Lessons from Computational Modelling of Reference Production in Mandarin and English

Guanyi Chen and **Kees van Deemter**
Department of Information and Computing Sciences
Utrecht University
{g.chen, c.j.vandeemter}@uu.nl

Abstract

Referring expression generation (REG) algorithms offer computational models of the production of referring expressions. In earlier work, a corpus of referring expressions (REs) in Mandarin was introduced. In the present paper, we annotate this corpus, evaluate classic REG algorithms on it, and compare the results with earlier results on the evaluation of REG for English referring expressions. Next, we offer an in-depth analysis of the corpus, focusing on issues that arise from the grammar of Mandarin. We discuss shortcomings of previous REG evaluations that came to light during our investigation and we highlight some surprising results. Perhaps most strikingly, we found a much higher proportion of under-specified expressions than previous studies had suggested, not just in Mandarin but in English as well.

1 Introduction

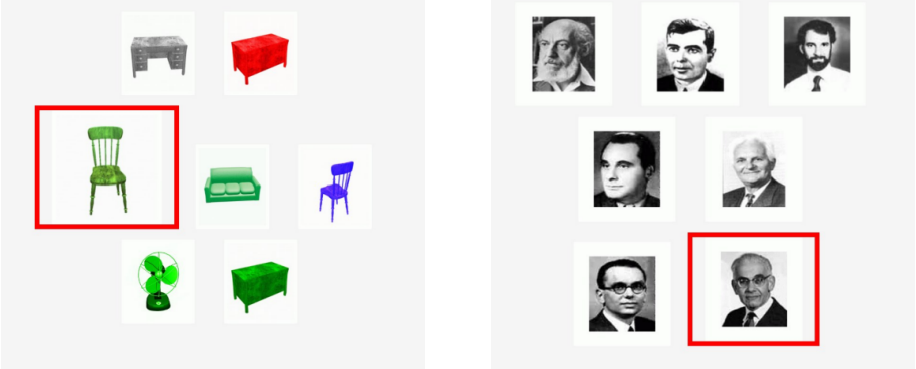
Referring expression generation (REG) originated as a sub-task of traditional natural language generation systems (NLG, Reiter and Dale, 2000). The task is to generate expressions that help hearers to identify the referent that a speaker is thinking about. REG has important practical value in natural language generation (Gatt and Krahmer, 2018), computer vision (Mao et al., 2016), and robotics (Fang et al., 2015). Additionally, REG algorithms can be seen as models of human language use (van Deemter, 2016).

In line with this second angle, and unlike REG studies which have started to use black-box Neural Network based models (e.g., Mao et al. (2016); Ferreira et al. (2018) and Cao and Cheung (2019)), we focus on two aspects (cf., Krahmer and van Deemter (2012)): 1) designing and conducting controlled elicitation experiments, yielding corpora which are then used for analysing and evaluating REG algorithms to gain insight into linguistic phenomena, e.g., GRE3D3 (Dale and Viethen, 2009), TUNA (Gatt et al., 2007; van Deemter et al., 2012), COCONUT (Jordan and Walker, 2005), and MAPTASK (Gupta and Stent, 2005). 2) designing algorithms 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 and 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 and 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. So far, only a preliminary analysis has been performed on MTUNA, and this analysis has focussed on issues of Linguistic Realisation (van Deemter et al., 2017): the REs in the corpus have not yet been compared with those in other languages, and the performance of REG algorithms on the corpus has not been evaluated.

To fill this gap, we provide a more detailed analysis of the use of Mandarin REs on the basis of the MTUNA corpus. We 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 ETUNA corpus. Since it has been claimed that Mandarin favours brevity over clarity – the idea that Mandarin is "cooler" than these other languages (Newnham, 1971; Huang, 1984) – relying more on communicative context for disambiguation than western languages, we concentrated on the use of over- and under-specification. After all, if Mandarin favours brevity over clarity to a greater extent than English and Dutch, then one would expect to see less over-specification and

Proceedings of The 13th International Conference on Natural Language Generation, pages 263–272,
Dublin, Ireland, 15-18 December, 2020.



(a) (b)

Figure 1: Two scenes from the TUNA experiment, in which (a) is a situation from the furniture domain while (b) is from the people domain.

more over-specification in Mandarin.

