Translation Mining: Definiteness across Languages (A Reply to Jenks 2018)

David Bremmers
Jianan Liu
Martijn van der Klis
Bert Le Bruyn

We present a parallel corpus study that compares the distribution of German contracted/uncontracted articles and Mandarin bare nouns/demonstratives. Work by Schwarz (2009) and Jenks (2018) leads us to predict that German contracted articles pattern with Mandarin bare nouns and German uncontracted articles with Mandarin demonstratives. We show that these predictions are only partly borne out and argue for a more fine-grained typology of definiteness.

Keywords: semantics, definiteness, strong/weak distinction, German, Mandarin

The distinction between weak and strong definiteness was originally proposed by Schwarz (2009) to account for the difference between German contracted and uncontracted definite articles. Jenks (2018) extends it to Mandarin, linking bare nouns to weak definiteness and demonstratives to strong definiteness.

We present a parallel-corpus study that compares the distribution of German contracted/uncontracted articles and Mandarin bare nouns/demonstratives. The work by Schwarz and Jenks leads us to predict that German contracted articles pattern with Mandarin bare nouns and German uncontracted articles with Mandarin demonstratives. We show that these predictions are only partly borne out and argue for a more fine-grained typology of definiteness and of strong definiteness in particular.

The article is structured as follows. After giving relevant background on weak and strong definiteness (section 1), we present our parallel-corpus study (sections 2 and 3) and argue for a new dimension in the typology of definiteness (section 4). We conclude with a brief summary of the main findings (section 5).

1 Weak and Strong Definites across Languages: Setting the Stage

Here, we give a brief overview of Schwarz 2009 and Jenks 2018, distinguishing between data (section 1.1) and analysis (section 1.2). We conclude with a summary and outlook (section 1.3).

We wish to thank our two *LI* reviewers and the members of the Time in Translation project for very constructive and valuable feedback. We also thank our Mandarin consultants: Xiaoli Dong, Shuangshuang Hu, Siyang Kong, Zhenghao Li, Chou Mo, and Feifei Zhao. Financial support from NWO is gratefully acknowledged (grant 360-80-070). The final version of the article is the outcome of discussions among all coauthors. Initial responsibilities were as follows. German data collection and analysis: David; Mandarin data collection and analysis: Jianan; design of the Translation Mining interface and implementation: Martijn; concept and writing: Bert.

Linguistic Inquiry, Volume 53, Number 4, Fall 2022
735–752

https://doi.org/10.1162/ling_a_00423

1.1 Data

Schwarz (2009) brings together insights from Hawkins's (1978) seminal work on English definites and from a rich descriptive tradition on definites in other Germanic languages and dialects (e.g., Ebert 1971a,b). Schwarz's main data are taken from Standard German (see Wiltschko 2013 on Austro-Bavarian German and Meier 2019 on Zurich German).

What sets Standard German apart from English is that it has two forms for the definite article: one rendering uniqueness—or "weak" definiteness—and one rendering familiarity—or "strong" definiteness. Weak/Uniqueness definites are primarily the immediate- and larger-situation definites in Hawkins's typology, and strong/familiarity definites predominantly correspond to Hawkins's anaphoric definites. As for the associative (or bridging) uses Hawkins discusses, Schwarz argues that some qualify as strong and others as weak.

The weak vs. strong distinction in Standard German manifests itself formally in that weak definite articles contract with certain prepositions, whereas strong definite articles resist contraction. Outside the prepositional domain, no formal difference can be detected. Key examples from Schwarz 2009 are given in (1) and (2).

(1) Der Empfang wurde **vom / #von dem Bürgermeister** eröffnet.
the reception was by.the / by the mayor opened
'The reception was opened **by the mayor**.'
(Schwarz 2009:40)

(2) In der New Yorker Bibliothek gibt es ein Buch über Topinambur. Neulich
in the New York library exists there a book about topinambur recently
war ich dort und habe **#im / in dem Buch** nach einer Antwort auf die Frage
was I there and have in.the / in the book for an answer to the question
gesucht, ob man Topinambur grillen kann.
searched if one topinambur grill can
'In the New York Public Library, there is a book about topinambur. Recently, I was there and searched **in the book** for an answer to the question of whether one can grill topinambur.'
(Schwarz 2009:30)

In (1), the mayor has not been introduced before but is the unique mayor of the contextually salient town. This is a weak/uniqueness context, and the definite article contracts with the preposition. In (2), a book is introduced in the first sentence and referred back to in the second. This is a case of strong/familiarity definiteness, and contraction is not allowed.

Several studies have followed up on Schwarz 2009 and have argued that the weak vs. strong distinction underlies definiteness paradigms in typologically diverse languages (see, e.g., Arkoh and Matthewson 2013 for an extension to Akan, and Aguilar-Guevara, Pozas Loyo, and Vázquez-Rojas Maldonado 2019 for an overview). We focus on the case of Mandarin as presented by Jenks (2018).

Jenks provides the following key examples:

(3) (**#Nà / #Zhè ge**) **táiwān (de) zǒngtǒng** hěn shēngqì.
that / this CLF Taiwan ('s) president very angry
'**The president of Taiwan** is very angry.'
(Jenks 2018:507)

(4) Jiàoshì lǐ zuò-zhe yī gè nánshēng hé yī gè nǚshēng. Wǒ zuótiān yùdào
classroom in sit-ASP one CLF boy and one CLF girl I yesterday meet
#(**nà gè**) **nánshēng**.
that CLF boy
'There are a boy and a girl sitting in the classroom. I met **the boy** yesterday.'
(Jenks 2018:510)

(3) shows that nouns referring to unique individuals like the president of Taiwan typically occur bare and that demonstratives are not allowed with them. (4) shows that the bare noun *nánshēng* cannot refer back to the boy in the first sentence and that the demonstrative is required. These facts build a strong case in favor of an active weak/strong distinction for Mandarin bare nouns and demonstratives, parallel to what we find with contracted and uncontracted definite articles in Standard German.

1.2 Analysis

The basics of an analysis of weak and strong definites go back to the semantics Schwarz (2009) proposes for each of them.

