Investigating the Content and Form of Referring Expressions in Mandarin: Introducing the Mtuna Corpus

Kees van Deemter
University of Aberdeen

Le Sun
Institute of Software
Chinese Academy of Sciences

Rint Sybesma
Leiden University

Xiao Li
University of Aberdeen

Bo Chen
Institute of Software
Chinese Academy of Sciences

Muyun Yang
Harbin Institute of Technology

Abstract

East Asian languages are thought to handle reference differently from English, particularly in terms of the marking of definiteness and number. We present the first Data-Text corpus for Referring Expressions in Mandarin, and we use this corpus to test some initial hypotheses inspired by the theoretical linguistics literature. Our findings suggest that function words deserve more attention in Referring Expression Generation than they have so far received, and they have a bearing on the debate about whether different languages make different trade-offs between clarity and brevity.

1 Introduction

East Asian languages can differ considerably from the languages of Western Europe, which have often dominated formal and computational studies of language. One phenomenon where these differences are obvious is Referring Expressions (REs), where languages such as Mandarin differ markedly from, for example, English, in terms of their expression of *number* (e.g., Am I referring to 1 thing or more?), *maximality* (Am I talking about all the things that have a certain combination of properties, or only some of them?), and *givenness* status (Am I talking about something that the hearer is familiar with?).

To gain an insight in these matters, and to assist future research, we have embarked on a data gathering enterprise focussing on East Asian languages, starting with a language elicitation experiment in which speakers of Mandarin were asked to produce one-shot REs in a carefully balanced range of situations. The present paper introduces the corpus and offers an initial assessment of some of our research questions. The Mtuna data-text corpus is freely available from `homepages.abdn.ac.uk/k.vdeemter-/pages/mtuna-webpage/`, containing the original Chinese characters, their transcription into (phonetic) pinyin notation, and an informal English gloss. Each RE is coupled with a pictorial scene that shows the referent and its distractors in the same way as participants in the experiment saw it.

2 Initial Research Questions

According to the linguistics literature, REs without a numeral can take three different shapes, namely (1) Demonstrative + Classifier + Noun Group (e.g., *Na ge laoren*, "That (person) old person"), (2) Demonstrative + Noun Group (*Na laoren*, "That old person") , and (3) (bare) Noun Group *Laoren*, "Old person") (Cheng and Sybesma 1999, 2015).[1] We call these the DCN pattern, the DN pattern, and the N pattern respectively. Together we call them the *Canonical* Patterns of reference in Mandarin.

Noun Groups (the third pattern) can be strikingly open to interpretation: they can be understood as indefinite, generic, or definite; moreover, they are not marked for number. Thus, a bare Noun Group like *lüse de yizi* (lit: *green colour chair*) can mean *the green chair*, but equally, *the green chairs*, *green chairs* (in general), and *a green chair*. We are in-

[1] Classifiers are words that attribute entities to ontological classes; in certain contexts classifiers are obligatory. Noun Groups are combinations of Nouns and their modifiers (e.g., adjectives).

Proceedings of The 10th International Natural Language Generation conference, pages 213–217,
Santiago de Compostela, Spain, September 4-7 2017.

terested how frequently each of these NPs occur because it will give us a first insight into the role of underspecification in Mandarin. Following a small pilot experiment with 10 speakers, we set out to address the following questions:

Research Questions: *Are the three Canonical patterns the ones that are used predominantly when people refer? Is this only true for sentence positions where definiteness is the norm (in Mandarin this is the pre-verbal position), or is it equally true for other positions? How frequently are REs underspecified for number, maximality,*[2] *and givenness?*

The original English TUNA corpora were collected in 2006 and used for multiple shared tasks (Gatt and Belz 2010) on Referring Expression Generation (REG) and other work in this area (van Deemter et al. 2012). Each corpus consists of REs produced by humans presented with a target item (1 or 2 pieces of furniture, or 1 or 2 people's faces) and a set of distractors (other pieces of furniture or faces), in a web-based elicitation paradigm. A Dutch TUNA was conducted in 2011 (DTuna, Koolen et al. 2011) and an Arabic one in 2015 (Khan 2015).

