‘Articleless’ languages are not created equal

Jianan LIU — *Utrecht University*
Shravani PATIL — *Tilburg University*
Daria SERES — *Humboldt University of Berlin* and *Universitat Autònoma de Barcelona*
Olga BORIK — *Universidad Nacional de Educación a Distancia*
Bert LE BRUYN — *Utrecht University*

Abstract. We adopt a translation corpus approach based on the first chapter of *Harry Potter and the Philosopher’s Stone* to evaluate Dayal’s updated version of the neo-Carlsonian framework and the predictions it makes for bare nouns in Hindi, Russian and Mandarin (Dayal 2004). Our Hindi data turn out to be overall in line with Dayal’s predictions but the same does not hold for our Russian and Mandarin data, leading us to explore a number of extensions and modifications of Dayal’s analysis. For Mandarin, our data lead us to hypothesize a role for the numeral *yi* (‘one’) as an indefinite article and for demonstratives as definite articles. For our Hindi and Russian bare noun data, we argue that the only way to account for them is to reverse at least part of Dayal’s updates to the neo-Carlsonian framework and to hypothesize that Hindi – unlike Russian – is developing an indefinite article.

Keywords: definiteness, indefiniteness, bare nouns, Hindi, Russian, Mandarin.

1. Introduction

In his seminal paper on reference to kinds across languages, Chierchia defends the intuition that languages that do not have articles freely allow their bare nouns (henceforth BNs) to give rise to definite and indefinite interpretations (Chierchia 1998). On the basis of fine-grained data from Hindi, Russian and Mandarin, Dayal (2004) argues that this generalization holds for bare plurals (henceforth BPs) and for BNs in classifier languages, but not for bare singulars (henceforth BSs), the latter being restricted to definite interpretations. Dayal accounts for this new generalization in an updated version of Chierchia’s neo-Carlsonian framework.

In this paper, we retrace the data underlying Dayal’s argumentation and sketch the way she accounts for them (Section 2). Following up on problematic data from Russian, we propose a translation corpus study based on the first chapter of *Harry Potter and the Philosopher’s Stone* and its translations to the three languages studied by Dayal (2004) (Section 3). Our Hindi data turn out to be overall in line with Dayal’s predictions but the same does not hold for our Russian and Mandarin data (Section 4), leading us to explore a number of extensions and modifications of Dayal’s analysis (Sections 5 and 6). The overall conclusion we will arrive at is that BNs do not behave in the same way across so-called ‘articleless’ languages and that the explanation might lie in the fact that some of these languages are less articleless than the literature has suggested up till now.

2. BNs in ‘articleless’ languages: from Hindi to Russian and Mandarin

Dayal (2004) argues that Hindi BNs display a singular/plural asymmetry. Whereas plural BNs (henceforth *bare plurals* or *BPs*) straightforwardly allow for narrow scope indefinite readings, singular BNs (henceforth *bare singulars* or *BSs*) turn out to be more restricted. We illustrate this asymmetry with the minimal pair in (1) (see Dayal 2004):

(1) a. #*caroN taraf* ***cuuha*** *hai*
everywhere mouse is
b. *caroN taraf* ***cuuhe*** *haiN*
everywhere mice are

The intended readings of *cuuha* (1a) and *cuuhe* (1b) are those in which they take scope under *caroN taraf*, leading to the assertion that there were mice everywhere. As Dayal points out, this reading is available for (1b) but not for (1a), the latter only leading to the pragmatically odd assertion that the same mouse was everywhere. With Dayal, we conclude that the opposition between (1a) and (1b) shows that Hindi BSs do not have the same range of indefinite readings as Hindi BPs.

Even though the contrast in (1) seems to bear on scope in the indefinite domain, Dayal (2004) gives it a definiteness twist, arguing that the data follow if we assume BPs allow for indefinite interpretations but BSs do not. In 2.2, we develop this intuition in more detail, sketch how Dayal derives it in an extended version of the neo-Carlsonian framework and explore the implications for Hindi, Russian and Mandarin. To properly frame the extensions Dayal proposes, we however start by taking another look at the original version of the neo-Carlsonian framework (Chierchia 1998) and the predictions it makes about Hindi BSs and BPs.

2.1. Chierchia’s predictions for Hindi BSs and BPs

Chierchia (1998) does not explicitly treat Hindi, but we can generate the predictions he makes for its BSs and BPs. For presentation purposes, we work out the predictions under Dayal’s assumption that Hindi is an articleless language, an assumption that Chierchia (1998) does not commit to.

For Hindi BSs, Chierchia’s neo-Carlsonian framework presents two ways to derive indefinite readings. The first is a simple existential shift (∃): under Dayal’s assumption that Hindi is an articleless language, BSs are predicted to be able to undergo a covert ∃-shift and end up with an indefinite interpretation. The second way is a more involved one, building on Chierchia’s Derived Kind Predication (henceforth *DKP*) and the fact that Hindi BSs can refer to kinds, as illustrated in (2) (see Dayal 2004:402):

(2) ***kutta*** *aam janvar hai*
dog common animal is
‘The dog is a common animal.’