(5) *The weak/strong distinction in Schwarz 2009*
a. Weak definite: $\lambda s_r.\lambda P{:}\exists!x\ (P(x)(s_r)).\ \iota x\ [P(x)(s_r)]$
b. Strong definite: $\lambda s_r.\lambda P.\lambda y{:}\exists!x\ (P(x)(s_r)\&x{=}y).\ \iota x\ [P(x)(s_r)\&x{=}y]$

Both weak and strong definites are linked to a pragmatically supplied resource situation formalized as the situation pronoun s_r in which the referent is unique. This uniqueness is spelled out in both the presuppositional and the asserted content. Strong definites are special in that they come with a pragmatically supplied index y that the referent of the definite is said to be identical to. This formalizes anaphoricity.

Schwarz assumes the situation pronoun s_r can stand for a contextually salient situation but can also be identified with the (Austinian) topic situation or be bound by a quantifier over situations. In the remainder of this article, we will focus on examples in which s_r is identified with the topic situation—that is, the situation an utterance is about. We follow McKenzie (2012, 2015) in assuming that a situation can consist of multiple eventualities as long as a coherent relation can be established between them. In line with McKenzie's observations on Kiowa, we take spatiotemporal contiguity to be a good predictor for which eventualities can be considered to belong to the same situation.

Jenks's analysis of the weak/strong distinction builds on Schwarz's. We compare Jenks's entries in (6) with Schwarz's in (5).

(6) *The weak/strong distinction in Jenks 2018*
a. Weak definite: $\lambda s_r.\lambda P{:}\exists!x\ (P(x)(s_r)).\ \iota x\ [P(x)(s_r)]$
b. Strong definite: $\lambda s_r.\lambda P.\lambda Q{:}\exists!x\ (P(x)(s_r)\&Q(x)).\ \iota x\ [P(x)(s_r)]$

There are two differences between the entries in (5) and (6). The first is minor and is concerned with the semantic type of the index argument in the strong definite: in Schwarz's analysis, the index argument y is of type e whereas in Jenks's analysis, the index argument Q is of type $\langle e,t\rangle$. This move is motivated under the assumption that the index occupies a predicative position in Mandarin (Zhang 2015).

The second difference is more fundamental. Whereas in (5b) the index argument is active in both the presupposition and the assertion, in (6b) it only appears in the presupposition. The main consequence of this move is that the weak and strong definites become identical in their assertive content while a stronger presupposition is retained for the strong definite. Under Maximize Presupposition (Heim 1991), this means that the strong definite has to be used as soon as its presuppositions are met. Jenks (2018:524) formalizes this in a principle he terms *Index!*.

(7) *Index!*
Represent and bind all possible indices.

Index! requires the use of strong definites as soon as an anaphoric relation can be established, effectively blocking the use of weak definites in anaphoric contexts. With Index! in place, Jenks's analysis is more restrictive than Schwarz's. Whereas Schwarz assumes the difference in acceptability between the contracted and uncontracted forms in (2) is a matter of preference, Jenks hardwires the difference into his analysis.

1.3 Summary and Outlook

Schwarz (2009) and Jenks (2018) define two types of definiteness and argue that (Standard) German and Mandarin formally distinguish between them. Weak definiteness is concerned with uniqueness and is marked by contracted definites in German and bare nouns in Mandarin. Strong definiteness adds dynamicity by requiring referents to be familiar from previous discourse. Strong definiteness is marked with uncontracted definites in German, whereas Mandarin relies on demonstratives. At the level of analysis, the main difference between Schwarz's and Jenks's analyses lies in the way they deal with the competition between the two types of definiteness. Whereas Schwarz leaves open how the competition should play out, Jenks's constraint Index! categorically bans the use of markers of weak definiteness for previously introduced referents.

In sections 2 and 3, we present a parallel-corpus study that puts the predictions of Schwarz's and Jenks's analyses to the test. Parallel—or translation—corpora render the same content in different languages. They are thus particularly suited for checking whether a semantic category that is active in one language is replicated in another. For weak and strong definiteness, the main prediction that follows from Schwarz's and Jenks's work combined is that German contracted

definites pattern with Mandarin bare nouns and that German uncontracted definites pattern with Mandarin demonstratives.[1]

The parallel-corpus study we will present reveals that this prediction is only partly borne out. We will argue that the main problem lies with Mandarin bare nouns, and we will develop a more fine-grained typology of strong definiteness. Furthermore, we will argue that the predictions Index! makes are too strict, even for German.

2 Corpus and Methodology

As indicated above, the main prediction that follows from Schwarz's and Jenks's work combined is that German contracted definites pattern with Mandarin bare nouns and that German uncontracted definites pattern with Mandarin demonstratives. Parallel corpora allow us to test this prediction. We present our corpus and methodology in sections 2.1 and 2.2, respectively, and provide results and discussion in section 3.

2.1 Corpus

The corpus we selected is the first volume of the Harry Potter series and its translations into German and Mandarin (see bibliographic details following the reference list). We chose this corpus for a number of reasons. We highlight two. First, we want this research to lead to a broader exploration of article systems in languages that are traditionally considered articleless. To do so, we need a corpus that can easily be extended with more languages. The availability of the Harry Potter series in multiple typologically diverse languages provides exactly that. Second, we wanted to have a source language that does not formally distinguish between weak and strong definiteness. This guarantees that the translator is not biased by a formal distinction in the source text and focuses on rendering the meaning in the best way possible. For this reason, we decided to study a corpus with an English source text rather than one with a Mandarin or a German source text.

2.2 Methodology

2.2.1 Data Selection We selected our data on the basis of the German translation. We did this because German is known to mark the weak/strong definiteness distinction in a restricted domain—namely, prepositional phrases (PPs). Starting from Mandarin bare nouns and demonstratives or starting from English definites would have led to a high number of data points with no relevant comparative material in German.