2 Background

The analysis reported in paper is based on the MTUNA (for Mandarin) and ETUNA (for English) corpus. We start by briefly introducing the TUNA experiments in general, and we highlight some special features of MTUNA together with its initial findings.

2.1 The TUNA Experiments

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 linguist's 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 1 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 referring expressions that identify the target referents from the other objects in the scene (their "distractors"). For example, for the scene in Figure 1, 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 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 and 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 and Krahmer, 2010), German (GTUNA, Howcroft et al., 2017), and Mandarin (van Deemter et al., 2017).

2.2 The Mandarin TUNA

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 typewritten 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 DTUNA, where there are totally 40 trials. Different from other TUNAs which always asked subjects essentially the same question, namely *Which object/objects appears/appear in a red window?*, MTUNA distinguished between referring expressions in subject and object position.[1]

[1]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.

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 referring expression 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 1) most definite REs are bare nouns. 2) indefinite REs also appear quite often, especially in subject position.

3 Research Questions

Analogous to studies of earlier TUNA corpora, our primary research question (**RQ1**) 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 algorithms 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 minimally number of properties), is expected to have higher performance on MTUNA than on ETUNA. The expected effect on other classic algorithms is less clear.

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.[2] For this reason, algorithms such as the Incremental Algorithm (Dale and 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 second research question (**RQ2**) asks *to what extent the role of* `TYPE` *differs between English and Mandarin.* Connected with this, we were curious to what extent this issue affects the performance of the classic REG algorithms.

As discussed in section 1, the coolness hypothesis stated that Chinese relies more on the communicative context for disambiguation than western languages, 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 Chen et al. (2018), indicting that they sometimes even drop REs. In this study, we were curious about (**RQ3**) *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.

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, we investigated (**RQ4**) *how syntactic position influences the use of over-/under-specification and the performance of REG algorithms.*

4 Method

Before we address the four research questions in section 3, we explain how we annotated the corpus. The annotated corpus is available at `github.com/a-quei/mtuna-annotated`

4.1 Annotating the Corpus

1650 REs were semantically annotated (after omitting some unfinished REs from the corpus) following the scheme of van der Sluis et al. (2006).[3] For simplicity, instead of XML we use the 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

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

[3] 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. The annotated corpus is public available at: xxx.

	Domain	Total	Mini.	Real	Nom.	Num.	Wrong	Other	Under
MTUNA	furniture	377	46	117	132	2	11	5	64
	people	371	16	216	68	13	4	6	48
MTUNA-OL	furniture	264	9	83	104	0	8	4	56
	people	222	14	144	36	2	1	3	22
ETUNA	furniture	158	1	58	62	0	0	0	37
	people	132	3	75	37	0	0	0	7

Table 1: Frequencies of referring expressions that fall in each type specifications in MTUNA, 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, **wrong** is the duplicate-attribute over-specification, **other** stands for the RE that cannot be classified into any of these categories, and **under** is the under-specification.

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 2: 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.

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 for distinguishing the target. Therefore, we decided to use hairColour and beardColour as separate attributes. As pointed out in van Deemter et al. (2012), since the attribute hairColour is depend 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 we also re-annotated the ETUNA corpus, using the same annotators. Details about which properties were annotated and examples of annotated REs can be found in Appendix A.

4.2 Annotating Over-/Under-specifications

To gain an insightful analysis of the speakers' use of over- and under-specification, and to ensure that our annotations are well defined, we will offer some definitions. In addition, given our interest in the role of TYPE, we will sub-categorise by distinguishing different types of over-specifications. Concretely, we asked the annotators to consider the following types of specifications:

Minimal Description. an RE that successfully singles out the target referent and does this by using the minimum possible number of properties. These are the REs that match Dale and Reiter's Full Brevity;

Numerical Over-specification. an RE that uses more properties than the corresponding minimal description uses, yet the removal of any of them results in a referential confusion. For instance, for the scene in 1(a), the RE *the green chair* is a numerical over-specification as it uses more properties than the minimal description *the large one*;

Nominal Over-specification. an RE from which only one of its properties is removable, namely the TYPE of the target;

Under-specification. an RE all of whose properties are true of the referent but that causes referential confusion (i.e., it is not a distinguishing description in the sense of Dale (1992));

Wrong Description. an RE whose properties use one or more incorrect values for a given attribute. In line with previous TUNA evaluations, we only consider a value to be wrong if it could prevent a hearer from recognising the target. For example, the RE *the pink chair* is not called wrong if the referent is a red chair.