Though a number of other REG data gathering exercises have followed (see e.g., van Deemter 2016), the TUNA setup suits our research questions well. However, the analysis of the corpus is very different this time. For whereas earlier TUNAs focussed on the properties expressed by a given RE (chair, green, etc.), our research questions mean that function words (English: *the, a, one, two, this, those, both*) are key. Although these are sometimes considered to be part of Linguistic Realisation, they are not just "syntactic sugar", since they contribute much to the information conveyed by these REs (e.g., Kamp and Reyle 1993, and many other treatments of the semantics and pragmatics of English).

3 The Mtuna Experiment and Corpus

The 44 stimuli of our experiment (40 + 4 items originally used for training only) resemble closely those of earlier TUNAs; like these, they were semantically balanced (e.g. the number of cases when the target could be identified by means of colour was identical to the number of cases when the target could be identified by means of size). They include references to sets as well as individual items. Instructions to participants were translated from Dtuna, except that the new instructions did not include examples of actual REs in the target language (i.e., Mandarin), since these could have biased participants towards particular syntactic patterns. Also, where earlier TUNAs had always asked subjects essentially the same question, namely "Which object/objects appears/appear in a red window?", the new experiment distinguished between REs in pre-verbal and post-verbal position.

Figure 1: *A furniture trial, with the RE in pre-verbal position.*

Participants were recruited from the Chinese Academy of Sciences (Institute of Software) and the Harbin Institute of Technology. Data from 37 participants have been obtained. 35/37 were self-assessed native speakers of Mandarin, 2/37 were merely fluent. 29/37 were from the North of China and 8/37 from the South. Subjects were discouraged from using location in their REs, being told that the recipient might view the scenes on a page that uses a different layout. Items were presented in random order and with random layout where all entities were allotted to cells in a 3-by-5 grid invisible to participants.

Sentence position was varied in a between-

[2] An occurrence of the above-mentioned NP *lüse de yizi* would be *maximal* if it denoted *all* the green chairs in the scene.

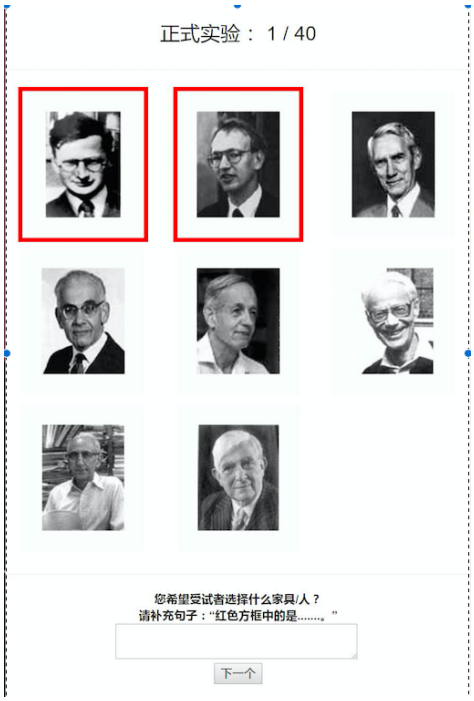

Figure 2: *A people trial with RE in post-verbal position.*

subjects design: Participants who were asked to produce REs in pre-verbal position were asked this trigger question: *Shenme jiaju/ren chuxian zai le hongkuang zhong?* (What furniture / person occur(s) in red frame(s)?) Immediately below this question, the page continues: *Qing buchong juzi: ..... zai hongse fangkuang zhong* (Please complete the sentence: ".....is in the red frame(s)".) Participants who were asked to produce REs in post-verbal position were asked: *Nin xiwang shoushizhe xuanze shenme jiaju/ren?* (What furniture / person do you want the participants to choose?) The page continues: *Qing buchong juzi: Hongse fangkuang zhong de shi.....* (Please complete the sentence: "What's in the red frame is .....")

4 Initial Analysis of the Mtuna corpus

The corpus was subjected to an initial analysis of all descriptions elicited. Our conclusions need to be handled with care, because further analysis is needed and the narrowness of our participant base (recruited from two Language Technology groups) may have biassed our results.