The goal of our selection procedure was to end up with a dataset with a more or less even distribution of contracted and uncontracted PPs, at the same time maximizing the likelihood of including minimal pairs. To do so, we extracted all PPs with a definite article, divided them into contracted and uncontracted ones, and then selected those that contained prepositions that appeared

[1] Strictly speaking, Schwarz does not make claims about Mandarin and Jenks does not make claims about German. This is why the crosslinguistic prediction we look into is attributed to Schwarz's and Jenks's work combined.

in both lists. A further selection was done for contracted PPs, as these greatly outnumbered uncontracted PPs. In particular, we included all contracted PPs from the first three chapters and restricted the contracted PPs from the other chapters to those that had uncontracted counterparts in the novel—that is, uncontracted PPs involving the same preposition and noun.[2] Our selection procedure gave rise to a total of 96 data points, including 40 contracted and 56 uncontracted PPs.

2.2.2 Data Processing Once the set of German PPs was established, we aligned them with the English original and the Mandarin translation. We also annotated all data points for the forms that were used in the three languages. After annotation, each of the 96 data points could be characterized as a triple ⟨German form, English form, Mandarin form⟩ (e.g., ⟨contracted, definite, bare⟩). Alignment and annotation were done by two of the authors, one a native speaker of German, the other a native speaker of Mandarin. All data processing was done in a custom-made online interface that links data and annotation to a number of parallel-corpus analysis tools.

2.2.3 Analysis We calculated basic descriptive statistics for all data points, including the frequency of forms per language and the frequency of the correspondences between German and Mandarin. Given that the number of data points and languages is limited, this could have sufficed. However, as indicated above, we hope this research will lead to a broader exploration of articles in languages that are traditionally considered articleless. Given that correspondences between more than two languages become difficult to process with basic descriptive statistics, we need another type of analysis that can help us do so. One such family of analyses is known as *proximity maps* (see Georgakopoulos and Polis 2018 for discussion). These have gained traction in the typological literature and have been shown to hold promise for crosslinguistic work at the syntax-semantics interface as well (see van der Klis, Le Bruyn, and de Swart 2020). We introduce them in this article as a proof of concept.

We rely on a specific implementation of proximity maps known as *probabilistic semantic maps* (Wälchli and Cysouw 2012). They are a powerful tool for analyzing the use of language-specific forms across data points drawn from a parallel corpus. Here, we provide an intuitive explanation of how probabilistic semantic maps are built and how they can be interpreted.[3] The example maps we treat are based on our data and will recur in section 3.

2.2.4 Probabilistic Semantic Maps Probabilistic semantic maps are generated through Multidimensional Scaling (MDS), a dimensionality reduction algorithm. Each data point is represented by a dot in a two-dimensional space, as in figure 1. All other things being equal, data points are closer to one another if they use the same form in a given language. The more forms that correspond between two data points, the closer they are. We illustrate with the triples in (8).

[2] The automated scripts we used for extraction and selection are available at https://github.com/time-in-translation/conll-extractor and https://github.com/time-in-translation/conll-extractor/blob/master/conll_extractor/prepositions/data.py.

[3] For technical details, see Wälchli and Cysouw 2012, van der Klis, Le Bruyn, and de Swart 2017, van der Klis and Tellings 2020.

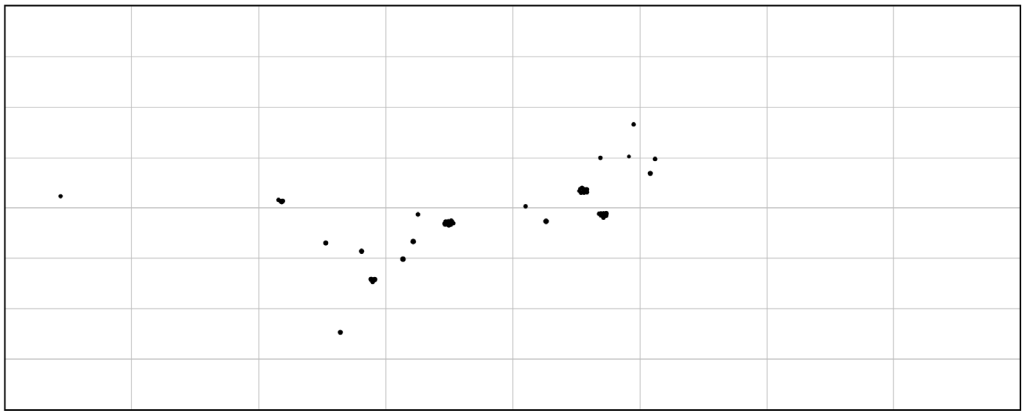

Figure 1
A probabilistic semantic map

(8) *Examples of triples*
a. ⟨contracted, definite, bare⟩
b. ⟨contracted, definite, bare⟩
c. ⟨uncontracted, definite, bare⟩
d. ⟨uncontracted, definite, demonstrative⟩

All other things being equal, a dot corresponding to a data point like (8a) will be closer to one corresponding to (8b) than to one corresponding to (8c): (8a) and (8b) share all forms, whereas they differ from (8c) in their first position. At the same time, a dot corresponding to a data point like (8d) will be even farther away from (8a)/(8b) than (8c) is, because it differs from (8a)/(8b) in even more forms.

Probabilistic semantic maps that are faithful to all distances between data points are rare. This is because they are limited to two dimensions. MDS can, however, be run with as many dimensions as we like. Dimensions will try to be faithful to the distances between as many data points as possible, but they will progressively also try to be faithful to distances that earlier dimensions were not yet faithful to. This allows us to choose the dimensions that best allow us to study the correspondences (or lack thereof) between forms of different languages.

By default, we run MDS with five dimensions. In figure 1, we have represented the first two. Figure 2 shows how we can use these to visualize correspondences between forms. Each shaded cluster in figure 2 represents the distribution of a form across the data points in the corpus. For convenience, we have limited ourselves to two forms from two languages: the clusters with dotted lines represent the two forms from language A; the clusters with solid lines represent the two forms from language B.