DKP is an operation that kicks in when a kind combines with a predicate requiring reference to regular individuals (see Chierchia 1998:364):

(3) Derived Kind Predication (DKP):
If P applies to regular individuals and k denotes a kind, then
$P(k) = \exists x[{}^{\cup}k(x) \wedge P(x)]$

On its definition in (3), DKP leads to existential quantification over instantiations of a kind, effectively giving rise to a (derived) indefinite reading. With BSs referring to kinds in Hindi, Chierchia’s DKP thus constitutes a second path to indefinite readings for BSs.

Moving to Hindi BPs, Chierchia predicts them to give rise to indefinite readings on a par with Hindi BSs. The one difference with BSs resides in the fact that the latter have two paths that lead to indefinite readings – the existential and the DKP path – whereas BPs only have the DKP path at their disposal. The full derivation of the DKP path starts with a shift from predicates to their corresponding kinds (the *down*-shift,$^{\cap}$) and is followed by DKP. The reason BPs do not have the existential path at their disposal is that Chierchia ranks the shift from predicates to their corresponding kinds (the *down*-shift, $^{\cap}$) above the $\exists$- and the iota (ι)-shifts, and argues that the $^{\cap}$-shift is defined for plurals but not for singulars. Given that the $^{\cap}$-shift is defined for BPs, its ranking above the $\exists$-shift blocks the latter from applying and cuts off the existential route to indefinite readings for BPs. For BSs, the $^{\cap}$-shift is undefined, and its higher ranking has no effect on the availability of the $\exists$-shift, maintaining the latter as a path towards indefinite readings.

Summarizing Chierchia’s predictions for Hindi, we have worked out how BSs can get indefinite interpretations through the $\exists$-shift and DKP whereas BPs get indefinite readings through DKP alone. Importantly, though, the opposition in the availability of the $\exists$-shift has no bearing on the asymmetry we find in (1). Indeed, both the $\exists$-shift and DKP are expected to allow for narrow scope readings, leaving the unavailability of the narrow scope reading of the BS in (1a) and its asymmetry with the BP in (1b) unaccounted for.

2.2. Dayal’s account for Hindi and its predictions

Dayal’s extensions of Chierchia’s neo-Carlsonian framework are mainly targeting the singular paradigm. We present the underlying intuition and discuss the extensions Dayal proposes, focusing on BSs but also briefly looking into BPs. Dayal’s account is inspired by the intuition that Hindi BSs cannot get indefinite readings but only definite ones, straightforwardly explaining why *cuuha* in (1a) cannot but refer to a unique mouse and lead to the pragmatically odd assertion that the same mouse was everywhere. To derive this restriction to definite readings for BSs, Dayal introduces two extensions to Chierchia’s neo-Carlsonian framework. The first is to not only rank the $^{\cap}$-shift above the $\exists$-shift but to do the same for the ι-shift, leading to the ranking $^{\cap}, \iota{>}\exists$. The effect of this move is that the $\exists$-shift no longer constitutes a viable path to indefinite readings for Hindi BSs – independently of the fact that the $^{\cap}$-shift is not defined for them. The second extension Dayal proposes is to restrict the availability of DKP to kinds that have a 'semantically transparent relation to their instantiations' (Dayal 2004:430), a property that Dayal associates with kind reference of plural nouns but not of singular nouns. The effect of this second extension is that DKP is also cut off as a viable path to indefinite readings for Hindi BSs.

With the two extensions she proposes, Dayal makes sure that there are no paths to indefinite readings for Hindi BSs in her updated version of Chierchia's neo-Carlsonian framework. She thus guarantees that the only non-kind referring readings BSs can get in regular argument position are definite ones, deriving the pragmatically odd reading of *cuuha* in (1a). For BPs, Dayal's extensions have no impact on the availability of DKP-generated indefinite readings. The narrow scope indefinite reading Chierchia predicts for *cuuhe* in (1b) is thus maintained and the contrast with (1a) accounted for.

Dayal's extensions of Chierchia's neo-Carlsonian framework make a number of predictions. First, for Hindi BSs, the prediction is that they should never give rise to indefinite readings in regular argument position. Second, given that the extensions are defined at the level of type-shift rankings and DKP, they are intended to be language independent and the predictions for Hindi BSs should extend to BSs in any other articleless language. Finally, under the assumption that articleless languages without a grammaticalized singular/plural distinction in the nominal domain do not impose restrictions on the application of DKP (Dayal 2004:413), Dayal predicts them to differ from languages like Hindi and always allow for indefinite readings of their BNs. In what follows, we present Dayal's take on these predictions for Hindi, Russian and Mandarin and discuss how they have been received in the literature.

For Hindi, Dayal admits that there are cases in which BSs seem to get an indefinite reading (see Dayal 2011):

(4) *anu* ***kitaab*** *paRhegii*
Anu book read-FUT
'Anu will read a book.'

To account for cases like (4), Dayal argues that *kitaab* does not appear in regular argument position but rather in a pseudo-incorporated position. Crucially, pseudo-incorporated nouns can be argued not to type-shift, their apparent indefiniteness stemming from the construction they appear in. As such, examples like (4) do not need to pose a threat for Dayal's prediction that Hindi BSs in regular argument position only take on definite readings.