1. Are the three Canonical Patterns the ones that are actually used? Table 2 shows the numbers for references to singular referents only. The N Pattern dominated, whereas no DN or DCN Patterns were found. We did found a small number of Ordinal Patterns, all of which came from a small number of subjects who had ignored the instruction to avoid mentioning the location of the referent (saying things like "the first ... from the left"). Indefinite NPs occurred quite often, even in pre-verbal position, where we had expected not to see them (though the bulk of these REs were produced by just 3 participants). These results appear to be at odds with linguists' views about the dominant patterns; one possible explanation is that Demonstratives are restricted to situations in which the antecedent is either pointed at or mentioned in earlier text (Jenks 2015).

Pattern	Pre-verbal	Post-verbal
N	315	326
DN	0	0
DCN	0	0
Other D	5	4
Indefinite	98	54
Ordinal	4	0

Table 2: Raw frequencies of referential patterns for singular (i.e., non-set) references. *Other D* are structures such as *Lan yizi, zui da de nage* ("The blue chair, that largest one"), where the Demonstrative takes a different position than in DN and DCN. Indefinites tended to be of the form *"Yi ..."* ("One ...") Ordinals were expressions such as ("First from the left"). Not all participants answered all the questions, with different numbers of entries for Pre-verbal and Post-verbal.

2. Was the choice of RE pattern influenced by sentence position? A Chi-Square calculation on the figures of Table 2 suggests a cautiously affirmative answer ($p < .05$), caused by the larger number of indefinites in pre-verbal position. We had expected to see fewer N Patterns in post- than in pre-verbal position, but this expectation was not borne out.

3. How often were REs non-specific as to number, maximality, and givenness? Mandarin's explicit markers for *maximality* were used very rarely (14 occurrences of *dou* ("all")). No explicit markers for *givenness* were found. In both cases, we may have missed out on less obvious markers (e.g., syntactic position may play a role), therefore we plan a new experiment that will investigate readers' or listeners' interpretation of the REs produced in the corpus.

Mandarin (transcribed into pinyin)	Approximate English Gloss
Furniture	
Xiaode lüsede xiangqian de zhuozi	Small green [sub] forward table
Yizhang lüse de shuzhuo	One[cla] green [sub] desk
Zhengmian chao qian de xiezitai, chicun jiaoxiaode nage	Front facing [sub] writing desk, the smaller size [sub] of those
Yige chouti chaowai de lüse xiao shuzhuo	One[cla] small desk with a drawer facing out [sub]
Lüse zhuozi	Green table
Yige zhengmian chaoxiang guancezhe de xiangdui xiao de zhuozi	A[cla] relatively small table facing observer [sub]
People	
Dai yanjing hei toufa de liang ge ren	Wear glasses black hair [sub] two [cla] people
Liang ge dai yanjing de nianqing nanxing	Two [cla] wear glasses [sub] young men
Hei toufa liang ge ren	Black hair two [cla] people
Liang ge dai yanjing chuan heise xifu de heise toufa de nanren zai hongse fangkuang zhong	Two [cla] wear glasses wear black clothes [sub] black hair man in red box
Dai yanjing hei toufa de liangwei kexuejia	Wear glasses black hair [sub] two[cla] scientists
Yige zhengmian de chaowai de dai yanjing, chuan xizhuang da lingdai hei toufa de nanren	One[cla] face outward [sub] wear glasses, wear suit and tie black-haired man

Table 1: Some REs as found in the corpus referring to the target referents in Figures 1 and 2. [cla] denotes a classifier, [sub] denotes a subordinating *de*. The modelling of classifier and subordinator use is a topic to which we will turn in later research.

	Number was marked	Number was not marked
singular post-verbal	59	325
singular pre-verbal	106	312
plural post-verbal	231	157
plural pre-verbal	297	121

Table 3: A singular RE was counted as marked for number if it was of the form *Yi* ... ("One ..."). A plural RE was marked for number if it contained the numeral *liang* ("two") or an ordinal or if it used a conjunction.

Number can be marked by numerals, by ordinals or by the use of logical conjunction (e.g., *he* ("and"), as in *hongse yizi he lüse dianfengshan* ("red chair AND green fan")). Table 3 suggests that number tended to be marked when the referent was plural but not when it was singular; number was marked more often in pre-verbal than in post-verbal position.