If there had been a clear correspondence between forms in languages A and B, we would have expected the clusters from the two languages to resemble each other. This is clearly not

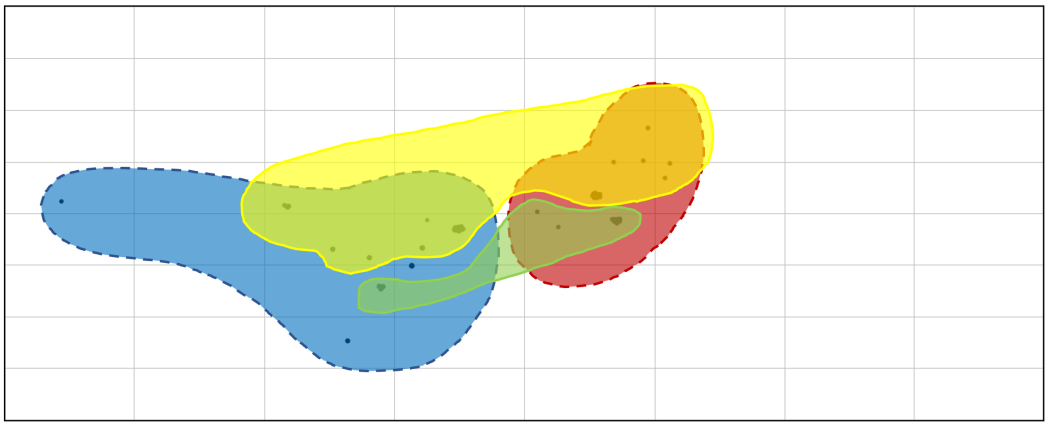

Figure 2
A probabilistic semantic map with language-specific form-based clusters

what happens in figure 2: rather than revealing a similar distribution, the map suggests that the distribution of the forms in the two languages is orthogonal.

3 Results and Discussion

3.1 Results

3.1.1 Descriptive Statistics We indicated above that the German dataset consists of 40 (41.5%) contracted and 56 (58.5%) uncontracted cases. For English and Mandarin, we report on the forms that appeared at least three times as counterparts of one of these. For English, these are the definite (n=80, 83%), the bare singular (n=5, 5%), and the demonstrative (n=4, 4%). For Mandarin, they are the bare noun (n=79, 82%) and the demonstrative (n=13, 13.5%).

As for the correspondences between German and Mandarin, we also restrict ourselves to those forms that appear at least three times as counterparts. Among the 40 German contracted forms, 3 (7.5%) correspond to demonstratives in Mandarin and 34 (85%) correspond to bare nouns. Among the 56 German uncontracted forms, 10 (18%) correspond to demonstratives in Mandarin and 45 (80.5%) correspond to bare nouns.

3.1.2 A Probabilistic Semantic Map Even though the descriptive statistics already give a good idea of the data, probabilistic semantic maps allow us to visualize them in a format that is easier to process, in particular when more languages are added. Figure 3 is identical to figure 2, but the language-specific form-based clusters have now been identified. Crucially, we find that Mandarin bare nouns are the counterparts of both contracted and uncontracted cases in German and that the same holds for Mandarin demonstratives. We thus find that the division of labor between bare nouns and demonstratives is not parallel but orthogonal to the one between contracted and uncontracted definites in German.

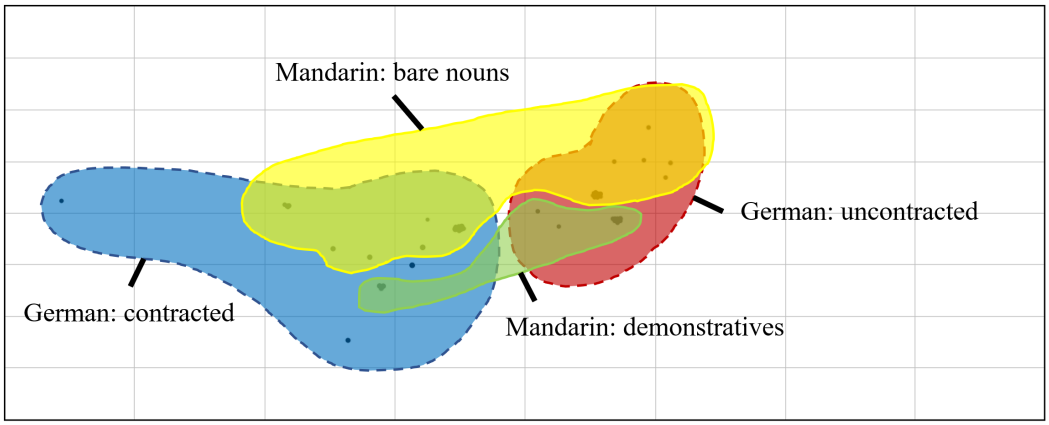

Figure 3
A probabilistic semantic map with the main form-based clusters for German and Mandarin

3.2 Discussion

The data strongly suggest that the main prediction that stems from Schwarz's and Jenks's work combined is not borne out. The one-to-one mappings we expect between German contracted definites and Mandarin bare nouns on the one hand, and between German uncontracted definites and Mandarin demonstratives on the other hand, are not there. A more fine-grained discussion of the data is required, though. We start with the German data (section 3.2.1) and then move to Mandarin demonstratives (section 3.2.2) and bare nouns (section 3.2.3). We end with a summary (section 3.2.4).

3.2.1 German: Confirmation of Schwarz's Analysis and Problems for Index! The German data show that unique referents are contracted and that familiar referents are uncontracted. This is in line with the basic predictions of Schwarz's analysis.

(9) **E:** "I suppose we could take him to the zoo," said Aunt Petunia slowly, " . . . and leave him **in the car** . . . "
G: "Ich denke, wir könnten ihn in den Zoo mitnehmen," sagte Tante Petunia
I think we could him to the zoo take said Aunt Petunia
langsam, " . . . und ihn **im** **Wagen** lassen . . . "
slowly and him in.the car leave

(10) [Context: As the owls flooded into the Great Hall as usual, everyone's attention was caught at once by a long thin package carried by six large screech owls. Harry was just as interested as everyone else to see what was in this large parcel and was amazed when the owls soared down and dropped it right in front of him, knocking his bacon to the floor.]