We annotated each RE in both corpora [4], and we annotated each scene in each corpus. Thus, for each RE, we annotate which of the above specification types it falls in, and how many over-specified/under-specified properties the RE contains. In Appendix B, Table 7 records, for each scene, how many different minimal descriptions the scene permits (most often just 1, but sometimes 2 or 3). The results per RE are depicted in Table 1[5] and the results per scene are in Appendix B.

5 Analysis

Before reporting results and analysis, we explain what datasets and algorithms were analysed, and how evaluation was performed.

Dataset. The sources of our dataset are the MTUNA and ETUNA corpora. For Mandarin, we used the whole MTUNA dataset. For comparing between languages fairly, we only used REs for scenes that were shared between MTUNA and ETUNA; we call this set of shared scenes MTUNA-OL. The original MTUNA has 20 trials, with 10 trials for each domain. The MTUNA-OL and ETUNA contains 13 trials, in which there are 7 and 6 trials from furniture and people domain respectively. (More details of which scene is used can be found in the Appendix.)

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 and Reiter, 1995).

Evaluation Metrics. We used what are still the most commonly used metrics for evaluating attribute choice in REG. One is 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. We also report the "perfect recall percentage" (PRP), the proportion of times the algorithm achieves a DICE score of 1, which is seen as a indicator of the recall of an algorithm.

5.1 Performance of Algorithms on MTUNA

We report the evaluation results on MTUNA and MTUNA-OL in the Table 2 and 3. For the FB algorithm, we tested both the original version and the version that always appends 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 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 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 1, a Chi Squared Test suggests that, in MTUNA, there are more real over-specifications ($\chi^2(1, 747) = 55.95, p < .001$) but fewer nominal over-specifications ($\chi^2(1, 747) = 26.57, p <$

[4] When applying this annotation scheme to REs that have multiple targets, adaptations need to be made. But since the focus of this paper is on singular REs, we will not offer details.

[5] We observed a large number of minimal descriptions in the furniture domain of MTUNA. This is a result of the fact that some trials in MTUNA use `TYPE` in their minimal descriptions.

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

.001) in the people domain than in the furniture domain[6]. 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, FB outperforms many IAs on PRP, which does not happen in the furniture domain.

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 (Appendix B), 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 added 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 miminal descriptions: *the large one* and *the green one*. Our FB produce the second minimal description (as it can only produce one RE at a time). When we computing 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 attributes. This implies that when a corpus contains multiple minimal REs, this will artificially lower the DICE scores.[7] 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.

5.2 Cross-linguistic Comparison

Table 3 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 on the order of the their performance. An interesting observation is that, after correcting a few errors in the annotation of ETUNA (cf. section 4.1), 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 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 suggest that, in fact, it is English speakers who

[6]This highlights the importance of sub-categorising the different kinds of over-specifications, as we have done in section 4

[7]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.

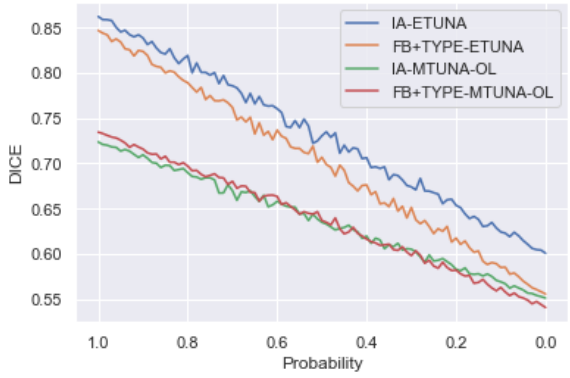


Figure 2: Change of the performance with respected to different probabilities of inserting superfluous TYPE for either the FB+TYPE and IA on either the people domain of MTUNA-OL and ETUNA.

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, 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 deviations for every model is higher.

