, that *Many A are B* is true if less than 20% of A are B, and false otherwise. Although this contradicts received wisdom about the meaning of these quantifiers, our evaluation suggests that, for the type of generation task at hand, our algorithms "get away" with this simplification. While this outcome gives rise to interesting questions – Could an NLG algorithm that models vague expressions as if they were crisp pass the Turing test? – we believe that it would be interesting to experiment with alternative assumptions that do more justice to what is known about these quantifiers.

For instance, one could represent the meaning of quantifiers probabilistically (Moxey & Sanford, 1993), or using a version of Fuzzy Logic. In both cases, the representations in question would tell us to what extent a given QE is applicable in a given situation: let's call this its *degree of applicability*. Such a move could even benefit quantifiers that linguists generally consider to be crisp. For example, Degen and Tanenhaus (2011) and van Tiel (2014) pointed out that, when reading QEs like *Some of A are B*, readers' acceptability is lower than 1 (though often higher than 0) if the target set is either too small or too large. A similar approach is taken in the Bayesian quantifier models of (Frank & Goodman, 2012), Franke (2014) and Qing (2014, Chapter 4), which are learned from experimental data. The resulting non-crisp meaning representations could be fed into our generation algorithms in a number of ways. For example, in the Incremental Algorithm, the choice of the next QD to be included in the description (which was done in lines 2 and 3 of Algorithm 6.2) could be made on the basis of the degree of applicability of the expression in combination with its preference degree. It would be interesting to see whether, as a result of this move, the quality of the resulting QDs (as measured by evaluation studies such as the one reported in the current study) will improve. Since the present work focuses on the production of a wide range of quantifiers rather than on sophisticated models for specific quantifiers, this exploration was left for future research.

Generating Mandarin Quantified Descriptions

We believe the QDG framework we proposed is universal across different languages. Nonetheless, the are still a number of issues that need to be aware of when building Mandarin QDG systems. First, in Mandarin, the plurality can be expressed either explicitly or implicitly.

(82) a. 图片 中 有 红色方块 和 蓝色圆圈 。
túpiàn zhōng yǒu hóngsèfāngkuài hé lánsèyuánquān
There are red squares and blue circles.

b. 图片 中 有 一些 红色 方块 和 一些 蓝色圆圈 。
túpiàn zhōng yǒu yìxiē hóngsèfāngkuài hé yìxiē lánsèyuánquān
There are some red squares and some blue circles.

Description (82-a) and (82-b) show examples for the two situations, respectively, and note that, in MQTUNA, the implicit version is more frequent. This requires that when the algorithm calls the `Pluralise` function, it needs to consider the surface form of the current QE in advance. If the final QE is in the form analogous to the description (82-a), then the algorithm should not call the `Pluralise` function.

Second, given the conclusion that Mandarin speakers use more vague quantifiers than English speakers, a Mandarin QDG algorithm needs to be powerful on handling vagueness. To this end, we have discussed the potential ways to include non-crisp meaning representations above. Additionally, this also matters the construction of the quantifier preference order of the incremental algorithm. As we can see by comparing the top-10 quantifiers in English (Table 6.1) and in Mandarin (Table 6.5), the list of frequent quantifiers in Mandarin is very different from that in English, and there are more vague quantifiers.

Therefore, to build a Mandarin QDG system, we need to adopt the quantifier preference order accordingly: moving some vague quantifiers upward and, meanwhile, moving some crisp quantifiers downward.

Third, in the §6.3, we also found that Mandarin QDs in MQTUNA are generally longer than those in QTUNA, especially when domain size is large. To reproduce such a characteristic, we probably need to adopt the stop criteria. Both QDG-GREEDY and QDG-IA stop producing QEs with respect to the parameter θ. When building Mandarin QDG systems, we may need to increase the value of θ to produce longer QDs.

Last, Mandarin QDs also pose challenges for surface realisations (more issues of realisation in Mandarin can be found in Chapter 7). This is because: 1) most QEs can be expressed in three different ways (see the example (76)); (2) in addition to handling anaphora, the realiser for Mandarin QDs also need to handle A-drop; (3) the realiser needs to decide whether to express plurality explicitly or implicitly. As aforesaid, this decision is made in accordance with whether the algorithm calls the `Pluralise` function or not.

6.5 Summary

In this chapter, we investigated how English and Mandarin speakers use quantifiers if they are free to describe a visual scene in whatever way they want and its computational models.

We decided to look at English first. We conducted the QTUNA experiment where participants were asked to describe a series of visual scenes using any quantifier they want, using as many sentences as they choose, and using any sentence pattern they like. To see how the quantifier use changes with respect to the domain size (i.e., the number of objects in the scene), during the experiment, we test three different domain sizes (i.e., 4, 9, and 20) The experiment yielded the QTUNA corpus. By analysing the corpus, we found that all the completeness, the correctness, and the frequency of crisp quantifiers are reduced with respect to the rise of domain size. We also found that there is no clear correlation between the length of QDs.

Subsequently, we conducted the same experiments on Mandarin speakers, which yields the MQTUNA corpus. All the conclusions we made in the QTUNA experiment were still held in MQTUNA. To assess the coolness hypothesis, we compare the completeness, the length, and the use of vague quantifiers of QDs in MQTUNA and in QTUNA. We found that Mandarin QDs are less complete and use more vague quantifiers than English QDs which are in line with the fact that Mandarin is "Cooler" than English. Nevertheless, inconsistent with the Coolness hypothesis, QDs in MQTUNA are generally longer than those in QTUNA.

Building our findings in the elicitation experiments, we built two algorithmic production models that aim at mimicking the language production behaviours of human beings, namely `qdg-greedy` and `qdg-ia`. We then evaluated these algorithms on producing English QD in two alternative ways: human judgement (i.e., asking readers to judge whether a QD is natural or not) and re-construction (asking readers to reconstruct the scene given a QD). The evaluation results suggested our algorithms produce descriptions that are both natural and useful. At length, we listed some issues that need to be aware when applying our algorithms to Mandarin QDG.

CHAPTER 7

Surface Realisation

Abstract. *We aim at realising noun phrases in Mandarin. In this chapter, we report three studies about Mandarin linguistic realisation. In the first study, we introduce simpleNLG-ZH, a realisation engine for Mandarin that follows the software design paradigm of simpleNLG. We explain the core grammar (morphology and syntax) and the lexicon of simpleNLG-ZH, which is very different from English and other languages for which simpleNLG engines have been built. We, then, evaluate the coverage and correctness of simpleNLG. In the second study, we zoom in on one realisation module, namely, classifier selection. In addition to the dictionary-based classifier selector in simpleNLG-ZH, we explore several data-driven alternatives. In the last study, we conduct a human experiment to assess how hard the task of classifier selection is for human beings.*

—

The publications related to this chapter are:

1. Chen, G., van Deemter, K., & Lin, C. (2018). SimpleNLG-ZH: A linguistic realisation engine for Mandarin. *Proceedings of the 11th International Conference on Natural Language Generation*, 57–66. https://doi.org/10.18653/v1/W18-6506

2. Jarnfors, J., Chen, G., van Deemter, K., & Sybesma, R. (2021). Using BERT for choosing classifiers in Mandarin. *Proceedings of the 14th International Conference on Natural Language Generation*, 172–176. https://aclanthology.org/2021.inlg-1.17

7.1 Introduction

In this chapter, we study the surface realisation related issues for Mandarin. The surface realisation is one of the major components in the classic natural language generation pipeline (see §2.1). From a viewpoint of practical NLG, surface realisation is the module responsible for mapping information produced by earlier components to well-formed output strings in the target language (Reiter & Dale, 2000). More specifically, it employs language-specific morpho-syntactic constraints to achieve proper word ordering, inflection, and selection of function words. From a linguistic viewpoint, surface realisation embodies our understanding of grammar and morphology. Naturally, building a complete surface realisation module for Mandarin is beyond the aim of this thesis. Our aim is twofold. On the one hand, we wanted to build a surface realisation system that is able to express, in at least one way, all the information that the other NLG modules on which we have worked have accumulated. For example, if previous modules have decided to refer to an object using a definite NP that ascribes 3 properties to the referent (e.g. "*it is a table, it is red, and it is large*"), then our surface realisation modules should be able to produce at least one well-formed NP that does this (e.g., in English, "*the large table that is red*"). On the other hand, we want to highlight some of the surface realisation decisions that are particularly difficult to make for Mandarin, such as the choice of classifiers.

Different types of realisers exist (Gatt & Krahmer, 2018). One line of work aims primarily for linguistic depth and coverage by acquiring probabilistic grammar from large corpora. For example, OPENCCG (White et al., 2007) built a grammar bank based on Combinatorial Categorial Grammar, extracted from the Penn Treebank (Marcus et al., 1993). When realising, OPENCCG applies a chart-based algorithm to generate all possible surface forms, which are then re-ranked by language models. Another line of work aims primarily for ease of use and extendibility. This includes one of the most popular realisation engines: simpleNLG (Gatt & Reiter, 2009). simpleNLG performs linearisation and morphological inflection by means of human-crafted grammar-based rules. Unlike the realiser following the first strategy, it is more controllable and extendable because it follows the principle of keeping a clear separation between morphological and syntactic operations. This may explain why SimpleNLG is more popular in practical applications. It has become the realisation method of choice in many practical NLG applications, such as BabyTalk (Portet et al., 2009) and Absum (Lapalme, 2013). To date, it has been adapted to German (Bollmann, 2011; Braun et al., 2019), French (Vaudry & Lapalme, 2013), Portuguese (de Oliveira & Sripada, 2014), Italian (Mazzei et al., 2016), Spanish (Ramos-Soto et al., 2017), Filipino (Ong et al., 2011), Telugu (Dokkara et al., 2015), and Tibetan (Kuanzhuo et al., 2020).

Therefore, in the first study of this chapter, we attempt to build a surface realiser based on the tradition of simpleNLG, propose simpleNLG-ZH ("Zhongwen" is Mandarin for "Chinese", G. Chen et al., 2018c), and evaluate the coverage and correctness of simpleNLG-ZH by means of the unit test as well as human evaluation. We started with focusing on realising NPs and then extended it to cover other constructions and phenomena in Mandarin. Before simpleNLG-ZH, there was no such adaptation work yet for Sinitic languages. Note that, there have been two Mandarin realisers following different traditions other than simpleNLG. One is the KPML (G. Yang & Bateman, 2009), a large-scale multilingual generation and development. It supports limited sentence structures in Mandarin (G. Yang & Bateman, 2009). He et al. (2009) introduced a data-driven generator, with dependency trees as input. They used divide-and-conquer to break the dependency tree into sub-trees, realising each sub-tree using a log-linear model recursively. However, their system needs a large amount of fully inflected dependency trees as training data.

Although we took existing simpleNLG systems as a source of inspiration, the system is, in many ways, a re-design. [1] For example, the morpho-syntactic structure of Mandarin is very different from the languages that previous simpleNLG has covered. Building on this, as a highly *analytical* language (see §3.2 for more details), Mandarin needs far fewer morphological operations but many more syntactic constraints than English (C.-T. J. Huang

[1] The German, Portuguese, and Spanish simpleNLG systems copied many features from the one for English (in the case of German) or French (in the other two cases).

et al., 2009).

As aforesaid, we are interested in some of the Linguistic Realisation decisions that are particular to Mandarin and that are, therefore, not covered by previous simpleNLG. One typical example is that the grammar of Mandarin requires that, in certain syntactic positions, a noun must be preceded by a *classifier*. Classifiers often give a rough indication of the kind of entity denoted by the noun (for more details about the grammar of classifiers see §A). For example, the classifier “只” (zhǐ) in the NP “一只狗” (yìzhīgǒu; *a dog*) indicates the head noun “狗” (gǒu; *dog*) is an animal. Mandarin contains a large number of classifiers. In the primary version of simpleNLG-zh, a classifier is inserted by looking up a dictionary, where each entry is a classifier-noun pair since the choice of classifiers is limited by the (head) noun with which the classifier is associated. However, this may still leave several options, which may sometimes produce a different meaning, for example,

(83) a. 一 个 电脑 / 一 台 电脑
yí gè diànnǎo / yí tái diànnǎo
‘a computer’

b. 一 个 老师 / 一 位 老师
yí gè lǎoshī / yí wèi lǎoshī
‘a teacher’

c. 一 个 人 / 一 群 人
yí gè rén / yí qún rén
‘a person / a group of people’

d. 一 杯 咖啡 / 一 听 咖啡
yì bēi kāfēi / yì tīng kāfēi
‘a cup/can of coffee’

Although each of these cases involves classifier choice, the problem of choosing a classifier is likely to be more challenging in those cases, such as (83-b)-(83-d), where the classifier adds information, for example, in terms of politeness ((83-b), neutral vs. polite), number ((83-c), singular vs. plural), or quantity ((83-d), a cup vs. a can of coffee). This is perhaps clearest in the case of (83-d), where “杯” (bēi; *cup*) and “听” (tīng; *can*) indicate different containers, and consequently different quantities, of coffee; these classifiers are known as measure words, as opposed to the “pure” classifiers of (83-a)-(83-c). simpleNLG-zh selects classifiers on the basis of a dictionary, where each noun is corresponding with a single classifier. However, clearly, choices as such cannot be accomplished by a dictionary-based approach. To explore classifier selection more closely, we conduct another study (i.e., the second study of this chapter), in which we try several data-driven approaches on this task and evaluate them on a large scale classifier selection dataset.

Nevertheless, the way we accomplish the classifier selection is to ask a model to decide the most proper classifier given its context (which is what a surface realiser should do). This is slightly different from how human beings produce classifiers. We are, therefore, curious, how hard this task is for human beings. To investigate this, we conduct a study asking human participants to accomplish the exact same task as realisers do, evaluate the outcomes, and compare the performance of participants with that of the models we tried in the second study.

To sum up, in this chapter, we conduct three studies. In the first study, we introduce simpleNLG-zh. It was developed as an adaptation from V4.4.8 of the original English

```
Phrase s1 = new SPhraseSpec(’leave’);
s1.setTense(PAST);
s1.setObject(new NPPhraseSpec(’the’, ’house’));
Phrase s2 = new StringPhraseSpec(’the boys’);
s1.setSubject(s2);
```

Figure 7.1: A simpleNLG-EN code snippet for realising the sentence *The boys left the house*.

```
Phrase s1 = new SPhraseSpec(’离开’);
s1.setParticle(’了’);
s1.setObject(new NPPhraseSpec(’房子’));
Phrase s2 = new NPPhraseSpec(’男孩’);
s1.setSubject(s2);
```

Figure 7.2: A simpleNLG-ZH code snippet for realising the sentence “男孩离开了房子”.

SimpleNLG[2] (abbreviated as simpleNLG-EN in the rest of this chapter). We show that SimpleNLG-ZH has wide coverage on test sentences, and on the human authored corpus MTUNA (van Deemter et al., 2017) as well. In the second study, on a large scale classifier selection corpus, we compare a number of classifier selectors, including not only the traditional rule-based methods but also the most recent deep learning-based methods. In the last study, we investigate how hard the task of classifier selection is for Mandarin speakers.

7.2 Study 1: Constructing a Mandarin Realisation Engine

We build simpleNLG-ZH on the basis of the simpleNLG framework. In this section, we start by introducing the simpleNLG framework. Subsequently, we introduce the operations in simpleNLG-ZH and evaluate simpleNLG-ZH.

7.2.1 The simplenlg Framework

simpleNLG is a realisation engine designed for practical use. The input format of simpleNLG is similar to a simplified dependency tree where the user should determine the specifiers, modifiers and complements of each input phrase using a set of features. simpleNLG encodes different constraints, regarding lexicon, morphology, syntax and orthography, as a feature set (combining the features from the input) and passes the resulting structure onto the next stage. Figure 7.1 and Figure 7.2 are two code snippets showing examples of an input for simpleNLG-EN and simpleNLG-ZH for generating the sentence “男孩离开了房子” (nánhái líkāile fángzi; *The boys left the house*), respectively. To

[2] The software is available at: https://github.com/simplenlg/

construct a sentence using simpleNLG, we need to establish a verb phrase object and set its object(s) and subject.

simpleNLG follows good software engineering design principles, clearly separating the modules for lexical and syntactic operations. The lexical component provides interfaces that handle the lexical features and apply morphological rules. Vital features such as `person`, `number` and `tense` are appended to target constituents or words for further realisation processes. The syntactic component takes over at the phrase and clause level, and provides `Java` classes for each phrasal sub-type (`PhraseSpecs`), where `SPhraseSpec` stands for the class that model clauses.

simpleNLG-EN offers significant coverage of English morphology and syntax and provides easy-to-use APIs with which the realisation process is programmatically controllable. It provides a well-established lexicon, the repository of the relevant items and their properties. The lexicon was constructed from the NIH specialist lexicon[3], which contains more than 300,000 entries. Each lexical entry was tagged with detailed lexical features as initial features of words. Simple shallow semantic features, like `COLOUR` and `QUANTITATIVE`, are appended for deciding word order.

7.2.2 Morphology Operators

Morphology in Mandarin is usually thought to be extremely simple (Jensen, 1990). Packard (2000) has challenged this view, arguing that more morphological operations are involved in the construction of Chinese words than is usually thought, which include, for example, word compounds (see §A for more details). However, key mechanisms such as subject-verb agreement (which is treated by simpleNLG-EN as part of morphology operations) are absent from Mandarin. We have therefore sided with mainstream linguistic opinion and kept our morphology component relatively simple. We use only two main rules for morphology: (1) mapping pronouns to their surface forms; and (2) appending the collective marker “们” (mén).

Pronoun

Realising the surface forms of pronouns in simpleNLG-ZH is similar to simpleNLG-EN in its use of the features `gender` (masculine, feminine or neuter), `number` (singular or plural), and `person` (first, second or third). However, written Mandarin has different *third person plural* forms for all three different genders, i.e., “他们” (masculine), “她们” (feminine) and “它们” (neuter) (all of them have the same pronunciation: *tāmén*) rather than the one plural form *they* in English.

Collective Marker

In Mandarin, to say how many entities there are in a set, *classifiers* must be used. This is typically done in a *number phrase* of the form [number + classifier + noun], for instance “一把椅子” (yìbǎyǐzi; *a chair*), “两张桌子” (liǎngzhāngzhuōzi; *two tables*). Since number phrases are typically used referentially (not as quantifiers), they have generally been regarded as indefinite expressions, and these cannot be placed in subject or topic position in Mandarin (C.-T. J. Huang et al., 2009).

[3] The lexicon of simpleNLG-EN can be found at https://github.com/simplenlg/simplenlg/blob/master/src/main/java/simplenlg/lexicon/default-lexicon.xml

```
NPPhraseSpec book =
    this.phraseFactory.createNounPhrase("一", "本", "书");
```

Figure 7.3: A simpleNLG-ZH code snippet for realising the phrase "一本书".

Unlike English, Mandarin bare nouns and number phrases with numbers larger than 1 can express plural meaning without the help of inflected plural markers. The morpheme "们" in plural nouns serves as a "collective" marker rather than a traditionally plural marker (Y.-h. A. Li, 2006); here a "plurality" is a number of individuals, whereas a "collective" is a group (of individuals) as a whole. Under that definition, adding a morpheme "们" makes a nominal phrase definite, which results in the morpheme "们" incompatible with a number phrases, so "们" cannot co-occur with number phrases. For example, the following phrase is not acceptable in Mandarin:

(84) 三 个 人们
sān gè rénmén
'three people'

Note that the rules discussed above do not apply to pronouns that follow the rules defined above.

It is hard to determine automatically whether a user wants to talk about a number of individuals or about a group as a whole. Moreover, "们" is always only optional. Therefore, in simpleNLG-ZH, "们" is only added if the feature `MEN` is set to true. In addition, the system will refuse to add a "们" to a number phrase. The way of constructing number phrases is discussed in §7.2.3.

7.2.3 Syntactic Operators

The syntax module inherits the basic structure of simpleNLG-ZH, dividing the syntactic operations into processors that handle NPs, adjective phrases, verb phrases, verb phrases, and clauses. Each processor is enriched based on the grammar of Mandarin. In this section, we start with introducing the processor for the focus of this thesis: NP, and then discuss other types of phrases.

Noun Phrase

The Noun Phrase module is the most complex phrase module in simpleNLG-ZH. Building on the grammar of Mandarin (see Appendix A), each noun phrase in simpleNLG-ZH contains multiple specifiers, pre-modifiers, post-modifiers, complements, and a head noun.

Number Phrase. Each number phrase is constructed by a number, a classifier and a head noun; both the numeral and the classifier function as *specifiers* of the NP (for more about specifiers, please see below).

As Number Phrases are very common in Mandarin, we designed a new constructor specifically for them. For instance, the number phrase "一本书" (yìběnshū; *a book*) can be constructed using the code in Figure 7.3. The choice of classifiers depends mainly on the head noun. Given a head noun, the current simpleNLG-ZH retrieve the corresponding classifier from a pre-defined dictionary. Additionally, as discussed in this chapter, for a given noun, the choice of classifiers may depend on its meaning. For example, the classifier of “房子” (fángzi; *house*) can be “座”, “幢”, “间”, and many other possible classifiers based on the size or the shape of the house. Therefore, simpleNLG-ZH also allow users to specify classifiers “by hand”. Attempts on automating this process by data-driven methods can be found in §7.3.

Specifier. simpleNLG-ZH allows multiple specifiers (compared to a single specifier in simpleNLG-EN) within one NP. For example, a number phrase needs two specifiers: a numeral and a classifier. All the following categories can be placed in specifier position: pronouns (with or without the collective marker “men"), proper names, classifiers, numerals and demonstratives. These specifiers appear in the following order (the $A > B$ means A should appear before B): `proper name` > `pronoun` > `demonstrative` > `numeral` > `classifier`. The decision of whether or not to realise each of these specifiers is subject to a number of constraints (C.-T. J. Huang et al., 2009).

1. Suppose the input specification asks for a pronoun in the specifier position. This pronoun must have a collective marker except in a structure that includes [demonstrative/numeral + classifier]. For instance, (85-a) contains the collective marker, but (85-b) does not;

   (85) a. 他们 学生
   tāmén xuéshēng
   ‘them students’
   b. 他 一 个 学生
   tā yígè xuéshēng
   ‘them students’

2. Proper names in specifier position can only be realised if the structure includes [pronoun + numeral + classifier], [demonstrative + classifier] or [demonstrative + numeral + classifier], for example:

   (86) 张三 那个 学生
   zhāngsān nàgè xuéshēng
   ‘the student called Zhangsan’

3. A demonstrative or a numeral will only be realised if there is a classifier in the same NP and vise versa:

   (87) (那/一) 个 学生
   (nà/yí) gè xuéshēng
   ‘that/a student’

As discussed in §7.2.2, number phrases are often seen as indefinite phrases but not always. When they are for quantification they can be placed in the subject/topic position. Therefore, simpleNLG-ZH permits a number phrase in the subject/topic position, e.g.,

(88) 三 个 人 吃 两 块 蛋糕
sān gè rén chī liǎng kuài dàngāo
'three people eat two piece of cakes'

For nouns (including bare nouns, pronouns and proper nouns), the feature possessive is also realised in the specifier position: simpleNLG-ZH adds a particle "的" (de) as an associative marker after the noun.

Localiser. Localisers (corresponding to English words such as "on", "above", etc.) form a special syntactic category. They are used in *location phrases*, which is a particular type of preposition phrases. The location information in a location phrase is expressed in the localiser rather than the head preposition. For example, in (89), localiser "上" (*on*) works as a supplement of the noun phrase in the proposition phrase (i.e., location phrase).

(89) $[_{PP}$ 在 $[_{NP}$ 桌子 上$]]$
zài zhuōzi shàng
'on the table'

In simpleNLG-ZH, the localiser itself is defined as a normal noun with a lexical feature LOCATIVE in the lexicon. When constructing a location phrase, if the localiser is a disyllabic word, such as "上面" (shàngmiàn), then a particle "的" is inserted before the localiser to construct the phrase (90-a). However, if such a prepositional phrase works as a pre-modifier of another noun, then that inserted particle will be disregarded, such as (90-b).

(90) a. 在 桌子 的 上面
zài zhuōzi shàngmiàn
'on the table'
b. 在 桌子 上面 的 书
zài zhuōzi shàngmiàn de shū
'the book on the table'

Pre-modifier. simpleNLG-EN handles the orders of multiple pre-modifiers based on their meanings, where the meanings are acquired from a huge lexicon that contains a series of tags (e.g., COLOUR, QUANTITATIVE) indicating the meaning of words. It adds pre-modifiers in the order of quantitative adjectives, colour adjectives, classifying adjectives and nouns. For simpleNLG-ZH, more categories of words can be placed in the pre-modifier position, other than just adjectives and nouns. It performs re-ordering based on pre-modifiers' part-of-speech and lexical features set by the users.

Our system handles two different types of adjectives, namely, predicative adjectives and non-predicate adjectives. For predicative adjectives, the system will automatically add a "的" (de) between the adjectives and the head noun, such as "绿的椅子" (lǜ de yǐzi; *green chair*). "的" can be omitted by setting the feature NO_DE to TRUE, which results in the phrase "绿椅子" (lǜ yǐzi; *green chair*). We leave whether to add a "的" for the users of simpleNLG-ZH because such an choice is not subject to strict rules and is with a certain degree of variation (Paul, 2010). Non-predicate adjectives, in contrast to predicative adjectives, are a special type of adjectives that cannot function as predicate on their own (e.g., “男” (ná; *male*) and “女” (nǚ; *female*)), in which the particle “的” (de) is always omitted. Thus, the particle “的” will not be appended if the adjective is non-predicate, such as “男人” (nánrén; *man*). The feature is set based on the information of the lexicon loaded into simpleNLG-ZH (details see §7.2.4).

Nouns and noun phrases, as pre-modifiers, can play two different roles: they can be concatenated with the head noun to construct a compound noun: for example, (91-a); or, they can be connected by means of a particle “的”, which works as an associative marker: for example, (91-b). To construct the latter, the feature `ASSOCIATIVE` should be set to `TRUE`. The order of the pre-modifiers is `localisers` > `verbs/clauses` > `adjectives with de` > `nouns with associative marker` > `adjectives without de` > `non-predicate adjectives` > `nouns`.

(91) a. 大学 教育
dàxuē jiàoyù
‘university education’
b. 黑 头发 的 人
hēitóufà de rén
‘the man with black hair’

Adjective Phrase

Adjective phrases in Mandarin differ from those in the languages for which previous simpleNLG engines were built. Most adjectives in Mandarin can act as the predicate of a clause without the help of a copula verb (see below). Such adjectives are called predicate adjectives.

Predicate Adjective. Although adjectives can act as predicates, it is necessary to distinguish them from verbs (C.-T. J. Huang et al., 2009). We implemented the realisation of a clause like (92-a) by specifying an empty copula. This is achieved by creating a new constructor which accepts a subject noun and a predicate adjective.

Predicate adjectives in simpleNLG-ZH also accept negative words and modal words. For example, the sentence (92-b) has both a negative word “不” (bù; *not*), and a modal word “应该” (yīnggāi; *could*).

(92) a. 他 很 高
tā hěn gāo
‘he is very tall’
b. 他 应该 不 高
tā yīnggāi bù gāo
‘he couldn’t be tall’

Non-predicate Adjective. As discussed when we introduced pre-modifiers, non-predicate adjectives always omit the particle “的” between the adjective and the head noun. However, when a non-predicate adjective functions as a predicate (with the help of a copula), such as

”他 是 男的” (tā shì nánde; *he is a man*), the copula “是” (shì; *is*) and the particle “的” (de) are obligatory (Paul, 2010).

“比” construction. In English, degree adjectives have comparative and superlative degrees, whose realisation is implemented in the morphology processor. In Mandarin, realisation is performed by modifying the syntax. The superlative degree is realised by adding an adverb pre-modifier “最” (zuì; *most*); the comparative is realised by the “比” (bǐ) construction.

simpleNLG-ZH implements the “比” construction as a prepositional phrase. For example, for the sentence (93-a), the word “比” itself is seen as the head of a preposition phrase, which is a pre-modifier of an adjective phrase. Such a construction (i.e., as an adjective phrase) can act as the pre-modifier of a noun phrase, for example, (93-b). Note that the head of this noun phrase can be omitted, but the particle “的” (de) should be maintained as a sentence-final marker, i.e. (93-c).

(93) a. 他 比 小明 高
tā bǐ xiǎomíng gāo
‘he is taller than Xiaoming’
b. 他们 班 没有 比 他 更高 的 人
tāmén bān méiyǒu bǐ tā gènggāode rén
‘none of his classmates is taller than he’
c. 他们 班 没有 比 他 更高 的
tāmén bān méiyǒu bǐ tā gènggāode
‘none of his classmates is taller than he’

Verb Phrase

Pre-modifier and Post-modifier. Verb phrases can contain the associative markers “得” (dé) and “地” (dè). The latter is appended to the pre-modifier if it is disyllabic, for example, “快速地跑” (kuàisù de pǎo; *fast run*). If the pre-modifier is monosyllabic, “快跑” (kuàipǎo; *fast run*) is constructed instead, with the particle “地” disregarded. The particle “得” connects head verbs with their complements: “跑得快” (pǎodekuài; *running fast*).

Aspect. KPML (G. Yang & Bateman, 2009) used templates with particles like “过” (guò), “了” (zhě) or “着” (zhe)to model aspect. However, KPML’s coverage of language variation is limited because it uses a limited number of templates. Since aspect in Mandarin is realised using post-verbal or post-clause particles, we took a more flexible strategy that enables users to add particles based on their need.

Particles can be in two positions: post-verbal and post-clausal. In “他吃着饭” (tā chīzhe fàn; *he is eating*), the particle “着” (zhe), which is a aspectual durative marker, is appended to a `VPPhraseSpec` object. Similarly, the class `SPhraseSpec`, which represents a clause, has the capability to append a particle to its end. For example, in “他吃饭了” (tā chī fànle; *he has eaten*), the particle “了” is appended to the clause “他吃饭” (tāchīfàn; *he eats*).

Clause

At the Clause level, apart from the issues related to negative and interrogative sentences inherited from simpleNLG-EN, we considered "把" (bǎ) and "被" (bèi) constructions which are two common constructions in Mandarin. We also discuss how topicalised sentences are realised using simpleNLG-ZH.

Negative Sentence. Negative sentences in simpleNLG-ZH are realised by inserting negative words before the predicate verb (or the predicate) and after a modal word. For example, the negation of (94-a) is the sentence with an inserted negative word "不" (bù; *not*) before "去" (qù; *go*) and after the modal word "应该" (yīngāi; *should*) resulting (94-b). simpleNLG-ZH can also realise negative modal by viewing the negative modal as a merged word, much like *haven't* or *shouldn't* in English (D. Xu, 1997), for example, the sentence (94-c).

(94) a. 他 应该 去 上学
tā yīnggāi qù shàngxué
'he should go to school'
b. 他 应该 不 去 上学
tā yīnggāi bū qù shàngxué
'he should haven't gone to school'
c. 他 不 应该 去 上学
tā bū yīnggāi qù shàngxué
'he should not go to school'

In addition, Mandarin has a number of different negative words, selected based on the head verb. For example, applied to the sentence "他有椅子" (tā yǒu yǐzi; *he has chairs*), instead of using "不" (bù), the word "没" (méi) should be used: "他没有椅子" (tā mēiyǒu yǐzi; *he doesn't have a chair*). simpleNLG-ZH allows users to specify by hand what negation word should be chosen in a specific case by using the feature `negative_word`, thus overruling the system's default choice.

"把" Construction. The "把" construction is a commonly seen and useful structure for focusing on the result or influence of an action, which does not exist in English. For example, considering the sentence (95-a), with the "把" construction, the influence of "打" (dǎ; beat) is highlighted. The natural phrase order of this example is the sentence (95-b), which is the basic structure that simpleNLG-ZH can handle, i.e., [`subject` + `predicate verb` + `object`]. In the "把" construction, however, the marker adverb "把" is added after the subject, and the object is moved to the position right before the predicate verb phrase: [`subject` + "把" + `object` + `predicate verb`].

Note that the positions of modal words and negative words do not follow the movement of the verb phrases (Y. Liu et al., 2001). In other words, in the resulting "把" construction, the modal words and negative words are placed before the object in their own order, as in (95-c). simpleNLG-ZH realises a sentence with the "把" construction if the user set the feature `BA` to `TRUE`.