E: They had hardly fluttered out of the way when another owl dropped a letter **on top of the parcel**.

G: Sie waren kaum aus dem Weg geflattert, als eine andere Eule einen Brief
they were hardly out the way fluttered when a other owl a letter
auf das Paket warf.
on the parcel threw

The car in (9) does not refer back to a previously introduced car; rather, it refers to the unique family car. It consequently counts as a weak definite. *The parcel* in (10) refers back to the package that was introduced before and therefore counts as a strong definite. As Schwarz's analysis predicts, German relies on a contracted definite in (9) and an uncontracted definite in (10).

There are two types of contexts that deserve special mention. Both combine a dimension of uniqueness with a dimension of familiarity. They thus count as in-between cases, and we expect there to be some variation in the markers that appear. The first type is concerned with reference to a familiar but one-of-a-kind stone known as *the Philosopher's Stone.* The second type involves bridging. We find that both contracted and uncontracted definites can be used to refer to the Philosopher's Stone and that bridging is equally variable.

(11) **E:** "I'm going out of here tonight and I'm going to try and get **to the Stone** first."
G: "Ich gehe heute Nacht raus und versuche als Erster **zum Stein** zu kommen."
I go today night out and try as first to.the stone to get

(12) **E:** "How do you think you'd get **to the Stone** without us?"
G: "Wie glaubst du eigentlich, dass du ohne uns **zu dem Stein** kommst?"
how believe you actually that you without us to the stone get

(13) [Context: "OUT!" roared Uncle Vernon, and he took both Harry and Dudley by the scruffs of their necks and threw them into the hall, slamming the kitchen door behind them.]
E: Harry and Dudley promptly had a furious but silent fight over who would listen **at the keyhole**.
G: Prompt lieferten sich Harry und Dudley einen erbitterten, aber stummen
promptly gave themselves Harry and Dudley a furious but silent
Kampf darum, wer **am Schlüsselloch** lauschen durfte.
fight about who at.the keyhole listen could

(14) [Context: Ducking under Peeves they ran for their lives, right to the end of the corridor, where they slammed into a door—and it was locked.]
E: "Oh, move over," Hermione snarled. She grabbed Harry's wand, tapped **the lock** and whispered, "Alohomora!"
G: "Ach, geh mal beiseite," fauchte Hermine. Sie packte Harrys Zauberstab,
oh go once away snarled Hermione she took Harry's wand
klopfte **auf das Türschloss** und flüsterte: "Alohomora!"
tapped on the doorlock and whispered alohomora

The Philosopher's Stone is referred to with a contracted definite in (11) and with an uncontracted definite in (12). Neither of them counts as the first reference to the stone. In (13) and (14), *Schlüsselloch* and *Türschloss* refer to the unique lock of a previously introduced door, but the former appears with a contracted definite whereas the latter combines with an uncontracted definite.

The data in (11)–(14) show that as soon as a dimension of uniqueness is combined with a dimension of familiarity, both contracted and uncontracted definites become available. This is compatible with Schwarz's basic analysis but undermines the validity of Jenks's constraint Index!. This constraint predicts that the use of weak definites is proscribed as soon as familiarity comes into play. The data in (11) and (13) show that this prediction is not borne out.[4] We assume that the choice between contracted and uncontracted definites in (11)–(14) is not free, but that the constraints at work go beyond the basic distinction between weak and strong definiteness. For a discussion of some of the factors involved, see Aguilar-Guevara and Zwarts 2010.

We conclude that the German data are in line with the predictions Schwarz (2009) makes and that they argue against the stricter competition between weak and strong definites that follows from Jenks's constraint Index!.

3.2.2 Mandarin: The Case of Demonstratives With German following Schwarz's predictions, we turn to the Mandarin data to understand the orthogonality between German contracted/uncontracted definites and Mandarin bare nouns/demonstratives. In this section, we focus on Mandarin demonstratives.

Our Mandarin dataset contains 13 demonstratives. The vast majority of them (n=10, 77%) appear in contexts that take an uncontracted definite in German and thus behave the way Jenks predicts; that is, they are used in strong definiteness contexts. We argue that the three remaining demonstratives are not to be considered counterexamples to Jenks's predictions. (15) is a representative example.

(15) **E:** "I'm not having one **in the house**, Petunia!"
M: "Pèinī, wǒ juébù ràng tāmen rènhérén jìn **zhè dòng fángzi**."
Petunia I not have them anyone enter this CLF house

(15) is uttered by a husband who assures his wife that certain people will never be welcome in their house. One can argue that the demonstrative is used to refer to the unique family home, but a more plausible analysis is that the demonstrative retains its full deictic force and refers to the house the speaker and listener are in at the moment of speech. Our consultants confirm that the latter analysis is the one that corresponds best to their intuitions. This extends to the two other cases of demonstratives appearing as counterparts of German contracted definites.

We conclude that the Mandarin demonstratives in our corpus mark strong and not weak definiteness. This is in line with Jenks's analysis. If demonstratives do appear as counterparts of

[4] Data like those in (13) and (14) are also interesting for a discussion of the more involved claims Schwarz makes about the relation between weak and strong definites in different types of bridging. A reviewer notes that part of the explanation might lie in the fact that *door* is part of *Türschloss* (lit. 'door lock') but not of *Schlüsselloch* (lit. 'keyhole'). A similar effect of compounding is mentioned by Schwarz (2009:283).

contracted definites, they retain their full deictic force and receive a slightly different interpretation from the one conveyed by the German translation.[5]

3.2.3 Mandarin: The Case of Bare Nouns Having argued that the few Mandarin demonstratives that do not follow Jenks's predictions can be considered special cases and do not challenge his analysis, we now turn to Mandarin bare nouns.

Our Mandarin dataset contains 79 bare nouns. They are the majority option, both in contexts in which German uses contracted definites (34 (85%) are rendered as bare nouns) and in contexts in which German uses uncontracted definites (45 (80.5%) are rendered as bare nouns). The use of bare nouns in the former contexts is in line with Jenks's predictions, but their use in the latter poses a serious challenge. Examples like (16) and (17) show that this challenge is real.