5.3 RQ2: the Role of TYPE

On the use of TYPE, we first 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. Second, for Lv's hypothesis introduced in section 3, we observed a smaller proportion of uses of TYPE in the people domain ($\chi^2(1, 485) = 24.16, p < .001$), where all the objects share the same value of TYPE. 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.

To further assess the role of TYPE and to find more evidence regarding Lv's hypothesis, 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 2.

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: 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.

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 maybe 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.

5.4 RQ3: Over-/Under-specification

In light of Table 1, some obvious conclusions can be drawn. For example, more "real" over-specifications are used for more complex domains (i.e., the people domain) than for simple ones. Focusing on RQ3 in section 3, its two hypotheses are both rejected: no significant difference has been found in the use of over-specifications ($\chi^2(1, 775) = 0.82, p = 0.052$) or in the use of under-specifications ($\chi^2(1, 775) = 0.745, p = 0.105$). Focusing on the people domain, where FB+TYPE performed better in English than in Mandarin, we found no significant difference ($\chi^2(1, 354) = 2.53, p = 0.112$).

5.5 RQ4: Syntactic Position

For RQ4, we counted the number of real over-specifications and under-specification in subject and object position. In the MTUNA-OL corpus, there are 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 detail.

As for the computational modelling, generally speaking, all algorithms performed better for REs in subject position than for REs is object position, with one exception, namely the GR algorithm for the people domain; the difference is significant in the Furniture domain, but is not in the people domain, possibly because the furniture domain contains more under-specifications.

6 Discussion

6.1 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.

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 significant influence of the syntactic position of the RE on the use of under-specification 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). On the other hand, 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 position? 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.

Perhaps our most surprising findings regard the use of under-specification: firstly (deviating from what van Deemter et al. (2017) hypothesised), we did not find significantly more under-specifications in MTUNA than in ETUNA. We found 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 and 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) report that only 5% of REs in DTUNA were under-specifications.[8]

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 ignored 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?

6.2 Lessons about REG Evaluation

Most REG evaluations so far have made use of the DICE score (Dice, 1945). However, on top of the discussions of van Deemter and Gatt (2007) and of section 5, we identify the following three

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

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 (section 2.2) 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 for 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 section 5, these issues impact evaluation results and this might cause the conclusions from evaluating algorithms with TUNA not to be reproducable.

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.[9] 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.

Acknowledgements

We thank the anonymous reviewers for their helpful comments. Guanyi Chen is supported by China Scholarship Council (No.201907720022).

[9]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.

References

Meng Cao and Jackie Chi Kit Cheung. 2019. Referring expression generation using entity profiles. *arXiv preprint arXiv:1909.01528*.

Yuen Ren Chao. 1965. *A grammar of spoken Chinese*. Univ of California Press.

Guanyi Chen, Kees van Deemter, and Chenghua Lin. 2018. Modelling pro-drop with the rational speech acts model. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 57–66. Association for Computational Linguistics (ACL).

Robert Dale. 1989. Cooking up referring expressions. In *27th Annual Meeting of the association for Computational Linguistics*, pages 68–75.

Robert Dale. 1992. *Generating referring expressions: Constructing descriptions in a domain of objects and processes.* The MIT Press.

Robert Dale and Ehud Reiter. 1995. Computational interpretations of the gricean maxims in the generation of referring expressions. *Cognitive science*, 19(2):233–263.

Robert Dale and Jette Viethen. 2009. Referring expression generation through attribute-based heuristics. In *Proceedings of the 12th European workshop on natural language generation (ENLG 2009)*, pages 58–65.

Kees van Deemter. 2016. *Computational models of referring: a study in cognitive science*. MIT Press.

Kees van Deemter and Albert Gatt. 2007. Content determination in GRE: Evaluating the evaluator. In *Using Corpora for Natural Language Generation: Language Generation and Machine Translation*.

Kees van Deemter, Albert Gatt, Ielka van der Sluis, and Richard Power. 2012. Generation of referring expressions: Assessing the incremental algorithm. *Cognitive science*, 36(5):799–836.

Kees van Deemter, Le Sun, Rint Sybesma, Xiao Li, Bo Chen, and Muyun Yang. 2017. Investigating the content and form of referring expressions in Mandarin: introducing the mtuna corpus. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 213–217, Santiago de Compostela, Spain. Association for Computational Linguistics.