(95) a. 他 把 小明 重重 地 打
tā bǎ xiǎomíng zhòngzhòng de dǎ

‘he beat Xiaoming heavily’

b. 他 重重 地 打 小明
tā zhòngzhòng de dǎ xiǎomíng
‘he beat Xiaoming heavily’

c. 他应该没把小明打疼
tā yīnggāi méi bǎ xiǎomīng dǎ téng
‘he should haven’t beaten xiaoming heavily’

“被” Construction. The “被” construction in Mandarin is one of the ways to express the passive, using the basic syntactic structure: [`object` + “被” + `subject` + `predicate verb`]. Using the same example for the “把” construction, the transformed sentence would be “小明 被 他 重重 地 打” (xiǎomīng bèitā zhòngzhòng de dǎ; *Xiaoming is beaten heavily by him*). simpleNLG-ZH chooses between active and passive based on the value of the feature `PASSIVE`, which is inherited from simpleNLG-EN.

Interrogative. simpleNLG-ZH inherits and adapts all its interrogative patterns from simpleNLG-EN, including “有没有” (yǒuméiyǒu; *Yes-or-no*) and wh-questions: “怎么” (zěnmè; *How*), “什么” (shénmè; *What*), “哪里” (nǎlǐ; *Where*), “谁” (shuí; *Who*)， “为什么” (wèishénmè; *Why*)， “多少” (duōshǎo; *How Many*). simpleNLG-ZH adds two further types, namely “哪个” (nǎgè; *Which*) and “什么时候” (shénmèshíshòu; *When*). For Yes-or-no sentences, simpleNLG-ZH appends the interrogative particle “吗” at the end of a sentence; for instance, “你去上学吗？ ” (nǐ qù shàngxué ma; *Will you go to school?*).

In simpleNLG-EN, for wh-questions, only *What* and *Who* made a difference between placing the interrogative marker in subject and object position. In simpleNLG-ZH, however, nearly all wh-question markers can be placed in both positions. Here we use a “什么” (*What*) sentence as an example: For the sentence (96-a), if we set the feature `INTERROGATIVE_TYPE` to `what_object`, then the sentence is changed to (96-b). Setting the feature to `what_subject` results in (96-c). In interrogated “把” constructions and ‘被” constructions, the wh-question markers are placed *in situ*, i.e., replacing the phrases in the original subject or object position, according to the value of `INTERROGATIVE_TYPE`.

(96) a. 台风 摧毁 了 他 的 房子
táifēng cuīhuǐ le fángzi
‘the typhoon destroyed his house’

b. 台风 摧毁 了 什么
táifēng cuīhuǐ le shénme
‘what did the typhoon destroy?’

c. 什么 摧毁 了 他 的 房子
shénme cuīhuǐ le tādefángzi
‘what destroyed his house’

Topicalisation. Topic structures, especially gapped topic structures, are a very common syntactic structure in Mandarin (L. Xu & Langendoen, 1985). For example, (97-a) is a gapped topicalised sentence, in which the constituent after the “的” in the phrase ”那把大号的” (nàbǎ dàhào de; *the large one*) moved into the topic position and left a gap.

Lexical Category	Universal POS Tag
adverb	ADV, PART
noun	NOUN, PROPN
preposition	ADP
demonstrative	DET
conjunction	SCONJ, CCONJ
pronoun	PRONOUN
adjective	ADJ
modal	AUX
verb	VERB

Table 7.1: Relationship between Universal POS tags and lexical categories in simpleNLG-ZH

In the current version of simpleNLG-ZH, we realise a gapped topicalised sentence by viewing it as two coordinated noun phrases, in which the second noun phrase has an empty head noun. For (97-a), the two noun phrases are (97-b) and (97-c). In the current version of our system, there is no guarantee that the empty head of the second clause is bounded by the first clause. We also consider orthography in topicalisation, i.e., a conjunction of words between two phrases should be changed to a comma. In our system, the topicalised sentence, as a `CoordinatedPhraseElement` object, calls the `topicalise()` function to take care of the punctuation.

(97) a. 绿色的椅子，那把大号的
lǜsè de yǐzi, nà bǎ dàhào de
'(as for) the green chair, it is the large one'

b. 绿色的椅子
lǜsè de yǐzi
'the green chair'

c. 那把大号的
nàbǎ dàhào de
'the large one'

7.2.4 Lexicon

Unlike simpleNLG-EN, we did not have a ready-to-use elaborate lexicon for simpleNLG-ZH. Instead, we extracted a primary lexicon from the Chinese as a Foreign Language (CFL) corpus[4] (J. Lee et al., 2017), which is a sub-corpus of the Universal Dependencies corpus. The CFL corpus has 451 human tagged dependency trees and 7,256 tokens in total. Each word in CFL was primarily mapped to one of the lexical categories in simpleNLG-ZH based on the relations in Table 7.1 as well as the following rules:

1. The tag `<proper/>` is appended for `PROPN`s;
2. The tag `<nonpredicate/>` is appended for non-predicate adjectives manually, which is based on the non-predicate adjective list in Y. Liu et al. (2001);

[4] The dataset is available at https://github.com/UniversalDependencies/UD_Chinese-CFL/tree/master

3. The tag `<locative/>` is appended for localisers manually;
4. The words that serve as a dependent of a `clf` (classifier) dependency relation are given the category `classifier`.

The constructed lexicon has 1,639 lexical entries in total.

7.2.5 Evaluating SimpleNLG-ZH

To assess the coverage and the correctness of simpleNLG-ZH, we decided to evaluate it in two ways. Firstly, following Ramos-Soto et al. (2017) and Bollmann (2011), we applied a set of unit tests to each module of the system, using the test cases from simpleNLG-EN plus a set of newly constructed test cases that address some of the peculiarities of Mandarin (e.g., the "把" construct).

Secondly, we evaluated the system using a set of expressions from a corpus of actual language use; this was reminiscent of Mazzei et al. (2016) and Bollmann (2011), but using a larger set of expressions. In all cases, when faced with an input expression (i.e., from a test set or corpus), we used this expression to construct a formatted input that was then passed to simpleNLG-ZH to produce an output expression which was then compared to the input expression.

Evaluation with Tests Cases

The test cases consist of 144 sentences manually translated and adapted from simpleNLG V4.4.8 `JUnit Tests` and two reference grammar books (C.-T. J. Huang et al., 2009; Y. Liu et al., 2001). The test cases cover all the linguistic features discussed in previous sections and all possible syntactic structures of referring expressions in Mandarin introduced in van Deemter et al. (2017). All the tests were passed by simpleNLG-ZH, that is, the generated sentences were all identical *verbatim* to the inputs.

Corpus-based Evaluation.

We picked 100 noun phrases at random from the MTUNA corpus (van Deemter et al., 2017), which is the corpus that the first version of simpleNLG-ZH focuses on and also which this thesis focus on (i.e., NPs). MTUNA is a corpus that has totally 1,650 referring expressions. We then re-generated these expressions using simpleNLG-ZH. Not all re-generated NPs were identical verbatim to the original MTUNA NPs. 35 noun phrases did not match completely (i.e., *verbatim*) with the original noun phrases. Table 7.2 lists some typical examples, showing differences in word ordering, punctuation, and so on. We ran a human evaluation to find out whether the realised sentences were acceptable (i.e., are they fluent and do they have the same meaning as their inputs). Two native speakers annotated the outputs; they reached a good inter-annotator agreement ($\kappa = 0.77$) and were asked to produce a consensus annotation, which was then used for our evaluation. It turned out that 90 out of 100 sentences were judged to be acceptable, which we consider a very encouraging result.

We classified the unmatched sentences into three types. The first one is where punctuation was different, as in Example 1 in Table 7.2. The reason is that some sentences used commas to separate modifiers but simpleNLG-ZH does not. These cases were generally judged to be acceptable.

T	ID	Noun Phrases from MTuna	Realised Sentence	Acc.
1	1	黑头发，络腮胡，黑西服，浅色衬衣 hēitóufà, luòsāihú, hēixīfú, qiǎnsèchènyī *black hair, whiskers, black suit and light shirt*	黑 头发 络腮 胡 黑 西服 浅色 衬衣 hēitóufà luòsāihú hēixīfú qiǎnsèchènyī *black hair, whiskers, black suit and light shirt*	Yes
2	2	一张大的红色的沙发 yìzhāng dà de hóngsè de shāfā the large red sofa	一 张 红色 的 大 的 沙发 yìzhāng hóngsè de dà de shāfā the red large sofa	Yes
	3	戴眼镜的两个人 dài yǎnjìng de liǎng gè rēn *the two people who wear glasses*	两 个 戴 眼镜 的 人 liǎng gè dài yǎnjìng de rēn *the two people who wear glasses*	Yes
	4	红色正面朝向屏幕小椅子或者绿色背向屏幕的大风扇 hóngsè zhèngmiàn cháoxiàng píngmù xiǎo yǐzì huòzhě lǜsè bèixiàng píngmù de dà fēngshàn *the red fronting small chair and the green backing large fan*	正面 朝向 屏幕 小 红色 椅子 或者 背向 屏幕 的 绿色 大 风扇 zhèngmiàn cháoxiàng píngmù xiǎo hóngsè yǐzì huòzhě bèixiàng píngmù de lǜsè dà fēngshàn *the fronting red small chair and the backing green large fan*	No
	5	黑色头发戴眼镜的 hēsè tóufà dài yǎnjìng de *the person with black hair and glasses*	戴 眼镜 的 黑色 头发 dài yǎnjìng de hēsè tóufà *the person with glasses and black hair*	No
3	6	红色椅子，椅子背朝向右边，可以看到椅子背的正面 hóngsè yǐzì, yǐzìbèi cháo yòubiān, kěyǐ kàndào yǐzìbèi de zhèngmiàn *It is a red chair whose back is facing right and we could see the front of its back.*	*(failed)*	No
	7	正朝向我们的小的椅子和正朝向我们的大的风扇 zhèng cháoxiàng wǒmén de xiǎo de yǐzì hé zhèng cháoxiàng wǒmén de dà de fēngshàn *the small chair facing us and the large fan facing us*	正 朝向 我 的 小 的 椅子 和 正 朝向 我 的 大 的 风扇 zhèng cháoxiàng wǒ de xiǎo de yǐzì hé zhèng cháoxiàng wǒ de dà de fēngshàn *the small chair facing me and the large fan facing me*	No

Table 7.2: Example sentences (with their Pinyin and translations) that were not identical to the inputs from MTUNA *(unmatched sentences)*. The last column says whether the output was judged to be acceptable by our annotators. T and Acc. represents "type" and "acceptable", respectively.

The second type is where the word order of the realised sentences was different from the input. There are three sub-types:

- The order of adjective pre-modifiers was different, as in Examples 2 and 4. Most of these deviations were judged to be acceptable, but sentence 4 shows an unacceptable example, where the word "红色" (hóngsè; *red*) before "小" (xiǎo; *little*) accidentally produced a new word, "小红色" (*light red*), which has different meaning;

- simpleNLG-ZH enforces the pre-modifiers to appear following the specifiers. However, in the MTUNA corpus, there are expressions, like Example 3, that switch the place of specifiers and pre-modifiers. All such re-orderings were judged to be acceptable;
- There is a special syntactic pattern of noun phrases in Mandarin, where a Noun is omitted that is recoverable from the context. For example, in Example 5, the head is omitted in the original sentence to construct a free relative (Teng, 1979) where the particle "的" works as a sentence-final marker. However, simpleNLG-ZH cannot recognise the functionality of the particle, thus it switches two pre-modifiers according to the orders defined in §7.2.3, which results in a noun phrase with a different meaning. We found 6 unacceptable cases of the second type.

simpleNLG-ZH failed to reproduce some types of language use that are highly colloquial and not strictly grammatical. We found 4 such cases, as in Example 6 in Table 7.2, and in Example 7, where the pronoun "我们" (wǒmén; *us*) in the sentence actually refers to the subject himself (but using the plural form); simpleNLG-ZH realises this as a singular pronoun.

Comparing these results with earlier evaluations of simpleNLG-like systems, our results on the tests sets were perfect (with system input constructed by hand from the input expressions), which was also the case for most earlier studies (Bollmann, 2011; Ramos-Soto et al., 2017). Four of the previous evaluations involved a corpus. Bollmann (2011) and Dokkara et al. (2015) evaluated their system on 152 sentences from five Wikipedia articles and 738 sentences randomly picked from a book, respectively. The linguistic variation of their test set is greater than ours (which focused on referring expressions), but the quality of their output may have been lower: Dokkara et al. (2015) reported 57% of exact matches, lower than our 65%. Bollmann (2011) reported 76% of the sentences "could be generated", though what this meant is not entirely clear. Mazzei et al. (2016) tested the coverage and scalability of their system by automatically mapping 20 dependency trees from the Universal Dependency corpus. They reported only 10% exact matching sentences (2/20) and their discussion suggests that their results for declarative and interrogative sentences may have been disappointing. Braun et al. (2019) used a larger (compared to other evaluations that involves corpora and the current study) test set, which contains 3800 sentences. Since German is a morphology rich language, most of these test cases are for assessing inflection operators.

7.2.6 Discussion

The realisation has turned out to be non-trivial in all the languages addressed in the simpleNLG tradition so far, but *where* the most challenging problems are (i.e., in which components of the system), and what the optimal balance between hand-crafting and data-driven method should lie, is something that differs per language.

As for the former issue, we have seen that Mandarin appears to require only a small set of morphological operators, but a much-enhanced set of syntactic processing rules.

As for the latter issue, our study of errors in simpleNLG-ZH offers support for the idea that some issues in realisation are best handled using data-driven methods (Langkilde, 2000; White et al., 2007). As it stands, simpleNLG-ZH makes all its decisions based on a combination of handcrafted rules and explicit stipulations. It would be preferable if the role of the developer in making these decisions could be reduced. This is true for the choice of classifiers, for the use of particles (such as "的" and "了"), for the choice between different negation words ("不" or "没"), and for ordering the modifiers and specifiers (as mentioned in §7.2.5). In all these cases, simpleNLG-ZH assumes that the choice is made outside the system (i.e., by a person or by another component of the NLG system). It would be useful if these choices were made by simpleNLG-ZH itself, but it is difficult to see how a rule-based approach could accomplish this.

7.3 Study 2: Selecting Classifiers using Data-Driven Methods

As discussed, the current simpleNLG-ZH's treatment of classifier selection is imperfect, because such a selection is context-dependent. We, therefore, need clever algorithmic solutions. These algorithms are of predicting what classifier suits a given discourse context. Before developing new algorithms, we need to redefine the task to fit the potential data-driven methods. The most sophisticated model we are aware of is Peinelt et al. (2017). Ambitiously, these authors decided to deal with classifiers of all different types, also including measure words for instance, which are difficult to predict because they add information. They approached the problem as follows: Given a sentence in which a classifier is yet to be realised, and the head noun is flagged, predict the missing classifier. For example, in the input:

(98) 一 ⟨CL⟩ 精彩 的 ⟨h⟩球赛⟨/h⟩
yì ⟨CL⟩ jīngcǎi de ⟨h⟩qiúsài⟨/h⟩
'a wonderful ball game'

⟨CL⟩ indicates where the missing classifier is and the ⟨h⟩ tag pair flags the head noun. The authors construct a large-scale classifier dataset, namely ChineseClassifierDataset[5] (henceforth, CCD) by extracting and filtering data from three publicly available Chinese corpora (including the Lancaster Corpus of Mandarin Chinese, the UCLA Corpus of Written Chinese, and the Leiden Weibo Corpus). They did experiments on their CCD corpus with several baselines, including a rule-based system, two machine learning based systems, and an LSTM-based system (Hochreiter & Schmidhuber, 1997). An initial valuation study indicated that the LSTM achieved the best performance.

Our own work takes the same perspective as Peinelt et al. (2017). But although the *performance* of the model of Peinelt et al. is encouraging, it still leaves considerable room for improvement; in particular, the question comes up whether BERT (Devlin et al., 2019), with its superior ability to take context into account, might do better. In addition, the model of Peinelt et al. offers only limited *insight*, because it does not distinguish between different types of classifiers. In other words, the performance of the model may mask important differences between different types of classifier choices. A good way to address this limitation would be to make use of an existing categorisation of classifier types. But although linguists generally agree that "true" (or "sortal") classifiers should be distinguished from measure words (L. L.-S. Cheng & Sybesma, 1999; Croft, 1994), there is some disagreement on how these sub-types should be defined and what further divisions between sub-types should be taken into account. Sub-types are often described by example, without computationally implementable criteria or explicit lists of classifiers (N. N. Zhang,

[5] The dataset is available at: github.com/wuningxi/ChineseClassifierDataset

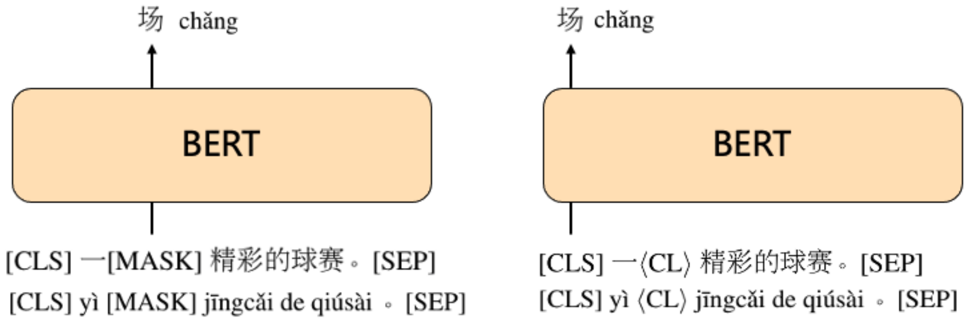

场 chǎng
BERT
[CLS] 一[MASK] 精彩的球赛。[SEP]
[CLS] yì [MASK] jīngcǎi de qiúsài 。 [SEP]
场 chǎng
BERT
[CLS] 一⟨CL⟩ 精彩的球赛。[SEP]
[CLS] yì ⟨CL⟩ jīngcǎi de qiúsài 。 [SEP]

Figure 7.4: Sketch of our BERT-based Classifier selection models: predicting the classifier by unmasking the [MASK] (left); predicting the classifier as classification (right).

2013). To our knowledge, Her and Lai (2012) are the only ones to provide comprehensive lists of classifiers of various sub-types, and in what follows we will make use of these lists.

In this section, we start by introducing two different BERT-based models, one of which uses word masking and one of which performs classification. Subsequently, we report on our comprehensive evaluation experiments, in which we compare our BERT-based models with each other and with several baselines, using the CCD dataset.

7.3.1 Choosing Classifiers using BERT

To test whether the most recent contextual pre-trained language models help choose classifiers, we decided to try BERT. Specifically, we use BERT to accomplish the task of choosing classifiers in two ways: an unsupervised way (i.e., predicting classifiers by unmasking masked tokens) and a supervised way (i.e., fine-tuning BERT on the task of classifier prediction).

Unmasking Masked Classifiers

In order to assess how well BERT, as a masked language model, can model classifiers, we tried to use BERT without any fine-tuning on the task of classifier selection. Specifically, as shown in Figure 7.4 (left), we replace the classifier indicator ⟨CL⟩ with the [MASK] symbol of BERT and ask BERT to unmask it. [6] The unmasked token serves as the predicted classifier. (Note that addressing the classifier selection task in this way will sometimes produce words that are not classifiers.) We refer to this model as MLM.

Classifying Classifiers

Additionally, we test BERT in its classic use. To do this, we fine-tune BERT on the CCD as a multi-class classification, where there are 172 classes (i.e., 172 classifier words) in total, and make a prediction with the help of the [CLS] symbol (see Figure 7.4 (right)). We refer to this model as BERT.

[6] Since our experiments suggested that the head flag (i.e., ⟨h⟩ and ⟨/h⟩) makes no contribution to classifier selection, we drop it to speed up the prediction.

Research Questions

At the start of our research, we formulated the following hypotheses and research questions.

1. Since BERT models context closely and is pre-trained on large scale corpora, we expect it to outperform other models;

2. How do the two BERT-based models compare? Although we expect BERT to outperform MLM, we were curious to see how well MLM performs;

3. We are curious how well BERT can handle classifiers that add information (concretely: measure words, plurality, and politeness).

7.3.2 Evaluating Classifier Selectors

Setup

Dataset. In total, there are 681,102 sentences in the CCD dataset. We split the dataset into training (60%), development (20%), and test (20%) sets following Peinelt et al. (2017).

Baselines. We tried several baseline models proposed in Peinelt et al. (2017), including:

1. a rule-based model (Rule): given a head noun, assign the most frequent classifier associated with it in the training data. If two or more classifiers are equally frequent, one of the classifiers is randomly assigned. If the head noun does not appear in the training data, then the classifier "gè" (which is particularly frequent and often seen as a "default" classifier) is assigned;

2. an LSTM model: For this model, we use a bi-directional LSTM (Hochreiter & Schmidhuber, 1997; Schuster & Paliwal, 1997) to encode the input; it makes predictions using the hidden representation of the last time step.

Metrics. We evaluate each model in terms of accuracy, macro-averaged precision, recall, and F1. Additionally, since the distribution of the CCD is skewed (e.g., more than 25% of the sentences use "个"), we also report the weighted averaged precision, recall, and F1.

Implementation Details. For BERT, we use the "bert-base-chinese" version. [7] When fine-tuning, we set the learning rate to 2e-5 and batch size to 150. For the LSTM, we set the batch size to 256, the hidden size to 300, and the learning rate to 2e-5. We use pre-trained Chinese word embeddings from S. Li et al. (2018)[8].

Results and Analysis

Table 7.3 charts the performance of each model. The results confirm the assumption of our first research question that BERT performs the best, defeating all models on all metrics with large margins. For example, for accuracy, compared to the second best model LSTM, BERT boosts performance from 70.44% to 81.71%. Considering its simplicity, the rule-based

[7] The "bert-base-chinese" can be found at: huggingface.co/bert-base-chinese

[8] These are word embeddings trained by skip-gram on 9 large Chinese corpora with 300 dimensions. It is available at: github.com/Embedding/Chinese-Word-Vectors

		Macro-averaged			Weighted-averaged		
Model	Acc.	Precision	Recall	F1	Precision	Recall	F1
Rule	61.89	34.87	20.50	23.39	58.23	61.90	58.24
LSTM	70.44	33.11	20.12	22.48	67.90	70.44	68.12
MLM	62.22	51.91	33.40	37.68	77.28	62.23	68.21
BERT	**81.71**	**52.86**	**38.10**	**40.77**	**80.70**	**81.71**	**80.77**

Table 7.3: Evaluation Results of each model on CCD. The best results are **boldfaced**, whereas the second best are underlined. MLM is the model that uses BERT as a masked language model, while BERT is the fine-tuned BERT. Acc. stands for Accuracy.

Category	Frequency	Accuracy
True Classifier	85,917	87.8
Dual Classifier	10,817	65.2
Measure Words	11,317	61.1

Table 7.4: BERT's performance on different types of classifiers; frequency of each type in the CCD test set.

system achieved considerably good performance, with higher macro-averaged precision, recall, and F1 than LSTM. This also confirms the viability of a dictionary-based classifier selector, such as the one embedded in simpleNLG-ZH.

MLM, as a model without any training on CCD, performs remarkably well. It receives the second best weighted average as well as micro-averaged F1 (in line with our second research question). Note that, as mentioned, there is no guarantee that the outputs of MLM are classifiers. Concretely, during testing, MLM produces 1566 word types that are not classifiers. This is one of the reasons why its fine-tuned version, BERT, has a major improvement on the (macro-averaged and weighted averaged) recall scores. Nonetheless, it surprised us that MLM can produce a greater *variety* of classifiers than all other models. More specifically, out of 172 classifiers available in CCD, MLM has correctly produced 160 different classifiers, compared to the 140 of Rule, 108 of LSTM, and 136 of BERT. This suggests MLM can sometimes handle rarely seen classifiers.

Regarding the last research question, we looked into measure words, plurality, and politeness respectively. First, we categorise classifiers in CCD into three sub-categories: true classifiers, measure words, and dual classifiers (i.e., classifiers that can function either as true classifiers or as measure words) based on the lists provided by Her and Lai (2012)[9]. Table 7.4 breaks down the performance into different sub-types of classifiers. As we can see, although measure words appear more frequently in CCD than dual classifiers, they still receive a significantly lower accuracy.

Second, for politeness, the only frequent enough[10] politeness classifier is "位" (wèi), which expresses politeness when referring to a person. "位" appears 6737 times in the training data, but only obtains a recall score of 59.87%, which is low compared to equally

[9] These classifier lists were constructed on the basis of the Mandarin Daily Dictionary of Chinese Classifiers (MDDCC).

[10] We define a classifier as *frequent enough* if it appears more than 50 times in the training set.

frequent classifiers (classifiers with frequencies in the range of [5000, 8000) have an average recall score of 77.84%). The confusion matrix (which is too large to print here but will be made available), shows that it is highly likely to be confused with its neutral alternative "个" (gè).

Third, regarding plurality, we pick out frequent-enough classifiers that only convey the meaning of plurality[11], including "群" (qún), "堆" (duī), "些" (xiē), "套" (tào), "对" (duì), and "双" (shuāng). Their recall scores are 52.51% (2453), 52.12% (1914), 56.51% (1910), 34.57% (1308), 62.39% (1321), and 76.49% (806), respectively, where the number in brackets is the frequency of that classifier in the training set. Meanwhile, the average recall of the range [800, 1500) and [1500, 3000) are 61.48% and 76.97%. It is interesting that BERT does a relatively good job for handling plural classifiers meaning "pair" (i.e., "对", and "双") while failing to handle plural classifiers meaning "multiple" (i.e., "群", "堆", "些", and "套"). All in all, classifiers that add information regarding measurement, plurality and politeness could not be properly selected. One explanation is that their context cannot provide enough information to pick the right classifier. Thus, for the last research question, BERT does not work well in handling classifiers that add information.

Distance between the Classifier and the Head Noun. We also explore factors that might influence the decisions of BERT. First, we consider the *distance* between the classifier and the head noun. For instance, for example (98), there is a pre-modifier consisting of two words between the classifier "场" (chǎng) and the head noun "球赛" (qiúsài). Thus, the distance for example (98) is 2. We expect that the larger the distance is, the worse BERT performs. In our experiments, for correct predictions, the average distance (in terms of the number of words) is 1.08 while for incorrect predictions it is 1.21. An unpaired t-test confirms that distance has a negative effect on the model's performance ($p < .001$).

7.3.3 Discussion

In this study, we accomplish the task by means of the state-of-the-art machine learning technique. From an evaluation on a large scale classifier selection corpus (namely CCD), we found that 1) a contextualised pre-trained model (i.e., BERT and MLM) performs remarkably well on the task of choosing classifiers in Mandarin, and fine-tuning helps improve the recall of choosing classifiers; 2) a simple rule-based system has respectable performance; but 3) in terms of accuracy, a pre-trained masked language model (i.e., MLM) was able to select proper classifiers about equally well as the above rule-based system; 4) BERT struggles to predict classifiers that add information such as measurement, plurality, and politeness.

7.4 Study 3: How well can Human Beings Choose Classifier from its Context?

The study in §7.3 suggested that BERT struggles to predict classifiers that add information such as measurement, plurality, and politeness. It confirmed our expectation that some classifier occurrences cannot be predicted from their linguistic context alone since they themselves carry additional information. This makes us aware of how hard this task (i.e., filling Mandarin classifiers given their context) is? In response to this, we conduct a series

[11] Some classifiers have multiple meanings, one of which expresses plurality.

of experiments involving human participants. In these human experiments, we ask several participants to choose classifiers given a linguistic context. By comparing the outcomes of this experiment with the corresponding "references" from the CCD corpus, we will obtain a better understanding of the difficulty of the task.

7.4.1 Experiment Setup

A natural experiment setting is to simply sample a number of items from the CCD corpus and conduct human experiments on the sampled set. Although this setting can provide a good enough picture of how well human beings can accomplish the same task as our model in the previous study did, as figured out in §7.3.2, models like BERT can perform remarkably well on frequent classifiers while relatively badly on infrequent classifiers. We are also aware of how well human beings can do in choosing classifiers if we look at a wider range of classifiers, especially those that are rarely seen. We, therefore, propose another setting, where we first sample a certain amount of classifiers and then sample their context accordingly. We refer to the experiments with the above two settings as Experiment A and Experiment B, respectively, and detail each of them below.

Experiment A

In the first experiment, we randomly sampled 200 items from the test set of CCD. Subsequently, we manually filtered the noise from the sampled set, which results in a corpus containing 186 items. We say a sample is a noise if:

1. the classifier word in the given context is not a classifier. For example, the word "个" in the sentence "他 整 个 暑假 都 在 他 奶奶 家。" (tā zhěng gè shǔjià dōu zài tā nǎinǎi jiā; *He spent the whole holiday in his grandma's house.*) is a component of the word "整个" (zhěnggè; *whole*) rather than a classifier;

2. the sentence is not readable. Since one resource of the CCD dataset is the Chinese social media, there is a certain amount of unreadable content that has failed to be filtered out during pre-processing (see Peinelt et al. (2017) for more details).

We then recruited 4 native Mandarin speakers to fill classifiers given these 186 items. Three of them have background in statistics and the rest one has background in computer science.

Experiment B

In the second experiment, we first sampled 100 distinct classifiers also from the test set of CCD and, then, sampled 2 items for each classifier. After filtering the data by means of the same way of experiment A, we obtained a corpus with 162 items. We observed that there is more noise in this sampled corpus than the one in experiment A. This suggests that the data for less frequent classifiers are of lower quality. Similar to experiment A, we asked 4 native Mandarin speakers to accomplish the task. Two of them have background in statistics and the other two have background in computer science.

	Accuracy (SD)	Percent Agreement
Experiment A	70.97 (2.28)	67.92
Experiment B	41.82 (2.16)	47.22

Table 7.5: The results of human experiments of classifier selection. “SD” is the abbreviation for “standard deviation”.

Metrics

We calculated the accuracy[12] of each participant to compare the human chosen classifiers and those in the sampled corpora. This also helps us to “evaluate” the human performance in the same way as our models. In addition, to quantify how well each participant agrees with each other, we also report the *Percent Agreement* (McHugh, 2012).

7.4.2 Research Questions

Recall that our primary goal is to obtain an impression on how well human beings can accomplish the same task. Analogous to many NLP tasks, it is natural to expect that humans can outperform any of our models. However, meanwhile, since we believe the task of classifier selection is non-trivial, we did not expect humans could approach an extremely high accuracy (e.g., 98%).

Additionally, considering that experiment B looks at infrequent classifiers, which might also be hard for human participants. We expected human participants would perform worse in experiment B than in experiment A in terms of accuracy. Moreover, we also foresaw that participants have less agreement in experiment B than in experiment A.

7.4.3 Experiment Results

Table 7.5 charts the results of both experiment A and experiment B. To compare these results with the performance of our models, we compared the accuracy in experiment A with the accuracy in Table 7.3. Nevertheless, the accuracy in experiment B is not comparable to those accuracy numbers in Table 7.3 as the data in the corpus of experiment B follows a very different distribution from that of CCD. One metric that makes the comparison more meaningful is the macro-averaged recall since it averages over all classifiers in the test set and, for each classifier, it computes the fraction of classifiers that are chosen correctly by a model.

For experiment A, the accuracy is 70.97% and it is surprising to see BERT receives 81.71% (cf. Table 7.3) – a higher accuracy than humans. It embodies that our first expectation, about humans defeating BERT, is rejected. One possible reason is that BERT was fine-tuned on a subset of the target corpus while humans might not be familiar with the specific genre language used in the corpus.

The mean accuracy of experiment B is 41.82%, which, in line with our expectation, is much lower than that of experiment A. It also confirms that the task of classifier selection is

[12] The term “accuracy” here means the matches between the classifiers chosen by the participants and those in the corpora. This does not imply that a “mismatch” means “inaccurate”. In other words, a lower accuracy does not always mean the participant did a worse job since there are possibilities that, given a context, multiple proper classifiers exist.

non-trivial for both algorithms and human beings. If we further compare this number (i.e., 41.82%) with the macro-averaged recall of BERT, it is slightly higher than BERT's 38.10%. Although this is not a fair comparison, it somehow suggests that humans and BERT perform at a similar level when it comes to choose proper infrequent classifiers. Our last research question has also been validated to be true. That is, participants in experiment A have much higher agreement than participants in experiment B.