(16) [Context: As the owls flooded into the Great Hall as usual, everyone's attention was caught at once by a long thin package carried by six large screech owls. Harry was just as interested as everyone else to see what was in this large parcel and was amazed when the owls soared down and dropped it right in front of him, knocking his bacon to the floor.]

M: Tāmen pūshan-zhe chìbǎng gānggāng fēi zǒu, yòu yǒu yī zhǐ māotóuyīng
they flutter-ASP wings right fly away and have one CLF owl
xié lái yī fēng xìn, rēng zài **bāoguǒ** shàngmiàn.
bring come one CLF letter throw to parcel on
'They had hardly fluttered out of the way when another owl dropped a letter on top of the parcel.'

(17) [Context: Dudley quickly found the largest snake in the place. It could have wrapped its body twice around Uncle Vernon's car and crushed it into a dustbin—but at the moment it didn't look in the mood. In fact, it was fast asleep.]

E: He looked back **at the snake** and winked, too.
G: Er drehte sich wieder **zu der Schlange** um und zwinkerte zurück.
he turned himself again to the snake around and winked back
M: Tā huí-guò tóu lái kàn-zhe **jù mǎng**, yě duì tā zhǎ-le-zhǎ yǎn.
he back-ASP head to stare-ASP huge snake too to it wink-ASP-wink eye

(16) is the Mandarin version of (10). In (16) and (17), reference is made, respectively, to a parcel and a snake that were introduced before. This anaphoric reading is also the reading our consultants find most natural. (16) and (17) thus unambiguously show that bare nouns can be used in strong definiteness contexts.

We conclude that the Mandarin bare nouns in our corpus mark both weak and strong definiteness. These data are incompatible with Jenks's analysis of the weak/strong opposition in Mandarin.

[5] Jenks (2018) proposes a unified analysis of demonstratives that covers both deictic and strong definite uses. We remain neutral as to whether a unified or an ambiguity analysis should be pursued.

3.2.4 Summary Our corpus data show that there is no one-to-one correspondence between German contracted/uncontracted definites and Mandarin bare nouns/demonstratives. This runs counter to the predictions that stem from Schwarz's and Jenks's work combined. A closer analysis of the data reveals that German contracted and uncontracted definites, as well as Mandarin demonstratives, by and large behave as expected (sections 3.2.1–3.2.2). The main problem lies with Mandarin bare nouns, as these appear in both weak and strong definiteness contexts (section 3.2.3). A further issue our data raise is that Jenks's constraint Index! is too strong, not only for the Mandarin data but also for the German data. In section 4, we reflect on the acceptability of bare nouns in weak and strong definiteness contexts.

4 Two Types of Strong Definiteness

The results in section 3 show that Mandarin bare nouns appear in weak and strong definiteness contexts alike. Here, we consider how these facts affect our understanding of definiteness. At the heart of our discussion lies a discrepancy between our data and Jenks's: Jenks uses examples like (4) to argue that bare nouns are ungrammatical as anaphors, whereas we find examples like (16) in which bare nouns are perfectly fine as anaphors. We hypothesize that the opposition indicates two types of strong definiteness: text-level and situation-level familiarity. We start, however, by briefly discussing two competing hypotheses.

4.1 Dialectal Variation

An obvious hypothesis about the opposing judgments for (4) and (16) is that they stem from dialectal differences. This could be the case, as our consultants are from mainland China whereas Jenks's are from Taiwan. However, our consultants agree with Jenks's that (4) is unacceptable. The opposition between the unacceptability of the bare noun in (4) and its acceptability (16) is thus real.

4.2 Pragmatic Coreference

Another hypothesis one could entertain is that the anaphoric reading of (16) is pragmatically induced rather than semantically encoded. Under this hypothesis, the fact that the package in (16) is identified with the package that was introduced before is driven by context and not by semantics. Even though this hypothesis could explain why the bare noun in (16) can have an anaphoric reading, it would need to be supplemented with an explanation for the fact that pragmatic coreference is not an option for the bare noun in (4). We do not see what this explanation could look like; instead, we turn to an alternative hypothesis in which anaphoricity is semantically encoded in both (4) and (16).

4.3 Text-Level vs. Situation-Level Familiarity

The hypothesis we argue for is based on a comparison of Jenks's data with our corpus data. The contexts in our corpus typically display a classical narrative style in which events are presented in chronological order. (16) is a good example, chronologically relating the coming and going

of a group of owls followed by an event involving a single owl. (4) is crucially different in the sense that the event referred to in the second sentence chronologically precedes the one in the first. We hypothesize that the discrepancy in judgments about contexts like those in (4) and (16) is related to this difference in narrative structure and indicates a difference between two types of strong definiteness: text-level and situation-level familiarity. We introduce the two types, make explicit how we take Mandarin bare nouns and demonstratives to relate to them, and discuss the predictions our hypothesis makes.

4.3.1 Two Types of Familiarity We start by introducing the two types of familiarity on the basis of the English versions of (4) and (16) (repeated here).

(18) There are a boy and a girl sitting in the classroom. I met **the boy** yesterday. (= (4))

(19) [Context: As the owls flooded into the Great Hall as usual, everyone's attention was caught at once by a long thin package carried by six large screech owls. Harry was just as interested as everyone else to see what was in this large parcel and was amazed when the owls soared down and dropped it right in front of him, knocking his bacon to the floor.]

They had hardly fluttered out of the way when another owl dropped a letter on top of **the parcel**. (= (16))

In (18), *the boy* refers back to the previously introduced boy, and in (19), *the parcel* refers back to the previously introduced long thin package. Both meet the requirement of previous introduction traditionally associated with strong definiteness. In Schwarz's and Jenks's analyses of strong definiteness, this requirement is formalized through an identity relation with a pragmatically supplied index. We refer to this type of strong definiteness as *text-level familiarity.*

Situation-level familiarity is stricter and requires the anaphor to be introduced in the same topic situation as its antecedent. We argue that *the parcel* in (19) meets this requirement but *the boy* in (18) does not. As we indicated in section 1.2, we build on McKenzie's (2012, 2015) work and take spatiotemporal contiguity to be a good indicator of which eventualities can be considered part of a single situation. In (19), the sentences linking *a long thin package* and *the parcel* describe spatiotemporally contiguous eventualities, and they can thus be assumed to be part of a single overarching topic situation. *The parcel* thus meets the requirements of situation-level familiarity. (18) is different: the adverb *yesterday* introduces a clear temporal break between the situations described by the first and second sentences. *The boy* consequently does not meet the requirements of situation-level familiarity.

