Definiteness: Towards a Global Perspective*

Kees van Deemter
University of Aberdeen
k.vdeemter@abdn.ac.uk

Le Sun
Chinese Academy of Sciences
sunle@iscas.ac.cn

Rint Sybesma
Leiden University
R.P.E.Sybesma@hum.leidenuniv.nl

Abstract

In this paper, we outline a plan for investigating the expression of definiteness in East-Asian languages, by means of a series of language elicitation and interpretation experiments followed by computational modelling. Next, we report on the first experiment in this series, in which we elicitated referring expressions in Mandarin. Our approach in this experiment is similar to earlier experiments in the TUNA tradition (which focussed on English and later on Dutch and Arabic), but unlike these earlier experiments, our focus here is on language use, and especially the use of function words, rather than on attribute choice (i.e., the choice of semantic properties to be expressed in the referring expression).

1 Definiteness: the standard view

Many theorists believe that the distinction between definiteness and indefiniteness lies at the heart of communication (Lewis 1969, Kamp and Reyle 1993). Though different views exist concerning the proper treatment of definiteness (with different roles for uniqueness and familiarity, see e.g. Jenks (2015), and with different perspectives on context, salience, and presupposition) theorists tend to believe that whereas indefinite NPs introduce new entities into the Common Ground (in the sense of Clark and Marshall 1981) of the interlocutors, definite NPs rely on previously shared information to identify uniquely particular entities within that Common Ground, a process sometimes known as *information sharing* (van Deemter 2016, Chapter 1). If these theories are correct, then it is important that speakers can mark an NP as definite or as indefinite, and that listeners can tell whether an NP is definite or not.

In English, it is thought that definiteness is usually marked by devices such as the definite determiner and the genitive, and indefiniteness by devices such as the indefinite determiner. Examples exist of English NPs that appear to be marked as definite yet express neither uniqueness nor familiarity; for example, “My daughter is a GP” may not imply that the speaker has only one daughter. However, such examples have often been put aside as either somehow exceptional, or as involving some highly qualified notion of uniqueness after all.[1]

*Presented at the ESSLLI 2017 workshop QUAD: Quantifiers and Determiners (Retoré, Steedman). This work is partly supported by the National Natural Science Foundation of China under Grant no. 61433015.

[1] See Egli and von Heusinger (1995) and Ludlow and Segal (2004) for departures from this view.

2 Definiteness around the world

This story may be plausible for Germanic and Romance languages, but definiteness works differently elsewhere, including some Slavonic languages, and especially East Asian ones (e.g. Huang 1984, Chierchia 1998, Lyons 1999). Cheng and Sybesma (1999, 2015) propose that definiteness in Mandarin can be expressed in three different ways, namely (1) Demonstrative + Classifier + Noun, (2) Demonstrative + Noun, and (3) (bare) Noun. However, bare Nouns (the third pattern) can be indefinite and generic as well as definite, and this raises this question: When a listener hears a bare Noun in Mandarin, how can she tell whether the speaker is identifying (uniquely) an entity in Common Ground? Possible answers include: (a) She cannot, therefore the utterance is underspecified in a crucial way, or (b) Speakers employ other, additional methods for marking their NPs as definite, such as sentence position, situational context, or intonation; there are indications, for instance, that sentence position might be relevant because, in Mandarin, the Subject and other preverbal positions favour definiteness (Chao 1968).

These questions go to the heart of a time-honoured discussion, in which authors in both East and West have argued that East-Asian languages are more ambiguous than the languages of the West, whereas others have argued against this position.[2]