Case Studies

To understand why human participants cannot defeat BERT and why experiment B obtained a lower percentage of agreement among participants' decisions, we looked into the sampled test cases and did an error analysis. First, there are a certain amount of mis-uses of classifiers, especially when on the general classifier "个". For example, in the following example (99), the subject mis-used the classifier "个" for the head noun "钱" (qián; *money*).

(99) a. 我们 的 网站，从未 做 一 分 钱 的 广告 。
wǒmén dè wǎngzhàn, cóngwèi zuò yì fēn qián de guǎnggào
'Our website have never spent money on advertisement.'

b. * 我们 的 网站，从未 做 一 个 钱 的 广告 。
wǒmén dè wǎngzhàn, cóngwèi zuò yì gè qián de guǎnggào

As we have discussed in §7.3.2, our models encountered difficulties when predicting classifiers that add information since the provided contexts do not offer enough information for the selector to make predictions. We found that human beings have similar difficulties. For instance, for the same head noun "衣服" (yīfú; *clothes*), we found the following two items from the two sampled corpora.

(100) a. 我 穿 了 那 套 衣服 。
wǒ chuān lè nà tào yīfú
'I wore that suit of clothes.'

b. 我 穿 了 那 件 衣服 。
wǒ chuān lè nà jiǎn yīfú
'I wore that piece clothes.'

(101) a. 他 送 了 我 满满 一 大 袋 衣服 。
tā gěi le wǒ mǎnmǎn yí dà dài yīfú
'He gave a package of clothes.'

b. 他 送 了 我 满满 一 大 筐 衣服 。
tā gěi le wǒ mǎnmǎn yí dà kuāng yīfú
'He gave a basket of clothes.'

c. 他 送 了 我 满满 一 大 箱 衣服 。
tā gěi le wǒ mǎnmǎn yí dà xiāng yīfú
'He gave a box of clothes.'

d. 他 送 了 我 满满 一 大 包 衣服 。
tā gěi le wǒ mǎnmǎn yí dà bāo yīfú
'He gave a bag of clothes.'

In the example (100), both "那 套 衣服" and "那 件 衣服" are grammatically correct and fluent Mandarin expressions. To decide whether it should be "a suite of" or "a piece of", we need information beyond the given context. Likewise, all classifiers in example (101) are grammatically correct and fluent, but, without knowing what the container is, it is hard to choose a classifier from "package", "basket", "box", and "bag". These two examples, in aggregate, suggest that depending on what information is lacking, the candidate classifiers could also be different. In other words, as what we have found in §7.3, the context adds a certain amount of information but it is still not enough for deciding the exact classifier. For instance, in example (101), as what we have found in §7.3, the context tells us someone gives me a "set" of clothes, which singles out classifiers that indicate singularity (e.g., "件"), but does not tell us what kind of container this set of clothes are in.

At length, we have also observed a certain amount of cases, where multiple classifiers can be used and these classifiers have a similar meaning. For example, the head noun "早餐" (zǎocān; *breakfast*) accepts multiple alternative classifiers, including "顿" (dùn) as well as "餐" (cān). There are possibilities that using these two classifiers could result in different meanings, but the following two sentences have the same meaning:

(102) a. 我 吃 了 在北京 的 最后 一 顿 早餐 。
wǒ chī le zàiběijīng de zuìhòu yí dùn zǎocān
'I had my last meal in Beijing'
b. 我 吃 了 在北京 的 最后 一 餐 早餐 。
wǒ chī le zàiběijīng de zuìhòu yí cān zǎocān

This also implies that our current "evaluation" of algorithms and human performance is not entirely unproblematic. A mismatch between the chosen one and the one in corpus might because any of the following reasons:

1. the choice is inaccurate;
2. the context does not provide enough information for deciding a single classifier and the information added by the chosen classifier is different from the reference one, but both of them are grammatically correct;
3. there are multiple classifiers available for conveying the same meaning.

Apparently, it is problematic to view a mismatch caused by the second or the third reason an "incorrect" choice. Therefore, in future, in order to better assess the performance of each algorithm (as well as human beings), we are currently planning two large-scale human experiments. One is another reader experiment to sufficiently explore, for each context, how many alternative classifiers there are and what is the "preference" of these classifiers. To assess an algorithm, it is reasonable to evaluate how well an algorithm can mimic such a preference. A similar paradigm has been done in the task of referential form selection (Castro Ferreira et al., 2016). The other is a reader experiment. Given a context and multiple alternative classifiers, each participant will be asked to decide the readability or the acceptability of each combination (i.e., a combination of a context and a classifier).

7.4.4 Discussion

In this study, to understand how hard the classifier selection task is for human beings, we conducted a human experiment to ask subjects to accomplish the same task as the models in §7.3. We found that the best performed model BERT defeats human beings in terms of accuracy. Just like BERT, human beings also met difficulties on classifiers that add information. Finally we argued that the current evaluation using accuracy is problematic because many contexts could have multiple "proper" classifiers whereas the current evaluation views the choices that are different from corpus as "incorrect" choices.

7.5 Summary

In this chapter, we first introduced and evaluated a realisation engine for Mandarin in the tradition of simpleNLG (a simple-to-use, controllable and extendable realisation engine developed by Gatt and Reiter (2009)). In the course of building simpleNLG-ZH we found that, due to the fact that Mandarin is an analytic language, realisation in Mandarin needs much less morphological operations and much more syntactic operations compared to English. We hope simpleNLG-ZH can be a good starting point for work on other Sinitic languages, such as Cantonese.

Another characteristic we found is that, while conducting surface realisation, many elements have multiple alternatives, such as particles, classifiers, aspect markers, etc. Therefore, in the last two studies in this chapter, we picked one of these elements as an example, namely classifier. Specifically, in the second study, we attempted to tackle the task of classifier selection using a number of data-driven techniques. As expected, BERT achieved remarkably good performance even it was not fine-tuned on the classifier selection dataset. Its performance was further boosted once fine-tuning was done. Moreover, we also found that all these data-driven models met a similar problem: it is hard for them to decide classifiers that add information to the resulting NPs.

Lastly, we examined how hard the classifier selection is for human beings. We asked participants to accomplish the same tasks as the one the above models were aiming at. Surprisingly, we found that BERT performs better than human beings. To understand why, more experiments are needed.

As discussed in §3.4 and §7.2, Mandarin poses many difficult choices for the construction of a good surface realiser. The current simpleNLG-ZH leave these choices for users. In this chapter, we took the choice of classifiers as an example, but there are still many other issues we have not addressed. For example, simpleNLG-ZH provides an API for the "把" construction. It allows users of simpleNLG-ZH to decide whether to use it or not. On the one hand, the current version of simpleNLG-ZH lacks restrictions of when the "把" construction should be avoided. In light of C. N. Li and Thompson (1989), "把" construction may only be used in the context where the verb expresses "settlement" of, or action upon, the object. Building on this idea, Ye et al. (2007) found that it is generally used with verbs that are high in transitivity, but is not used with verbs that express states or emotions (e.g., *love* and *miss*). On the other hand, in many cases, for expressing a certain meaning, using or not using "把" construction are both grammatically correct. For example, the following two sentences are expressing the same meaning, but the referring expressions "这三本书" (zhèsānběnshū; *the three books*) are placed in different positions:

(103) a. 我把这三本书卖了
wǒbǎzhèsānběnshūmàile
I sold the three books.
b. 我卖了这三本书

wǒmàilezhèsānběnshū
I sold the three books.

Nevertheless, it has been suggested that, analogous to the choice of classifiers, Mandarin speakers would choose between (103-a) and (103-b) differently given different contexts. A corpus study by J. Chen et al. (2021) suggested that the "把" construction is preferred if the referring expression is discourse-old (i.e., it has been mentioned in previous discourse) or animate or long (3 syllables or more). How the use of the "把" construction can be decided by the realiser is worth exploring.

Another example is the decision of the order of multiple pre-modifiers. The current simpleNLG-ZH roughly breaks pre-modifiers into quantitative adjectives, colour adjectives and classifying adjectives, and orders them in the order of quantitative adjectives, colour adjectives and classifying adjectives. This simple rule works fine if the number of pre-modifiers is relatively small. It is interesting to see whether it still works if there is a large number of pre-modifiers and how well data-driven approaches can work on this task.

CHAPTER 8

Conclusion

8.1 What have we learnt?

We summarise the lessons that we have been able to learn from the studies described in this thesis. Some of these lessons are specifically about Mandarin, whereas others are universal across languages (or at least, Mandarin and English). Rather than aiming for completeness, our summary will focus on a few of the more striking findings.

8.1.1 How do Mandarin Speakers Use Noun Phrases?

We have looked at two kinds of NPs: Referring Expressions (REs) and Quantified Expressions (QEs). We briefly address each of these in turn.

One-shot Referring Expressions

One-shot REs aim to identify their intended referent entirely within one NP, that is, without relying on the linguistic context of this NP. Although one-shot REs is a topic to which a lot of theoretical and computational work has been devoted (e.g. Dale and Reiter (1995), Engelhardt et al. (2006), Engelhardt et al. (2011), Koolen et al. (2011), Krahmer and van Deemter (2012), Paraboni et al. (2017), and Pechmann (1989), van Deemter (2016)), we have argued in Chapter 4 that the existing conceptual apparatus for thinking about reference is not as precise and complete as it should be.

To fill this gap, we proposed a formal perspective on reference, in the form of a set of definitions and an accompanying annotation scheme. Many of these definitions concern specific kinds of over-specification. For example, we define *numerical over-specifications* to be NPs that are longer than necessary without containing any superfluous property (see §4.4.2). Furthermore, we define *nominal over-specifications* to be NPs in which the TYPE attribute is the only superfluous attribute (also see §4.4.2).

To show the benefits of this approach, we analysed a Mandarin RE corpus MTUNA and an English RE corpus ETUNA and compared the use of REs in them. We found that there was no difference between Chinese and English in the use of over-specification. However, when differences between kinds of over-specification were taken into account, then various differences between these languages came to light. For example, Mandarin speakers use more TYPEless REs than English speakers, possibly because Mandarin allows zero head nouns in noun phrases.

In addition to various kinds of over-specification, the new scheme contains concepts such as wrong (i.e., incorrect) specification and under-specification. Focusing on the latter, previous studies had reported that only about 5% of human-produced REs are instances of under-specification (Ferreira et al., 2005; Koolen et al., 2011; Pechmann, 1989). Consequently, none of the algorithms we examined (and many previous work, e.g., van Deemter, Gatt, Sluis, et al. (2012) examined) for the generation of one-shot REs (e.g. Dale (1989) and Dale and Reiter (1995)) ever produce under-specified REs (when these can be avoided). In stark contrast with this received wisdom, we found that both MTUNA and ETUNA contained as many as 15% REs under-specifications. This suggests that future work on REG should start to pay proper attention to this hitherto overlooked phenomenon. One approach that is worth trying is the rational speech act model which could produce under-specification if a referent is salient enough (see §2.2).

Referring Expressions in Context

When REs are placed in (linguistic) contexts, Mandarin speakers often choose to not express them overtly. For example, in the conversation:

(104) a. 你看见张三了吗?
Did you see Zhangsan?
b. 看见了。
[I] saw [him].

REs in both subject position (i.e., *I*) and object position (i.e., *him*) are dropped, which, for Mandarin speakers, is more pragmatically natural than the one where both pronouns are not dropped (i.e., 我看见他了; *I saw him*). REs of this kind is named Zero Pronouns (ZPs; C.-T. J. Huang (1984)). In NLP, ZPs have been widely studied from the perspective of resolution (Yin, Zhang, et al., 2017; Yin, Zhang, Zhang, et al., 2017; Yin et al., 2018). Nevertheless, modelling the use of ZPs has not attracted much attention. In §5.2, we built computational models for investigating contributing factors for the use of ZPs. We found that factors, such as recency, syntactic position, and discourse status, that proved to affect pronominalisation also affect the use of ZPs in Mandarin.

Quantified Expressions

In addition to referring, another prime function of NPs is quantifying, such as the following QEs:

(105) a. some chairs
b. most students

To understand how English and Mandarin speakers use QEs, we conducted a series of elicitation experiments. In these experiments, our participants were free to describe a visual scene in whichever way they want, using as many sentences as they want, and using any sentence pattern that they choose. Concretely, given a visual scene, we asked each participant to say, for example, "*All objects are square. Half of the squares are blue.*". We called these descriptions Quantified Descriptions (QD), each of which consists of multiple QEs.

These experiments result in an English QD corpus QTUNA as well as a Mandarin QD corpus MQTUNA. We analysed the QDs in QTUNA and MQTUNA. Generally speaking, we found that the domain size is a key factor that affects the completeness (i.e., whether a description can let a reader fully reconstruct the scene), the correctness (i.e., whether a description says everything correct), and the vagueness (i.e., whether a QD uses any vague quantifier, e.g., *most* and *many*) of the generated QDs. To be more specific, there were fewer logically complete QDs, fewer correct QDs and more vague quantifiers in larger domains than smaller domains. However, we found that speakers did not produce longer QDs in larger domains. Additionally, both English and Mandarin speakers were likely to mention shape in position A and to mention colour in position B. For example, they were more likely to utter "*half of the squares are blue*" rather than "*half of the blue objects are squared*".

By comparing QDs in two corpora, we observed that Mandarin speakers were more likely to produce longer QDs, incomplete QDs, incorrect QDs, and QDs that contain more vague quantifiers than English speakers. Additionally, for each QE $Q(A, B)$, Interestingly, we also found Mandarin speakers frequently drop the phrases in position A, i.e., saying "一半是红的" (yíbànshìhóngsède; *half are red*) instead of saying "一半的图形是红的" (yíbàndètúxíngshìhóngsède; *half of the objects are red*) if what this phrase describes has been mentioned in the previous discourse or is the property shared by all objects.

8.1.2 Lessons for Modelling of Noun Phrases in Mandarin

We have tested computational models for one-shot Referring Expression Generation (REG), Referential Form Selection (RFS), the use of Zero Pronouns (ZPs), surface realisation, and the use of classifiers in Mandarin. Although we have not yet tested quantified description generation (QDG) models in Mandarin, we have discussed potential ways of doing so.

One-shot REG

The task of one-shot REG asks algorithms to produce REs that are human-like. Although Mandarin RE corpora, such as MTUNA, have been built, no work has been done for building and assessing Mandarin one-shot REG models. In this thesis, we tested three classic REG algorithms, including the Full Brevity Algorithm (FB), the Greedy Algorithm (GR), and the Incremental Algorithm (IA) on the MTUNA dataset and compared the results with their results on ETUNA. We found that, unlike English, IA is no longer always the winner. Precisely, IA merely won in the simple domain (i.e., the furniture domain), but, in the complex domain (i.e., the people domain), it did not work better than the FB algorithm, which is an algorithm that takes brevity as its priority. We also leant that a good Mandarin REG algorithm should include probabilities to model TYPEless REs and take the syntactic position into consideration.

RFS and Modelling ZPs

Both RFS and Modelling ZPs are sub-tasks of REG in context. REG in context is a task to generate REs given their linguistic context. In Chapter 5, we started with building models for ZPs based on the Rational Speech Act (RSA) model, which had proved to work well on pronominalisation in English (Orita et al., 2015). It assumes that Mandarin speakers tend to use a ZP if the referent is salient enough for successful communication. The experimental results suggested that the RSA works respectably on modelling the use of ZPs in Mandarin compared to a strong rule-based baseline (Yeh & Mellish, 1997).

When modelling full speakers' behaviour (i.e., RFS which asks models to choose from ZP, pronoun, proper name, and description), we examined both neural-based models and feature-based machine learning models and found that neural-based methods work better. By comparing different neural models on RFS, we found that, for both English and Mandarin RFS, using a single RNN can often obtain a remarkably good performance compared to those with much more complex neural architectures and, for only Mandarin RFS, incorporating pre-trained word embeddings (i.e., SGNS) or language models (e.g., BERT) can significantly help models learn more useful linguistic information, and, consequently, improve the performance. One major disadvantage of using neural models is that they are considered to be black-boxes and it is hard to link their behaviours to corresponding linguistic theories. To address this issue, we conducted interpretability studies using probing classifiers. From the probing studies, we learnt that neural models can learn certain information, including referential status, syntactic position, and recency, but failed to learn information that requires the model to have an overall understanding of the whole document or the whole corpus.

Quantified Description Generation

As aforesaid, we built two QD datasets: QTUNA and MQTUNA. To mimic the QD in these corpora, we designed two QDG models following a similar paradigm as modelling the production of REs, i.e., viewing the task as a step-wise addition of descriptive information that narrows down an initial set of possibilities. To examine them on the task of English QDG, we designed two evaluation protocols. One is to ask human judges to judge the quality of the machine-generated QDs. The other is to ask participants to re-produce scenes given machine-generated QDs. The evaluation results suggested that our models can produce QDs that are both natural and useful.

Although we believe these algorithms should be universal across different languages, to make them produce Mandarin QDs, adaptation is needed. We argued that a workable Mandarin QDG model should have abilities to decide whether to express plurality explicitly or implicitly, to handle more vague quantifiers, and to decide when to stop in accordance with the characteristics of Mandarin QDs.

Surface Realisation

Surface realisation is the very last stage of an NLG system. It is responsible for mapping the plan (i.e., the "plan" from earlier stages) to its well-formed surface form. No extendable, wide-coverage, and easy-to-use surface realiser has been developed. To fill this gap, we built a Mandarin realisation engine, namely simpleNLG-ZH, following the tradition of simpleNLG. In the course of building this software, we learnt that, compared to realisers for western languages, a Mandarin realiser should have fewer morphological operators and more syntactic operators due to the fact that Mandarin is an analytic language.

In addition, we also argue that a good Mandarin realiser should be able to handle situations where a component (e.g., plural maker, particle, head noun, etc.) is optional and where a component (e.g., classifier) has multiple alternatives. To understand this better, we dived into one specific component: classifier. We then built classifier selection models based on either rule-based or neural-based (i.e., BERT) models. We learnt that, although the rule-based solution could already achieve good performance, BERT can further dramatically increase the performance, which performed even better than that of human beings.

8.1.3 Coolness

One of our research questions in this thesis is to validate the idea of coolness. Recall that we introduced three interpretations of coolness in §3.1.

1. The first interpretation is from C.-T. J. Huang (1984), where the concept of "cool-hot" division was first introduced. It is only about anaphora and suggested that anaphora in Mandarin is often pragmatic naturally dropped. In this section, this interpretation is referred as [1];

2. The second one suggests that, in addition to anaphora, many other categories in Mandarin are also not expressed obligatorily, e.g., definite markers, plural markers, and aspect markers and so on. We refer it as [2].

3. The last interpretation is linked to the idea of the clarity-brevity trade-off in NLG. It hypotheses that Mandarin speakers prefer brevity to clarity. We refer it as [3].

In what follows, we list pieces of evidence we found that support coolness and that oppose coolness. For each item, we use the notations introduced above to link each piece of evidence to one of the three interpretations. Additionally, we also codify the subject matter with each piece of evidence links to. Concretely, [RE-one] is the one-shot RE, [RE-context] is the RE in context, [QD] is the QD, and [R] is any issue that is related to surface realisation. We then marry the code that refers to an interpretation to the code that refers to a subject matter. For example, [1-OR] means the current evidence is related to the first interpretation of coolness and is about one-shot REs. Also, note that we also include findings from previous research that relates to coolness. Evidence that supports coolness includes:

- [2-RE-one] van Deemter et al. (2017) found that determiners and number markers are not used frequently in Mandarin REs, which makes 76.18% of REs in MTUNA are bare nouns;

- [3-RE-one] Mandarin speakers used more TYPEless REs than English speakers;

- [3-RE-one] Unlike English, for Mandarin, the Incremental Algorithm did not always outperform the Full Brevity (an algorithm that always produces the shortest RE) algorithm;

- [1-RE-context] 13.6% of REs in the OntoNotes dataset (a commonly used Mandarin corpus) are ZPs. The task of RFS needs to take this new category (in addition to the pronoun, proper name and description) into consideration. [1]

- [1-QD] The anaphora in the position A of QEs in the form of $Q(A, B)$ are often omitted;

- [2-QD] The singularity/plurality of a QE was often expressed implicitly.

[1] Note that although in our task definition of RFS, we considered not only anaphora but also cataphora, we still regard it as evidence for the first interpretation of coolness.

- [3-QD] Mandarin speakers produced less complete QDs than English speakers;
- [3-QD] Mandarin speakers used more vague quantifiers than English speakers;
- [2-R] Many components of Mandarin NPs (e.g., plural markers, definite makers, and aspect markers) are optional. Our surface realiser is able to handle them.

Evidence that opposes coolness includes:

- [3-RE-one] There was no significant difference between the use of over-specifications and under-specifications;
- [3-QD] On average, Mandarin speakers produced longer QDs than English speakers.

As we can see from the above two lists, all the evidence that we collected appears to support the first two interpretations of coolness ([1] and [2]). As for the third interpretation (i.e., [3]), it consists of two parts: Mandarin speakers prefer brevity and violate clarity. Interestingly, "preferring brevity" is valid for the use of REs but is not valid for the use of QDs. Whereas, "violating clarity" is valid for the use of QDs, but is not valid for the use of REs. Recall that the primary assumption behind coolness is that Mandarin relies more on communicative context for disambiguation compared to languages like English. Therefore, one possible explanation of the lack of evidence for [3] is that since both the above two pieces of counter-evidence are about one-shot production of NPs, as we have discussed in §4.4.7, contexts of them[2] are overly simple for speakers to rely on.

We need to note that one limitation of our discussion in this section is that the way in which our experiments were set up was not perfect from a point of view of comparing languages. In all cases, the English corpus was collected and studied first, and the Mandarin corpus was collected after that. Although both endeavours always followed the same elicitation methodology and were based on broadly the same set of stimuli (i.e., set of input scenes), there nonetheless were experimental details that differed across the two elicitation experiments. These details include the precise choice of situations (i.e., which exact scenes were shown to participants?), the order in which these situations were presented to participants, and the demographics of participants. It is possible, though perhaps not likely, that if these conditions are fully controlled, a different picture may emerge, potentially also enabling a different verdict on questions surrounding coolness.

8.1.4 The Choice of Computational Methodology

As was explained in the Introduction to this thesis, our primary aim has been to use computational algorithms to shed light on the ways in which Mandarin speakers use Noun Phrases. Thus, our main contributions have been in what we have called Theoretical NLG (see §1.5, where "T" stood for theoretical NLG and "P" for practical NLG).

Building computational models of the production of Noun Phrases involves different tasks (§1), depending on the kind of Noun Phrase involved. In building such models, we have used whatever computational method seemed best suited for the task. For example, we have used *rule-based* algorithms for one-shot referring expression generation, modelling the use of zero pronouns, quantified description generation, and surface realisation; we have used *feature-based* Machine Learning for modelling the use of zero pronouns and

[2] The context of a one-shot RE is the distractors in the given scene.

referring expression generation in context; we have used (neural) *Deep Learning* for referring expression generation in context, and for modelling the choice of classifiers. Our reasons for choosing these particular methods for these particular tasks are broadly familiar: neural methods, for example, are particularly called for when a huge amount of data is available, and when the problem is not very well understood yet.

In our own work, we have often constructed systems that *combine* different computational methods. This is perhaps clearest in our work on Surface Realisation (§7). Although this work has mostly resulted in rule-based algorithms, our study of classifiers (which is a part of Surface Realisation) has shown how Deep Learning, judiciously used, can add further sophistication to an otherwise rule-based system. Our reason for using neural methods for modelling classifier choice was that classifiers are highly frequent, so a very large amount of relevant data was available to us, and this helped neural models to do well; and although sensible rules for classifier choice had been suggested by linguists, it was difficult to make such rules sufficiently precise that a computer algorithm can use them.

In other areas of our work, the balance between computational methods was different. In particular, we found that neural models did not always do very well when performing End2End generation. For example, although our work on referential form selection (§5.3 and §5.4) suggested that neural models perform better than rule-based systems, our recent work on generating full referring expressions (Same et al., 2022) suggested the opposite: we showed there that existing rule-based systems tend to perform at least as well as End2End neural REG models. Probably, neural methods did not work so well for REG because REG is closely tied up with choosing the semantic content of a text, (e.g., choosing what properties of a referent to mention), causing neural models to do less well. The idea that neural methods struggle to choose appropriate semantic content became commonplace when Reiter (2018a) and Rohrbach et al. (2018) pointed out that neural NLG (Dusek & Jurcicek, 2016) systems sometimes "hallucinate", i.e., producing contents that are not present in the inputs or that are inconsistent with these inputs.

The general question of what scientific method best suits a particular research question or research task is still far from completely resolved, of course. However, in view of the limited area of work described in this thesis, we expect that the kind of "hybrid" approaches that we proposed in our chapter on Surface Realisation, which combines symbolic and sub-symbolic methods, will increasingly be used in both practical and theoretical NLG, and in Natural Language Processing more generally.

8.2 Future Work and Open Questions

Given the lessons summarised above, we first came up with two recommendations for potential future work. As has been pointed out in the previous section, one recommendation is to conduct fully controlled language comparison studies on the subject matters we have looked at in this thesis. The other is to conduct elicitation experiments where NPs are placed in richer contexts and to analyse the results with the focus on validating the third interpretation of coolness (i.e., [3]).

In what follows, we discuss potential future work and open questions for each task in this thesis.

8.2.1 One-shot REG

In §4.4, we distinguished numerical over-specification from other types of over-specification as none of the properties of a numerical over-specification is superfluous. In future, it would be interesting to conduct experiments to explore the following questions:

1. When speakers over-specify, are they more likely to produce expressions that are “built around" (see §4.4 for its formal definition) minimal description or ones that are built around numerical over-specification?

2. Since numerical over-specifications use more properties, will they help readers to identify the target referent faster than by means of real over-specifications (Paraboni et al., 2017)?

3. What if numerical over-specifications use TYPE for distinguishing the target objects?

4. Do our new perspective and findings in §4 still stand on more complex and realistic references than those in MTUNA and ETUNA?

In §4.5, we explained the reason why there are many TYPEless REs in the people domain of MTUNA by citing Lv’s hypothesis (Lv, 1979), suggesting that if omitting the head noun results in a distinguishing description, then the head noun is omissible. Nonetheless, we also mentioned an alternative explanation: animacy, suggesting the TYPE is more likely to be dropped for animates than inanimates. We plan an experiment to confirm which explanation is accurate.

In §4.4.5, we sketched possibilities for extending our new perspective of specifications to plural REs. Building on this, we will detail an annotation scheme for over- and under-specified plural REs and use it to analyse plural REs in ETUNA and MTUNA.

Regarding building a better Mandarin one-shot REG model, we plan to

1. Consider advanced non-deterministic REG models, such as RSA and PRO (see §2.2.1 for more details), to model non-negligible TYPEless REs;

2. Model plural REG in Mandarin. One factor that needs to be considered is that a bare noun in a Mandarin RE could either refer to a single referent or multiple referents since Mandarin can express plurality implicitly (see §3.4 for more discussions);

3. Develop an evaluation metric that can overcome the shortcomings of DICE (see §4.5.5).

8.2.2 REG in Context

When modelling RFS selection in §5.3, we pointed out that the WEBNLG might not be a good corpus to study the use of REs in context. In future, we will explore or construct REG (recall that RFS is a sub-task of REG) datasets that contain natural and realistic use of REs in English so that we can make a better comparison between REs in Mandarin and in English.

Regarding constructing better Mandarin (and English) REG in context model, we are interested in models that:

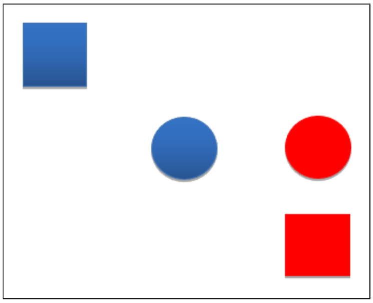


Figure 8.1: A scene of domain size $N = 4$ from QTUNA.

1. Take deictic pronouns (i.e., *I* and *you*) and their referents into account. Deictic pronouns are often referring to participants of conversations or writers of documents. There is often no proper name or description in their reference chains;

2. Marry neural techniques with the RSA. Recently, there has been work on building so-called NeuralRSA models (Andreas & Klein, 2016; Fried et al., 2018; Monroe et al., 2017; Monroe et al., 2018). We plan to apply this idea to the task of REG in context;

3. In this thesis, we only examined the RFS task. We plan to move our focus from RFS to REG.

Regarding interpreting NeuralRFS (and NeuralREG) models, we plan to:

1. Conduct a more fine-grained analysis of how these models handle anaphora and cataphora (i.e., REs whose meaning are determined or specified by later expressions);

2. Consider other model interpretation techniques other than probing classifier;

3. Design new probing tasks on the basis of other factors that could influence RFS, such as animacy, competition and positional attributes (see Same and van Deemter (2020) for an overview).

8.2.3 Generation of QDs and QEs

Our primary plan is to build Mandarin QD generation systems using our quantified description generation algorithms in §6.4 and evaluate the systems on MQTUNA. We listed a number of potential issues that need to be aware of when applying the algorithms to Mandarin in §6.4.8. Later on, we consider the following open questions. Note that since this subject matter has not been sufficiently explored before, the open questions we discussed here are less language-dependent.

How efficiently do speakers use quantification? Speakers in our corpus were frequently less than optimally "efficient" in their use of quantification, saying more than was strictly necessary for describing the scene. An extreme example is a QD for the scene in Figure 8.1 where some speakers use as many as three quantifiers (i.e., "*Half the objects are squares. Half the squares are red. Half the circles are red.*"), whereas others use only one (i.e., "*All possible combinations are shown.*") Another type of example arises when a scene of size

$N = 4$ can be described saying "*There are red circles and blue squares*" (using two plural noun phrases), in which case the description "*There are two red circles and two blue squares*" could be regarded as inefficient. Investigating the mechanisms that allow speakers to be maximally efficient – and the conditions under which these mechanisms are actually deployed – is a rich area for further research. Once again, there is an analogy here with research on the production of REs, where researchers have studied under what circumstances speakers tend to "over-specify" a referent (see Chapter 4). Perhaps the main question raised by these phenomena is whether speakers are "inefficient" because they cannot help themselves, or to help the reader understand the description (i.e., Bell (1984) and Coupland and Jaworski (2008)). Analogous questions regarding quantification have yet to be answered.

How to capture variation in the corpus? Substantial differences between speakers are known to exist in many other areas of language production (e.g., Gibbs and Van Orden (2012), Holden et al. (2009), Horton and Keysar (1996), and van Deemter (2016)). Such differences are likely to affect all the issues discussed in Chapter 6. One approach would be to investigate how key properties of the descriptions vary between different types of speakers, looking at differences in level or type of education for example. A different approach would be to design a probabilistic generator, which generates all the different types of descriptions that are seen in the corpus but take into account their frequencies. The degree of fit between such a probabilistic model and the corpus could be measured using the *generalisation criterion* methodology of Busemeyer and Wang (2000), analogous to the probabilistic modelling of reference in van Gompel et al. (2019).

How to quantify over more challenging types of scenes? The scenes on which our work has focused are relatively simple. How does quantification work if the domain size is further increased? For example, one might expect to find that, similar to the findings of Chapter 6, the participants would produce even more vague quantifiers, more incompleteness, and so on. Scenes could also be populated by more naturalistic objects, standing in more naturalistic situations (e.g. a person walking a dog). Evidently, naturalistic scenes permit many more than 2 attributes, each of which will tend to have more than 2 values, and so on. Naturalistic scenes threaten to undermine one of our ideas on which our algorithm rests, namely to start computing the set of all possible scenes (i.e., constructing $\mathcal{S}$), and to work by chipping away from that set. Suppose one wants to describe the people in a football stadium, saying something like:

(106) Nearly everyone in the stadium was wearing the Liverpool colours.

It is unclear what were all the possibilities that this description is trying to rule out since it is difficult to determine all the things people might be wearing. Furthermore, it seems likely that the aim of the utterance is to state that the situation in the stadium runs counter to normal expectations – an aspect of quantification that has been noted widely in the literature (Moxey & Sanford, 1993), but was not covered by our models so far.