4.3.2 Bare Nouns, Demonstratives, and Strong Definiteness With the two types of strong definiteness in place, we can present the way we assume Mandarin bare nouns and demonstratives relate to them. We hypothesize that bare nouns can be used for situation-level familiarity but not for text-level familiarity. Demonstratives, on the other hand, can be freely used for both types.

The underlying intuition is that indices are only available in the topic situation in which they have been introduced. The difference between bare nouns and demonstratives lies in the deictic component of the latter: demonstratives are able to refer to situations other than the topic situation

of the sentence they appear in and thus to access indices from other topic situations.[6] A full formalization lies beyond the scope of this article, but the crucial step lies in enriching the analyses Schwarz and Jenks propose for strong definites with a mechanism that allows us to keep track of how the topic situations of different sentences relate to one another. This means we need a dynamic interpretation not only of the indices in (5b) and (6b) but also of the situation pronouns. With this mechanism in place, we can work out the relationship between indices and situation pronouns to derive the difference between text- and situation-level familiarity.

The Mandarin versions of (18) and (19) illustrate our hypothesis. *Nánshēng* 'boy' in (4) meets the requirements of text-level familiarity but not those of situation-level familiarity. This explains Jenks's observation that Mandarin requires the demonstrative in this context. *Bāoguǒ* 'parcel' in (16) does meet the requirements of situation-level familiarity, and this explains our finding that Mandarin allows the use of the bare noun in this context.

4.3.3 Predictions If our hypothesis is on the right track, we expect that manipulating the spatiotemporal contiguity of eventualities in examples like (18) and (19) leads to changes in the acceptability of bare nouns. Corpora do not allow us to check the outcomes of these manipulations, so we turn to consultants and their judgment.

Above, we presented (18) as involving two eventualities that are spatiotemporally disjoint. We hypothesized that this is why the bare noun is unacceptable in the Mandarin version. (20) is a minimally different variant.

(20) There were a boy and a girl in the classroom. I entered and hit **the boy**.

(20) is different from (18) in that the state of there being a boy and a girl in the classroom and the event of the speaker going in and hitting the boy spatiotemporally overlap and can straightforwardly be thought of as being part of a single overarching topic situation. We thus have set up a context in which we no longer have to rely on text-level familiarity but can also resort to situation-level familiarity. In line with our hypothesis, we predict the demonstrative to remain available but the bare noun to become a viable option as well.

(21) is the Mandarin version of (20).

(21) **M:** Jiàoshì lǐ yǒu yī gè nánshēng hé yī gè nǚshēng. Wǒ jìn jiàoshì
classroom in have one CLF boy and one CLF girl I enter classroom
dǎ-le **nánshēng**.
hit-ASP boy
'There were a boy and a girl in the classroom. I entered and hit **the boy**.'

Our consultants report that a demonstrative can be added to *nánshēng* 'boy' in the second sentence of (21) but that this is not required. When asked to compare the acceptability of (21) with the Mandarin version of (18) (i.e., Jenks's original example), they indicate that there is a clear difference between the two: (21) is acceptable without the demonstrative whereas (18) is not. These

[6] Wolters (2006) argues that English demonstratives need to be interpreted with respect to nondefault situations. A full comparison between the conditions of use of Mandarin and English demonstratives regrettably lies beyond the scope of this article.

judgments are in line with our predictions: if the antecedent is introduced in the same topic situation as its anaphor, the latter can be realized as a bare noun.

Let us turn to (19). Above, we presented it as a context in which all eventualities are part of a single overarching topic situation. In (22), we present a slight modification.

(22) [Context: As the owls flooded into the Great Hall as usual, everyone's attention was caught at once by a long thin package carried by six large screech owls. Harry was just as interested as everyone else to see what was in this large parcel and was amazed when the owls soared down and dropped it right in front of him, knocking his bacon to the floor.]

M: Màigé jiàoshòu qián yī tiān jì gěi hālì #(**zhè ge**) **bāoguǒ**.
McGonagall Professor before one day send to Harry this CLF package
'Professor McGonagall had sent the package to Harry the day before.'

(22) is different from the Mandarin version of (19) in that the eventuality of sending the package is spatiotemporally disjoint from all other eventualities in the context. Situation-level familiarity is consequently no longer available. In line with our hypothesis, we predict (22) to differ from the Mandarin version of (19) in that the bare noun is no longer a viable option to mark familiarity. This prediction is borne out, as our consultants report that *bāoguǒ* 'package' in (22) requires the demonstrative.

4.4 Summary

In this section, we defended the hypothesis that the acceptability of bare nouns in strong definiteness environments in Mandarin indicates that there are two subtypes of strong definiteness: text-level and situation-level familiarity. Demonstratives can mark both, but bare nouns are limited to the latter. We showed how the hypothesis explains our data and generalizes to new contexts.

5 Conclusion

Schwarz (2009) and Jenks (2018) argue that the weak/strong definiteness distinction is active in German and Mandarin, respectively. We carried out a parallel-corpus study to check the crosslinguistic predictions that follow. Our corpus data show that the distributions of German contracted/uncontracted definites and Mandarin bare nouns/demonstratives are orthogonal (section 2). Closer scrutiny of the data reveals that the problem lies with Mandarin bare nouns, as they appear in both weak and strong definiteness contexts (section 3). We argued that the discrepancy between our data and Jenks's indicates that there are two types of strong definiteness: text-level and situation-level familiarity. Bare nouns turn out to be compatible only with situation-level familiarity (section 4).