Dayal takes Russian to be a good example of another articleless language with a grammaticalized singular/plural distinction in the nominal domain and argues that Russian BSs align with their Hindi counterparts. (5) replicates the BS/BP asymmetry we saw in (1) (see Dayal 2004):

(5) a. #***Sobaka*** *byla vezde.*
dog was everywhere

b. ***Sobaki*** *byli vezde.*
dogs were everywhere

Whereas (5b) straightforwardly allows for the reading according to which there were dogs everywhere, the singular *sobaka* only leads to the same pragmatically odd reading as (1a), according to which the same dog was everywhere.

For articleless languages that allow for BNs but do not have a grammaticalized singular/plural distinction, Dayal discusses Mandarin and points out that Mandarin BNs are on a par with Hindi BPs rather than with Hindi BSs in allowing for narrow scope readings in contexts like (1) (see Dayal 2004):

(6) ***Gou*** *zai meigeren-de houyuan-li jiao.*
dog at everyone-DE backyard-inside bark

(6) is compatible with a reading in which different dogs are barking in different people’s backyards. This reading is similar to the one we get for Hindi BPs in (1b), in line with Dayal’s predictions.

In the formal semantics literature, Dayal’s account has been the predominant one for Hindi BNs and the literature on Mandarin has not called into question the predictions Dayal makes. For Russian, the story is different and multiple authors have argued that Russian BSs do not show any signs of inherent definiteness (e.g., Bronnikov 2006; Šimík & Demian 2020; Seres & Borik 2021). (7) illustrates this (see Seres & Borik 2021):

(7) V každom dome igral **rebënok**.
in every house played child.NOM

(7) straightforwardly allows for a reading according to which different children were playing in different houses, showing that the BS *rebënok* can take narrow scope under the universal *každom dome*. We concede that the structure of (7) is possibly different from the one in (5a) but this should not affect Dayal’s prediction, and we conclude that (7) constitutes a clear counterexample.

2.3. Towards a cross-linguistic re-assessment of Dayal’s account

Although the Russian facts have an immediate impact on the validity of Dayal’s analysis, we are not aware of any attempt at re-evaluating Dayal’s account for other languages than Russian. We assume that this is because the literature – up till recently – lacked the right tools to compare the distributions of BNs across languages and properly assess the empirical scope of counterexamples like (7). In this paper, we propose a translation corpus study and assess the predictions Dayal makes by analyzing translations of the same text to Hindi, Russian and Mandarin, allowing for a broad parallel evaluation of Dayal’s predictions for these three languages.

3. Methodology

Translation corpus research has recently been argued to constitute a valuable addition to the toolbox of semanticists who study cross-linguistic variation. The phenomena that have been studied include – among others – tense and aspect (Fuchs & Gonzalez 2022; van der Klis et al. 2022; Mulder et al. 2022; de Swart et al. 2022a; Tellings et al. 2021), negation (de Swart 2020) and reference (Bremmers et al. 2022). As for languages, the main focus has been on Romance and Germanic, but we also find extensions to a broader set of European languages (Gehrke 2022; de Swart et al. 2022b) as well as to Mandarin (Bogaards 2022; Bremmers et al. 2022; Mo 2022). Parallel to theory-oriented research, recent work is also covering the methodological side of translation corpus research from a semantics perspective (Le Bruyn et al. 2022; Le Bruyn et al. 2023; Le Bruyn & de Swart *submitted*).

The main advantage of translation corpora is that they present the same semantic content in a maximally similar way in different languages. For research into reference, this means that we can neatly trace the ways different languages deal with reference in the same (or maximally similar) contexts. By choosing a source language that makes a formal distinction between definiteness and indefiniteness, we can furthermore use this distinction as an independently motivated criterion to distinguish between definite and indefinite reference in languages that have been argued not to mark this distinction.

The source corpus we selected is the first chapter of *Harry Potter and the Philosopher's Stone*, a fairly recent novel that has been translated to Russian, Hindi, and Mandarin but also to an impressive array of other typologically diverse languages, allowing for the easy scaling up of the approach we pursue. We extracted all referential expressions from the English source text (N=1210) but for the current research, we focus on *a*(*n*) + N_{sg} (n=90), *the* + N_{sg} (n=140) and N_{pl} (n=52) and look into how they are rendered in the Russian, Hindi and Mandarin translations of the novel. The choice of these referential expressions is inspired by the dimensions that play a role in Dayal's analysis: number and (in)definiteness.

For *a*(*n*) + N_{sg}, Dayal predicts BS translations to be rare in Hindi and Russian and for BN translations to be perfectly fine in Mandarin. For Hindi, BSs should only be allowed to occur in pseudo-incorporation constructions (as in (4)) whereas, in other contexts, the Hindi translator is predicted to rely on overt determiners, the ∃-shift and DKP both being cut off as viable paths to indefiniteness for BSs. In line with the examples Dayal presents herself, the default way of rendering a singular indefinite in Hindi is to rely on *ek* ('one'):

(8) *bahut saal pahle, yehaaN* *(**ek**) **aurat** *rahtii thii.*
many years ago here one woman lived
'Once upon a time, a woman used to live here.'

For Russian, we expect to find the same empirical picture as for Hindi, BSs being the minority option and determiners like *odin* ('one') being the default option for rendering singular indefinites.