One possible solution is to abandon the idea of starting from the complete set of all possibilities, starting instead from a suitably sized *sample* of possible scenes, possibly gleaned from other football matches in the same stadium, proceeding as before in other ways (e.g., terminating when all distractor scenes from the sample have been ruled out). Note that this approach would be sensitive to constraints and statistical regularities that the speaker and hearer are attuned to. For instance, the sample would tend to bear out the regularity that if one's left shoe is brown then so is one's right shoe. More interestingly, a large-enough sample of scenes could go a long way towards embodying our "normal expectations" regarding the outfits that people in stadiums normally wear, including that expectation that the Liverpool colours do not normally dominate to such an extent.

8.2.4 Surface Realisation

As we have pointed out that simpleNLG-ZH assumes that many choices are made outside the system (i.e., by a person or by another component of the NLG system). These include the choice of classifiers, the use of particles (such as "的" (de) and "了" (le)), the choice between different negation words ("不" (bù) or "没" (méi)), the use of aspect markers, and ordering the modifiers and specifiers (see more details in §7.2.3). We tried to let data-driven methods help the choice of classifiers in §7.3. The evaluation suggested that BERT can accomplish the task with a remarkably good performance and can even defeat human beings (see §7.4). In future, we will be concerned with other choices, including the use of particles, negation words, aspect markers, and so on.

8.3 Concluding Remarks

Focusing on the "coolness" hypothesis, this thesis has studied three noun phrase generation tasks, i.e., one-shot referring expression generation (§4), referring expression generation in context (§5) and the quantified description generation (§6), and the realisation of noun phrases (§7). We identified considerable evidence supporting the coolness hypothesis. Nevertheless, we also found a handful of evidence against the hypothesis, suggesting that Chinese speakers are not always more brief than English speakers (see §8.1.3). We argued that, to say the last word in the coolness hypothesis, there is a need for further language comparison experiments on multiple different languages (rather than merely English and Chinese).

We hope this thesis can pave the way for further computational research on other phenomena link to the coolness hypothesis, such as discourse markers, aspect markers, definiteness and so on (see §3.1 and §8.2) because *prima facie* evidence suggests that these, too, are linguistic phenomena where coolness plays an important role, but where one PhD project did not suffice to do them justice.

APPENDIX A

Syntax of the Chinese Noun Phrase: a Brief Synopsis

In this section, we introduce the grammar of Mandarin Noun Phrases (NPs). Our review mostly follows the book: *The Syntax of Chinese* (C.-T. J. Huang et al., 2009), emphasising points that are of particular relevance to this thesis. Note that since the focus of this thesis is the NP, the grammar of other types of phrase (e.g., verb phrase) or structures beyond phrase level (e.g., passive construction, ba construction, and so on) will not be covered in this section. [1] Also note that since most concepts discussed in this section are about all Chinese languages rather than Mandarin specific, in what follows, we will use the term "Chinese" to refer to the language we are discussing.

A.1 Categories

We start by talking about the "units" in Chinese NPs. Then, the question is *what is the units in Chinese?* There has been a considerable amount of evidence suggesting that Chinese is a *morpheme-based* language rather than a *word-based* language. Note that, unlike Western Languages, in Chinese, a morpheme is often a character. The meaning of a Chinese word is often the composition of its morphemes. For example, the meaning of the word "紫花" (zǐhuā; *purple flower*) is the composition of "紫" (zǐ; *purple*) and "花" (huā; *flower*). Indeed, there are exceptions, where the compositionality is lacking. The combination of "红" (hóng; *red*) and "花" (huā; *flower*) is not "*red flower*", but "*saffron*" (i.e., a kind of medicine). Nevertheless, in this thesis, we use "word" as the unit of Chinese. This is because Chinese has derivational morphemes but lacks inflectional morphemes. [2] This results in the fact that Chinese is an analytic language, i.e., a language that has no inflectional morpheme to convey grammatical relationships, such as grammatical agreement or morphophonemic and paradigmatic alternative (Arcodia & Basciano, 2017; Packard, 2000). For example, in

[1] §7 provides a brief introduction of structures other than the NP structure in order to build a Mandarin surface realiser.

[2] English is mostly analytical but less analytic than Chinese.

Gender/Person	Singular	Plural
1st Person	我 wǒ	我们 wǒmén
2nd Person	你 nǐ	你们 nǐmén
3rd Person + Male	他 tā	他们 tāmén
3rd Person + Female	她 tā	她们 tāmén
3rd Person + Neutral	它 tā	它们 tāmén

Table A.1: List of pronouns in Chinese.

Chinese, there is no need to mark verbs that are third-person singular. Therefore, in this subsection, we introduce the basic word categories in Chinese NPs. It is also worth noting that we hereby only introduce the categories that are parts of NPs, categories such as verb, aspectuality, and clause-typer will not be covered.

A.1.1 Noun

The Noun plays a central role in an NP. One characteristic of Nouns in Chinese is that they cannot be modified by the negation morpheme "不" (bù; *not*), such as:

(107) * 不 新闻
bù xīnwén
not news

Such a characteristic has been used to tell the difference between Nouns and Verbs, namely *bu*-test (Y. Li, 1997).

A.1.2 Pronoun

Pronouns in Chinese are used in a similar way as in English. The form of a pronoun changes with respect to gender, plurality, and person. Table A.1 lists all pronouns in Chinese. Note that the gender is marked only in the written form.

A.1.3 Localiser

One special type of word in Chinese appears when describing the location of an object:

(108) 在 桌子 上
zài zhuōzì shàng
on the table

Such information is always expressed by using the prepositional phrase in English (i.e., "*on the table*"), but, in Chinese, the meaning of "on" is expressed by a *localiser* "上" (shàng) rather than the preposition "在" (zài). There has been a long-term argumentation towards the true category of localisers. Candidates include nouns, sub-class of nouns, or being as a separate category. Following the suggestion of C.-T. J. Huang et al. (2009), in this thesis, we view localiser as a separate category.

There are two types of localisers: monosyllabic localisers and disyllabic localisers. For example, the disyllabic version of “上” is “上面”. Although they have the same meaning, practically, disyllabic localisers allow particle “的” (de) (e.g., expression (109-a) and (109-b)) while monosyllabic localisers do not allow it (e.g., expression (109-c) and (109-d)).

(109) a. 在 桌子 上面
zài zhuōzì shàngmiàn
on the table
b. 在 桌子 的 上面
zài zhuōzì de shàngmiàn
on the table
c. 在 桌子 上
zài zhuōzì shàng
on the table
d. * 在 桌子 的 上
zài zhuōzì de shàng
on the table

A.1.4 Adjective

Unlike most western languages, a Chinese adjective can function as a predicate without the help of a copular verb. [3] Or more precisely, it rejects the copular “是” (shì; *be*). For example, the grammatically correct sentence (110-a) contains no verb where the adjective phrase “很 漂亮” (hěn piàoliàng; *very beautiful*) acts as a predicate on its own, and it becomes un-grammatical when a copular is inserted (i.e., the sentence (110-b)).

(110) a. 她 很 漂亮
tā hěn piàoliàng
She is beautiful.
b. * 她 是 很 漂亮
tā shì hěn piàoliàng
(lit.) She is very beautiful.

Although, as shown in example (110-a), a Chinese adjective can act similarly to a verb (phrase), there are still significant differences between adjectives and verbs. One major difference is related to the following use of adjectives: when a sentence is describing a situation where there are two participants, in order to modify one of the two participants, an adjective needs to be introduced by a particle “对” (duì). For example, the use of the verb “适合” (shìhé; *suit*) and the adjective “合适” (héshì; *suitable*) is different:

(111) a. 这个 工作 对 你 很 合适
zhègè gōngzuò duì nǐ hěn shìhé
This job is very suitable for you.
b. 这个 工作 很 适合 你
zhègè gōngzuò hěn héshì nǐ

[3] Note that not all adjectives are predicative.

This job suits you very much.

c. * 这个 工作 很 合适 你
zhègè gōngzuò hěn shìhé nǐ
This job suitable you very much.

Either example (111-a) or example (111-b) is grammatically correct, but the example (111-c) is not.

Another interesting use of adjectives relates to the use of reduplication patterns among disyllabic predicative words in Chinese (Zhu, 1982). Specifically, the reduplication pattern for verbs is:

$$AB \rightarrow ABAB,$$

such as "检查" (jiǎnchá; *check*) → "检查检查" (jiǎnchájiǎnchá; *check*), while that for adjectives is:

$$AB \rightarrow AABB,$$

such as "简单" (jiǎndān; *simple*) → "简简单单" (jiǎnjiǎndāndān; *simple*). [4]

A.1.5 Preposition

The prepositions in Chinese include "至于" (zhìyú; *as for*), "关于" (guānyú; *about*), "从" (cóng; *from*), "给" (gěi; *to/for*), "在" (zài; *at*), "向" (xiàng; *toward*), "把" (bǎ), "被" (bèi)[5], and so on. In addition to the usages that are similar to English prepositions, propositions in Chinese have three other characteristics.

First, "至于" and "关于" have to occur with the NP in a pre-subject position:

(112) 关于这件事， 他们已经讨论过了。
guānyú zhè jiànshì, tāměn yǐjìng tǎolùn guòlè
Regarding this issue, they already discussed (it).

Second, "给", "在" and "向" can act as verbs:

(113) 他 给了 我 一把 剑。
tā gěilè wǒ yìbǎ jiàn
He gave me a sword.

In this example, "给" (gěi; *to/for*) is expressing the meaning of "*give*". Last, "把" and "被" are used to build "把" construction and passive construction.

A.1.6 Functional Categories

In line with many languages, Chinese makes use of function words to construct phrases/-clauses/sentences. We introduce 4 function categories that are regularly used in Chinese NPs. The following example NP from C.-T. J. Huang et al. (2009):

(114) 那 一 杆 枪
nà yì gǎn qiāng
that gun

[4] Note that modifier-head adjectival compounds are exception. They reduplicate as ABAB rather than AABB.

[5] Note that the classification of "把" (bǎ) and "被" as prepositions is questioned by many linguists.

contains three different categories of function words:

1. A *Determinor* “那” (nà; *that*), which explicitly mark the current NP as a definite NP[6];
2. A *Numeral* “一” (yì; *one*), indicating there is only one gun; and
3. A *Classifier* “杆”， indicting “*gun*” belongs to a class of objects with the general shape and texture of a thin shaft. This category resembles the word “*piece*” in the context of “*a piece of*” in English, but the major difference is that, in Chinese, every noun could associate with a classifier.

The last one is the *Particle*, a typical example of which is the word “的” (dè). Practically, it appears in the syntactic context [X 的 Y] (C.-T. J. Huang et al., 2009), where if Y is a noun, then X could be any of a noun (phrase), an adjective (phrase), a preposition (phrase), or a full clause. It turns a phrase inside a larger NP into a modifier, for example:

(115) a. 这位学者的观点
zhè wèi xuézhě de guāndiǎn
this scholar’s opinion
b. 十分诱人的条件
shífēn yòurén dè tiáojiàn very enticing term
c. 关于战争的传言
guānyú zhànzhēng dè chuányán
gossip about war
d. 我去国外的理由
wǒ qù guówài dè lǐyóu
the reason for my going abroad

In turn, these examples have Xs that is a noun phrase, an adjective phrase, a preposition phrase, and a clause.

A.2 Structure of NPs in Chinese

One important characteristic of Chinese NPs is their “simplicity”. [7] Consider the bare noun “狗” (gǒu; *dog*) in the following example sentences from C.-T. J. Huang et al. (2009):

(116) a. 狗很聪明。
gǒu hěn cōngmíng
Dogs are intelligent.
b. 我看到狗。
wǒ kàn dào gǒu
I saw a dog/dogs.
c. 狗跑走了。
gǒu pǎo zǒu le
The dog(s) ran away.

[6] The definitness of Chinese NPs is not always expressed explicitly.

[7] A more precise terminology of the subject matter we aim at here is “nominal phrase”, but, for simplicity, we keep using the abbreviation “NP”.

From this example, we could easily find that the bare noun “狗” could be either a definite NP (sentence (116-c)) or an indefinite NP (sentence (116-a) and (116-b)); or either a singular noun or a plural noun (sentence (116-b) and (116-c)). In other words, a bare noun in Chinese equals to [(definite/indefinite) article + (singular/plural) noun] in English. Conversely, in some of the situations, Chinese NPs might also appear to be more “complex”. For example, when counting, a Chinese NP needs classifier (e.g., the classifier “本” in the NP “三本书” (sānběnshū; *three books*)) while a English NP only combines the number with a noun in plural form: “*three books*”.

Also note that in this review, we call a determiner phrase (DP) an NP, which is not fully correct in theoretical linguistics. For example, “*the book*” is a DP containing a NP “*book*”. Interestingly, as discussed above, a Chinese NP can act as an English DP. Although in light of Hong and Shi (2013) and C.-T. J. Huang et al. (2009), there are benefits to interpreting every nominal phrase in Chinese as a DP, for this thesis, given our subject matter (i.e., NLG), there is no risk to call all of them (i.e., Chinese nominal phrase, Chinese NP, Chinese DP, and English DP) as NPs. In what follows, we start to introduce the structure and sub-structures of Chinese NPs.

A.2.1 Number Phrase

One major constituent of the Chinese NP is the number phrase, which is constructed through the form [number + classifier + noun], such as “三本书” (sānběnshū; *three books*). Generally speaking, Chinese number phrases are regarded as non-definite expressions (e.g., sentence (117-b)) and do not occur in subject or topic positions (e.g., sentence (117-a)). On the contrary, bare nouns in subject and topic positions are definite expressions.

However, such a statement (i.e., number phrases cannot be placed in subject or topic positions) is not always true. Y.-h. A. Li (2006) argued that number phrases can be allowed in subject or topic positions if they involve the notion of “quantity". For example, the sentence (117-a) is un-grammatical while the sentence (117-c) is grammatical. This is because the “大概” (da4gài; *probably*) in (117-c) expresses the sufficiency of a certain amount, indicting the the number phrase in its subject position (i.e., “三 个 学生” (sān gè xuéshēng; *three students*)) denotes quantity.

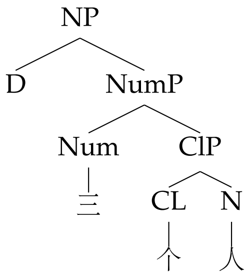
(117) a. * 三 个 学生, 我 以为 吃了 蛋糕。
sān gè xuéshēng, wǒ yǐwéi chīlè dàngāo
Three students, I thought (they) ate the cake.
b. 学生 吃了 蛋糕。
xuéshēng chílè dàngāo
The students ate the cake.
c. 三 个 学生， 我 想 大概 吃不完 两 个 蛋糕。
sān gè xuéshēng, wǒ xiǎng dàgài chībùwán liǎng gè dàngāo
Three students, I think probably cannot finish two cakes.

C.-T. J. Huang et al. (2009) labelled example (117-c) as a *Quantity-denoting Expression* and example (117-a) as *Indefinite Individual-denoting Expression* to highlight the fact that they refer to some entities/individuals (indefinite referents). Pragmatically, a quantity-denoting expression does not co-refer with or bind a pronoun while an in-definite individual denoting expression can be co-indexed with referential or bound pronouns. For example, the sentence (118) is not acceptable since the number phrase “三 个 人” (sān gè rén; *three people*) cannot be referred by the pronoun “他们”. We need a definite expression to replace it.

(118) * 三 个 人 抬不起 两架 你 给 他们 的 钢琴。
sān gè rén táibùqǐ liángjià nǐ gěi tāmén de gāngqín
Three people cannot lift two (of the) pianos that you gave to them.

The actual position of a number phrase in an NP is shown in (119), where D is “determiner”, NumP is “number phrase”, ClP is “classifier phrase” and N is noun.

(119)



From now on, we head to the rest components in the tree.

A.2.2 Demonstrative

If demonstratives are in D position of (119), we should find [demonstrative + number + classifier + noun], for example:

(120) 这/那 三 个 人
zhè/nà sān gè rén
this/that three people

A demonstrative is sometimes followed by a classifier directly, without a number, although one may argue that the number "one" is present underlying because the interpretation is singular:

(121) 这/那 个 人
zhè/nà gè rén
this/that person

A.2.3 Pronoun

Both of the pattern [pronoun + number + classifier (+ noun)] and the pattern [pronoun + noun] are possible:

(122) a. 他们 两 个 人/学生
tāmén liǎng gè rén/xuéshēng
(lit.) them two people/students

b. 他们 学生
tāmén xuéshēng

(lit.) them students

c. 他们 三 个

tāmén sān gè

(lit.) them three

However, when the number and classifier expressions do not occur, the pronoun must be plural:

(123) * ta xuesheng

tā xuéshēng

(lit.) he student

These expressions can occur in all argument positions.

Given the structure in (119), since pronouns are similar to the definite article in English, pronoun occupies the D position. As a matter of fact, pronouns and demonstratives, which have both been claimed to occupy the D position, can occur together:

(124) 我 喜欢 你们 这些 乖 孩子

wǒ xǐhuān nǐmén zhěxiē guāi háizì

(lit.) I like you these good kids.

In this case, we say they are in a double-headed D position or two separate D positions.

A.2.4 Proper Names

Proper names in Chinese could also occur in the D position, followed by a pronoun or a demonstrative in the D position and a number expression. This said, the following two structures are acceptable: [proper name + pronoun/demonstrative + number + classifier + noun] (example (125-a)) or [proper name + pronoun + demonstrative (+ noun phrase)] (example (125-b)).

(125) a. 我 喜欢 张三 和 李四 他们 几个 乖 孩子 。

wǒ xǐhuān zhāngsān hé lǐsì tāmén jǐge guāi háizi

(lit.) I like Zhangsan, Lisi those several good kids.

b. 我 喜欢 张三 他 这个 用功 的 学生 。

wǒ xǐhuān zhāngsān tā zhègè yònggōng de xuéshēng

(lit.) I like Zhangsan him this diligent student.

In the same NP, the pronoun does not need to agree with the proper name in number, but the pronoun needs to be plural if the number following the pronoun is more than one:

(126) a. 我 喜欢 张三 他们 (那) 三个 。

wǒ xǐhuān zhāngsān tāmén (nà) sāngè

(lit.) I like Zhangsan those three.

b. * 我 喜欢 张三 他 (那) 三个 。

wǒ xǐhuān zhāngsān tā (nà) sāngè

(lit.) I like Zhangsan him those three

Unlike pronouns, proper names cannot precede nouns directly:

(127) × 我 喜欢 张三 和 李四 学生 。
wǒ xǐhuān zhāngsān hé lǐsì xuéshēng
(lit.) I like Zhangsan and Lisi students.

To fix (127), a number expression or a pronoun/demonstrative is required.

A.2.5 Common Nouns

Common nouns in Chinese can sometimes function as proper names, and proper names can sometimes function as common nouns. In other words, in addition to the N position in Figure (119), a common noun, acting as a proper name, can also precede [(pronoun/demonstrative) + number + classifier] (i.e., being in the D position). The common noun "弟弟" (dìdì; *little brother*) takes the D position in the following sentence and acts as a proper name.

(128) 弟弟 他 一个 人 就 解决了 问题 。
dìdì tā yígè rén jiù jiějuéle wèntí
My little brother solved the whole problem on his own.

A.3 Other Issues in Chinese NP

In this sub-section, we discuss a few remaining issues related to Chinese NPs.

A.3.1 Plurality

For expressing plurality, on the one hand, as discussed, a bare noun in Chinese can denote plurality. On the other hand, the Chinese do not have much inflectional morphology. One specific plural morpheme that is worth mentioning here is the morpheme "们" (mén), which inflects merely pronouns as well as human nouns. However, Y.-h. A. Li (1999) argued that "们" is somehow more like a "collective" marker rather than the traditionally understood plural morpheme. A number phrase with a common noun as the head is incompatible with "们":

(129) * 三 个 学生们
sán gè xuéshēng mén
three students

Conversely, consider the following example, "孩子们" (házimèn; *children*) with a "们" morpheme to refer to a definite group. Without "们", this NP becomes indefinite and vague.

(130) 我 去 找 孩子们
wǒ qù zhǎo háizimèn
I will go find the children.

In nutshell, C.-T. J. Huang et al. (2009) summarised the use of "们" using the following rules:

- “们” can be suffixed to pronoun, proper name, and some common nouns;
- Common nouns with “们” must be interpreted as definite;
- The “们” attached to proper names can sometimes be interpreted as a plural marker if it refers to individuals with same/similar name/property;
- A pronoun/proper name with “们” can be followed, but not preceded, by a number phrase. In the cases with proper names, it only has a collective reading.

A.3.2 Distributive Marker

As discussed above, a pronoun with “们” followed by a number phrase could have a plural reading. However, it can only be done with the help of a distributive marker “都” (dōu):

(131) 他们 两个 都 结婚 了 。
tāmén liǎnggè dōu jiéhūn le
Both of them got married.

It must be about two marriages, rather than the two of them being married to each other). This concludes our brief synopsis of the syntax of the Chinese NPs.

APPENDIX B

Samenvatting

In het baanbrekende werk van de taalkundige James Huang worden menselijke talen onderverdeeld in “koele” talen (d.w.z. talen die meer afhankelijk zijn van context) en “warme” talen (d.w.z. talen die minder afhankelijk zijn van context). Mandarijn wordt beschouwd als een schoolvoorbeeld van koele talen, veel koeler dan westerse talen zoals het Engels en het Nederlands; met andere woorden, Huang veronderstelde dat de beoogde betekenis van uitdrukkingen in het Mandarijn meer afhangt van de context dan die van hun Engelse tegenhangers. Voortbouwend op dit idee lijkt het aannemelijk dat woorden en zinsdelen in het Mandarijn eerder worden weggelaten of ondergespecificeerd dan in het Engels, mits hun context de lezers voldoende informatie kan bieden omde bedoelde betekenis af te leiden.

James Huang introduceerde oorspronkelijk “koelte” in verband met het gebruik van anafora in het Mandarijn en het Engels. Concreet voerde hij aan dat Engels een warme taal is omdat Engelse voornaamwoorden over het algemeen niet kunnen worden weggelaten, terwijl Mandarijn koel is, omdat de voornaamwoorden meestal op heel natuurlijke wijze kunnen worden weggelaten. Als iemand bijvoorbeeld vraagt: “Heeft John Tom gisteren gezien?”, dan kan een Mandarijn-spreker eenvoudig “看见了” (kanjian le, saw) antwoorden om de betekenis uit te drukken “Hij zag hem”. In dit voorbeeld worden voornaamwoorden in zowel de onderwerppositie als de objectpositie verwijderd. Daarentegen kan het Engelse woord “saw” (en evenzo het Nederlandse word “keek”) op zichzelf geen volledige en grammaticaal correcte zin vormen. Het weglaten van voornaamwoorden wordt “pro-drop” genoemd. Voornaamwoorden (d.w.z. woorden zoals hij en hem) en weggelaten voornaamwoorden (d.w.z. pro-drop) zijn twee verschillende soorten anaforische (d.w.z. verkorte) verwijzingen.

In later werk is het begrip koelte soms in een bredere zin opgevat en omvat het andere fenomenen dan anafora. Het is bijvoorbeeld in verband gebracht met de afweging tussen duidelijkheid en beknoptheid in taalgebruik. Er is gesuggereerd dat sprekers van “koele” talen de neiging hebben om hun uitingen korter maar minder duidelijk te houden dan sprekers van “warme” talen. Deze suggestie zou de impact van koelte op het taalgebruik heel breed maken. In dit proefschrift hebben we besloten om ons te concentreren op zelfstandige naamwoorden en ons te richten op het begrijpen en valideren van de koelte-hypothese op Mandarijn zelfstandige naamwoorden.

We hebben dit probleem aangepakt met behulp van Natural Language Generation Technieken. In het bijzonder voeren we experimenten uit om erachter te komen wat Mandarijnsprekers in een bepaalde situatie zeggen, vergelijken dit met wat Engelssprekenden zeggen, we controleren of de uitkomsten in overeenstemming zijn met de "coolness"-hypothese, en we bouwen modellen voor het genereren van natuurlijke taal om het gedrag van menselijk sprekers na te bootsen. De resulterend computermodellen helpen ons beter te begrijpen hoe mensen spreken. Omgekeerd kan het begrijpen van menselijke spraakpatronen ons helpen om betere systemen voor het genereren van natuurlijke taal voor het Mandarijn te bouwen, ook voor praktische doelen.

We waren benieuwd naar twee soorten Noun Phrases. De eerste soort is die van refererende expressies. Stel, bij voorbeeld, dat Tom de enige student in de bus is en hij draagt een bril. In de bus zitten 20 andere mensen en 3 van hen dragen een bril. Om naar Tom te refereren (i.e., te verwijzen), kan men de verwijzende uitdrukking "de student" gebruiken die alle andere objecten in de bus (bijvoorbeeld andere mensen, stoelen, enz.) uitsluit. We zijn geïnteresseerd in twee onderzoeksvragen:

- Onder welke omstandigheden kunnen verwijzende uitdrukkingen worden ondergespecificeerd of zelfs geheel worden weggelaten (d.w.z. wanneer pro-drop plaatsvindt);
- Onder welke omstandigheden hebben sprekers de neiging om verwijzingen uitvoeriger te specificeren dan strikt nodig is? Dit gebeurt, in bovenstaand voorbeeld, als we zeggen "de student die een bril draagt", omdat "die een bril draagt" niet helpt om eventuele "afleiders" uit te sluiten.
- Onder welke omstandigheden hebben sprekers de neiging om te weinig te specificeren? Bijvoorbeeld, "de persoon die een bril draagt" zou niet uniek naar Tom verwijzen, omdat er andere mensen in de bus zijn die ook een bril dragen.

Het andere type zelfstandige naamwoorden waarop we ons concentreren, zijn kwantificerence uitdrukkingen. Om de situatie van mensen in de bovengenoemde bus te beschrijven, zou je kunnen zeggen: "Er zit maar één student in de bus." Of "Een paar mensen in de bus dragen een bril". We onderzoeken welke soorten gekwantificeerde uitdrukkingen Mandarijn en Engelssprekenden gebruiken en hoe ze gekwantificeerde uitdrukkingen anders gebruiken.

Bij het modelleren van de productie van deze twee soorten zelfstandige naamwoorden hebben we verschillende rekenmodellen gebruikt: van klassieke op regels gebaseerde modellen tot op neurale netwerken gebaseerde modellen. Hoewel state-of-the-art neurale modellen vaak beter presteren dan klassieke, op regels gebaseerde modellen, gedragen deze modellen zich vaak als 'zwarte dozen' die moeilijk te koppelen zijn aan linguïstische of andere inzichten. Bovendien vonden we dat neurale modellen het soms niet goed doen op end2end Natural Language Generation. Samengevat stellen we dat de toekomst toebehoort aan hybride systemen die verschillende rekenmethoden combineren.

We koppelen de verschijnselen die we tijdens onze studies hebben waargenomen aan de koelte-hypothese. Als de koelte-hypothese juist is, dan is het aannemelijk dat we in het Mandarijn minder overspecificaties en meer onderspecificaties kunnen vinden dan in het Engels. Hoewel het meeste bewijsmateriaal dat we gevonden hebben "koelte" ondersteunt, hebben we ook soms evidentie voor het omgekeerde gevonden. Dit suggereert dat de koelte-hypothese in bepaalde situaties opgaat, maar niet altijd. We hopen dat ons werk de weg zal effenen voor computationeel onderzoek naar andere verschillen tussen talen, en dat het zal leiden tot betere systemen voor het genereren van teksten in het Mandarijn.

APPENDIX C

Curriculum Vitae

Guanyi Chen was born on January 27, 1993, in Beijing. He did his undergraduate study in e-Commerce Engineering with Law at both the Beijing University of Posts and Telecommunications and the Queen Mary University of London from 2011 to 2015. In 2016, he received his Master's degree in Artificial Intelligence at the University of Edinburgh. He started his journey as a PhD student at the Department of Computing Science of the University of Aberdeen in 2017. He moved to the Department of Information and Computing Sciences of Utrecht University in 2018. From 2019 to the end of 2021, he was founded by the China Scholarship Council scholarship.

He has also interned in multiple research centres in the industry. These include Samsung Research Center Beijing (2017), Microsoft Research Asia (2019), and Samsung AI Lab (2021). Starting from 2022, he has worked as a lecturer/researcher at the Department of Information and Computing Sciences of Utrecht University.