Our results also shed light on the competition between weak and strong definiteness markers. Jenks's constraint Index! leads to a strict separation between contexts allowing for weak and strong definites. Setting the Mandarin facts aside, we found that this constraint is too strong, even for our German data.

Our data led us to maintain that German contracted/uncontracted definites are uniformly weak/strong. Mandarin demonstratives also turn out to be uniformly strong but Mandarin bare nouns turn out to be ambiguous between weak and strong definites, the latter being restricted to situation-level familiarity. Another way to go—suggested by a reviewer—is to assume that Mandarin bare nouns uniformly mark weak definiteness. On this analysis, their anaphoric uses would be indicative of the overlap between weak and strong definiteness contexts. We leave the exploration of this competing analysis for future work. The main challenge it faces is to explain why German consistently opts for its strong definite in these contexts whereas Mandarin can also rely on its weak definite.

We conclude with a methodological note. This article has shown the potential of parallel-corpus research for formal approaches to language. On the one hand, we have shown how a small study can lead to relevant results. On the other hand, we have laid the foundation for larger-scale studies, both at the level of corpus compilation and at the level of analysis.

References

Aguilar-Guevara, Ana, Julia Pozas Loyo, and Violeta Vázquez-Rojas Maldonado. 2019. *Definiteness across languages.* Berlin: Language Science Press.

Aguilar-Guevara, Ana, and Joost Zwarts. 2010. Weak definites and reference to kinds. In *Proceedings of SALT 20*, ed. by Nan Li and David Lutz, 179–196. https://journals.linguisticsociety.org/proceedings/index.php/SALT/issue/view/108.

Arkoh, Ruby, and Lisa Matthewson. 2013. A familiar definite article in Akan. *Lingua* 123:1–30. doi:http://dx.doi.org/10.1016/j.lingua.2012.09.012.

Ebert, Karen. 1971a. Referenz, Sprechsituation und die bestimmten Artikel in einem nordfriesischen Dialekt (Fering). Doctoral dissertation, Christian-Albrechts-Universität zu Kiel.

Ebert, Karen. 1971b. Zwei Formen des bestimmten Artikels. In *Probleme und Fortschritte der Transformationsgrammatik*, ed. by Dieter Wunderlich, 159–174. Munich: Hueber.

Georgakopoulos, Thanasis, and Stéphane Polis. 2018. The semantic map model: State of the art and future avenues for linguistic research. *Language and Linguistics Compass* 12(2), 1–33.

Hawkins, John A. 1978. *Definiteness and indefiniteness.* London: Croom Helm.

Heim, Irene. 1991. Artikel und Definitheit. In *Semantik: Ein internationales Handbuch der zeitgenössichen Forschung*, ed. by Arnim von Stechow and Dieter Wunderlich, 487–535. Berlin: Walter de Gruyter.

Jenks, Peter. 2018. Articulated definiteness without articles. *Linguistic Inquiry* 49:501–536.

van der Klis, Martijn, Bert Le Bruyn, and Henriëtte de Swart. 2017. Mapping the perfect via translation mining. In *15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017*, ed. by Mirella Lapata, Phil Blunsom, and Alexander Koller, 2:497–502. Association for Computational Linguistics.

van der Klis, Martijn, Bert Le Bruyn, and Henriëtte de Swart. 2020. De la sémantique des temps verbaux à la traductologie: Une comparaison multilingue de *L'Étranger* de Camus. In *Linguistic approaches to tense, aspect, modality, evidentiality, based on the novel* L'Etranger *("The Stranger") by Albert Camus, and its translations*, ed. by Eric Corre, Dan Thành Do-Hurinville, and Huy Linh Dao, 12–37. Amsterdam: John Benjamins.

van der Klis, Martijn, and Jos Tellings. 2020. Multidimensional scaling and linguistic theory. Ms., Utrecht University. https://arXiv:2012.04946.

McKenzie, Andrew. 2012. The role of contextual restriction in reference-tracking. Doctoral dissertation, University of Massachusetts Amherst.

McKenzie, Andrew. 2015. A survey of switch-reference in North America. *International Journal of American Linguistics* 81:409–448.
Meier, Cécile. 2019. Temporal information and definite descriptions. Abstract for poster presented at SALT 2019. http://salt.linguistics.ucla.edu/29/abstracts/meier-salt29-abstract.pdf.
Schwarz, Florian. 2009. Two types of definites in natural language. Doctoral dissertation, University of Massachusetts Amherst.
Wälchli, Bernhard, and Michael Cysouw. 2012. Lexical typology through similarity semantics: Toward a semantic map of motion verbs. *Linguistics* 50:671–710.
Wiltschko, Martina. 2013. Descriptive relative clauses in Austro-Bavarian German. *Canadian Journal of Linguistics/Revue canadienne de linguistique* 58:157–189.
Wolter, Lynsey. 2006. That's That: The semantics and pragmatics of demonstrative noun phrases. Doctoral dissertation, University of California, Santa Cruz.
Zhang, Niina. 2015. Nominal-internal phrasal movement in Mandarin Chinese. *The Linguistic Review* 32: 375–425.

Bibliographical details of the source text and its translations

Rowling, J. K. 1997. *Harry Potter and the philosopher's stone.* London: Bloomsbury.
Rowling, J. K. 1997/1998. *Harry Potter und der Stein der Weisen.* Trans. by Klaus Fritz. Hamburg: Carlsen Verlag.
Rowling, J. K. 2000. *Hā lì bō tè yǔ mó fǎ shí.* Trans. by Nong Su. Beijing: People's Literature Publishing House.

David Bremmers
Utrecht University
d.j.e.bremmers@students.uu.nl

Jianan Liu
Utrecht University
Utrecht Institute of Linguistics OTS (UiL OTS)
j.liu4@uu.nl

Martijn van der Klis
Utrecht University
Utrecht Institute of Linguistics OTS (UiL OTS)
m.h.vanderklis@uu.nl

Bert Le Bruyn
Utrecht University
Utrecht Institute of Linguistics OTS (Uil OTS)
b.s.w.lebruyn@uu.nl