Given that Dayal does not cut off the DKP path to indefiniteness for Mandarin BNs, she predicts the latter to be viable translations for singular indefinites.

For *the* + N_{sg} and for N_{pl}, Dayal predicts BNs/BSs/BPs to be the default options in all three languages. Given that she ranks the ι-shift at the same level as the $^{\cap}$-shift, each of the languages should straightforwardly allow its BSs/BNs to appear in singular definite contexts. Furthermore, given that Dayal takes DKP not only to be a viable path to indefiniteness for Mandarin BNs but also for Hindi and Russian BPs, we expect to find BPs/BNs as the default translations of N_{pl} in all three languages.

In Section 4, we organize the presentation of the results around the three types of contexts we have sketched above: singular indefinites, singular definites and plural indefinites. Because of this division of contexts, we can abstract away from number marking in Russian and Hindi, allowing us to resort to BNs as a general label and directly compare our Russian, Hindi, and Mandarin data. For each of the contexts, we compare the three languages and present descriptive and – where applicable – inferential statistics. The inferential statistic we will rely on is Fisher’s Exact Test, an alternative to the classic chi-square test that provides more reliable results for smaller datasets in which some expressions are far less frequent than others.

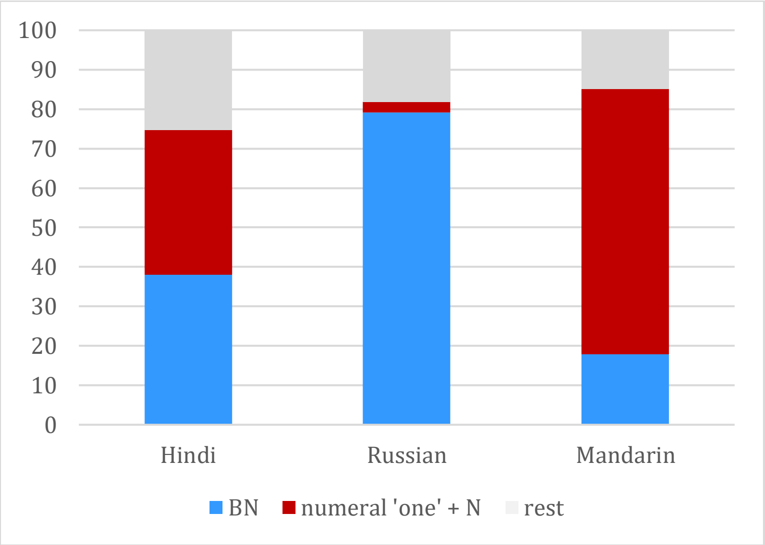
One final remark is in order before turning to the results. Even though translations render the same meaning as their original texts, it does happen that translators opt for different structures in which the referents of the original are not translated one-on-one. A concrete example from our corpus is ‘having a tantrum’ that is translated to Mandarin as *fā píqì* (litt. ‘lose temper’): the overall meaning is the same but there is no direct reference to a tantrum in the translation. We separately report on these cases but do not take them into account in our analyses.

4. Results

4.1. Singular indefinite contexts

For singular indefinite contexts (n=90), we found 23 cases of different constructions in Mandarin, 9 in Russian and 6 in Hindi. We report on three types of translations: (i) BNs, (ii) numeral ‘one’ + N, (iii) rest. For Hindi and Russian, BNs are restricted to BSs and for Mandarin, the numeral option includes a classifier. Graph *1* summarizes the data.

Graph *1* shows that there are big differences in how each language renders singular indefinites. Whereas Russian barely relies on the numeral, the latter is slightly more frequent than the BS in Hindi and is clearly the majority option in Mandarin. The differences in distribution of BNs and the numeral are also statistically significant (α=.05), Fisher’s Exact Tests leading to p-values smaller than 0.01 for the comparisons of the different language pairs. The rest category is varied in each of the languages but BNs and numeral ‘one’ + N clearly come out as the majority options for Hindi and Mandarin. In Russian, none of the rest options (proper names, pronouns, indefinite determiners, etc.) appear in more than two contexts.

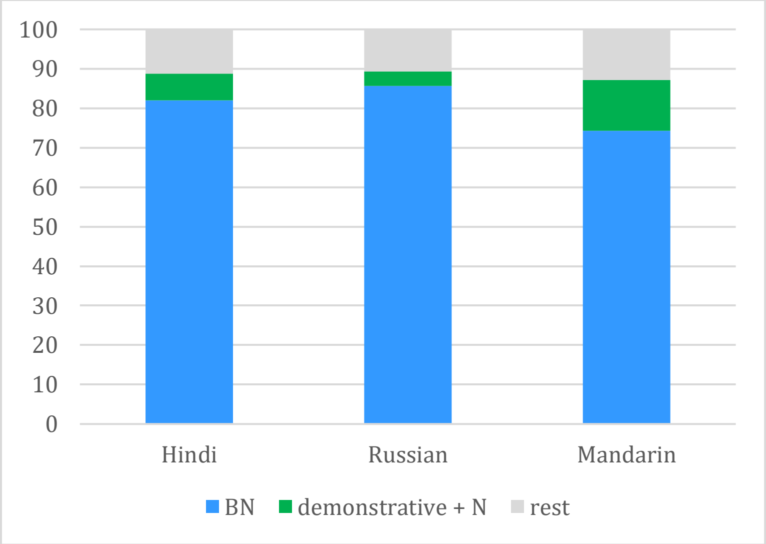


Graph 1: Relative frequencies of BN, numeral ‘one’ + N and rest translations of English indefinite singulars (*a*(*n*) + N_{sg}) in Hindi, Russian and Mandarin