3 An experimental/algorithmic approach

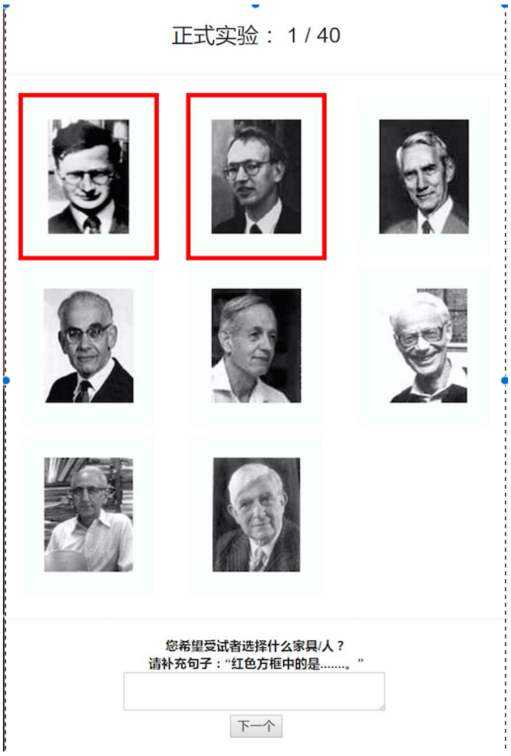
In Computational Linguistics there exists a well-established research tradition in which computational models of human speaking are tested quantitatively against specially designed data-text corpora which result from controlled elicitation experiments (Van Deemter 2016). We believe that this method lends itself well to addressing the questions of Section 2. We are therefore embarking on a new series of experiments, whose aim it is to find out how definiteness is expressed and understood in a given language. Our initial focus is on Mandarin, with plans for Cantonese (and possibly Tibetan) in the making. A small pilot experiment with 10 speakers of Mandarin suggested that this is fertile ground, with speaker-participants employing a wide variety of referential strategies. Next, we embarked on an experiment in which we addressed the following questions:

Are current theories correct in asserting that each of the three linguistic patterns of definiteness in Mandarin is used in situations where the speaker seeks to identify an entity for a hearer? Is this only true for sentence positions where definiteness is the norm (in Mandarin, this is the Subject, or more precise the pre-verbal position), or is it equally true in other positions? Is it only true when referents have been introduced textually, or also when they have been introduced through observation (e.g., of a picture), which might favour the use of a demonstrative?

As a result of our elicitation experiment (section 4), we now have a “data-text” corpus in which each of the elicited NPs is coupled with a stimulus. NPs will be annotated with relevant information (e.g., which syntactic pattern does the NP use?), after which the corpus will be analysed quantitatively. In a later interpretation experiment, we hope to expose hearer-participants to the utterances produced by speaker-participants, and to query them about their understanding of these utterances (e.g., for what proportion of descriptions in the corpus are hearers able to tell correctly whether the situation described contains 1 or 2 referents that meet the description?).

Combined, these elicitation experiments (with speakers) and interpretation experiments (with hearers) will teach us a lot about the way in which definiteness is expressed and under-

[2] Compare Lam (2010) for an analogous discussion on a different linguistic topic (i.e., presuppositions).



正式实验： 1 / 40

您希望受试者选择什么家具/人？
请补充句子："红色方框中的是……。"

Figure 1: *A people trial, with the referring expression in object position.*

stood; once a number of languages have been addressed, our understanding of the differences between them should be much improved. The Computational Modelling paradigm has not been applied to questions of this kind before, since it has tended to focus exclusively on the question of what attributes (e.g. colour, size, etc.) are expressed in a given NP.

4 A Data-Text Corpus of Referring Expressions in Mandarin

The setup of the Mtuna experiment resembles the original TUNA experiment (e.g., van Deemter et al. 2012) and its later incarnations (Gatt and Belz 2010, Koolen et al. 2011, Khan 2015), with 22 stimuli depicting configurations of furniture and 22 stimuli depicting people. However, the annotation of the corpus had to be very different this time. For whereas earlier TUNAs focussed on the properties expressed by a referring expression (chair, green, etc.), our research questions mean that function words (English: *the, a, one, two, this, those, both*) are key. Stimuli include references to sets as well as individual items. Instructions to participants were adapted from Koolen et al. (2011). Importantly, the new experiment distinguished between referring expressions in preverbal and post-verbal position. All participants were discouraged from using location in their referring expressions, 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, in which all entities were allotted to cells in a 3-by-5 grid invisible to participants.