Judith Degen, Robert D Hawkins, Caroline Graf, Elisa Kreiss, and Noah D Goodman. 2020. When redundancy is useful: A bayesian approach to "overinformative" referring expressions. *Psychological Review*.

Lee R Dice. 1945. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302.

Rui Fang, Malcolm Doering, and Joyce Y Chai. 2015. Embodied collaborative referring expression generation in situated human-robot interaction. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, pages 271–278.

Thiago Castro Ferreira, Diego Moussallem, Ákos Kádár, Sander Wubben, and Emiel Krahmer. 2018. Neuralreg: An end-to-end approach to referring expression generation. *arXiv preprint arXiv:1805.08093*.

Albert Gatt and Anja Belz. 2010. Introducing shared tasks to nlg: The tuna shared task evaluation challenges. In *Empirical methods in natural language generation*, pages 264–293. Springer.

Albert Gatt and Emiel Krahmer. 2018. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, 61:65–170.

Albert Gatt, Emiel Krahmer, Roger van Gompel, and Kees van Deemter. 2013. Production of referring expressions: Preference trumps discrimination. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 35.

Albert Gatt, Ielka van der Sluis, and Kees van Deemter. 2007. Evaluating algorithms for the generation of referring expressions using a balanced corpus. In *Proceedings of the 11th European Workshop on Natural Language Generation (ENLG'07)*, pages 49–56, Schloss Dagstuhl, Germany. Association for Computational Linguistics.

Albert Gatt, Ielka van der Sluis, and Kees van Deemter. 2008. Xml format guidelines for the tuna corpus. Technical report, Technical report, Dept of Computing Science, University of Aberdeen.

Roger van Gompel, Kees van Deemter, Albert Gatt, Rick Snoeren, and Emiel J Krahmer. 2019. Conceptualization in reference production: Probabilistic modeling and experimental testing. *Psychological review*, 126(3):345.

Surabhi Gupta and Amanda Stent. 2005. Automatic evaluation of referring expression generation using corpora. In *Proceedings of the Workshop on Using Corpora for Natural Language Generation*, pages 1–6. Citeseer.

William S Horton and Boaz Keysar. 1996. When do speakers take into account common ground? *Cognition*, 59(1):91–117.

David M Howcroft, Jorrig Vogels, and Vera Demberg. 2017. G-tuna: a corpus of referring expressions in german, including duration information. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 149–153.

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

Huang C-T James, Y-H Audrey Li, and Yafei Li. 2009. The syntax of chinese. *Cambridge, Cambridge). doi*, 10.

Pamela W Jordan and Marilyn A Walker. 2005. Learning content selection rules for generating object descriptions in dialogue. *Journal of Artificial Intelligence Research*, 24:157–194.

Ruud Koolen, Albert Gatt, Martijn Goudbeek, and Emiel Krahmer. 2011. Factors causing overspecification in definite descriptions. *Journal of Pragmatics*, 43(13):3231–3250.

Ruud Koolen and Emiel Krahmer. 2010. The d-tuna corpus: A dutch dataset for the evaluation of referring expression generation algorithms. In *LREC*.

Emiel Krahmer and Kees van Deemter. 2012. Computational generation of referring expressions: A survey. *Computational Linguistics*, 38(1):173–218.

Willem JM Levelt. 1993. *Speaking: From intention to articulation*, volume 1. MIT press.

Shuxiang Lv. 1979. Problems in the analysis of chinese grammar.

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. 2016. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20.

Richard Newnham. 1971. *About Chinese*. Penguin Books Ltd.

Ehud Reiter and Robert Dale. 2000. *Building natural language generation systems*. Cambridge university press.

Ielka van der Sluis, Albert Gatt, and Kees van Deemter. 2006. Manual for the tuna corpus: Referring expressions in two domains. Technical Report AUCS/TR0705, Department of Computing Science, Univ. of Aberdeen.

Ielka van der Sluis, Albert Gatt, and Kees van Deemter. 2007. Evaluating algorithms for the generation of referring expressions: Going beyond toy domains. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP'07)*, Borovets, Bulgaria. RANLP.

Yanzhi Yi, Hangyu Deng, and Jinglu Hu. 2020. Improving image captioning evaluation by considering inter references variance. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 985–994, Online. Association for Computational Linguistics.