4.2. Singular definite contexts

For singular definite contexts (n=140), we found 18 cases of different constructions in Russian, 12 in Mandarin and 5 in Hindi. Across the three languages, there was one construction that – despite remaining a distant second overall – stood out: the demonstrative. Even though Dayal makes no explicit predictions about the competition between BNs and demonstratives, our data do suggest that there is an interaction between the two and we consequently report on (i) BNs, (ii) demonstrative + N, (iii) rest. As for singular indefinites, BNs are restricted to BSs for Hindi and Russian. For Mandarin, the ‘demonstrative + N’ option typically contains a classifier. We summarize the data in Graph *2*.

Graph *2* shows that BNs are the majority option in all three languages. At the same time, we see that demonstratives are gaining ground, in particular in Mandarin. Pairwise comparisons between the languages show that the differences in distribution of BNs and demonstratives are significant for Russian-Mandarin ($p < 0.01$, Fisher’s Exact Test) but not for Russian-Hindi (p=0.15) nor for Hindi-Mandarin (p=0.14).



Graph 2: Relative frequencies of BN, demonstrative + N and rest translations of English definite singulars (*the* + N_{sg}) in Hindi, Russian and Mandarin

4.3. Plural indefinite contexts

For plural indefinite contexts (n=52), we found 5 cases of different constructions for Russian, 4 for Mandarin and 4 for Hindi. No constructions involving plural determiners appeared in more than two contexts in any of the languages, leaving us with no clear competitors to compare BNs to. In the absence of obvious competitors, we refrain from presenting graphs with relative frequencies and running inferential statistics. Our data show that BNs/BPs come out as the main category for translating plural indefinites in all of the languages (n=31 in Hindi, n=32 in Russian, n=39 in Mandarin).

5. Discussion

In Sections 2 and 3, we worked out the predictions Dayal makes for the translation of singular indefinites, singular definites and plural indefinites to Hindi, Russian and Mandarin. For singular definites and plural indefinites, we argued that Dayal predicts BNs/BSs/BPs to be the default options. For singular indefinites, however, Mandarin would have BNs as the default option whereas Hindi and Russian should both show a clear preference for nouns preceded by indefinite determiners like the numerals *ek* and *odin* (‘one’).

The picture that emerges from our results in Section 4 is different from the one predicted by Dayal. In this section, we zoom in on singular definite and singular indefinite contexts, discuss in how far our data are in line or at least compatible with Dayal’s predictions and explore extensions and modifications where relevant. Throughout, we will argue that Dayal’s analysis has to be extended and ultimately modified. The alternative analysis we move towards is one in which so-called ‘articleless’ languages do have articles that compete with BNs in varying ways. For reasons of space, we do not treat plural indefinite contexts separately. The general plural indefinite results are in line with Dayal’s predictions and even though the data deserve to be unpacked further, we identified no tendencies that would go against Dayal’s analysis.

5.1. Singular definite contexts

Our singular definite data are overall in line with Dayal’s predictions in the sense that BSs/BNs clearly constitute the majority option in all three languages. The one surprise in our data is the special role of demonstratives that leads to a statistically traceable difference between Russian and Mandarin. Given that Hindi demonstratives do not lead to significant differences with Russian or Mandarin, we focus here on the Mandarin case.

The role of demonstratives in the referential system of Mandarin is not predicted by Dayal in her 2004 paper but has received attention in the more recent literature. Jenks (2018) argues that Mandarin demonstratives function as grammaticalized markers of familiarity and block BNs from marking this subtype of definiteness. In what follows, we argue that there is a division of labor between BNs and demonstratives, that it is different from the one proposed by Jenks and that it does not jeopardize the core of Dayal’s analysis.

Our data show that demonstratives are used in familiarity contexts (10), but at the same time, we find that they are not obligatory in these contexts (9), *contra* Jenks (2018). Both (9) and (10) are part of a bigger context in which a cat is introduced and referred back to, (9) occurring before (10).

(9) **English**
Mr Dursley blinked and stared at **the cat**. It stared back.
Mandarin
Désīlǐ xiānshēng zhă. le zhă yăn, dīng zhe ***māo*** *kàn*
Dursley Mr blink LE blink eye stare ASP cat look

(10) **English**
[…] he watched the cat in his mirror.
Mandarin
[…] *tā cóng hòushìjìng lǐ kànkàn* ***nà zhī māo***.
he from rear-view-mirror inside look that CL cat

Both *māo* in (9) and *nà zhī māo* in (10) refer back to the same cat that was introduced earlier. They thus count as familiar definites and show that Mandarin resorts both to BNs and to demonstratives in familiarity contexts. The exact division of labor between the two is an empirical puzzle that has been tackled in several recent papers (Bremmers et al. 2022; Dayal & Jiang 2021; Simpson & Wu 2022). The data in (9) and (10) are in line with Bremmers et al.’s (2022) proposal that Mandarin is sensitive to situation-level familiarity, allowing for familiar readings of BNs if they are introduced in the same situation as their antecedent and requiring the use of demonstratives to refer back to referents introduced in different situations. We refer the reader to Bremmers et al. (2022)

for further details but the intuition for (9) is that it is part of the same scene in which the cat is introduced through the eyes of Mr Dursley whereas (10) is part of another scene in which Mr Dursley drives off to work and looks back at the cat through his rear-view mirror. In line with Bremmers et al. (2022), we find that the BN can felicitously refer back to the cat within the same scene it was introduced in but that the translator resorts to the demonstrative when referring back to the cat in a separate scene.