Publications

1. Chen, G., van Deemter, K., & Lin, C. (2018a). Modelling pro-drop with the rational speech acts model. *Proceedings of the 11th International Conference on Natural Language Generation*, 159–164. https://doi.org/10.18653/v1/W18-6519

2. Chen, G., van Deemter, K., & Lin, C. (2018b). SimpleNLG-ZH: A linguistic realisation engine for Mandarin. *Proceedings of the 11th International Conference on Natural Language Generation*, 57–66. https://doi.org/10.18653/v1/W18-6506

3. Mao, R., Chen, G., Li, R., & Lin, C. (2018). ABDN at SemEval-2018 task 10: Recognising discriminative attributes using context embeddings and WordNet. *Proceedings of The 12th International Workshop on Semantic Evaluation*, 1017–1021. https://doi.org/10.18653/v1/S18-1169

4. Li, R., Lin, C., Collinson, M., Li, X., & Chen, G. (2019). A dual-attention hierarchical recurrent neural network for dialogue act classification. *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, 383–392. https://doi.org/10.18653/v1/K19-1036

5. Chen, G., & Yao, J.-G. (2019). A closer look at recent results of verb selection for data-to-text NLG. *Proceedings of the 12th International Conference on Natural Language Generation*, 158–163. https://doi.org/10.18653/v1/W19-8622

6. Chen, G., van Deemter, K., & Lin, C. (2019). Generating quantified descriptions of abstract visual scenes. *Proceedings of the 12th International Conference on Natural Language Generation*, 529–539. https://doi.org/10.18653/v1/W19-8667

7. Chen, G., van Deemter, K., Pagliaro, S., Smalbil, L., & Lin, C. (2019). QTUNA: A corpus for understanding how speakers use quantification. *Proceedings of the 12th International Conference on Natural Language Generation*, 124–129. https://doi.org/10.18653/v1/W19-8616

8. Zheng, Y., Chen, G., Huang, M., Liu, S., & Zhu, X. (2019). Persona-aware dialogue generation with enriched profile. *NeurIPS 2019 on Conversational AI Workshop, Vancouver*

9. Chen, G., & van Deemter, K. (2020). Lessons from computational modelling of reference production in Mandarin and English. *Proceedings of the 13th International Conference on Natural Language Generation*, 263–272. https://www.aclweb.org/anthology/2020.inlg-1.33

10. Zheng, Y., Chen, G., & Huang, M. (2020). Out-of-domain detection for natural language understanding in dialog systems. *IEEE/ACM Transactions Audio, Speech, and Language Processing*

11. Li, X., Chen, G., Lin, C., & Li, R. (2020). DGST: A dual-generator network for text style transfer. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 7131–7136. https://doi.org/10.18653/v1/2020.emnlp-main.578

12. Li, R., Li, X., Chen, G., & Lin, C. (2020). Improving variational autoencoder for text modelling with timestep-wise regularisation. *Proceedings of the 28th International Conference on Computational Linguistics*, 2381–2397. https://doi.org/10.18653/v1/2020.coling-main.216

13. van Miltenburg, E., Lu, W.-T., Krahmer, E., Gatt, A., Chen, G., Li, L., & van Deemter, K. (2020). Gradations of error severity in automatic image descriptions. *Proceedings of the 13th International Conference on Natural Language Generation*, 398–411. https://www.aclweb.org/anthology/2020.inlg-1.45

14. Chen, G., Zheng, Y., & Du, Y. (2020). Listener's social identity matters in personalised response generation. *Proceedings of the 13th International Conference on Natural Language Generation*, 205–215. https://www.aclweb.org/anthology/2020.inlg-1.26

15. Chen, G., & van Deemter, K. (2021a). Computational modeling of quantifier use: Elicitation experiments, models, and evaluation. *Journal Paper in Preparation*

16. Chen, G., & van Deemter, K. (2021b). Varieties of specification: Redefining over- and under-specification for an enhanced understanding of referring expressions. *Journal Paper in Preparation*

17. Peng, X., Chen, G., Lin, C., & Stevenson, M. (2021). Highly efficient knowledge graph embedding learning with Orthogonal Procrustes Analysis. *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2364–2375. https://doi.org/10.18653/v1/2021.naacl-main.187

18. Zeng, C., Chen, G., Lin, C., Li, R., & Chen, Z. (2021). Affective decoding for empathetic response generation. *Proceedings of the 14th International Conference on Natural Language Generation*, 331–340. https://aclanthology.org/2021.inlg-1.37

19. Zheng, Y., Chen, G., Liu, X., & Lin, K. (2021). Mmchat: Multi-modal chat dataset on social media. *CoRR*, *abs/2108.07154*. https://arxiv.org/abs/2108.07154

20. Chen, G., Same, F., & van Deemter, K. (2021). What can neural referential form selectors learn? *Proceedings of the 14th International Conference on Natural Language Generation*, 154–166. https://aclanthology.org/2021.inlg-1.15

21. Jarnfors, J., Chen, G., van Deemter, K., & Sybesma, R. (2021). Using BERT for choosing classifiers in Mandarin. *Proceedings of the 14th International Conference on Natural Language Generation*, 172–176. https://aclanthology.org/2021.inlg-1.17

22. Same, F., Chen, G., & van Deemter, K. (2022). Non-neural models can matter: A re-evaluation of neural referring expression generation systems. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*

Acronyms

OntoNotes A dataset that was constructed in the OntoNotes project, which was built on two time-tested resources, following the Penn Treebank for syntax and the Penn PropBank for predicate-argument structure. 96, 99–101, 113–116, 118, 119, 121–123, 199

OntoNotes-c Character based OntoNotes where samples longer than 512 characters are removed. 115–117, 120

E2ENLG The End2End NLG Challenge Corpus. 21, 23, 26

ETUNA English TUNA. 5, 6, 38, 64–67, 69, 79, 81, 82, 84–90, 92, 93, 195–197, 202

MQTUNA Mandarin Quantified TUNA. 6, 8, 128, 129, 140–148, 165, 166, 197, 198, 203

MTUNA Mandarin TUNA. 5, 6, 8, 58, 62–67, 69, 72, 73, 79, 80, 82, 84–89, 92, 93, 170, 180–182, 195–197, 199, 202

QTUNA Quantified TUNA. 6, 8, 65, 128, 129, 132–136, 139–141, 144–148, 151, 153, 157, 160, 166, 197, 198, 203

simpleNLG-EN The original SimpleNLG system. A suffix is used for distinguishing from SimpleNLG-ZH.. 170, 171, 173, 174, 177–180

simpleNLG-ZH A Mandarin realisation engine following the tradition of SimpleNLG. 7, 167–180, 182, 183, 186, 192, 193, 198, 205

simpleNLG The original SimpleNLG system.. 7, 9, 167–171, 175, 180, 182, 192, 198

TUNA Towards a UNified Algorithm for the Generation of Referring Expressions. 8, 37–39, 58, 62, 64, 71, 73, 75, 84, 92, 93, 129, 132

webNLG The WebNLG corpus. 43, 96, 97, 102, 103, 106–110, 112–114, 118, 121–123, 202

BERT Bidirectional Encoder Representations from Transformers. 113, 115–117, 121–123, 184–190, 192, 198, 199, 205

FB The Full Brevity Algorithm for Referring Expression Generation. 33, 34, 87–90, 197

GR The Greedy Algorithm for Referring Expression Generation. 34, 35, 87, 91

IA The Incremental Algorithm for Referring Expression Generation. 34–37, 39, 87–90, 93, 197

LSTM Long and Short Term Memory. 19, 25, 41, 44, 185, 186

MLM Masked Language Model. 184, 186, 187

NNLG	Neural Natural Language Generation. 19, 20, 25–27
PRO	Probabilistic Referential Over-specification. 202
RNN	Recurrent Neural Network. 19, 20, 23, 24, 198
RSA	Rational Speech Act. 48, 97–99, 102, 123, 198, 202, 203
SGNS	Skip-Gram with Negative-Sampling. 117, 121, 198
AZP	Anaphoric Zero Pronoun. 97, 99, 101, 102
DICE	Sorensen–Dice coefficient. 87, 88, 90, 92, 202
DNZP	Deictic Non-anaphoric Zero Pronoun. 97, 99–102
MR	Meaning Representation. 21
NLG	Natural Language Generation. 1, 3–5, 11–14, 16–19, 21, 23, 25–30, 35, 37, 43, 48, 51, 53, 55–58, 69, 127, 146, 167, 168, 183, 198–201, 205, 212
NP	Noun Phrase. 2–5, 44, 66, 83, 84, 123, 126, 168, 169, 172, 173, 180, 192, 195, 196, 200, 201, 207, 208, 210–216
NZRE	Non-Zero from of Referring Expression. 99–102
QD	Quantified Description. 6–8, 129, 130, 139, 141, 144–150, 152, 153, 155, 157, 159, 162, 163, 165, 166, 197–200, 203
QDG	Quantified Description Generation. 149–152, 165, 166, 198
QE	Quantified Expression. 2, 3, 5, 6, 44–48, 129, 132, 133, 136–142, 145–147, 149, 150, 152–157, 162, 164–166, 195–197, 199
RDF	Resource Description Framework. 43
RE	Referring Expression. 2, 3, 5, 7, 8, 29–44, 55, 58, 59, 61–73, 75–80, 82–93, 96–102, 108, 109, 112–114, 117, 118, 121–123, 149, 154, 163, 195–200, 202–204
REG	Referring Expression Generation. 5, 6, 16–18, 29–31, 33, 35, 37–44, 53, 58, 59, 61–66, 68, 69, 79, 84–93, 95–97, 102–105, 108, 113, 123, 154, 159, 160, 163, 197, 201–203
RFS	Referential Form Selection. 6, 8, 95–97, 102, 103, 105, 107–110, 112, 113, 118, 121–123, 197–199, 202, 203
ZP	Zero Pronoun. 4, 6, 56, 96, 97, 99–102, 114, 117, 118, 121–123, 196–199

Bibliography

Alt, C., Gabryszak, A., & Hennig, L. (2020). Probing linguistic features of sentence-level representations in neural relation extraction. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 1534–1545. https://doi.org/10.18653/v1/2020.acl-main.140

Althaus, E., Karamanis, N., & Koller, A. (2004). Computing locally coherent discourses. *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, 399–406. https://doi.org/10.3115/1218955.1219006

Andreas, J., & Klein, D. (2016). Reasoning about pragmatics with neural listeners and speakers. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 1173–1182. https://doi.org/10.18653/v1/D16-1125

Angeli, G., Manning, C., & Jurafsky, D. (2012). Parsing time: Learning to interpret time expressions. *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 446–455. https://aclanthology.org/N12-1049

Appelt, D. E., & Appelt, D. E. (1992). *Planning english sentences*. Cambridge University Press.

Arcodia, G. F., & Basciano, B. (2017). Morphology, modern (R. Sybesma, Ed.). *Encyclopedia of Chinese Language and Linguistics*.

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

Arnold, J. E. (2001). The effect of thematic roles on pronoun use and frequency of reference continuation. *Discourse processes*, *31*(2), 137–162.

Arnold, J. E. (2010). How speakers refer: The role of accessibility. *Language and Linguistics Compass*, *4*(4), 187–203.

Arts, A. (2004). *Overspecification in instructive texts*. Nijmegen: Wolf.

Arts, A., Maes, A., Noordman, L., & Jansen, C. (2011). Overspecification facilitates object identification. *Journal of Pragmatics*, *43*(1), 361–374.

Attali, N., Scontras, G., & Pearl, L. S. (2021). Every quantifier isn't the same: Informativity matters for ambiguity resolution in quantifier-negation sentences. *Proceedings of the Society for Computation in Linguistics*, *4*(1), 394–395.

Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate (Y. Bengio & Y. LeCun, Eds.). http://arxiv.org/abs/1409.0473

Balakrishnan, A., Rao, J., Upasani, K., White, M., & Subba, R. (2019). Constrained decoding for neural NLG from compositional representations in task-oriented dialogue. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 831–844. https://doi.org/10.18653/v1/P19-1080

Banerjee, S., & Lavie, A. (2005). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, 65–72. https://www.aclweb.org/anthology/W05-0909

Bard, E. G., Anderson, A. H., Chen, Y., Nicholson, H. B., Havard, C., & Dalzel-Job, S. (2007). Let's you do that: Sharing the cognitive burdens of dialogue. *Journal of Memory and Language*, *57*(4), 616–641.

Bard, E. G., Aylett, M. P., Trueswell, J., & Tanenhaus, M. (2004). Referential form, word duration, and modeling the listener in spoken dialogue. *Approaches to studying world-situated language use: Bridging the language-as-product and language-as-action traditions*, 173–191.

Barr, D., van Deemter, K., & Fernández, R. (2013). Generation of quantified referring expressions: Evidence from experimental data. *Proceedings of the 14th European Workshop on Natural Language Generation*, 157–161. https://aclanthology.org/W13-2120

Barwise, J., & Cooper, R. (1981). Generalized quantifiers and natural language. *Philosophy, language, and artificial intelligence* (pp. 241–301). Springer.

Barwise, J., & Perry, J. (1981). Situations and attitudes. *The Journal of Philosophy*, *78*(11), 668–691. http://www.jstor.org/stable/2026578

Barzilay, R., & Lapata, M. (2005). Collective content selection for concept-to-text generation. *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, 331–338. https://aclanthology.org/H05-1042

Barzilay, R., & Lapata, M. (2006). Aggregation via set partitioning for natural language generation. *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, 359–366. https://aclanthology.org/N06-1046

Barzilay, R., & Lee, L. (2004). Catching the drift: Probabilistic content models, with applications to generation and summarization. *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, 113–120. https://aclanthology.org/N04-1015

Bateman, J. A. (1997). Enabling technology for multilingual natural language generation: The kpml development environment. *Natural Language Engineering*, *3*(1), 15–55.

Beck, D., Haffari, G., & Cohn, T. (2018). Graph-to-sequence learning using gated graph neural networks. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 273–283. https://doi.org/10.18653/v1/P18-1026

Belinkov, Y., Durrani, N., Dalvi, F., Sajjad, H., & Glass, J. (2017). What do neural machine translation models learn about morphology? *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 861–872. https://doi.org/10.18653/v1/P17-1080

Belinkov, Y., Màrquez, L., Sajjad, H., Durrani, N., Dalvi, F., & Glass, J. (2017). Evaluating layers of representation in neural machine translation on part-of-speech and semantic tagging tasks. *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 1–10. https://aclanthology.org/I17-1001

Bell, A. (1984). Language style as audience design. *Language in society*, *13*(2), 145–204.

Belz, A. (2008). Automatic generation of weather forecast texts using comprehensive probabilistic generation-space models. *Natural Language Engineering*, *14*(4), 431–455.

Belz, A., Kow, E., Viethen, J., & Gatt, A. (2010). Generating referring expressions in context: The GREC task evaluation challenges. In E. Krahmer & M. Theune (Eds.), *Empirical methods in natural language generation: Data-oriented methods and empirical evaluation* (pp. 294–327). Springer. https://doi.org/10.1007/978-3-642-15573-4_15

Belz, A., & Reiter, E. (2006). Comparing automatic and human evaluation of NLG systems. *11th Conference of the European Chapter of the Association for Computational Linguistics*, 313–320. https://aclanthology.org/E06-1040

Belz, A., & Varges, S. (2007). Generation of repeated references to discourse entities. *Proceedings of the Eleventh European Workshop on Natural Language Generation (ENLG 07)*, 9–16. https://aclanthology.org/W07-2302

Biecek, P., & Burzykowski, T. (2021). *Explanatory model analysis: Explore, explain, and examine predictive models*. CRC Press.

Bollmann, M. (2011). Adapting SimpleNLG to German. *Proceedings of the 13th European Workshop on Natural Language Generation*, 133–138. https://aclanthology.org/W11-2817

Bott, L., Bailey, T. M., & Grodner, D. (2012). Distinguishing speed from accuracy in scalar implicatures. *Journal of Memory and Language*, *66*(1), 123–142.

Bott, L., & Noveck, I. A. (2004). Some utterances are underinformative: The onset and time course of scalar inferences. *Journal of memory and language*, *51*(3), 437–457.

Braun, D., Klimt, K., Schneider, D., & Matthes, F. (2019). SimpleNLG-DE: Adapting SimpleNLG 4 to German. *Proceedings of the 12th International Conference on Natural Language Generation*, 415–420. https://doi.org/10.18653/v1/W19-8651

Brennan, S. E. (1995). Centering attention in discourse. *Language and Cognitive processes*, *10*(2), 137–167.

Briggs, G., & Harner, H. (2019). Generating quantified referring expressions with perceptual cost pruning. *Proceedings of the 12th International Conference on Natural Language Generation*, 11–18. https://doi.org/10.18653/v1/W19-8602

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., ... Amodei, D. (2020). Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, & H. Lin (Eds.), *Advances in neural information processing systems* (pp. 1877–1901). Curran Associates, Inc. https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

Busemeyer, J. R., & Wang, Y.-M. (2000). Model comparisons and model selections based on generalization criterion methodology. *Journal of Mathematical Psychology*, *44*(1), 171–189.

Cahill, A., & van Genabith, J. (2006). Robust PCFG-based generation using automatically acquired LFG approximations. *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, 1033–1040. https://doi.org/10.3115/1220175.1220305

Campbell, C. P. (1998). Rhetorical ethos: A bridge between high-context. *The Cultural Context in Business Communication. John Benjamins Publishing Company*, 31–48.

Cao, M., & Cheung, J. C. K. (2019). Referring expression generation using entity profiles. *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. https://doi.org/10.18653/v1/D19-1312

Carcassi, F., & Szymanik, J. (2021). Most vs more than half: An alternatives explanation. *Proceedings of the Society for Computation in Linguistics*, *4*(1), 334–343.

Carroll, J., & Oepen, S. (2005). High efficiency realization for a wide-coverage unification grammar. *Second International Joint Conference on Natural Language Processing: Full Papers*. https://doi.org/10.1007/11562214_15

Carston, R. (2008). *Thoughts and utterances: The pragmatics of explicit communication*. John Wiley & Sons.

Castro Ferreira, T., Krahmer, E., & Wubben, S. (2016). Towards more variation in text generation: Developing and evaluating variation models for choice of referential form. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 568–577. https://doi.org/10.18653/v1/P16-1054

Castro Ferreira, T., Moussallem, D., Kádár, Á., Wubben, S., & Krahmer, E. (2018). NeuralREG: An end-to-end approach to referring expression generation, 1959–1969. https://doi.org/10.18653/v1/P18-1182

Castro Ferreira, T., Moussallem, D., Krahmer, E., & Wubben, S. (2018). Enriching the WebNLG corpus. *Proceedings of the 11th International Conference on Natural Language Generation*, 171–176. https://doi.org/10.18653/v1/W18-6521

Castro Ferreira, T., van der Lee, C., van Miltenburg, E., & Krahmer, E. (2019). Neural data-to-text generation: A comparison between pipeline and end-to-end architectures. *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. https://doi.org/10.18653/v1/D19-1052

Castro Ferreira, T., Wubben, S., & Krahmer, E. (2018). Surface realization shared task 2018 (SR18): The Tilburg University approach. *Proceedings of the First Workshop on Multilingual Surface Realisation*, 35–38. https://doi.org/10.18653/v1/W18-3604

Chafe, W. (1976). Givenness, contrastiveness, definiteness, subjects, topics, and point of view. *Subject and topic*.

Chafe, W. (1994). Discourse, consciousness, and time. *Discourse*, *2*(1).

Chao, Y. R. (1965). *A grammar of spoken chinese*. Univ of California Press.

Chen, C., & Ng, V. (2014). Chinese zero pronoun resolution: An unsupervised approach combining ranking and integer linear programming. *AAAI*, 1622–1628.

Chen, C., & Ng, V. (2015). Chinese zero pronoun resolution: A joint unsupervised discourse-aware model rivaling state-of-the-art resolvers. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, 320–326. https://doi.org/10.3115/v1/P15-2053

Chen, C., & Ng, V. (2016). Chinese zero pronoun resolution with deep neural networks. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 778–788. https://doi.org/10.18653/v1/P16-1074

Chen, D. L., & Mooney, R. J. (2008). Learning to sportscast: A test of grounded language acquisition. *Proceedings of the 25th international conference on Machine learning*, 128–135.

Chen, G., Same, F., & van Deemter, K. (2021). What can neural referential form selectors learn? *Proceedings of the 14th International Conference on Natural Language Generation*, 154–166. https://aclanthology.org/2021.inlg-1.15

Chen, G., & van Deemter, K. (2020). Lessons from computational modelling of reference production in Mandarin and English. *Proceedings of the 13th International Conference* *on Natural Language Generation*, 263–272. https://www.aclweb.org/anthology/2020.inlg-1.33

Chen, G., & van Deemter, K. (2021). Varieties of specification: Redefining over- and under-specification for an enhanced understanding of referring expressions. *Journal Paper in Preparation*.

Chen, G., van Deemter, K., & Lin, C. (2018a). Modelling pro-drop with the rational speech acts model. *Proceedings of the 11th International Conference on Natural Language Generation*, 159–164. https://doi.org/10.18653/v1/W18-6519

Chen, G., van Deemter, K., & Lin, C. (2018b). Modelling pro-drop with the rational speech acts model. *Proceedings of the 11th International Conference on Natural Language Generation*, 57–66.

Chen, G., van Deemter, K., & Lin, C. (2018c). SimpleNLG-ZH: A linguistic realisation engine for Mandarin. *Proceedings of the 11th International Conference on Natural Language Generation*, 57–66. https://doi.org/10.18653/v1/W18-6506

Chen, G., van Deemter, K., & Lin, C. (2019). Generating quantified descriptions of abstract visual scenes. *Proceedings of the 12th International Conference on Natural Language Generation*, 529–539. https://doi.org/10.18653/v1/W19-8667

Chen, G., van Deemter, K., Pagliaro, S., Smalbil, L., & Lin, C. (2019). QTUNA: A corpus for understanding how speakers use quantification. *Proceedings of the 12th International Conference on Natural Language Generation*, 124–129. https://doi.org/10.18653/v1/W19-8616

Chen, G., & Yao, J.-G. (2019). A closer look at recent results of verb selection for data-to-text NLG. *Proceedings of the 12th International Conference on Natural Language Generation*, 158–163. https://doi.org/10.18653/v1/W19-8622

Chen, J., Fu, G., Yang, S., & Narasimhan, B. (2021). Processing factors and syntactic choice in mandarin child and caregiver speech.

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

Cheng, H., & Mellish, C. (2000). Capturing the interaction between aggregation and text planning in two generation systems. *INLG'2000 Proceedings of the First International Conference on Natural Language Generation*, 186–193. https://doi.org/10.3115/1118253.1118279

Cheng, J., Dong, L., & Lapata, M. (2016). Long short-term memory-networks for machine reading. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 551–561. https://doi.org/10.18653/v1/D16-1053

Cheng, L. L.-S., & Sybesma, R. (1999). Bare and not-so-bare nouns and the structure of np. *Linguistic inquiry*, *30*(4), 509–542.

Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder–decoder for statistical machine translation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1724–1734. https://doi.org/10.3115/v1/D14-1179

Clark, H. H. (1996). *Using language*. Cambridge university press.

Coupland, N., & Jaworski, A. (2008). *Sociolinguistics: A reader and coursebook*. Palgrave.

Coventry, K. R., Cangelosi, A., Newstead, S. E., & Bugmann, D. (2010). Talking about quantities in space: Vague quantifiers, context and similarity. *Language and Cognition*, *2*(2), 221–241.

Creaney, N. (1996). An algorithm for generating quantifiers. *Eighth International Natural Language Generation Workshop*. https://www.aclweb.org/anthology/W96-0413

Croft, W. (1994). Semantic universals in classifier systems. *Word*, *45*(2), 145–171.

Cunha, R., Castro Ferreira, T., Pagano, A., & Alves, F. (2020). Referring to what you know and do not know: Making referring expression generation models generalize to unseen entities. *Proceedings of the 28th International Conference on Computational Linguistics*, 2261–2272. https://doi.org/10.18653/v1/2020.coling-main.205

Dale, R. (1989). Cooking up referring expressions. *27th Annual Meeting of the Association for Computational Linguistics*, 68–75. https://doi.org/10.3115/981623.981632

Dale, R. (1992). *Generating referring expressions: Constructing descriptions in a domain of objects and processes*. The MIT Press.

Dale, R. (2020). Natural language generation: The commercial state of the art in 2020. *Natural Language Engineering*, *26*(4), 481–487. https://doi.org/10.1017/S135132492000025X

Dale, R., & Reiter, E. (1995). Computational interpretations of the gricean maxims in the generation of referring expressions. *Cognitive science*, *19*(2), 233–263.

Dale, R., & Reiter, E. (1996). The role of the gricean maxims in the generation of referring expressions. *arXiv preprint cmp-lg/9604006*.

Dale, R., & Viethen, J. (2009). Referring expression generation through attribute-based heuristics. *Proceedings of the 12th European Workshop on Natural Language Generation (ENLG 2009)*, 58–65. https://aclanthology.org/W09-0609

Dalianis, H. (1999). Aggregation in natural language generation. *Computational Intelligence*, *15*(4), 384–414.

de Carvalho, A., Reboul, A. C., Van der Henst, J.-B., Cheylus, A., & Nazir, T. (2016). Scalar implicatures: The psychological reality of scales. *Frontiers in Psychology*, *7*, 1500. https://doi.org/10.3389/fpsyg.2016.01500

Degen, J., Hawkins, R. D., Graf, C., Kreiss, E., & Goodman, N. D. (2020). When redundancy is useful: A bayesian approach to overinformative referring expressions. *Psychological review*.

Degen, J., & Tanenhaus, M. K. (2011). Making inferences: The case of scalar implicature processing. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *33*(33).

de Oliveira, R., & Sripada, S. (2014). Adapting SimpleNLG for Brazilian Portuguese realisation. *Proceedings of the 8th International Natural Language Generation Conference (INLG)*, 93–94. https://doi.org/10.3115/v1/W14-4412

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *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. https://doi.org/10.18653/v1/N19-1423

Dice, L. R. (1945). Measures of the amount of ecologic association between species. *Ecology*, *26*(3), 297–302.

Dietz, R. (2017). Vagueness and probability: Introduction. *Synthese*, *194*(10), 3693–3698. https://doi.org/10.1007/s11229-017-1347-6

Dimitromanolaki, A., & Androutsopoulos, I. (2003). Learning to order facts for discourse planning in natural language generation. https://aclanthology.org/W03-2304

Ding, S., Xu, H., & Koehn, P. (2019). Saliency-driven word alignment interpretation for neural machine translation. *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, 1–12. https://doi.org/10.18653/v1/W19-5201

Dokkara, S. R. S., Penumathsa, S. V., & Sripada, S. G. (2015). A simple surface realization engine for Telugu. *Proceedings of the 15th European Workshop on Natural Language Generation (ENLG)*, 1–8. https://doi.org/10.18653/v1/W15-4701

Duboue, P. A., & McKeown, K. R. (2003). Statistical acquisition of content selection rules for natural language generation. *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, 121–128. https://aclanthology.org/W03-1016

Dupuy, L. E., Van der Henst, J.-B., Cheylus, A., & Reboul, A. C. (2016). Context in generalized conversational implicatures: The case of some. *Frontiers in Psychology*, *7*, 381. https://doi.org/10.3389/fpsyg.2016.00381

Dusek, O., & Jurcicek, F. (2016). Sequence-to-sequence generation for spoken dialogue via deep syntax trees and strings. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 45–51. https://doi.org/10.18653/v1/P16-2008

Dusek, O., Novikova, J., & Rieser, V. (2018). Findings of the E2E NLG challenge. *Proceedings of the 11th International Conference on Natural Language Generation*, 322–328. https://doi.org/10.18653/v1/W18-6539

Dušek, O., Howcroft, D. M., & Rieser, V. (2019). Semantic noise matters for neural natural language generation. *Proceedings of the 12th International Conference on Natural Language Generation*, 421–426. https://doi.org/10.18653/v1/W19-8652

Dyer, C., Kuncoro, A., Ballesteros, M., & Smith, N. A. (2016). Recurrent neural network grammars. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 199–209. https://doi.org/10.18653/v1/N16-1024

Dziri, N., Madotto, A., Zaıane, O., & Bose, A. J. (2021). Neural path hunter: Reducing hallucination in dialogue systems via path grounding, 2197–2214. https://aclanthology.org/2021.emnlp-main.168

Edmonds, P., & Hirst, G. (2002). Near-synonymy and lexical choice. *Computational Linguistics*, *28*(2), 105–144. https://doi.org/10.1162/089120102760173625

Eikmeyer, H.-J., & Ahlsén, E. (1996). The cognitive process of referring to an object: A comparative study of german and swedish. *Proceedings of the 16th Scandinavian Conference on Linguistics, Turku, Finland*.

Engelhardt, P. E., Bailey, K. G., & Ferreira, F. (2006). Do speakers and listeners observe the gricean maxim of quantity? *Journal of Memory and Language*, *54*(4), 554–573.

Engelhardt, P. E., Demiral, Ş. B., & Ferreira, F. (2011). Over-specified referring expressions impair comprehension: An erp study. *Brain and cognition*, *77*(2), 304–314.

Espinosa, D., White, M., & Mehay, D. (2008). Hypertagging: Supertagging for surface realization with CCG. *Proceedings of ACL-08: HLT*, 183–191. https://aclanthology.org/P08-1022

Faille, J., Gatt, A., & Gardent, C. (2020). The natural language pipeline, neural text generation and explainability. *2nd Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence*, 16–21. https://aclanthology.org/2020.nl4xai-1.5

Fang, R., Doering, M., & Chai, J. Y. (2015). Embodied collaborative referring expression generation in situated human-robot interaction. *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, 271–278.

Feeney, A., Scrafton, S., Duckworth, A., & Handley, S. J. (2004). The story of some: Everyday pragmatic inference by children and adults. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, *58*(2), 121.

Ferreira, V. S., Slevc, L. R., & Rogers, E. S. (2005). How do speakers avoid ambiguous linguistic expressions? *Cognition*, *96*(3), 263–284.

Ficler, J., & Goldberg, Y. (2017). Controlling linguistic style aspects in neural language generation, 94–104. https://doi.org/10.18653/v1/W17-4912

Filippova, K., & Strube, M. (2007). Generating constituent order in German clauses. *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, 320–327. https://aclanthology.org/P07-1041

FitzGerald, N., Artzi, Y., & Zettlemoyer, L. (2013). Learning distributions over logical forms for referring expression generation. *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 1914–1925. https://aclanthology.org/D13-1197

Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, *336*(6084), 998–998.

Franke, M. (2014). Typical use of quantifiers: A probabilistic speaker model. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *36*(36).

Fried, D., Andreas, J., & Klein, D. (2018). Unified pragmatic models for generating and following instructions. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 1951–1963. https://doi.org/10.18653/v1/N18-1177

Fukumura, K., & van Gompel, R. P. (2011). The effect of animacy on the choice of referring expression. *Language and cognitive processes*, *26*(10), 1472–1504.

Gardent, C., & Narayan, S. (2015). Multiple adjunction in feature-based Tree-Adjoining Grammar. *Computational Linguistics*, *41*(1), 41–70. https://doi.org/10.1162/COLI_a_00217

Gardent, C., & Perez-Beltrachini, L. (2017). A statistical, grammar-based approach to microplanning. *Computational Linguistics*, *43*(1), 1–30. https://doi.org/10.1162/COLI_a_00273

Gardent, C., Shimorina, A., Narayan, S., & Perez-Beltrachini, L. (2017). Creating training corpora for NLG micro-planners. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 179–188. https://doi.org/10.18653/v1/P17-1017

Garey, M. R., & Johnson, D. S. (1990). *Computers and intractability; a guide to the theory of np-completeness*. W. H. Freeman & Co.

Gatt, A., & van Deemter, K. (2007a). Lexical choice and conceptual perspective in the generation of plural referring expressions. *Journal of Logic, Language and Information*, *16*(4), 423–443. http://staff.um.edu.mt/albert.gatt/pubs/jolli2007.pdf

Gatt, A., & Belz, A. (2010). Introducing shared tasks to nlg: The tuna shared task evaluation challenges. *Empirical methods in natural language generation* (pp. 264–293). Springer.

Gatt, A., & Krahmer, E. (2018). Survey of the state of the art in Natural Language Generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research (JAIR)*, *61*, 65–170. http://jair.org/media/5477/live-5477-10398-jair.pdf

Gatt, A., Krahmer, E., van Gompel, R., & van Deemter, K. (2013). Production of referring expressions: Preference trumps discrimination. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *35*(35).

Gatt, A., & Reiter, E. (2009). SimpleNLG: A realisation engine for practical applications. *Proceedings of the 12th European Workshop on Natural Language Generation (ENLG 2009)*, 90–93. https://aclanthology.org/W09-0613

Gatt, A., van der Sluis, I., & van Deemter, K. (2007). Evaluating algorithms for the generation of referring expressions using a balanced corpus. *Proceedings of the Eleventh European Workshop on Natural Language Generation (ENLG 07)*, 49–56. https://www.aclweb.org/anthology/W07-2307

Gatt, A., van der Sluis, I., & van Deemter, K. (2008). *Xml format guidelines for the tuna corpus* (tech. rep.). Technical report, Dept of Computing Science, University of Aberdeen. http://www.csd.abdn.ac.uk/~%20agatt/home/pubs/tunaFormat.pdf

Gatt, A., & van Deemter, K. (2007b). Incremental generation of plural descriptions: Similarity and partitioning. *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, 102–111. https://aclanthology.org/D07-1011

Gehrmann, S., Strobelt, H., & Rush, A. (2019). GLTR: Statistical detection and visualization of generated text. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 111–116. https://doi.org/10.18653/v1/P19-3019

Geurts, B., & Nouwen, R. (2007). 'at least'et al.: The semantics of scalar modifiers. *Language*, 533–559.

Gibbs, R. W., & Van Orden, G. (2012). Pragmatic choice in conversation. *Topics in Cognitive Science*, *4*(1), 7–20.

Giulianelli, M., Harding, J., Mohnert, F., Hupkes, D., & Zuidema, W. (2018). Under the hood: Using diagnostic classifiers to investigate and improve how language models track agreement information. *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, 240–248. https://doi.org/10.18653/v1/W18-5426

Givón, T. (1983). *Topic continuity in discourse*. John Benjamins Publishing Company.

Goldberg, Y. (2017). Neural network methods for natural language processing. *Synthesis lectures on human language technologies*, *10*(1), 1–309.

Gong, H., Sun, Y., Feng, X., Qin, B., Bi, W., Liu, X., & Liu, T. (2020). TableGPT: Few-shot table-to-text generation with table structure reconstruction and content matching. *Proceedings of the 28th International Conference on Computational Linguistics*, 1978–1988. https://doi.org/10.18653/v1/2020.coling-main.179

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT press.

Goodman, N. D., & Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. *Trends in cognitive sciences*, *20*(11), 818–829.

Green, M. J., & van Deemter, K. (2011). Vagueness as cost reduction: An empirical test. *Proceedings ofProduction of Referring Expressions' workshop at 33rd Annual Meeting of the Cognitive Science Society*.

Greenbacker, C., & McCoy, K. (2009). UDel: Generating referring expressions guided by psycholinguistc findings. *Proceedings of the 2009 Workshop on Language Generation and Summarisation (UCNLG+Sum 2009)*, 101–102. https://aclanthology.org/W09-2819

Grice, H. P. (1975). Logic and conversation. *Speech acts* (pp. 41–58). Brill.

Grosz, B. J., Joshi, A. K., & Weinstein, S. (1995). Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, *21*(2), 203–225. https://aclanthology.org/J95-2003

Gu, J., Bradbury, J., Xiong, C., Li, V. O., & Socher, R. (2018). Non-autoregressive neural machine translation. *International Conference on Learning Representations*. https://openreview.net/forum?id=B1l8BtlCb

Gu, J., Lu, Z., Li, H., & Li, V. O. (2016). Incorporating copying mechanism in sequence-to-sequence learning. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1631–1640. https://doi.org/10.18653/v1/P16-1154

Gu, J., Wang, C., & Zhao, J. (2019). Levenshtein transformer. *arXiv preprint arXiv:1905.11006*.

Gundel, J. K., Hedberg, N., & Zacharski, R. (1993). Cognitive status and the form of referring expressions in discourse. *Language*, 274–307.

Guo, Q., Qiu, X., Xue, X., & Zhang, Z. (2021). Syntax-guided text generation via graph neural network. *Science China Information Sciences*.