Sentence position was varied in a between-subjects design: Participants who were asked to produce referring expressions 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)?) 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 referring expressions in post-verbal position were asked: *Nin xiwang shoushizhe xuanze shenme jia-*

Mandarin (transcribed into pinyin)	Approximate English Gloss
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 expressions found in the corpus, referring to the target referents in Figure 1. [cla] denotes a classifier, [sub] denotes a subordinating *de*. All expressions were automatically transcribed into (phonetic) Pinyin notation without any further modifications.

ju/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 is in the red frame is .....”)

The corpus was subjected to an initial analysis of all descriptions elicited. The distribution of Patterns differred notably from what we had expected. First of all, Bare Nouns dominate, with demonstratives occurring very rarely. Perhaps Demonstratives are restricted to situations in which the antecedent is either pointed at or mentioned in earlier text (as proposed by Jenks 2015); this might explain the scarcity of demonstratives. Moreover, differing from our prediction, indefinite NPs occurred quite often, even in pre-verbal position, where we had expected not to see them. Perhaps some participants had an unintended understanding of their task, which may have led them to describe objects without thinking particularly about identifying them for a reader (i.e., without actually referring); this might explain their use of indefinite NPs.

Based on a first look at our data, where numerals were remarkably frequent, a nuanced picture is starting to emerge, where Mandarin may be less fully specified than English with respect to singulars, yet *more* fully specified (e.g., for number) with respect to plurals. Furthermore, 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.

5 Discussion

It has been suggested that East Asian languages handle the trade-off between brevity and clarity differently to those of Western Europe, (e.g., Newnham 1971, Huang 1984), with the former allegedly leaning more towards brevity, and relying more on communicative context for disambiguation. While our preliminary findings confirm these ideas to some extent, the more important part of this issue is still unresolved. For example, if these ideas are correct, one might expect to see in Mandarin referring expressions less over-specification (where one or more properties can be removed from the referring expression without causing confusion) and more under-specification than in English and Dutch, for example. In future, we want to investigate these ideas and their implications for the computational modelling of reference.

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, T. Kiss and A. Alexiadou (Eds.), 42.1-3, Berlin: Mouton de Gruyter.

Chierchia, 1989. G. Chierchia. Reference to Kinds Across Languages. *Natural Language Semantics* **6**, p.339 – 405.

Clark and Marshall, 1981. H. H. Clark and C. R. Marshall (1981). Definite reference and mutual knowledge. In Joshi, A. K., Sag, I. A., and Webber, B. L., (Eds), *Elements of Discourse Understanding*, p.10 – 63. Cambridge University Press, Cambridge, UK.

Gatt and Belz, 2010. A. Gatt, A. and A. Belz (2010). 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.

Egli and von Heusinger, 1995. U.Egli and K. von Heusinger. The epsilon operator and E-type pronouns. In Egli, Pause, Schwarze, von Stechow and Wienold (Eds), Lexical Knowledge in the Organization of Language, p.121–141.

Huang. C. -T. J. Huang 1984. 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 Poceedings of Semantics and Linguistic Theory (SALT) 25, Stanford (Ca.).

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

Lam, 2010. B. Lam. Are Cantonese-speakers really descriptivists? Revisiting cross-cultural semantics. *Cognition*, **115** (2), 320329.

Lewis, 1969. D. Lewis. *Convention: A Philosophical Study.* Harvard University Press, Cambridge, Mass.

Ludlow and Segal, 2004. P. Ludlow and G. Segal. On a unitary semantical analysis for definite and indefinite descriptions. In Reimer and Bezuidenhout (Eds) *Descriptions and Beyond*, p. 420 – 436. Oxford University Press, Oxford, UK.

Lyons 1999. Ch. Lyons. *Definiteness.* Cambridge University Press, Cambridge, UK.

van Deemter et al., 2012. K. van Deemter, K., A. Gatt, I. van der Sluis, and R. Power. Generation of referring expressions: Assessing the incremental algorithm. *Cognitive Science*, **36** (5): 799836.

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