Clearly, further research is needed to unpack the Mandarin data further and compare the different proposals on the division of labor between BNs and demonstratives. Crucially, though, our data suggest that such a division of labor exists and argue in favor of analyzing Mandarin demonstratives as article-like expressions that compete with BNs. Dayal does not foresee a role for definite articles in Mandarin but adding them does not impact on the core of her analysis: in neo-Carlsonian analyses, articles constitute an additional layer that is independent of the rankings of type-shifts and of number constraints on DKP. If we extend Dayal’s analysis with the assumption that Mandarin demonstratives function as definite articles, the prediction that follows is that BNs freely take on definite readings except for the subtype that demonstratives specialize in. This prediction is in line with our data, and we conclude that singular definite contexts do not pose a threat to Dayal’s analysis.

5.2. Singular indefinite contexts

Up till now, we have argued that our data in singular definite and plural indefinite contexts are in line or at least compatible with the predictions Dayal makes. The only additional hypothesis we have proposed is that Mandarin demonstratives function as a specific type of definite articles. Crucially, though, our singular definite and plural indefinite data have not led us to propose deep modifications of Dayal’s analysis. In this section, we discuss the singular indefinite contexts in our data and argue that they lead both to extensions and modifications.

We remind the reader that the prediction Dayal makes is that singular indefinites are straightforwardly translated as BNs in Mandarin and that Hindi and Russian should disallow BS translations, preferring nouns preceded by indefinite determiners like *ek* and *odin* (‘one’) instead. We argue that the empirical picture we get for Hindi is in line with this prediction but that the same does not hold for Mandarin and Russian. We go through the three languages in turn and use the translations of *a map* in (11) and of *a new word* in (12) as our running examples.

(11) **English**
It was on the corner of the street that he noticed the first sign of something peculiar - a cat reading **a map**.
Hindi
Sadak-ke mod par dursley ko pehli ajib chiz dikh-i – ek billi, jo ***naksha***
Street-GEN corner on Dursley to first strange thing.F see-PST.F a cat.F who map
padh rahi thi.
read PROG be.PST

Russian
Tol'ko na uglu ulicy mister Dursley nakonec zametil, čto proisxodit čto-to
only on corner street-GEN mister Dursley finally noticed that happens something
strannoe, – a zametil on košku, vnimatel'no izučavšuju ležaščuju pered nej
strange and noticed he cat-ACC attentively examining lying in.front.of her
***kartu**.*
map-ACC

Mandarin
zài jiē.jiǎo shàng, tā kàn-dào-le dì-yī-gè yìcháng-de xìnhào
at street.corner on, he see-RVC-ASP ORD-one-CL peculiar-DE sign
*yì-zhī-māo zài kàn **dìtú**.*
one-CL-cat PROG read map

(12) **English**
She told him over dinner all about Mrs Next Door 's problems with her daughter and how Dudley had learnt **a new word** (‘Shan 't !’) .

Hindi
Unho-ne dinner par apne pati ko bata-ya ki padosan ki apni beti ke.sath
She-ERG dinner on her husband to told-PFV that neighbor of own daughter with
kya samasyaye chal rahi hai aur Dudley-ne **ek naya vakya** sikh-a
what problems go PROG be.PRES and Dudley-ERG a new sentence learn-PFV
hai ‘nahi karu-n-ga’.
be.PRES ‘no do-FUT-M’

Russian
Za obedom ona oxotno spletničala, rasskazav misteru Dursley o tom, čto u
at lunch she gladly gossiped having.told mister-DAT Dursley about that that at
ix. sosedki ser'ëznye problemy s dočer'ju, i naposledok soobščiv,
their neighbour serious problems with daughter and finally having.informed
*čto Dudley vyučil **novoe slovo** “xačču!”.*
that Dudley learnt new word I.wanna

Mandarin
Wǎnfàn zhuō shàng, désīlǐ tàitài xiàng tā jiǎngshùle línjū jiā de
Dinner table-on Durseley Mrs to he tell-ASP neighbour family DE
*mǔ-nǚ máodùn, hái shuō dálì yòu xuéhuì **yīgè xīncí***
mom-daughter conflict also say Dudley again learn-RVC one-CL new-word
(“*jué bù*”).
(never)

The translations of *a map* and *a new word* neatly illustrate the two major patterns that emerge from our data: one in which the Mandarin and Hindi translators both opt for a BN/BS (*a map*) and one in which they both opt for a construction with a numeral, Hindi *ek* and Mandarin *yi* ‘one’ (*a new word*), the Russian translator choosing a BS in both cases. We comment on the third pattern – one in which Hindi and Russian opt for a BS but Mandarin resorts to a construction with *yi* – in due course. The attentive reader will have noticed that the English original in (11) also contains a second indefinite singular – *a cat*. We leave it aside as the structures of its translations vary slightly.