Guo, Z., Zhang, Y., Teng, Z., & Lu, W. (2019). Densely connected graph convolutional networks for graph-to-sequence learning. *Transactions of the Association for Computational Linguistics*, *7*, 297–312. https://doi.org/10.1162/tacl_a_00269

Gupta, S., & Stent, A. (2005). Automatic evaluation of referring expression generation using corpora. *Proceedings of the Workshop on Using Corpora for Natural Language Generation*, 1–6.

Hall, E. T. (1989). *Beyond culture*. Anchor.

Halliday, M. A. K., Matthiessen, C. M., Halliday, M., & Matthiessen, C. (2014). *An introduction to functional grammar*. Routledge.

Harbusch, K., & Kempen, G. (2009). Generating clausal coordinate ellipsis multilingually: A uniform approach based on postediting. *Proceedings of the 12th European Workshop on Natural Language Generation (ENLG 2009)*, 138–145. https://aclanthology.org/W09-0624

Hartshorne, J. K., Snedeker, J., Liem Azar, S. Y.-M., & Kim, A. E. (2015). The neural computation of scalar implicature. *Language, cognition and neuroscience*, *30*(5), 620–634.

He, W., Wang, H., Guo, Y., & Liu, T. (2009). Dependency based Chinese sentence realization. *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, 809–816. https://aclanthology.org/P09-1091

Heeman, P. A., & Hirst, G. (1995). Collaborating on referring expressions. *Computational Linguistics*, *21*(3), 351–382. https://aclanthology.org/J95-3003

Hendrickx, I., Daelemans, W., Luyckx, K., Morante, R., & Van Asch, V. (2008). CNTS: Memory-based learning of generating repeated references. *Proceedings of the Fifth International Natural Language Generation Conference*, 194–95. https://aclanthology.org/W08-1129

Henschel, R., Cheng, H., & Poesio, M. (2000). Pronominalization revisited. *COLING 2000 Volume 1: The 18th International Conference on Computational Linguistics*. https://aclanthology.org/C00-1045

Her, O.-S., & Lai, W.-J. (2012). Classifiers: The many ways to profile one - a case study of taiwan mandarin. *International Journal of Computer Processing Of Languages*, *24*(01), 79–94.

Herbelot, A., & Vecchi, E. M. (2015). Building a shared world: Mapping distributional to model-theoretic semantic spaces. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 22–32. https://doi.org/10.18653/v1/D15-1003

Hessel, J., Holtzman, A., Forbes, M., Bras, R. L., & Choi, Y. (2021). Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*.

Hewitt, J., & Liang, P. (2019). Designing and interpreting probes with control tasks. *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. https://doi.org/10.18653/v1/D19-1275

Hinterwimmer, S. (2019). Prominent protagonists. *Journal of Pragmatics*, *154*, 79–91.

Hobbs, J. R., & Shieber, S. M. (1987). An algorithm for generating quantifier scopings. *Computational Linguistics*, *13*, 47–63.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735–1780.

Holden, J. G., Van Orden, G. C., & Turvey, M. T. (2009). Dispersion of response times reveals cognitive dynamics. *Psychological review*, *116*(2), 318.

Holtzman, A., Buys, J., Du, L., Forbes, M., & Choi, Y. (2020). The curious case of neural text degeneration. *International Conference on Learning Representations*. https://openreview.net/forum?id=rygGQyrFvH

Hong, S., & Shi, D. (2013). Dp-analysis and appositive structure. 汉语学习 *(Chinese language learning)*.

Horn, L. R. (1972). *On the semantic properties of logical operators in english*. University of California, Los Angeles.

Horton, W. S., & Keysar, B. (1996). When do speakers take into account common ground? *Cognition*, *59*(1), 91–117.

Hovy, E. (1987). Generating natural language under pragmatic constraints. *Journal of Pragmatics*, *11*(6), 689–719.

Hovy, E., Marcus, M., Palmer, M., Ramshaw, L., & Weischedel, R. (2006). OntoNotes: The 90% solution. *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, 57–60. https://aclanthology.org/N06-2015

Hovy, E. H. (1993). Automated discourse generation using discourse structure relations. *Artificial intelligence*, *63*(1-2), 341–385.

Howcroft, D., Vogels, J., & Demberg, V. (2017). G-TUNA: A corpus of referring expressions in German, including duration information. *Proceedings of the 10th International Conference on Natural Language Generation*, 149–153. https://doi.org/10.18653/v1/W17-3522

Huang, C.-T. J., Li, Y.-h. A., & Li, Y. (2009). *The syntax of Chinese* (Vol. 8). Cambridge University Press Cambridge.

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

Huang, C.-T. J. (1987). Remarks on empty categories in chinese. *Linguistic inquiry*, 321–337.

Huang, T.-H. K., Ferraro, F., Mostafazadeh, N., Misra, I., Agrawal, A., Devlin, J., Girshick, R., He, X., Kohli, P., Batra, D., Zitnick, C. L., Parikh, D., Vanderwende, L., Galley, M., & Mitchell, M. (2016). Visual storytelling. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1233–1239. https://doi.org/10.18653/v1/N16-1147

Huang, Y., Zheng, F., Su, Y., Li, F., & Wu, W. (2001). A theme structure method for the ellipsis resolution. *Seventh European Conference on Speech Communication and Technology*.

Inui, K., Tokunaga, T., & Tanaka, H. (1992). Text revision: A model and its implementation. *NLG*.

Jaeger, T. F., & Levy, R. P. (2007). Speakers optimize information density through syntactic reduction. *Advances in neural information processing systems*, 849–856.

James, H. C.-T., Li, Y.-H. A., & Li, Y. (2009). The syntax of chinese. *Cambridge, Cambridge). doi, 10*.

Janarthanam, S., & Lemon, O. (2014). Adaptive generation in dialogue systems using dynamic user modeling. *Computational Linguistics*, *40*(4), 883–920.

Jensen, J. T. (1990). *Morphology: Word structure in generative grammar* (Vol. 70). John Benjamins Publishing.

Jin, D., Jin, Z., Hu, Z., Vechtomova, O., & Mihalcea, R. (2020). Deep learning for text style transfer: A survey. *arXiv preprint arXiv:2011.00416*.

Jordan, P. W., & Walker, M. A. (2005). Learning content selection rules for generating object descriptions in dialogue. *Journal of Artificial Intelligence Research*, *24*, 157–194.

Kacprzyk, J., Wilbik, A., & Zadrożny, S. (2008). Linguistic summarization of time series using a fuzzy quantifier driven aggregation [Advances in Intelligent Databases and Information Systems]. *Fuzzy Sets and Systems*, *159*(12), 1485–1499. https://doi.org/https://doi.org/10.1016/j.fss.2008.01.025

Kamp, H., & Reyle, U. (1993). *From discourse to logic: Introduction to modeltheoretic semantics of natural language, formal logic and discourse representation theory* (Vol. 42). Springer Science & Business Media.

Kang, D., & Hashimoto, T. (2020). Improved natural language generation via loss truncation. *arXiv preprint arXiv:2004.14589*.

Kaplan, R. B. (1966). Cultural thought patterns in intercultural education. *Language learning*, *16*(1), 1–20.

Kaufman, E. L., Lord, M. W., Reese, T. W., & Volkmann, J. (1949). The discrimination of visual number. *The American journal of psychology*, *62*(4), 498–525.

Kazemzadeh, S., Ordonez, V., Matten, M., & Berg, T. (2014). ReferItGame: Referring to objects in photographs of natural scenes. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 787–798. https://doi.org/10.3115/v1/D14-1086

Keenan, E. L., & Moss, L. S. (1985). Generalized quantifiers and the expressive power of natural language. *Generalized quantifiers in natural language* (pp. 73–124). Foris Dordrecht.

Kempen, G. (2009). Clausal coordination and coordinative ellipsis in a model of the speaker. *Linguistics*.

Kennedy, C., & McNally, L. (2005). Scale structure, degree modification, and the semantics of gradable predicates. *Language*, 345–381.

Kenney, R., & Smith, P. (1996). *Vagueness: A reader*. MIT press.

Khan, I. H., Ritchie, G., & van Deemter, K. (2006). The clarity-brevity trade-off in generating referring expressions. *Proceedings of the Fourth International Natural Language Generation Conference*, 89–91. https://aclanthology.org/W06-1413

Kibrik, A. A., Khudyakova, M. V., Dobrov, G. B., Linnik, A., & Zalmanov, D. A. (2016). Referential choice: Predictability and its limits. *Frontiers in Psychology*, *7*, 1429. https://doi.org/10.3389/fpsyg.2016.01429

Kobayashi, G., Kuribayashi, T., Yokoi, S., & Inui, K. (2020). Attention is not only a weight: Analyzing transformers with vector norms. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 7057–7075. https://doi.org/10.18653/v1/2020.emnlp-main.574

Koller, A., & Petrick, R. P. (2011). Experiences with planning for natural language generation. *Computational Intelligence*, *27*(1), 23–40.

Koncel-Kedziorski, R., Hajishirzi, H., & Farhadi, A. (2014). Multi-resolution language grounding with weak supervision. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 386–396. https://doi.org/10.3115/v1/D14-1043

Koncel-Kedziorski, R., Bekal, D., Luan, Y., Lapata, M., & Hajishirzi, H. (2019). Text Generation from Knowledge Graphs with Graph Transformers. *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)*, 2284–2293. https://doi.org/10.18653/v1/N19-1238

Kondadadi, R., Howald, B., & Schilder, F. (2013). A statistical NLG framework for aggregated planning and realization. *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1406–1415. https://aclanthology.org/P13-1138

Konstas, I., & Lapata, M. (2013a). A global model for concept-to-text generation. *Journal of Artificial Intelligence Research*, *48*, 305–346.

Konstas, I., & Lapata, M. (2013b). Inducing document plans for concept-to-text generation. *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 1503–1514. https://aclanthology.org/D13-1157

Koolen, R., Gatt, A., Goudbeek, M., & Krahmer, E. (2011). Factors causing overspecification in definite descriptions. *Journal of Pragmatics*, *43*(13), 3231–3250.

Koolen, R., & Krahmer, E. (2010). The d-tuna corpus: A dutch dataset for the evaluation of referring expression generation algorithms. *LREC*.

Kotek, H., Sudo, Y., & Hackl, M. (2015). Experimental investigations of ambiguity: The case of most. *Natural Language Semantics*, *23*(2), 119–156.

Krahmer, E., & Theune, M. (2002). Efficient context-sensitive generation of referring expressions. *Information sharing: Reference and presupposition in language generation and interpretation*, *143*, 223–263.

Krahmer, E., & van Deemter, K. (2012). Computational generation of referring expressions: A survey. *Computational Linguistics*, *38*(1), 173–218. https://doi.org/10.1162/COLI_a_00088

Kuanzhuo, Z., Lin, L., & Weina, Z. (2020). Simplenlg-ti: Adapting simplenlg to tibetan. *Proceedings of the 13th International Conference on Natural Language Generation*, 86–90.

Kunz, J., & Kuhlmann, M. (2020). Classifier probes may just learn from linear context features. *Proceedings of the 28th International Conference on Computational Linguistics*, 5136–5146. https://doi.org/10.18653/v1/2020.coling-main.450

Kurtzman, H. S., & MacDonald, M. C. (1993). Resolution of quantifier scope ambiguities. *Cognition*, *48*(3), 243–279.

Kusner, M., Sun, Y., Kolkin, N., & Weinberger, K. (2015). From word embeddings to document distances. In F. Bach & D. Blei (Eds.), *Proceedings of the 32nd international conference on machine learning* (pp. 957–966). PMLR. http://proceedings.mlr.press/v37/kusnerb15.html

Kutlak, R., van Deemter, K., & Mellish, C. (2016). Production of referring expressions for an unknown audience: A computational model of communal common ground. *Frontiers in psychology*, *7*, 1275.

Lampouras, G., & Androutsopoulos, I. (2013). Using integer linear programming in concept-to-text generation to produce more compact texts. *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 561–566. https://aclanthology.org/P13-2100

Lampouras, G., & Vlachos, A. (2016). Imitation learning for language generation from unaligned data. *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, 1101–1112. https://aclanthology.org/C16-1105

Langkilde, I. (2000). Forest-based statistical sentence generation. *1st Meeting of the North American Chapter of the Association for Computational Linguistics*. https://aclanthology.org/A00-2023

Langner, M. (2020). OMEGA : A probabilistic approach to referring expression generation in a virtual environment. *Proceedings of the 13th International Conference on Natural Language Generation*, 296–305. https://aclanthology.org/2020.inlg-1.36

Lapalme, G. (2013). Natural language generation and summarization at RALI. *Proceedings of the 14th European Workshop on Natural Language Generation*, 92–93.

Lapata, M. (2006). Automatic evaluation of information ordering: Kendall's tau. *Computational Linguistics*, *32*(4), 471–484. https://doi.org/10.1162/coli.2006.32.4.471

Lappin, S. (2000). An intensional parametric semantics for vague quantifiers. *Linguistics and philosophy*, *23*(6), 599–620.

Lassiter, D. (2009). Vagueness as probabilistic linguistic knowledge. *International workshop on vagueness in communication*, 127–150.

Lee, J., Mansimov, E., & Cho, K. (2018). Deterministic non-autoregressive neural sequence modeling by iterative refinement. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 1173–1182. https://doi.org/10.18653/v1/D18-1149

Lee, J., Leung, H., & Li, K. (2017). Towards universal dependencies for learner Chinese. *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*, 67–71.

Leech, G. (2016). *Principles of pragmatics*. Routledge.

Lemon, O. (2008). Adaptive natural language generation in dialogue using reinforcement learning. *Proc. SEM-dial*, 141–148.

Leppänen, L., Munezero, M., Granroth-Wilding, M., & Toivonen, H. (2017). Data-driven news generation for automated journalism. *Proceedings of the 10th International Conference on Natural Language Generation*, 188–197. https://doi.org/10.18653/v1/W17-3528

Levelt, W. J. (1993). *Speaking: From intention to articulation* (Vol. 1). MIT press.

Levinson, S. C. (1983). *Pragmatics*. Cambridge University Press. https://doi.org/10.1017/CBO9780511813313

Levinson, S. C., Stephen, C., & Levinson, S. C. (2000). *Presumptive meanings: The theory of generalized conversational implicature*. MIT press.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., & Zettlemoyer, L. (2020). BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 7871–7880. https://doi.org/10.18653/v1/2020.acl-main.703

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

Li, J., Chen, X., Hovy, E., & Jurafsky, D. (2016). Visualizing and understanding neural models in NLP. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 681–691. https://doi.org/10.18653/v1/N16-1082

Li, J., Galley, M., Brockett, C., Gao, J., & Dolan, B. (2016). A diversity-promoting objective function for neural conversation models. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 110–119. https://doi.org/10.18653/v1/N16-1014

Li, L., van Deemter, K., & Paperno, D. (2020). Chinese long and short form choice exploiting neural network language modeling approaches. *Proceedings of the 19th Chinese National Conference on Computational Linguistics*, 874–880. https://aclanthology.org/2020.ccl-1.81

Li, L., van Deemter, K., Paperno, D., & Fan, J. (2019). Choosing between long and short word forms in Mandarin. *Proceedings of the 12th International Conference on Natural Language Generation*, 34–39. https://doi.org/10.18653/v1/W19-8605

Li, S., Zhao, Z., Hu, R., Li, W., Liu, T., & Du, X. (2018). Analogical reasoning on Chinese morphological and semantic relations. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 138–143. https://doi.org/10.18653/v1/P18-2023

Li, X., van Deemter, K., & Lin, C. (2016). Statistics-based lexical choice for NLG from quantitative information. *Proceedings of the 9th International Natural Language Generation conference*, 104–108. https://doi.org/10.18653/v1/W16-6618

Li, X., van Deemter, K., & Lin, C. (2018). Statistical NLG for generating the content and form of referring expressions. *Proceedings of the 11th International Conference on Natural Language Generation*, 482–491. https://doi.org/10.18653/v1/W18-6561

Li, Y. (1997). Remarks on chinese word order. *Zhongguo Yuyanxue Luncong*, 1–29.

Li, Y.-h. A. (1999). Plurality in a classifier language. *Journal of East Asian Linguistics*, *8*(1), 75–99.

Li, Y.-h. A. (2006). Argument determiner phrases and number phrases. *Argument*, *29*(4).

Li, Y. (2008). Three sensitive positions and chinese complex sentences: A comparative perspective. *J. Chin. Lang. Comput.*, *18*(2), 47–60.

Liang, P., Jordan, M., & Klein, D. (2009). Learning semantic correspondences with less supervision. *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, 91–99. https://aclanthology.org/P09-1011

Lidz, J., Pietroski, P., Halberda, J., & Hunter, T. (2011). Interface transparency and the psychosemantics of most. *Natural Language Semantics*, *19*(3), 227–256.

Lin, C.-Y. (2004). ROUGE: A package for automatic evaluation of summaries. *Text Summarization Branches Out*, 74–81. https://www.aclweb.org/anthology/W04-1013

Lin, C.-Y., & Och, F. J. (2004). Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, 605–612. https://doi.org/10.3115/1218955.1219032

Linzen, T., Dupoux, E., & Goldberg, Y. (2016). Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, *4*, 521–535. https://doi.org/10.1162/tacl_a_00115

Liu, T., Wang, K., Sha, L., Chang, B., & Sui, Z. (2018). Table-to-text generation by structure-aware seq2seq learning. *Thirty-Second AAAI Conference on Artificial Intelligence*.

Liu, Y., Gu, W., & Pan, W. (2001). *Chinese grammar*. The Commercial Press.

Luong, T., Pham, H., & Manning, C. D. (2015). Effective approaches to attention-based neural machine translation. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 1412–1421. https://doi.org/10.18653/v1/D15-1166

Lv, S. (1979). Problems in the analysis of chinese grammar.

Ma, X., Zhou, C., Li, X., Neubig, G., & Hovy, E. (2019). FlowSeq: Non-autoregressive conditional sequence generation with generative flow. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 4282–4292. https://doi.org/10.18653/v1/D19-1437

Madotto, A., Wu, C.-S., & Fung, P. (2018). Mem2Seq: Effectively incorporating knowledge bases into end-to-end task-oriented dialog systems. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1468–1478. https://doi.org/10.18653/v1/P18-1136

Malmi, E., Takala, P., Toivonen, H., Raiko, T., & Gionis, A. (2016). Dopelearning: A computational approach to rap lyrics generation. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 195–204.

Mann, W. C., & Thompson, S. A. (1988). Rhetorical structure theory: Toward a functional theory of text organization. *Text-interdisciplinary Journal for the Study of Discourse*, *8*(3), 243–281.

Mao, J., Huang, J., Toshev, A., Camburu, O., Yuille, A. L., & Murphy, K. (2016). Generation and comprehension of unambiguous object descriptions. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 11–20.

Marcus, M. P., Marcinkiewicz, M. A., & Santorini, B. (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational linguistics*, *19*(2), 313–330.

Matuszek, C., FitzGerald, N., Zettlemoyer, L., Bo, L., & Fox, D. (2012). A joint model of language and perception for grounded attribute learning, 1435–1442.

Mazzei, A., Battaglino, C., & Bosco, C. (2016). SimpleNLG-IT: Adapting SimpleNLG to Italian. *Proceedings of the 9th International Natural Language Generation conference*, 184–192. https://doi.org/10.18653/v1/W16-6630

McDonald, D. D. (1983). Description directed control: Its implications for natural language generation. *Computers & Mathematics with Applications*, *9*(1), 111–129.

McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemia medica*, *22*(3), 276–282.

McKeown, K. (1992). *Text generation*. Cambridge University Press.

McLuhan, M. (1964). *Understanding media: The extensions of man*. MIT press.

Mehri, S., & Eskenazi, M. (2020). USR: An unsupervised and reference free evaluation metric for dialog generation. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 681–707. https://doi.org/10.18653/v1/2020.acl-main.64

Mellish, C., Scott, D., Cahill, L., Paiva, D., Evans, R., & Reape, M. (2006). A reference architecture for natural language generation systems. *Natural language engineering*, *12*(1), 1–34.

Meteer, M. W. (1991). Bridging the generation gap between text planning and linguistic realization. *Computational Intelligence*, *7*(4), 296–304.

Mikolov, T., Karafiat, M., Burget, L., Cernocky, J., & Khudanpur, S. (2010). Recurrent neural network based language model. *Interspeech*, *2*(3), 1045–1048.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 3111–3119.

Mitchell, M., Dodge, J., Goyal, A., Yamaguchi, K., Stratos, K., Han, X., Mensch, A., Berg, A., Berg, T., & Daumé III, H. (2012). Midge: Generating image descriptions from computer vision detections. *Proceedings of the 13th Conference of the European Chapter* *of the Association for Computational Linguistics*, 747–756. https://aclanthology.org/E12-1076

Monroe, W., Hawkins, R. X., Goodman, N. D., & Potts, C. (2017). Colors in context: A pragmatic neural model for grounded language understanding. *Transactions of the Association for Computational Linguistics*, *5*, 325–338. https://doi.org/10.1162/tacl_a_00064

Monroe, W., Hu, J., Jong, A., & Potts, C. (2018). Generating bilingual pragmatic color references. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 2155–2165. https://doi.org/10.18653/v1/N18-1196

Monroe, W., & Potts, C. (2015). Learning in the rational speech acts model. *CoRR*, *abs/1510.06807*. http://arxiv.org/abs/1510.06807

Montague, R. (1973). The proper treatment of quantification in ordinary english. *Approaches to natural language* (pp. 221–242). Springer.

Mostowski, A. (1957). On a generalization of quantifiers. *Fundamenta Mathematicae*, *44*(2), 12–36.

Moxey, L. M., & Sanford, A. J. (1993). *Communicating quantities: A psychological perspective.* Lawrence Erlbaum Associates, Inc.

Nakanishi, H., Miyao, Y., & Tsujii, J. (2005). Probabilistic models for disambiguation of an HPSG-based chart generator. *Proceedings of the Ninth International Workshop on Parsing Technology*, 93–102. https://aclanthology.org/W05-1510

Newnham, R. (1971). *About Chinese*. Penguin Books Ltd.

Nie, F., Yao, J.-G., Wang, J., Pan, R., & Lin, C.-Y. (2019). A simple recipe towards reducing hallucination in neural surface realisation. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2673–2679. https://doi.org/10.18653/v1/P19-1256

Nouwen, R. (2010). What is in a quantifier? *The Linguistics Enterprise: From knowledge of language to knowledge in linguistics*, *150*, 235.

Noveck, I. (2007). The why and how of experimental pragmatics: The case of 'scalar inferences'. *Advances in Pragmatics. Basingstoke: Palgrave.*

Noveck, I. A. (2001). When children are more logical than adults: Experimental investigations of scalar implicature. *Cognition*, *78*(2), 165–188.

Nunberg, G. D. (1978). *The pragmatics of reference.* City University of New York.

Odijk, J. (1995). Generation of coherent monologues. *CLIN V: Proceedings of the 5th CLIN Meeting*, 123–131.

Ong, E., Abella, S., Santos, L., & Tiu, D. (2011). A simple surface realizer for Filipino. *Proceedings of the 25th Pacific Asia Conference on Language, Information and Computation*, 51–59. https://aclanthology.org/Y11-1006

Orita, N., Vornov, E., Feldman, N., & Daumé III, H. (2015). Why discourse affects speakers' choice of referring expressions. *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. https://doi.org/10.3115/v1/P15-1158

Osborne, T., & Liang, J. (2015). A survey of ellipsis in chinese. *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*, 271–280.

Packard, J. L. (2000). *The morphology of Chinese: A linguistic and cognitive approach*. Cambridge University Press.

Pandit, O., & Hou, Y. (2021). Probing for bridging inference in transformer language models. *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4153–4163. https://doi.org/10.18653/v1/2021.naacl-main.327

Papafragou, A., & Musolino, J. (2003). Scalar implicatures: Experiments at the semantics–pragmatics interface. *Cognition*, *86*(3), 253–282.

Papafragou, A., & Tantalou, N. (2004). Children's computation of implicatures. *Language acquisition*, *12*(1), 71–82.

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). BLEU: A method for automatic evaluation of machine translation. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 311–318. https://doi.org/10.3115/1073083.1073135

Paraboni, I., Lan, A. G. J., de Sant'Ana, M. M., & Coutinho, F. L. (2017). Effects of cognitive effort on the resolution of overspecified descriptions. *Computational Linguistics*, *43*(2), 451–459.

Paraboni, I., & van Deemter, K. (2014). Reference and the facilitation of search in spatial domains. *Language, Cognition and Neuroscience*, *29*(8), 1002–1017.

Paraboni, I., van Deemter, K., & Masthoff, J. (2007). Generating referring expressions: Making referents easy to identify. *Computational linguistics*, *33*(2), 229–254.

Pascanu, R., Mikolov, T., & Bengio, Y. (2013). On the difficulty of training recurrent neural networks. *International conference on machine learning*, 1310–1318.

Passonneau, R. J. (1996). Using centering to relax gricean informational constraints on discourse anaphoric noun phrases. *Language and Speech*, *39*(2-3), 229–264. https://doi.org/10.1177/002383099603900305

Paul, W. (2010). Adjectives in Mandarin Chinese: The rehabilitation of a much ostracized category. *Adjectives: Formal analyses in syntax and semantics, ed. Patricia Cabredo Hofherr and Ora Matushansky*, *1*, 15–151.

Pechmann, T. (1989). Incremental speech production and referential overspecification. *Linguistics*, *27*(1), 89–110.

Peinelt, N., Liakata, M., & Hsieh, S.-K. (2017). ClassifierGuesser: A context-based classifier prediction system for Chinese language learners. *Proceedings of the IJCNLP 2017, System Demonstrations*, 41–44. https://aclanthology.org/I17-3011

Pennington, J., Socher, R., & Manning, C. (2014). GloVe: Global vectors for word representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543. https://doi.org/10.3115/v1/D14-1162

Peters, S., & Westerståhl, D. (2006). *Quantifiers in language and logic*. Oxford University Press.

Petroni, F., Rocktäschel, T., Riedel, S., Lewis, P., Bakhtin, A., Wu, Y., & Miller, A. (2019). Language models as knowledge bases? *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2463–2473. https://doi.org/10.18653/v1/D19-1250

Pezzelle, S., & Fernández, R. (2019). Is the red square big? MALeViC: Modeling adjectives leveraging visual contexts. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2865–2876. https://doi.org/10.18653/v1/D19-1285

Pezzelle, S., Steinert-Threlkeld, S., Bernardi, R., & Szymanik, J. (2018). Some of them can be guessed! exploring the effect of linguistic context in predicting quantifiers.

Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), 114–119. https://doi.org/10.18653/v1/P18-2019

Pimentel, T., Hall Maudslay, R., Blasi, D., & Cotterell, R. (2020). Speakers fill lexical semantic gaps with context. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 4004–4015. https://doi.org/10.18653/v1/2020.emnlp-main.328

Pogue, A., Kurumada, C., & Tanenhaus, M. K. (2016). Talker-specific generalization of pragmatic inferences based on under-and over-informative prenominal adjective use. *Frontiers in psychology*, *6*, 2035.

Portet, F., Reiter, E., Gatt, A., Hunter, J., Sripada, S., Freer, Y., & Sykes, C. (2009). Automatic generation of textual summaries from neonatal intensive care data. *Artificial Intelligence*, *173*(7-8), 789–816.

Pouscoulous, N., Noveck, I. A., Politzer, G., & Bastide, A. (2007). A developmental investigation of processing costs in implicature production. *Language acquisition*, *14*(4), 347–375.

Power, R., & Williams, S. (2012). Generating numerical approximations. *Computational Linguistics*, *38*(1), 113–134. https://doi.org/10.1162/COLI_a_00086

Pradhan, S., Moschitti, A., Xue, N., Uryupina, O., & Zhang, Y. (2012). CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes. *Joint Conference on EMNLP and CoNLL - Shared Task*, 1–40. https://aclanthology.org/W12-4501

Prince, E. F. (1981). Towards a taxonomy of given-new information. *Radical pragmatics*.

Puduppully, R., Dong, L., & Lapata, M. (2019a). Data-to-text generation with content selection and planning. *Proceedings of the AAAI conference on artificial intelligence*, *33*(01), 6908–6915.

Puduppully, R., Dong, L., & Lapata, M. (2019b). Data-to-text generation with entity modeling. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2023–2035. https://doi.org/10.18653/v1/P19-1195

Puduppully, R., & Lapata, M. (2021). Data-to-text Generation with Macro Planning. *Transactions of the Association for Computational Linguistics*, *9*, 510–527. https://doi.org/10.1162/tacl_a_00381

Qasim, S. R., Mahmood, H., & Shafait, F. (2019). Rethinking table recognition using graph neural networks. *2019 International Conference on Document Analysis and Recognition (ICDAR)*, 142–147.

Qing, C. (2014). *Quantiative social-cognitive experimental pragmatics* (Doctoral dissertation). Universiteit van Amsterdam.

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

Ramos, A., Alonso, J. M., Reiter, E., van Deemter, K., & Gatt, A. (2019). Fuzzy-based language grounding of geographical references: From writers to readers. *International Journal of Computational Intelligence Systems*, *12*(2), 970–983.

Ramos-Soto, A., Bugarın, A., & Barro, S. (2016). Fuzzy sets across the natural language generation pipeline. *Progress in artificial intelligence*, *5*(4), 261–276.

Ramos-Soto, A., Janeiro-Gallardo, J., & Bugarın Diz, A. (2017). Adapting SimpleNLG to Spanish. *Proceedings of the 10th International Conference on Natural Language Generation*, 144–148. https://doi.org/10.18653/v1/W17-3521

Rao, S., Ettinger, A., Daumé III, H., & Resnik, P. (2015). Dialogue focus tracking for zero pronoun resolution. *Proceedings of the 2015 Conference of the North American Chapter* *of the Association for Computational Linguistics: Human Language Technologies*, 494–503. https://doi.org/10.3115/v1/N15-1052

Reape, M., & Mellish, C. (1999). Just what is aggregation anyway. *Proceedings of the 7th European Workshop on Natural Language Generation*, 20–29.

Reiter, E. (1990). The computational complexity of avoiding conversational implicatures. *28th annual meeting of the association for computational linguistics*, 97–104.

Reiter, E. (2007). An architecture for data-to-text systems. *Proceedings of the Eleventh European Workshop on Natural Language Generation (ENLG 07)*, 97–104. https://aclanthology.org/W07-2315

Reiter, E. (2018a). Hallucination in neural nlg. https://ehudreiter.com/2018/11/12/hallucination-in-neural-nlg/

Reiter, E. (2018b). A structured review of the validity of bleu. *Computational Linguistics*, *44*(3), 393–401.

Reiter, E., & Dale, R. (1997). Building applied natural language generation systems. *Natural Language Engineering*, *3*(1), 57–87. https://doi.org/10.1017/S1351324997001502

Reiter, E., & Dale, R. (2000). *Building natural language generation systems*. Cambridge university press.

Reiter, E., Robertson, R., & Osman, L. M. (2003). Lessons from a failure: Generating tailored smoking cessation letters. *Artificial Intelligence*, *144*(1-2), 41–58.

Reiter, E., & Sripada, S. (2002). Human variation and lexical choice. *Computational Linguistics*, *28*(4), 545–553.

Reiter, E., Sripada, S., Hunter, J., Yu, J., & Davy, I. (2005). Choosing words in computer-generated weather forecasts. *Artificial Intelligence*, *167*(1-2), 137–169.

Riba, P., Dutta, A., Goldmann, L., Fornés, A., Ramos, O., & Lladós, J. (2019). Table detection in invoice documents by graph neural networks. *2019 International Conference on Document Analysis and Recognition (ICDAR)*, 122–127.

Rieser, V., & Lemon, O. (2009). Natural language generation as planning under uncertainty for spoken dialogue systems. *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, 683–691. https://aclanthology.org/E09-1078

Rieser, V., & Lemon, O. (2011). Learning and evaluation of dialogue strategies for new applications: Empirical methods for optimization from small data sets. *Computational Linguistics*, *37*(1), 153–196. https://doi.org/10.1162/coli_a_00038

Rogers, A., Kovaleva, O., & Rumshisky, A. (2020). A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, *8*, 842–866. https://doi.org/10.1162/tacl_a_00349

Rohde, H., Seyfarth, S., Clark, B., Jäger, G., & Kaufmann, S. (2012). Communicating with cost-based implicature: A game-theoretic approach to ambiguity. *Proceedings of the 16th Workshop on the Semantics and Pragmatics of Dialogue*, 107–116.