5 Discussion

It has often been suggested that East Asian languages handle the trade-off between brevity and clarity differently to those of Western Europe, with the former (as typical instances of languages that are "cool" rather than "hot") allegedly leaning more towards brevity, and relying more on communicative context for disambiguation (Newnham 1971, Huang 1984). If this was true, one would expect that Mandarin REs use less over-specification (i.e., REs from which one or more properties can be removed without causing referential confusion) and more under-specification than in English and Dutch; equally, one might expect that Mandarin REs are less fully specified in terms of number, maximality, and givenness. In future, we want to investigate these hypotheses and their implications for REG more thoroughly.

Based on a first look at our data, a nuanced picture is emerging, where defaults are likely to play a role. Based on the literature (e.g., Chao 1968), Mandarin NPs in pre-verbal position may be interpreted as definite unless there is information to the contrary; based on our data, it may be that a Mandarin NP denotes a singular entity by default, and that plural interpretations only arise when the context enforces this (e.g., by means of a numeral). These issues need to be investigated further.

Some aspects of the unexpected distribution of patterns in Mandarin reported in section 4 may have been caused by ususual features of the communicative situation in which we placed our participants, for instance because only written input was available to them. If this was true, then this would also cast doubt on earlier results that were obtained with the same, TUNA-style, data gathering method.

Acknowledgments

This work is partly supported by the National Natural Science Foundation of China, Grant no. 61433015. We thank Stephen Matthews, University of Hong Kong, for comments, and Albert Gatt, University of Malta, for access to Dutch TUNA.

6 References

Chao 1968. Y. R. Chao. *A Grammar of Spoken Chinese.* University of California Press.

Cheng and Sybesma, 1999. L. Cheng and R. Sybesma. Bare and Not-So-Bare Nouns and the Structure of NP. *Linguistic Inquiry* **30** (4).

Cheng and Sybesma, 2015. L. Cheng and R. Sybesma. [Syntactic sketch of] Mandarin, Syntax Theory and Analysis. An International Handbook. Handbooks of Linguistics and Communication Science, ed. Tibor Kiss and Artemis Alexiadou, 42.1-3, Berlin: Mouton de Gruyter.

Gatt and Belz, 2010. A. Gatt and A. Belz. Introducing shared task evaluation to NLG: The TUNA shared task evaluation challenges. In Krahmer, E. and Theune, M., editors, Empirical Methods in Natural Language Generation, p. 264 - 293. Springer Verlag, Berlin.

Huang 1984. C. -T. J. Huang. On the distribution and reference of empty pronouns. *Linguistic Inquiry* **15** (4), p.531 – 574.

Jenks, 2015. P. Jenks. Two kinds of definites in numeral classifier languages. In Proceedings of Semantics and Linguistic Theory (SALT) 25.

Kamp and Reyle, 1993. H. Kamp and U. Reyle. *From Discourse to Logic*. Kluwer, Dordrecht.

Khan, 2015. I.H. Khan. Do Speakers Produce Different Referring Expressions in Their Native Language Than A Non-native Language? *International Journal of Computational Linguistics Research* **6** (2), p. 41 - 47.

Koolen et al. 2011. R. Koolen, A.Gatt, M. Goudbeek, E. Krahmer. Factors causing overspecification in definite descriptions. *Journal of Pragmatics* **43**, p. 3231 - 3250.

Newnham, 1971. R. Newnham. About Chinese. Harmondsworth, Middlesex: Penguin Books.

van Deemter et al., 2012. K. van Deemter, A. Gatt, I. van der Sluis, and R. Power. Generation of Referring Expressions: Assessing the Incremental Algorithm. Cognitive Science, **36** (5): p. 799 - 836.

van Deemter, 2016. K. van Deemter. Computational Models of Referring: a Study in Cognitive Science. MIT Press, Cambridge Mass.

Westerbeek et al. 2015. H. Westerbeek, R. Koolen, and A. Maes. Stored object knowledge and the production of referring expressions: The case of color typicality. *Frontiers in Psychology* **6**.