5.2.1. Hindi

For Hindi, our singular indefinite data overall seem in line with Dayal’s predictions, especially if we compare the frequency of BNs as translations of singular definites (over 80%) to the frequency of BNs as translations of singular indefinites (below 40%). (11) and (12) nicely illustrate the alternation between BSs and nouns preceded by *ek* that we find in our singular indefinite data. To be fully in line with Dayal’s predictions, the argument should be that *naya vakya* occurs in regular argument position and therefore requires *ek*, but that *naksha* is pseudo-incorporated and can therefore occur without the numeral. Parallel to *kitaab* in (4), we assume that a pseudo-incorporation analysis for *naksha* is not implausible. We do note that the literature on pseudo-incorporation in Hindi does not give us any direct way to argue for it. In future work, we count on developing Le Bruyn et al. (2016)’s analysis of pseudo-incorporation and on exploiting the constraints it predicts on verb-noun combinations. According to Le Bruyn et al., pseudo-incorporation – in languages that allow for it – is possible in case the verb taps into the explicit or implicit relational semantics of the noun it combines with. This analysis gives us a handle on cases like *read book*, *read map* and *learn new word*. Indeed, whereas *book* and *map* both come with a telic role in their qualia structure (Pustejovsky 1995) and can thus be argued to come with an implicit use relation that can be picked up on by *read*, a noun like *word* arguably does not come with any explicit or implicit relational semantics that *learn* can pick up on. The prediction Le Bruyn et al.’s analysis makes then is that *read book* and *read map* allow for pseudo-incorporation and that *learn* (*new*) *word* does not. These predictions are in line with the data in (4), (11) and (12), *read book* and *read map* leading to BS translations and *learn new word* leading to a translation with *ek*. We submit that a full analysis of the Hindi data requires further theoretical and empirical work but conclude that on the basis of the Hindi singular indefinite data alone, we have no reason to believe that they are incompatible with Dayal’s analysis. We get back to this conclusion when we discuss the Russian indefinite singular data (Section 5.2.3.).

5.2.2. Mandarin

For Mandarin, Dayal’s prediction is that BNs should straightforwardly give rise to indefinite interpretations. With BNs occurring as translations of singular indefinites in fewer than twenty percent of the cases, this is arguably not the empirical picture we find. A counterargument one can entertain is that the low frequency of BNs in Mandarin does not say anything about the grammaticality of indefinite interpretations of BNs, but this argument makes little sense from Dayal’s perspective if the low frequency of BSs in Hindi is to be indicative of their ungrammaticality in regular argument position. We argue that our singular indefinite data are not in line with the predictions Dayal makes and that the analysis she proposes for Mandarin has to be adapted.

There are at least two options available to make sure that Mandarin BNs are not predicted to freely occur in singular indefinite contexts. One is to reconsider Dayal’s assumption that DKP is freely available for BNs in Mandarin. The disadvantage of this strategy is that it would make the prediction that BNs also have a hard time getting an indefinite interpretation in plural contexts, contrary to fact (see Section 4.2.). The other option is to assume that Mandarin is not only developing a definite article (see Section 5.1.) but also an indefinite one, specifically in the singular domain. Unlike the DKP strategy, the article strategy correctly targets the singular domain alone. It does raise the question what it means to be a developing indefinite article and how we can best formalize the division of labor between BNs and this indefinite article in synchrony. In this respect, a relevant tendency in our data is that there is only one case of a BN in Mandarin that is translated with the numeral in Hindi whereas there are fifteen cases of Hindi BSs that are translated with the numeral in Mandarin. What this tendency suggests is that Mandarin BNs are more restricted than Hindi BSs but that they do share a common set of contexts in which they appear. One route that deserves to be explored then is that Mandarin *dìtú* in (11) is pseudo-incorporated in the same way as Hindi *naksha* and that the division of labor between the Mandarin developing indefinite article and Mandarin BNs is to be formalized as a competition between the indefinite article and BNs in pseudo-incorporation constructions.

We conclude that our data point to the need to adapt Dayal's analysis for Mandarin and that extending it with the hypothesis that Mandarin is developing an indefinite article holds promise. We furthermore conclude that our data are suggestive of a scalar relation between contexts allowing for BSs in Hindi and BNs in Mandarin, arguing in favor of analyzing the division of labor between the Mandarin indefinite article and Mandarin BNs as a competition between the article and BNs in pseudo-incorporation constructions. Here too, we hope to develop Le Bruyn et al.'s (2016) analysis of pseudo-incorporation in our future work, as their analysis is explicitly set up in terms of a competition with singular indefinite articles.

5.2.3. Russian

For Russian, Dayal's predictions are similar to the ones for Hindi, BSs being clearly dispreferred and the translator opting for nouns preceded by an indefinite determiner like *odin* in the great majority of the cases. For Hindi, our data were overall in line with these predictions, but our Russian data show a completely different picture: unlike in Hindi, BSs are by far the predominant option to render singular indefinites in Russian. (11) and (12) are representative examples: where Hindi alternates BSs and nouns preceded by *ek*, Russian uniformly opts for BSs. We conclude that our Russian indefinite singular data are not in line with Dayal's predictions.