Rohrbach, A., Hendricks, L. A., Burns, K., Darrell, T., & Saenko, K. (2018). Object hallucination in image captioning. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 4035–4045. https://doi.org/10.18653/v1/D18-1437

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, *8*(3), 382–439. https://doi.org/https://doi.org/10.1016/0010-0285(76)90013-X

Ross, J. (1982). Pronoun deleting processes in german. *Annual Meeting of the Linguistics Society of America*.

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

Saba, W. S., & Corriveau, J.-P. (1997). A pragmatic treatment of quantification in natural language. *AAAI/IAAI*, 610–615.

Same, F., Chen, G., & van Deemter, K. (2022). Non-neural models can matter: A re-evaluation of neural referring expression generation systems. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Same, F., & van Deemter, K. (2020). A linguistic perspective on reference: Choosing a feature set for generating referring expressions in context. *Proceedings of the 28th International Conference on Computational Linguistics*, 4575–4586. https://doi.org/10.18653/v1/2020.coling-main.403

Schmuckler, M. A. (2001). What is ecological validity? a dimensional analysis. *Infancy*, *2*(4), 419–436.

Schriefers, H., & Pechmann, T. (1988). Incremental production of referential noun phrases by human speakers. *Advances in natural language generation*, *1*, 172–179.

Schuster, M., & Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, *45*(11), 2673–2681.

Scott, D., & de Souza, C. S. (1990). Getting the message across in rst-based text generation. *Current research in natural language generation*, *4*, 47–73.

Sennrich, R., Haddow, B., & Birch, A. (2016). Neural machine translation of rare words with subword units. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1715–1725. https://doi.org/10.18653/v1/P16-1162

Shao, Z., Huang, M., Wen, J., Xu, W., & Zhu, X. (2019). Long and diverse text generation with planning-based hierarchical variational model. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3257–3268. https://doi.org/10.18653/v1/D19-1321

Shaw, J. (1998). Clause aggregation using linguistic knowledge. *Natural Language Generation*.

Shen, X., Chang, E., Su, H., Niu, C., & Klakow, D. (2020). Neural data-to-text generation via jointly learning the segmentation and correspondence. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 7155–7165. https://doi.org/10.18653/v1/2020.acl-main.641

Sheng, E., Chang, K.-W., Natarajan, P., & Peng, N. (2021). Societal biases in language generation: Progress and challenges. *arXiv preprint arXiv:2105.04054*.

Siddharthan, A., Nenkova, A., & McKeown, K. (2011). Information status distinctions and referring expressions: An empirical study of references to people in news summaries. *Computational Linguistics*, *37*(4), 811–842.

Smiley, C., Plachouras, V., Schilder, F., Bretz, H., Leidner, J., & Song, D. (2016). When to plummet and when to soar: Corpus based verb selection for natural language generation. *Proceedings of the 9th International Natural Language Generation conference*, 36–39. https://doi.org/10.18653/v1/W16-6606

Solt, S. (2016). On measurement and quantification: The case of most and more than half. *Language*, *92*(1), 65–100.

Sorodoc, I., Lazaridou, A., Boleda, G., Herbelot, A., Pezzelle, S., & Bernardi, R. (2016). "look, some green circles!": Learning to quantify from images. *Proceedings of the 5th Workshop on Vision and Language*, 75–79. https://doi.org/10.18653/v1/W16-3211

Sorodoc, I.-T., Gulordava, K., & Boleda, G. (2020). Probing for referential information in language models. *Proceedings of the 58th Annual Meeting of the Association for*

Computational Linguistics, 4177–4189. https://doi.org/10.18653/v1/2020.acl-main.384

Sperber, D., & Wilson, D. (1986). *Relevance: Communication and cognition* (Vol. 142). Citeseer.

Srinivasan, P., & Yates, A. (2009). Quantifier scope disambiguation using extracted pragmatic knowledge: Preliminary results. *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, 1465–1474. https://aclanthology.org/D09-1152

Stede, M. (2000). The hyperonym problem revisited: Conceptual and lexical hierarchies in language generation. *INLG'2000 Proceedings of the First International Conference on Natural Language Generation*, 93–99. https://doi.org/10.3115/1118253.1118267

Steedman, M. (2000). *The syntactic process* (Vol. 24). MIT press Cambridge, MA.

Stern, M., Chan, W., Kiros, J., & Uszkoreit, J. (2019). Insertion transformer: Flexible sequence generation via insertion operations. *International Conference on Machine Learning*, 5976–5985.

Stone, M., & Webber, B. (1998). Textual economy through close coupling of syntax and semantics. *arXiv preprint cmp-lg/9806020*.

Sun, R. (2008). *The cambridge handbook of computational psychology*. Cambridge University Press.

Susanto, R. H., Chollampatt, S., & Tan, L. (2020). Lexically constrained neural machine translation with Levenshtein transformer. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 3536–3543. https://doi.org/10.18653/v1/2020.acl-main.325

Sutskever, I., Martens, J., & Hinton, G. E. (2011). Generating text with recurrent neural networks. *ICML*.

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 3104–3112.

Szymanik, J. et al. (2016). *Quantifiers and cognition: Logical and computational perspectives* (Vol. 96). Springer.

Tang, D., Qin, B., & Liu, T. (2015). Document modeling with gated recurrent neural network for sentiment classification. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 1422–1432. https://doi.org/10.18653/v1/D15-1167

Teng, S.-h. (1979). Remarks on cleft sentences in Chinese. *Journal of Chinese Linguistics*, *7*(1), 101–14.

Teresa Guasti, M., Chierchia, G., Crain, S., Foppolo, F., Gualmini, A., & Meroni, L. (2005). Why children and adults sometimes (but not always) compute implicatures. *Language and cognitive processes*, *20*(5), 667–696.

Testoni, A., Pezzelle, S., & Bernardi, R. (2019). Quantifiers in a multimodal world: Hallucinating vision with language and sound. *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, 105–116. https://doi.org/10.18653/v1/W19-2912

Theune, M., Hielkema, F., & Hendriks, P. (2006). Performing aggregation and ellipsis using discourse structures. *Research on Language and Computation*, *4*(4), 353–375.

Theune, M., Klabbers, E., de Pijper, J.-R., Krahmer, E., & Odijk, J. (2001). From data to speech: A general approach. *Natural Language Engineering*, *7*(1), 47–86.

van der Auwera, J., & Baoill, D. Ó. (1998). *Adverbial constructions in the languages of Europe* (Vol. 3). Walter de Gruyter.

van der Lee, C., Gatt, A., van Miltenburg, E., Wubben, S., & Krahmer, E. (2019). Best practices for the human evaluation of automatically generated text. *Proceedings* *of the 12th International Conference on Natural Language Generation*, 355–368. https://doi.org/10.18653/v1/W19-8643

van der Sluis, I., Gatt, A., & van Deemter, K. (2006). *Manual for the tuna corpus: Referring expressions in two domains* (Technical Report AUCS/TR0705). Department of Computing Science, Univ. of Aberdeen. http://homepages.abdn.ac.uk/k.vdeemter/pages/TunaCorpusManual/index.html

van der Sluis, I., Gatt, A., & van Deemter, K. (2007). Evaluating algorithms for the generation of referring expressions: Going beyond toy domains. *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP'07)*. http://staff.um.edu.mt/albert.gatt/pubs/ranlp2007.pdf

van der Sluis, I., & Krahmer, E. (2007). Generating multimodal references. *Discourse Processes*, *44*(3), 145–174.

van Benthem, J. et al. (1986). *Essays in logical semantics*. Springer.

van Benthem, J. (1983). Determiners and logic. *Linguistics and Philosophy*, *6*(4), 447–478.

van Deemter, K. (2002). Generating referring expressions: Boolean extensions of the incremental algorithm. *Computational Linguistics*, *28*(1), 37–52.

van Deemter, K. (2012). *Not exactly: In praise of vagueness*. Oxford University Press.

van Deemter, K. (2016). *Computational models of referring: A study in cognitive science*. MIT Press.

van Deemter, K., & Gatt, A. (2007). Content determination in GRE: Evaluating the evaluator. *Using Corpora for Natural Language Generation: Language Generation and Machine Translation*.

van Deemter, K., & Gatt, A. (2009). Beyond DICE: Measuring the quality of a referring expression.

van Deemter, K., Gatt, A., Sluis, I. v. d., & Power, R. (2012). Generation of referring expressions: Assessing the incremental algorithm. *Cognitive science*, *36*(5), 799–836.

van Deemter, K., Gatt, A., van Gompel, R. P., & Krahmer, E. (2012). Toward a computational psycholinguistics of reference production. *Topics in cognitive science*, *4*(2), 166–183.

van Deemter, K., & Reiter, E. (2018). Lying and computational linguistics. *The oxford handbook of lying*. Oxford University Press.

van Deemter, K., Sun, L., Sybesma, R., Li, X., Chen, B., & Yang, M. (2017). Investigating the content and form of referring expressions in Mandarin: Introducing the mtuna corpus. *Proceedings of the 10th International Conference on Natural Language Generation*, 213–217. https://doi.org/10.18653/v1/W17-3532

van Deemter, K., Theune, M., & Krahmer, E. (2005). Real versus template-based natural language generation: A false opposition? *Computational linguistics*, *31*(1), 15–24.

van Gompel, R. P., van Deemter, K., Gatt, A., Snoeren, R., & Krahmer, E. J. (2019). Conceptualization in reference production: Probabilistic modeling and experimental testing. *Psychological review*, *126*(3), 345.

van Miltenburg, E., Lu, W.-T., Krahmer, E., Gatt, A., Chen, G., Li, L., & van Deemter, K. (2020). Gradations of error severity in automatic image descriptions. *Proceedings of the 13th International Conference on Natural Language Generation*, 398–411. https://www.aclweb.org/anthology/2020.inlg-1.45

van Tiel, B. (2014). *Quantity matters: Implicatures, typicality, and truth* (Doctoral dissertation). [Sl: sn].

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 5998–6008.

Vaudry, P.-L., & Lapalme, G. (2013). Adapting SimpleNLG for bilingual English-French realisation. *Proceedings of the 14th European Workshop on Natural Language Generation*, 183–187. https://aclanthology.org/W13-2125

Vicente, K. J., & Wang, J. H. (1998). An ecological theory of expertise effects in memory recall. *Psychological review*, *105*(1), 33.

Viethen, J., & Dale, R. (2006). Algorithms for generating referring expressions: Do they do what people do? *Proceedings of the Fourth International Natural Language Generation Conference*, 63–70. https://aclanthology.org/W06-1410

Viethen, J., & Dale, R. (2008). The use of spatial relations in referring expression generation. *Proceedings of the Fifth International Natural Language Generation Conference*, 59–67. https://aclanthology.org/W08-1109

von Heusinger, K., & Schumacher, P. B. (2019). Discourse prominence: Definition and application. *Journal of Pragmatics*, *154*, 117–127.

Walker, M. A., Rambow, O., & Rogati, M. (2001). SPoT: A trainable sentence planner. *Second Meeting of the North American Chapter of the Association for Computational Linguistics*. https://aclanthology.org/N01-1003

Wang, A. Y., & Piao, S. (2007). Translating vagueness? a study on translations of vague quantifiers in an english-chinese parallel corpus. *Proceedings of the Corpus Linguistics Conference*.

Wang, H. (2019). Revisiting challenges in data-to-text generation with fact grounding. *Proceedings of the 12th International Conference on Natural Language Generation*, 311–322. https://doi.org/10.18653/v1/W19-8639

Wang, L., Tu, Z., Shi, S., Zhang, T., Graham, Y., & Liu, Q. (2018). Translating pro-drop languages with reconstruction models, 1–9.

Wang, L., Tu, Z., Zhang, X., Li, H., Way, A., & Liu, Q. (2016). A novel approach to dropped pronoun translation, 983–993. https://doi.org/10.18653/v1/N16-1113

Wang, Z., Wang, X., An, B., Yu, D., & Chen, C. (2020). Towards faithful neural table-to-text generation with content-matching constraints. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 1072–1086. https://doi.org/10.18653/v1/2020.acl-main.101

Wen, T.-H., Gašić, M., Kim, D., Mrkšić, N., Su, P.-H., Vandyke, D., & Young, S. (2015). Stochastic language generation in dialogue using recurrent neural networks with convolutional sentence reranking. *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 275–284. https://doi.org/10.18653/v1/W15-4639

White, M., & Rajkumar, R. (2009). Perceptron reranking for CCG realization. *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, 410–419. https://aclanthology.org/D09-1043

White, M., & Rajkumar, R. (2012). Minimal dependency length in realization ranking. *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 244–255. https://aclanthology.org/D12-1023

White, M., Rajkumar, R., & Martin, S. (2007). Towards broad coverage surface realization with ccg. *Proceedings of the Workshop on Using Corpora for NLG: Language Generation and Machine Translation (UCNLG+ MT)*, 267–276.

Williams, S., & Reiter, E. (2008). Generating basic skills reports for low-skilled readers. *Natural Language Engineering*, *14*(4), 495–525.

Wilson, D., & Sperber, D. (2002). Relevance theory.

Wiseman, S., Shieber, S., & Rush, A. (2018). Learning neural templates for text generation. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 3174–3187. https://doi.org/10.18653/v1/D18-1356

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., et al. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., & Philip, S. Y. (2020). A comprehensive survey on graph neural networks. *IEEE transactions on neural networks and learning systems*, *32*(1), 4–24.

Xing, X., Fan, X., & Wan, X. (2020). Automatic generation of citation texts in scholarly papers: A pilot study. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 6181–6190. https://doi.org/10.18653/v1/2020.acl-main.550

Xu, D. (1997). *Functional categories in Mandarin Chinese* (Vol. 26). Holland Academic Graphics.

Xu, L., & Langendoen, D. T. (1985). Topic structures in Chinese. *Language*, 1–27.

Yager, R. R. (1982). A new approach to the summarization of data. *Information Sciences*, *28*(1), 69–86.

Yamura-Takei, M., Fujiwara, M., & Aizawa, T. (2001). Centering as an anaphora generation algorithm: A language learning aid perspective. *NLPRS*, *2001*, 557–562.

Yang, G., & Bateman, J. (2009). The Chinese aspect generation based on aspect selection functions. *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, 629–637. https://aclanthology.org/P09-1071

Yang, L., & Cahill, D. (2008). The rhetorical organization of chinese and american students expository essays: A contrastive rhetoric study. *International Journal of English Studies*, *8*(2), 113–132.

Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., & Hovy, E. (2016). Hierarchical attention networks for document classification. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1480–1489. https://doi.org/10.18653/v1/N16-1174

Ye, Z., Zhan, W., & Zhou, X. (2007). The semantic processing of syntactic structure in sentence comprehension: An erp study. *Brain research*, *1142*, 135–145.

Yeh, C.-L., & Mellish, C. (1997). An empirical study on the generation of anaphora in Chinese. *Computational Linguistics*, *23*(1), 169–190. https://aclanthology.org/J97-1007

Yi, Y., Deng, H., & Hu, J. (2020). Improving image captioning evaluation by considering inter references variance. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 985–994. https://doi.org/10.18653/v1/2020.acl-main.93

Yildirim, I., Degen, J., Tanenhaus, M. K., & Jaeger, T. F. (2013). Linguistic variability and adaptation in quantifier meanings. *Proceedings of the 35th Annual Meeting of the Cognitive Science Society*.

Yin, Q., Zhang, W., Zhang, Y., & Liu, T. (2017). A deep neural network for chinese zero pronoun resolution, 3322–3328. https://doi.org/10.24963/ijcai.2017/464

Yin, Q., Zhang, Y., Zhang, W., & Liu, T. (2017). Chinese zero pronoun resolution with deep memory network. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 1309–1318. https://doi.org/10.18653/v1/D17-1135

Yin, Q., Zhang, Y., Zhang, W.-N., Liu, T., & Wang, W. Y. (2018). Deep reinforcement learning for Chinese zero pronoun resolution, 569–578. https://doi.org/10.18653/v1/P18-1053

Yu, L., Poirson, P., Yang, S., Berg, A. C., & Berg, T. L. (2016). Modeling context in referring expressions. *European Conference on Computer Vision*, 69–85.

Yu, L., Tan, H., Bansal, M., & Berg, T. L. (2017). A joint speaker-listener-reinforcer model for referring expressions. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7282–7290.

Yu, Z., & Wan, X. (2019). How to avoid sentences spelling boring? towards a neural approach to unsupervised metaphor generation. *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)*, 861–871. https://doi.org/10.18653/v1/N19-1092

Yu, Z., Zang, H., & Wan, X. (2020a). Homophonic pun generation with lexically constrained rewriting. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2870–2876. https://doi.org/10.18653/v1/2020.emnlp-main.229

Yu, Z., Zang, H., & Wan, X. (2020b). Routing enforced generative model for recipe generation. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 3797–3806. https://doi.org/10.18653/v1/2020.emnlp-main.311

Yunxia, Z. (2000). Structural moves reflected in english and chinese sales letters. *Discourse studies*, *2*(4), 473–496.

Zajenkowski, M., & Szymanik, J. (2013). Most intelligent people are accurate and some fast people are intelligent.: Intelligence, working memory, and semantic processing of quantifiers from a computational perspective. *Intelligence*, *41*(5), 456–466.

Zang, H., & Wan, X. (2017). Towards automatic generation of product reviews from aspect-sentiment scores. *Proceedings of the 10th International Conference on Natural Language Generation*, 168–177. https://doi.org/10.18653/v1/W17-3526

Zarrieß, S., & Schlangen, D. (2016). Towards generating colour terms for referents in photographs: Prefer the expected or the unexpected? *Proceedings of the 9th International Natural Language Generation conference*, 246–255. https://doi.org/10.18653/v1/W16-6642

Zhang, D., Yuan, J., Wang, X., & Foster, A. (2018). Probabilistic verb selection for data-to-text generation. *Transactions of the Association for Computational Linguistics*, *6*, 511–527. https://doi.org/10.1162/tacl_a_00038

Zhang, N. N. (2013). *Classifier structures in mandarin chinese* (Vol. 263). Walter de Gruyter.

Zhang, S., Dinan, E., Urbanek, J., Szlam, A., Kiela, D., & Weston, J. (2018). Personalizing dialogue agents: I have a dog, do you have pets too? *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2204–2213. https://doi.org/10.18653/v1/P18-1205

Zhang, X., & Lapata, M. (2014). Chinese poetry generation with recurrent neural networks. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 670–680. https://doi.org/10.3115/v1/D14-1074

Zhang, Y., Galley, M., Gao, J., Gan, Z., Li, X., Brockett, C., & Dolan, B. (2018). Generating informative and diverse conversational responses via adversarial information maximization. In S. Bengio, H. M. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, & R. Garnett (Eds.), *Advances in neural information processing systems 31:*

Annual conference on neural information processing systems 2018, neurips 2018, december 3-8, 2018, montreal, canada (pp. 1815–1825). https://proceedings.neurips.cc/paper/2018/hash/23ce1851341ec1fa9e0c259de10bf87c-Abstract.html

Zhao, S., & Ng, H. T. (2007). Identification and resolution of Chinese zero pronouns: A machine learning approach. *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, 541–550. https://aclanthology.org/D07-1057

Zheng, Y., Chen, G., Huang, M., Liu, S., & Zhu, X. (2019). Personalized dialogue generation with diversified traits. *arXiv preprint arXiv:1901.09672*.

Zheng, Y., Chen, Z., Zhang, R., Huang, S., Mao, X., & Huang, M. (2021). Stylized dialogue response generation using stylized unpaired texts. *Proceedings of the AAAI Conference on Artificial Intelligence*. https://arxiv.org/abs/2009.12719

Zhou, H., Huang, M., Zhang, T., Zhu, X., & Liu, B. (2018). Emotional chatting machine: Emotional conversation generation with internal and external memory. *Thirty-Second AAAI Conference on Artificial Intelligence*.

Zhu, D.-X. (1982). Yufa jiangyi [lectures on chinese syntax]. *Beijing: Shangwu Yinshuguan*.

Zipf, G. K. (2016). *Human behavior and the principle of least effort: An introduction to human ecology*. Ravenio Books.

SIKS Dissertation Series

2011 01 Botond Cseke (RUN), Variational Algorithms for Bayesian Inference in Latent Gaussian Models

02 Nick Tinnemeier (UU), Organizing Agent Organizations. Syntax and Operational Semantics of an Organization-Oriented Programming Language

03 Jan Martijn van der Werf (TUE), Compositional Design and Verification of Component-Based Information Systems

04 Hado van Hasselt (UU), Insights in Reinforcement Learning; Formal analysis and empirical evaluation of temporal-difference

05 Bas van der Raadt (VU), Enterprise Architecture Coming of Age - Increasing the Performance of an Emerging Discipline.

06 Yiwen Wang (TUE), Semantically-Enhanced Recommendations in Cultural Heritage

07 Yujia Cao (UT), Multimodal Information Presentation for High Load Human Computer Interaction

08 Nieske Vergunst (UU), BDI-based Generation of Robust Task-Oriented Dialogues

09 Tim de Jong (OU), Contextualised Mobile Media for Learning

10 Bart Bogaert (UvT), Cloud Content Contention

11 Dhaval Vyas (UT), Designing for Awareness: An Experience-focused HCI Perspective

12 Carmen Bratosin (TUE), Grid Architecture for Distributed Process Mining

13 Xiaoyu Mao (UvT), Airport under Control. Multiagent Scheduling for Airport Ground Handling

14 Milan Lovric (EUR), Behavioral Finance and Agent-Based Artificial Markets

15 Marijn Koolen (UvA), The Meaning of Structure: the Value of Link Evidence for Information Retrieval

16 Maarten Schadd (UM), Selective Search in Games of Different Complexity

17 Jiyin He (UVA), Exploring Topic Structure: Coherence, Diversity and Relatedness

18 Mark Ponsen (UM), Strategic Decision-Making in complex games

19 Ellen Rusman (OU), The Mind's Eye on Personal Profiles

20 Qing Gu (VU), Guiding service-oriented software engineering - A view-based approach

21 Linda Terlouw (TUD), Modularization and Specification of Service-Oriented Systems

22 Junte Zhang (UVA), System Evaluation of Archival Description and Access

23 Wouter Weerkamp (UVA), Finding People and their Utterances in Social Media

24 Herwin van Welbergen (UT), Behavior Generation for Interpersonal Coordination with Virtual Humans On Specifying, Scheduling and Realizing Multimodal Virtual Human Behavior

25 Syed Waqar ul Qounain Jaffry (VU), Analysis and Validation of Models for Trust Dynamics

26 Matthijs Aart Pontier (VU), Virtual Agents for Human Communication - Emotion Regulation and Involvement-Distance Trade-Offs in Embodied Conversational Agents and Robots

27 Aniel Bhulai (VU), Dynamic website optimization through autonomous management of design patterns

28 Rianne Kaptein (UVA), Effective Focused Retrieval by Exploiting Query Context and Document Structure

29 Faisal Kamiran (TUE), Discrimination-aware Classification

30 Egon van den Broek (UT), Affective Signal Processing (ASP): Unraveling the mystery of emotions

31 Ludo Waltman (EUR), Computational and Game-Theoretic Approaches for Modeling Bounded Rationality

32 Nees-Jan van Eck (EUR), Methodological Advances in Bibliometric Mapping of Science

33 Tom van der Weide (UU), Arguing to Motivate Decisions

34 Paolo Turrini (UU), Strategic Reasoning in Interdependence: Logical and Game-theoretical Investigations

35 Maaike Harbers (UU), Explaining Agent Behavior in Virtual Training

36 Erik van der Spek (UU), Experiments in serious game design: a cognitive approach

37 Adriana Burlutiu (RUN), Machine Learning for Pairwise Data, Applications for Preference Learning and Supervised Network Inference

38 Nyree Lemmens (UM), Bee-inspired Distributed Optimization

39 Joost Westra (UU), Organizing Adaptation using Agents in Serious Games

40 Viktor Clerc (VU), Architectural Knowledge Management in Global Software Development

41 Luan Ibraimi (UT), Cryptographically Enforced Distributed Data Access Control

42 Michal Sindlar (UU), Explaining Behavior through Mental State Attribution

43 Henk van der Schuur (UU), Process Improvement through Software Operation Knowledge

44 Boris Reuderink (UT), Robust Brain-Computer Interfaces

45 Herman Stehouwer (UvT), Statistical Language Models for Alternative Sequence Selection

46 Beibei Hu (TUD), Towards Contextualized Information Delivery: A Rule-based Architecture for the Domain of Mobile Police Work

47 Azizi Bin Ab Aziz (VU), Exploring Computational Models for Intelligent Support of Persons with Depression

48 Mark Ter Maat (UT), Response Selection and Turn-taking for a Sensitive Artificial Listening Agent

49 Andreea Niculescu (UT), Conversational interfaces for task-oriented spoken dialogues: design aspects influencing interaction quality

---

2012 01 Terry Kakeeto (UvT), Relationship Marketing for SMEs in Uganda

02 Muhammad Umair (VU), Adaptivity, emotion, and Rationality in Human and Ambient Agent Models

03 Adam Vanya (VU), Supporting Architecture Evolution by Mining Software Repositories

04 Jurriaan Souer (UU), Development of Content Management System-based Web Applications

05 Marijn Plomp (UU), Maturing Interorganisational Information Systems

06 Wolfgang Reinhardt (OU), Awareness Support for Knowledge Workers in Research Networks

07 Rianne van Lambalgen (VU), When the Going Gets Tough: Exploring Agent-based Models of Human Performance under Demanding Conditions

08 Gerben de Vries (UVA), Kernel Methods for Vessel Trajectories

09 Ricardo Neisse (UT), Trust and Privacy Management Support for Context-Aware Service Platforms

10 David Smits (TUE), Towards a Generic Distributed Adaptive Hypermedia Environment

11 J.C.B. Rantham Prabhakara (TUE), Process Mining in the Large: Preprocessing, Discovery, and Diagnostics

12 Kees van der Sluijs (TUE), Model Driven Design and Data Integration in Semantic Web Information Systems

13 Suleman Shahid (UvT), Fun and Face: Exploring non-verbal expressions of emotion during playful interactions

14 Evgeny Knutov (TUE), Generic Adaptation Framework for Unifying Adaptive Web-based Systems

15 Natalie van der Wal (VU), Social Agents. Agent-Based Modelling of Integrated Internal and Social Dynamics of Cognitive and Affective Processes.

16 Fiemke Both (VU), Helping people by understanding them - Ambient Agents supporting task execution and depression treatment

17 Amal Elgammal (UvT), Towards a Comprehensive Framework for Business Process Compliance

18 Eltjo Poort (VU), Improving Solution Architecting Practices

19 Helen Schonenberg (TUE), What's Next? Operational Support for Business Process Execution

20 Ali Bahramisharif (RUN), Covert Visual Spatial Attention, a Robust Paradigm for Brain-Computer Interfacing

21 Roberto Cornacchia (TUD), Querying Sparse Matrices for Information Retrieval

22 Thijs Vis (UvT), Intelligence, politie en veiligheidsdienst: verenigbare grootheden?

23 Christian Muehl (UT), Toward Affective Brain-Computer Interfaces: Exploring the Neurophysiology of Affect during Human Media Interaction

24 Laurens van der Werff (UT), Evaluation of Noisy Transcripts for Spoken Document Retrieval

25 Silja Eckartz (UT), Managing the Business Case Development in Inter-Organizational IT Projects: A Methodology and its Application

26 Emile de Maat (UVA), Making Sense of Legal Text

27 Hayrettin Gurkok (UT), Mind the Sheep! User Experience Evaluation & Brain-Computer Interface Games

28 Nancy Pascall (UvT), Engendering Technology Empowering Women

29 Almer Tigelaar (UT), Peer-to-Peer Information Retrieval

30 Alina Pommeranz (TUD), Designing Human-Centered Systems for Reflective Decision Making

31 Emily Bagarukayo (RUN), A Learning by Construction Approach for Higher Order Cognitive Skills Improvement, Building Capacity and Infrastructure
32 Wietske Visser (TUD), Qualitative multi-criteria preference representation and reasoning
33 Rory Sie (OUN), Coalitions in Cooperation Networks (COCOON)
34 Pavol Jancura (RUN), Evolutionary analysis in PPI networks and applications
35 Evert Haasdijk (VU), Never Too Old To Learn – On-line Evolution of Controllers in Swarm- and Modular Robotics
36 Denis Ssebugwawo (RUN), Analysis and Evaluation of Collaborative Modeling Processes
37 Agnes Nakakawa (RUN), A Collaboration Process for Enterprise Architecture Creation
38 Selmar Smit (VU), Parameter Tuning and Scientific Testing in Evolutionary Algorithms
39 Hassan Fatemi (UT), Risk-aware design of value and coordination networks
40 Agus Gunawan (UvT), Information Access for SMEs in Indonesia
41 Sebastian Kelle (OU), Game Design Patterns for Learning
42 Dominique Verpoorten (OU), Reflection Amplifiers in self-regulated Learning
43 Withdrawn
44 Anna Tordai (VU), On Combining Alignment Techniques
45 Benedikt Kratz (UvT), A Model and Language for Business-aware Transactions
46 Simon Carter (UVA), Exploration and Exploitation of Multilingual Data for Statistical Machine Translation
47 Manos Tsagkias (UVA), Mining Social Media: Tracking Content and Predicting Behavior
48 Jorn Bakker (TUE), Handling Abrupt Changes in Evolving Time-series Data
49 Michael Kaisers (UM), Learning against Learning - Evolutionary dynamics of reinforcement learning algorithms in strategic interactions
50 Steven van Kervel (TUD), Ontologogy driven Enterprise Information Systems Engineering
51 Jeroen de Jong (TUD), Heuristics in Dynamic Sceduling; a practical framework with a case study in elevator dispatching

2013 01 Viorel Milea (EUR), News Analytics for Financial Decision Support
02 Erietta Liarou (CWI), MonetDB/DataCell: Leveraging the Column-store Database Technology for Efficient and Scalable Stream Processing
03 Szymon Klarman (VU), Reasoning with Contexts in Description Logics
04 Chetan Yadati (TUD), Coordinating autonomous planning and scheduling
05 Dulce Pumareja (UT), Groupware Requirements Evolutions Patterns
06 Romulo Goncalves (CWI), The Data Cyclotron: Juggling Data and Queries for a Data Warehouse Audience
07 Giel van Lankveld (UvT), Quantifying Individual Player Differences
08 Robbert-Jan Merk (VU), Making enemies: cognitive modeling for opponent agents in fighter pilot simulators
09 Fabio Gori (RUN), Metagenomic Data Analysis: Computational Methods and Applications
10 Jeewanie Jayasinghe Arachchige (UvT), A Unified Modeling Framework for Service Design.
11 Evangelos Pournaras (TUD), Multi-level Reconfigurable Self-organization in Overlay Services

12 Marian Razavian (VU), Knowledge-driven Migration to Services

13 Mohammad Safiri (UT), Service Tailoring: User-centric creation of integrated IT-based homecare services to support independent living of elderly

14 Jafar Tanha (UVA), Ensemble Approaches to Semi-Supervised Learning Learning

15 Daniel Hennes (UM), Multiagent Learning - Dynamic Games and Applications

16 Eric Kok (UU), Exploring the practical benefits of argumentation in multi-agent deliberation

17 Koen Kok (VU), The PowerMatcher: Smart Coordination for the Smart Electricity Grid

18 Jeroen Janssens (UvT), Outlier Selection and One-Class Classification

19 Renze Steenhuizen (TUD), Coordinated Multi-Agent Planning and Scheduling

20 Katja Hofmann (UvA), Fast and Reliable Online Learning to Rank for Information Retrieval

21 Sander Wubben (UvT), Text-to-text generation by monolingual machine translation

22 Tom Claassen (RUN), Causal Discovery and Logic

23 Patricio de Alencar Silva (UvT), Value Activity Monitoring

24 Haitham Bou Ammar (UM), Automated Transfer in Reinforcement Learning

25 Agnieszka Anna Latoszek-Berendsen (UM), Intention-based Decision Support. A new way of representing and implementing clinical guidelines in a Decision Support System

26 Alireza Zarghami (UT), Architectural Support for Dynamic Homecare Service Provisioning

27 Mohammad Huq (UT), Inference-based Framework Managing Data Provenance

28 Frans van der Sluis (UT), When Complexity becomes Interesting: An Inquiry into the Information eXperience

29 Iwan de Kok (UT), Listening Heads

30 Joyce Nakatumba (TUE), Resource-Aware Business Process Management: Analysis and Support

31 Dinh Khoa Nguyen (UvT), Blueprint Model and Language for Engineering Cloud Applications

32 Kamakshi Rajagopal (OUN), Networking For Learning; The role of Networking in a Lifelong Learner's Professional Development

33 Qi Gao (TUD), User Modeling and Personalization in the Microblogging Sphere

34 Kien Tjin-Kam-Jet (UT), Distributed Deep Web Search

35 Abdallah El Ali (UvA), Minimal Mobile Human Computer Interaction

36 Than Lam Hoang (TUe), Pattern Mining in Data Streams

37 Dirk Börner (OUN), Ambient Learning Displays

38 Eelco den Heijer (VU), Autonomous Evolutionary Art

39 Joop de Jong (TUD), A Method for Enterprise Ontology based Design of Enterprise Information Systems

40 Pim Nijssen (UM), Monte-Carlo Tree Search for Multi-Player Games

41 Jochem Liem (UVA), Supporting the Conceptual Modelling of Dynamic Systems: A Knowledge Engineering Perspective on Qualitative Reasoning

42 Léon Planken (TUD), Algorithms for Simple Temporal Reasoning

43 Marc Bron (UVA), Exploration and Contextualization through Interaction and Concepts

---

2014 01 Nicola Barile (UU), Studies in Learning Monotone Models from Data

02 Fiona Tuliyano (RUN), Combining System Dynamics with a Domain Modeling Method
03 Sergio Raul Duarte Torres (UT), Information Retrieval for Children: Search Behavior and Solutions
04 Hanna Jochmann-Mannak (UT), Websites for children: search strategies and interface design - Three studies on children's search performance and evaluation
05 Jurriaan van Reijsen (UU), Knowledge Perspectives on Advancing Dynamic Capability
06 Damian Tamburri (VU), Supporting Networked Software Development
07 Arya Adriansyah (TUE), Aligning Observed and Modeled Behavior
08 Samur Araujo (TUD), Data Integration over Distributed and Heterogeneous Data Endpoints
09 Philip Jackson (UvT), Toward Human-Level Artificial Intelligence: Representation and Computation of Meaning in Natural Language
10 Ivan Salvador Razo Zapata (VU), Service Value Networks
11 Janneke van der Zwaan (TUD), An Empathic Virtual Buddy for Social Support
12 Willem van Willigen (VU), Look Ma, No Hands: Aspects of Autonomous Vehicle Control
13 Arlette van Wissen (VU), Agent-Based Support for Behavior Change: Models and Applications in Health and Safety Domains
14 Yangyang Shi (TUD), Language Models With Meta-information
15 Natalya Mogles (VU), Agent-Based Analysis and Support of Human Functioning in Complex Socio-Technical Systems: Applications in Safety and Healthcare
16 Krystyna Milian (VU), Supporting trial recruitment and design by automatically interpreting eligibility criteria
17 Kathrin Dentler (VU), Computing healthcare quality indicators automatically: Secondary Use of Patient Data and Semantic Interoperability
18 Mattijs Ghijsen (UVA), Methods and Models for the Design and Study of Dynamic Agent Organizations
19 Vinicius Ramos (TUE), Adaptive Hypermedia Courses: Qualitative and Quantitative Evaluation and Tool Support
20 Mena Habib (UT), Named Entity Extraction and Disambiguation for Informal Text: The Missing Link
21 Kassidy Clark (TUD), Negotiation and Monitoring in Open Environments
22 Marieke Peeters (UU), Personalized Educational Games - Developing agent-supported scenario-based training
23 Eleftherios Sidirourgos (UvA/CWI), Space Efficient Indexes for the Big Data Era
24 Davide Ceolin (VU), Trusting Semi-structured Web Data
25 Martijn Lappenschaar (RUN), New network models for the analysis of disease interaction
26 Tim Baarslag (TUD), What to Bid and When to Stop
27 Rui Jorge Almeida (EUR), Conditional Density Models Integrating Fuzzy and Probabilistic Representations of Uncertainty
28 Anna Chmielowiec (VU), Decentralized k-Clique Matching
29 Jaap Kabbedijk (UU), Variability in Multi-Tenant Enterprise Software
30 Peter de Cock (UvT), Anticipating Criminal Behaviour
31 Leo van Moergestel (UU), Agent Technology in Agile Multiparallel Manufacturing and Product Support
32 Naser Ayat (UvA), On Entity Resolution in Probabilistic Data

33 Tesfa Tegegne (RUN), Service Discovery in eHealth

34 Christina Manteli (VU), The Effect of Governance in Global Software Development: Analyzing Transactive Memory Systems.

35 Joost van Ooijen (UU), Cognitive Agents in Virtual Worlds: A Middleware Design Approach

36 Joos Buijs (TUE), Flexible Evolutionary Algorithms for Mining Structured Process Models

37 Maral Dadvar (UT), Experts and Machines United Against Cyberbullying

38 Danny Plass-Oude Bos (UT), Making brain-computer interfaces better: improving usability through post-processing.

39 Jasmina Maric (UvT), Web Communities, Immigration, and Social Capital

40 Walter Omona (RUN), A Framework for Knowledge Management Using ICT in Higher Education

41 Frederic Hogenboom (EUR), Automated Detection of Financial Events in News Text

42 Carsten Eijckhof (CWI/TUD), Contextual Multidimensional Relevance Models

43 Kevin Vlaanderen (UU), Supporting Process Improvement using Method Increments

44 Paulien Meesters (UvT), Intelligent Blauw. Met als ondertitel: Intelligence-gestuurde politiezorg in gebiedsgebonden eenheden.

45 Birgit Schmitz (OUN), Mobile Games for Learning: A Pattern-Based Approach

46 Ke Tao (TUD), Social Web Data Analytics: Relevance, Redundancy, Diversity

47 Shangsong Liang (UVA), Fusion and Diversification in Information Retrieval

---

2015 01 Niels Netten (UvA), Machine Learning for Relevance of Information in Crisis Response

02 Faiza Bukhsh (UvT), Smart auditing: Innovative Compliance Checking in Customs Controls

03 Twan van Laarhoven (RUN), Machine learning for network data

04 Howard Spoelstra (OUN), Collaborations in Open Learning Environments

05 Christoph Bösch (UT), Cryptographically Enforced Search Pattern Hiding

06 Farideh Heidari (TUD), Business Process Quality Computation - Computing Non-Functional Requirements to Improve Business Processes

07 Maria-Hendrike Peetz (UvA), Time-Aware Online Reputation Analysis

08 Jie Jiang (TUD), Organizational Compliance: An agent-based model for designing and evaluating organizational interactions

09 Randy Klaassen (UT), HCI Perspectives on Behavior Change Support Systems

10 Henry Hermans (OUN), OpenU: design of an integrated system to support lifelong learning

11 Yongming Luo (TUE), Designing algorithms for big graph datasets: A study of computing bisimulation and joins

12 Julie M. Birkholz (VU), Modi Operandi of Social Network Dynamics: The Effect of Context on Scientific Collaboration Networks

13 Giuseppe Procaccianti (VU), Energy-Efficient Software

14 Bart van Straalen (UT), A cognitive approach to modeling bad news conversations

15 Klaas Andries de Graaf (VU), Ontology-based Software Architecture Documentation

16 Changyun Wei (UT), Cognitive Coordination for Cooperative Multi-Robot Teamwork

17 André van Cleeff (UT), Physical and Digital Security Mechanisms: Properties, Combinations and Trade-offs

18 Holger Pirk (CWI), Waste Not, Want Not! - Managing Relational Data in Asymmetric Memories

19 Bernardo Tabuenca (OUN), Ubiquitous Technology for Lifelong Learners

20 Lois Vanhée (UU), Using Culture and Values to Support Flexible Coordination

21 Sibren Fetter (OUN), Using Peer-Support to Expand and Stabilize Online Learning

22 Zhemin Zhu (UT), Co-occurrence Rate Networks

23 Luit Gazendam (VU), Cataloguer Support in Cultural Heritage

24 Richard Berendsen (UVA), Finding People, Papers, and Posts: Vertical Search Algorithms and Evaluation

25 Steven Woudenberg (UU), Bayesian Tools for Early Disease Detection

26 Alexander Hogenboom (EUR), Sentiment Analysis of Text Guided by Semantics and Structure

27 Sándor Héman (CWI), Updating compressed colomn stores

28 Janet Bagorogoza (TiU), Knowledge Management and High Performance; The Uganda Financial Institutions Model for HPO

29 Hendrik Baier (UM), Monte-Carlo Tree Search Enhancements for One-Player and Two-Player Domains

30 Kiavash Bahreini (OU), Real-time Multimodal Emotion Recognition in E-Learning

31 Yakup Koç (TUD), On the robustness of Power Grids

32 Jerome Gard (UL), Corporate Venture Management in SMEs

33 Frederik Schadd (TUD), Ontology Mapping with Auxiliary Resources

34 Victor de Graaf (UT), Gesocial Recommender Systems

35 Jungxao Xu (TUD), Affective Body Language of Humanoid Robots: Perception and Effects in Human Robot Interaction

---

2016 01 Syed Saiden Abbas (RUN), Recognition of Shapes by Humans and Machines

02 Michiel Christiaan Meulendijk (UU), Optimizing medication reviews through decision support: prescribing a better pill to swallow

03 Maya Sappelli (RUN), Knowledge Work in Context: User Centered Knowledge Worker Support

04 Laurens Rietveld (VU), Publishing and Consuming Linked Data

05 Evgeny Sherkhonov (UVA), Expanded Acyclic Queries: Containment and an Application in Explaining Missing Answers

06 Michel Wilson (TUD), Robust scheduling in an uncertain environment

07 Jeroen de Man (VU), Measuring and modeling negative emotions for virtual training

08 Matje van de Camp (TiU), A Link to the Past: Constructing Historical Social Networks from Unstructured Data

09 Archana Nottamkandath (VU), Trusting Crowdsourced Information on Cultural Artefacts

10 George Karafotias (VUA), Parameter Control for Evolutionary Algorithms

11 Anne Schuth (UVA), Search Engines that Learn from Their Users

12 Max Knobbout (UU), Logics for Modelling and Verifying Normative Multi-Agent Systems

13 Nana Baah Gyan (VU), The Web, Speech Technologies and Rural Development in West Africa - An ICT4D Approach

14 Ravi Khadka (UU), Revisiting Legacy Software System Modernization

15 Steffen Michels (RUN), Hybrid Probabilistic Logics - Theoretical Aspects, Algorithms and Experiments

16 Guangliang Li (UVA), Socially Intelligent Autonomous Agents that Learn from Human Reward

17 Berend Weel (VU), Towards Embodied Evolution of Robot Organisms

18 Albert Meroño Peñuela (VU), Refining Statistical Data on the Web

19 Julia Efremova (Tu/e), Mining Social Structures from Genealogical Data

20 Daan Odijk (UVA), Context & Semantics in News & Web Search

21 Alejandro Moreno Célleri (UT), From Traditional to Interactive Playspaces: Automatic Analysis of Player Behavior in the Interactive Tag Playground

22 Grace Lewis (VU), Software Architecture Strategies for Cyber-Foraging Systems

23 Fei Cai (UVA), Query Auto Completion in Information Retrieval

24 Brend Wanders (UT), Repurposing and Probabilistic Integration of Data; An Iterative and data model independent approach

25 Julia Kiseleva (TU/e), Using Contextual Information to Understand Searching and Browsing Behavior

26 Dilhan Thilakarathne (VU), In or Out of Control: Exploring Computational Models to Study the Role of Human Awareness and Control in Behavioural Choices, with Applications in Aviation and Energy Management Domains

27 Wen Li (TUD), Understanding Geo-spatial Information on Social Media

28 Mingxin Zhang (TUD), Large-scale Agent-based Social Simulation - A study on epidemic prediction and control

29 Nicolas Höning (TUD), Peak reduction in decentralised electricity systems - Markets and prices for flexible planning

30 Ruud Mattheij (UvT), The Eyes Have It

31 Mohammad Khelghati (UT), Deep web content monitoring

32 Eelco Vriezekolk (UT), Assessing Telecommunication Service Availability Risks for Crisis Organisations

33 Peter Bloem (UVA), Single Sample Statistics, exercises in learning from just one example

34 Dennis Schunselaar (TUE), Configurable Process Trees: Elicitation, Analysis, and Enactment

35 Zhaochun Ren (UVA), Monitoring Social Media: Summarization, Classification and Recommendation

36 Daphne Karreman (UT), Beyond R2D2: The design of nonverbal interaction behavior optimized for robot-specific morphologies

37 Giovanni Sileno (UvA), Aligning Law and Action - a conceptual and computational inquiry

38 Andrea Minuto (UT), Materials that Matter - Smart Materials meet Art & Interaction Design

39 Merijn Bruijnes (UT), Believable Suspect Agents; Response and Interpersonal Style Selection for an Artificial Suspect

40 Christian Detweiler (TUD), Accounting for Values in Design

41 Thomas King (TUD), Governing Governance: A Formal Framework for Analysing Institutional Design and Enactment Governance

42 Spyros Martzoukos (UVA), Combinatorial and Compositional Aspects of Bilingual Aligned Corpora

43 Saskia Koldijk (RUN), Context-Aware Support for Stress Self-Management: From Theory to Practice
44 Thibault Sellam (UVA), Automatic Assistants for Database Exploration
45 Bram van de Laar (UT), Experiencing Brain-Computer Interface Control
46 Jorge Gallego Perez (UT), Robots to Make you Happy
47 Christina Weber (UL), Real-time foresight - Preparedness for dynamic innovation networks
48 Tanja Buttler (TUD), Collecting Lessons Learned
49 Gleb Polevoy (TUD), Participation and Interaction in Projects. A Game-Theoretic Analysis
50 Yan Wang (UVT), The Bridge of Dreams: Towards a Method for Operational Performance Alignment in IT-enabled Service Supply Chains

2017 01 Jan-Jaap Oerlemans (UL), Investigating Cybercrime
02 Sjoerd Timmer (UU), Designing and Understanding Forensic Bayesian Networks using Argumentation
03 Daniël Harold Telgen (UU), Grid Manufacturing; A Cyber-Physical Approach with Autonomous Products and Reconfigurable Manufacturing Machines
04 Mrunal Gawade (CWI), Multi-core Parallelism in a Column-store
05 Mahdieh Shadi (UVA), Collaboration Behavior
06 Damir Vandic (EUR), Intelligent Information Systems for Web Product Search
07 Roel Bertens (UU), Insight in Information: from Abstract to Anomaly
08 Rob Konijn (VU), Detecting Interesting Differences:Data Mining in Health Insurance Data using Outlier Detection and Subgroup Discovery
09 Dong Nguyen (UT), Text as Social and Cultural Data: A Computational Perspective on Variation in Text
10 Robby van Delden (UT), (Steering) Interactive Play Behavior
11 Florian Kunneman (RUN), Modelling patterns of time and emotion in Twitter #anticipointment
12 Sander Leemans (TUE), Robust Process Mining with Guarantees
13 Gijs Huisman (UT), Social Touch Technology - Extending the reach of social touch through haptic technology
14 Shoshannah Tekofsky (UvT), You Are Who You Play You Are: Modelling Player Traits from Video Game Behavior
15 Peter Berck (RUN), Memory-Based Text Correction
16 Aleksandr Chuklin (UVA), Understanding and Modeling Users of Modern Search Engines
17 Daniel Dimov (UL), Crowdsourced Online Dispute Resolution
18 Ridho Reinanda (UVA), Entity Associations for Search
19 Jeroen Vuurens (UT), Proximity of Terms, Texts and Semantic Vectors in Information Retrieval
20 Mohammadbashir Sedighi (TUD), Fostering Engagement in Knowledge Sharing: The Role of Perceived Benefits, Costs and Visibility
21 Jeroen Linssen (UT), Meta Matters in Interactive Storytelling and Serious Gaming (A Play on Worlds)
22 Sara Magliacane (VU), Logics for causal inference under uncertainty
23 David Graus (UVA), Entities of Interest — Discovery in Digital Traces
24 Chang Wang (TUD), Use of Affordances for Efficient Robot Learning
25 Veruska Zamborlini (VU), Knowledge Representation for Clinical Guidelines, with applications to Multimorbidity Analysis and Literature Search

26 Merel Jung (UT), Socially intelligent robots that understand and respond to human touch

27 Michiel Joosse (UT), Investigating Positioning and Gaze Behaviors of Social Robots: People’s Preferences, Perceptions and Behaviors

28 John Klein (VU), Architecture Practices for Complex Contexts

29 Adel Alhuraibi (UvT), From IT-BusinessStrategic Alignment to Performance: A Moderated Mediation Model of Social Innovation, and Enterprise Governance of IT

30 Wilma Latuny (UvT), The Power of Facial Expressions

31 Ben Ruijl (UL), Advances in computational methods for QFT calculations

32 Thaer Samar (RUN), Access to and Retrievability of Content in Web Archives

33 Brigit van Loggem (OU), Towards a Design Rationale for Software Documentation: A Model of Computer-Mediated Activity

34 Maren Scheffel (OU), The Evaluation Framework for Learning Analytics

35 Martine de Vos (VU), Interpreting natural science spreadsheets

36 Yuanhao Guo (UL), Shape Analysis for Phenotype Characterisation from High-throughput Imaging

37 Alejandro Montes Garcia (TUE), WiBAF: A Within Browser Adaptation Framework that Enables Control over Privacy

38 Alex Kayal (TUD), Normative Social Applications

39 Sara Ahmadi (RUN), Exploiting properties of the human auditory system and compressive sensing methods to increase noise robustness in ASR

40 Altaf Hussain Abro (VUA), Steer your Mind: Computational Exploration of Human Control in Relation to Emotions, Desires and Social Support For applications in human-aware support systems

41 Adnan Manzoor (VUA), Minding a Healthy Lifestyle: An Exploration of Mental Processes and a Smart Environment to Provide Support for a Healthy Lifestyle

42 Elena Sokolova (RUN), Causal discovery from mixed and missing data with applications on ADHD datasets

43 Maaike de Boer (RUN), Semantic Mapping in Video Retrieval

44 Garm Lucassen (UU), Understanding User Stories - Computational Linguistics in Agile Requirements Engineering

45 Bas Testerink (UU), Decentralized Runtime Norm Enforcement

46 Jan Schneider (OU), Sensor-based Learning Support

47 Jie Yang (TUD), Crowd Knowledge Creation Acceleration

48 Angel Suarez (OU), Collaborative inquiry-based learning

2018 01 Han van der Aa (VUA), Comparing and Aligning Process Representations

02 Felix Mannhardt (TUE), Multi-perspective Process Mining

03 Steven Bosems (UT), Causal Models For Well-Being: Knowledge Modeling, Model-Driven Development of Context-Aware Applications, and Behavior Prediction

04 Jordan Janeiro (TUD), Flexible Coordination Support for Diagnosis Teams in Data-Centric Engineering Tasks

05 Hugo Huurdeman (UVA), Supporting the Complex Dynamics of the Information Seeking Process

06 Dan Ionita (UT), Model-Driven Information Security Risk Assessment of Socio-Technical Systems

07 Jieting Luo (UU), A formal account of opportunism in multi-agent systems

08 Rick Smetsers (RUN), Advances in Model Learning for Software Systems

09 Xu Xie (TUD), Data Assimilation in Discrete Event Simulations

10 Julienka Mollee (VUA), Moving forward: supporting physical activity behavior change through intelligent technology

11 Mahdi Sargolzaei (UVA), Enabling Framework for Service-oriented Collaborative Networks

12 Xixi Lu (TUE), Using behavioral context in process mining

13 Seyed Amin Tabatabaei (VUA), Computing a Sustainable Future

14 Bart Joosten (UVT), Detecting Social Signals with Spatiotemporal Gabor Filters

15 Naser Davarzani (UM), Biomarker discovery in heart failure

16 Jaebok Kim (UT), Automatic recognition of engagement and emotion in a group of children

17 Jianpeng Zhang (TUE), On Graph Sample Clustering

18 Henriette Nakad (UL), De Notaris en Private Rechtspraak

19 Minh Duc Pham (VUA), Emergent relational schemas for RDF

20 Manxia Liu (RUN), Time and Bayesian Networks

21 Aad Slootmaker (OUN), EMERGO: a generic platform for authoring and playing scenario-based serious games

22 Eric Fernandes de Mello Araujo (VUA), Contagious: Modeling the Spread of Behaviours, Perceptions and Emotions in Social Networks

23 Kim Schouten (EUR), Semantics-driven Aspect-Based Sentiment Analysis

24 Jered Vroon (UT), Responsive Social Positioning Behaviour for Semi-Autonomous Telepresence Robots

25 Riste Gligorov (VUA), Serious Games in Audio-Visual Collections

26 Roelof Anne Jelle de Vries (UT),Theory-Based and Tailor-Made: Motivational Messages for Behavior Change Technology

27 Maikel Leemans (TUE), Hierarchical Process Mining for Scalable Software Analysis

28 Christian Willemse (UT), Social Touch Technologies: How they feel and how they make you feel

29 Yu Gu (UVT), Emotion Recognition from Mandarin Speech

30 Wouter Beek, The "K" in "semantic web" stands for "knowledge": scaling semantics to the web

2019 01 Rob van Eijk (UL), Web privacy measurement in real-time bidding systems. A graph-based approach to RTB system classification

02 Emmanuelle Beauxis Aussalet (CWI, UU), Statistics and Visualizations for Assessing Class Size Uncertainty

03 Eduardo Gonzalez Lopez de Murillas (TUE), Process Mining on Databases: Extracting Event Data from Real Life Data Sources

04 Ridho Rahmadi (RUN), Finding stable causal structures from clinical data

05 Sebastiaan van Zelst (TUE), Process Mining with Streaming Data

06 Chris Dijkshoorn (VU), Nichesourcing for Improving Access to Linked Cultural Heritage Datasets

07 Soude Fazeli (TUD), Recommender Systems in Social Learning Platforms

08 Frits de Nijs (TUD), Resource-constrained Multi-agent Markov Decision Processes

09 Fahimeh Alizadeh Moghaddam (UVA), Self-adaptation for energy efficiency in software systems

10 Qing Chuan Ye (EUR), Multi-objective Optimization Methods for Allocation and Prediction

11 Yue Zhao (TUD), Learning Analytics Technology to Understand Learner Behavioral Engagement in MOOCs

12 Jacqueline Heinerman (VU), Better Together

13 Guanliang Chen (TUD), MOOC Analytics: Learner Modeling and Content Generation

14 Daniel Davis (TUD), Large-Scale Learning Analytics: Modeling Learner Behavior & Improving Learning Outcomes in Massive Open Online Courses

15 Erwin Walraven (TUD), Planning under Uncertainty in Constrained and Partially Observable Environments

16 Guangming Li (TUE), Process Mining based on Object-Centric Behavioral Constraint (OCBC) Models

17 Ali Hurriyetoglu (RUN), Extracting actionable information from microtexts

18 Gerard Wagenaar (UU), Artefacts in Agile Team Communication

19 Vincent Koeman (TUD), Tools for Developing Cognitive Agents

20 Chide Groenouwe (UU), Fostering technically augmented human collective intelligence

21 Cong Liu (TUE), Software Data Analytics: Architectural Model Discovery and Design Pattern Detection

22 Martin van den Berg (VU), Improving IT Decisions with Enterprise Architecture

23 Qin Liu (TUD), Intelligent Control Systems: Learning, Interpreting, Verification

24 Anca Dumitrache (VU), Truth in Disagreement - Crowdsourcing Labeled Data for Natural Language Processing

25 Emiel van Miltenburg (VU), Pragmatic factors in (automatic) image description

26 Prince Singh (UT), An Integration Platform for Synchromodal Transport

27 Alessandra Antonaci (OUN), The Gamification Design Process applied to (Massive) Open Online Courses

28 Esther Kuindersma (UL), Cleared for take-off: Game-based learning to prepare airline pilots for critical situations

29 Daniel Formolo (VU), Using virtual agents for simulation and training of social skills in safety-critical circumstances

30 Vahid Yazdanpanah (UT), Multiagent Industrial Symbiosis Systems

31 Milan Jelisavcic (VU), Alive and Kicking: Baby Steps in Robotics

32 Chiara Sironi (UM), Monte-Carlo Tree Search for Artificial General Intelligence in Games

33 Anil Yaman (TUE), Evolution of Biologically Inspired Learning in Artificial Neural Networks

34 Negar Ahmadi (TUE), EEG Microstate and Functional Brain Network Features for Classification of Epilepsy and PNES

35 Lisa Facey-Shaw (OUN), Gamification with digital badges in learning programming

36 Kevin Ackermans (OUN), Designing Video-Enhanced Rubrics to Master Complex Skills

37 Jian Fang (TUD), Database Acceleration on FPGAs

38 Akos Kadar (OUN), Learning visually grounded and multilingual representations

---

2020 01 Armon Toubman (UL), Calculated Moves: Generating Air Combat Behaviour

02 Marcos de Paula Bueno (UL), Unraveling Temporal Processes using Probabilistic Graphical Models

03 Mostafa Deghani (UvA), Learning with Imperfect Supervision for Language Understanding
04 Maarten van Gompel (RUN), Context as Linguistic Bridges
05 Yulong Pei (TUE), On local and global structure mining
06 Preethu Rose Anish (UT), Stimulation Architectural Thinking during Requirements Elicitation - An Approach and Tool Support
07 Wim van der Vegt (OUN), Towards a software architecture for reusable game components
08 Ali Mirsoleimani (UL),Structured Parallel Programming for Monte Carlo Tree Search
09 Myriam Traub (UU), Measuring Tool Bias and Improving Data Quality for Digital Humanities Research
10 Alifah Syamsiyah (TUE), In-database Preprocessing for Process Mining
11 Sepideh Mesbah (TUD), Semantic-Enhanced Training Data AugmentationMethods for Long-Tail Entity Recognition Models
12 Ward van Breda (VU), Predictive Modeling in E-Mental Health: Exploring Applicability in Personalised Depression Treatment
13 Marco Virgolin (CWI), Design and Application of Gene-pool Optimal Mixing Evolutionary Algorithms for Genetic Programming
14 Mark Raasveldt (CWI/UL), Integrating Analytics with Relational Databases
15 Konstantinos Georgiadis (OUN), Smart CAT: Machine Learning for Configurable Assessments in Serious Games
16 Ilona Wilmont (RUN), Cognitive Aspects of Conceptual Modelling
17 Daniele Di Mitri (OUN), The Multimodal Tutor: Adaptive Feedback from Multimodal Experiences
18 Georgios Methenitis (TUD), Agent Interactions & Mechanisms in Markets with Uncertainties: Electricity Markets in Renewable Energy Systems
19 Guido van Capelleveen (UT), Industrial Symbiosis Recommender Systems
20 Albert Hankel (VU), Embedding Green ICT Maturity in Organisations
21 Karine da Silva Miras de Araujo (VU), Where is the robot?: Life as it could be
22 Maryam Masoud Khamis (RUN), Understanding complex systems implementation through a modeling approach: the case of e-government in Zanzibar
23 Rianne Conijn (UT), The Keys to Writing: A writing analytics approach to studying writing processes using keystroke logging
24 Lenin da Nobrega Medeiros (VUA/RUN), How are you feeling, human? Towards emotionally supportive chatbots
25 Xin Du (TUE), The Uncertainty in Exceptional Model Mining
26 Krzysztof Leszek Sadowski (UU), GAMBIT: Genetic Algorithm for Model-Based mixed-Integer opTimization
27 Ekaterina Muravyeva (TUD), Personal data and informed consent in an educational context
28 Bibeg Limbu (TUD), Multimodal interaction for deliberate practice: Training complex skills with augmented reality
29 Ioan Gabriel Bucur (RUN), Being Bayesian about Causal Inference
30 Bob Zadok Blok (UL), Creatief, Creatieve, Creatiefst
31 Gongjin Lan (VU), Learning better – From Baby to Better
32 Jason Rhuggenaath (TUE), Revenue management in online markets: pricing and online advertising
33 Rick Gilsing (TUE), Supporting service-dominant business model evaluation in the context of business model innovation

34 Anna Bon (MU), Intervention or Collaboration? Redesigning Information and Communication Technologies for Development

35 Siamak Farshidi (UU), Multi-Criteria Decision-Making in Software Production

---

2021 01 Francisco Xavier Dos Santos Fonseca (TUD), Location-based Games for Social Interaction in Public Space

02 Rijk Mercuur (TUD), Simulating Human Routines:Integrating Social Practice Theory in Agent-Based Models

03 Seyyed Hadi Hashemi (UVA), Modeling Users Interacting with Smart Devices

04 Ioana Jivet (OU), The Dashboard That Loved Me: Designing adaptive learning analytics for self-regulated learning

05 Davide Dell'Anna (UU), Data-Driven Supervision of Autonomous Systems

06 Daniel Davison (UT), "Hey robot, what do you think?" How children learn with a social robot

07 Armel Lefebvre (UU), Research data management for open science

08 Nardie Fanchamps (OU), The Influence of Sense-Reason-Act Programming on Computational Thinking

09 Cristina Zaga (UT), The Design of Robothings. Non-Anthropomorphic and Non-Verbal Robots to Promote Childrens Collaboration Through Play

10 Quinten Meertens (UvA), Misclassification Bias in Statistical Learning

11 Anne van Rossum (UL), Nonparametric Bayesian Methods in Robotic Vision

12 Lei Pi (UL), External Knowledge Absorption in Chinese SMEs

13 Bob R. Schadenberg (UT), Robots for Autistic Children: Understanding and Facilitating Predictability for Engagement in Learning

14 Negin Samaeemofrad (UL), Business Incubators: The Impact of Their Support

15 Onat Ege Adali (TU/e), Transformation of Value Propositions into Resource Re-Configurations through the Business Services Paradigm

16 Esam A. H. Ghaleb (UM), BIMODAL EMOTION RECOGNITION FROM AUDIO-VISUAL CUES

17 Dario Dotti (UM), Human Behavior Understanding from motion and bodily cues using deep neural networks

18 Remi Wieten (UU), Bridging the Gap Between Informal Sense-Making Tools and Formal Systems - Facilitating the Construction of Bayesian Networks and Argumentation Frameworks

19 Roberto Verdecchia (VU), Architectural Technical Debt: Identification and Management

20 Masoud Mansoury (TU/e), Understanding and Mitigating Multi-Sided Exposure Bias in Recommender Systems

21 Pedro Thiago Timbó Holanda (CWI), Progressive Indexes

22 Sihang Qiu (TUD), Conversational Crowdsourcing

23 Hugo Manuel Proença (LIACS), Robust rules for prediction and description

24 Kaijie Zhu (TUE), On Efficient Temporal Subgraph Query Processing

25 Eoin Martino Grua (VUA), The Future of E-Health is Mobile: Combining AI and Self-Adaptation to Create Adaptive E-Health Mobile Applications

26 Benno Kruit (CWI & VUA), Reading the Grid: Extending Knowledge Bases from Human-readable Tables

27 Jelte van Waterschoot (UT), Personalized and Personal Conversations: Designing Agents Who Want to Connect With You

28 Christoph Selig (UL), Understanding the Heterogeneity of Corporate Entrepreneurship Programs

2020 1 Judith van Stegeren (UT), Flavor text generation for role-playing video games

2 Paulo da Costa (TU/e), Data-driven Prognostics and Logistics Optimisation: A Deep Learning Journey

3 Ali el Hassouni (VUA), A Model A Day Keeps The Doctor Away: Reinforcement Learning For Personalized Healthcare

4 Ünal Aksu (UU), A Cross-Organizational Process Mining Framework

5 Shiwei Liu (TU/e), Sparse Neural Network Training with In-Time Over-Parameterization

6 Reza Refaei Afshar (TU/e), Machine Learning for Ad Publishers in Real Time Bidding

7 Sambit Praharaj (OU), Measuring the Unmeasurable? Towards Automatic Co-located Collaboration Analytics

8 Maikel L. van Eck (TU/e), Process Mining for Smart Product Design

9 Oana Andreea Inel (VUA), Understanding Events: A Diversity-driven Human-Machine Approach

10 Felipe Moraes Gomes (TUD), Examining the Effectiveness of Collaborative Search Engines

11 Mirjam de Haas (UT), Staying engaged in child-robot interaction, a quantitative approach to studying preschoolers engagement with robots and tasks during second-language tutoring