To accommodate our Mandarin singular indefinite data, it sufficed to extend Dayal's analysis with the hypothesis that Mandarin is developing an indefinite article. To accommodate our Russian data, we do not see how a simple extension could make do. The problem is that Dayal has meticulously closed off all routes to indefinite readings for regular BS arguments and those are exactly the ones we need. At the same time, it seems that we cannot re-open these routes without making the wrong predictions for Hindi. We are thus faced with a stalemate: either we get the Russian data right and the Hindi data wrong or *vice versa*.

To break the stalemate, we think it is instructive to zoom out and go back to the type of examples that originally motivated Dayal's analysis, viz. those in which Hindi BSs turn out to resist narrow scope indefinite interpretations (see (1a)). Crucially, we find the same resistance to narrow scope with unambiguously indefinite expressions like English *a* + N_{sg}:

(13) **A dog** was everywhere.

As noted by Carlson (1977), (13) only has the same bizarre reading as *cuuha* in (1a), viz. one in which the same animal is said to be everywhere. What this suggests is that the odd reading of (1a) is unlikely to be due to a definite interpretation of *cuuha* and that closing off the existential and the DKP route to indefinite readings for BSs is not offering a solution to the real puzzle (1a) raises, viz. one that is not concerned with the absence of indefinite interpretations of BSs but with missing narrow scope readings of indefinite singulars. Even though we will not try to attempt to solve the real puzzle, we submit that there are information-structural considerations at play, explaining why minor variations on the same sentence lead to different intuitions ((5a) vs. (7)).

Under the assumption that Dayal’s closing off of the existential and the DKP route is the wrong way to derive the odd reading of (1a), we argue that at least one of the routes should be reopened, either deciding that DKP can apply to singular kinds or returning to Chierchia’s original type-shift ranking. The upshot of this move is that – independently of our pick – our Russian data follow straightforwardly.

With the Russian data accounted for, the last step to be taken is to offer an alternative account for the fact that Hindi BSs in our corpus give rise to indefinite readings far less easily than in Russian. We hypothesize that Hindi, like Mandarin, is developing an indefinite singular article that competes with BNs in pseudo-incorporation constructions. Unlike the blocking of DKP for singular kinds and the re-ranking of type-shifts, the hypothesis of a developing indefinite article can be done at a language-specific level and – as such – allows us to account for the fact that BSs have a hard time getting indefinite readings in Hindi while at the same time making sure that they freely get these readings in Russian. We conclude that our article strategy allows us to break the stalemate we faced. The fact that articles can show different degrees of grammaticalization furthermore comes with the additional perk of creating the flexibility we need to account for the difference in distribution of Hindi and Mandarin numerals, another theoretical and empirical challenge we, however, have to leave for future work.

5.4. Recap

Throughout this section, we have argued that our data are compatible with many of Dayal’s predictions but that her analysis does go wrong on some crucial points, in particular for definite singular contexts in Mandarin (Section 5.1.) and for indefinite singular contexts in Mandarin and Russian (Sections 5.2.2. and 5.2.3.). To accommodate the Mandarin data, we argued that it suffices to extend Dayal’s analysis, and we hypothesized that Mandarin is developing an indefinite and a definite article. Accommodating the Russian data turned out to require a real modification of Dayal’s analysis, re-opening at least one of the two paths to indefinite readings for BSs that Dayal meticulously closed off. With the Mandarin and Russian data accounted for, we were left with the

Hindi data that originally motivated the closing off of the indefiniteness paths for BSs. Given that there was no way to accommodate the Russian data otherwise, we proposed an alternative analysis for Hindi, hypothesizing that – parallel to Mandarin – it is developing a singular indefinite article. Further theoretical and empirical work is needed but we do believe we have laid the necessary groundwork to build on in our future work.

The picture that has emerged throughout this section is that some so-called ‘articleless’ languages are less articleless than the literature has assumed up till now. We found that Hindi, Russian and Mandarin behave truly differently from each other and that capturing these differences requires us to abandon language-independent strategies to account for language-specific tendencies. By resorting to the hypothesis that Mandarin and Hindi are developing articles, we opted for a language-specific strategy that holds the promise of capturing the tendencies we found in our Mandarin and Hindi data without generating predictions that pose a problem for Russian, in which BSs really do turn out to freely allow for both definite and indefinite readings.

6. Conclusion

In this paper, we adopted a translation corpus approach based on the first chapter of *Harry Potter and the Philosopher’s Stone* to come to a broad parallel evaluation of Dayal’s seminal work on reference in Hindi and the predictions it makes for Russian and Mandarin. Dayal’s core intuition is that Hindi BSs are different from Hindi BPs in that they do not allow for indefinite readings. In Section 2, we worked out how Dayal accounts for this intuition by closing off the two paths to indefinite readings that Chierchia’s original version of the neo-Carlsonian framework left for BSs in articleless languages. In Sections 3 to 5, we worked out how Dayal’s predictions can be operationalized for translation corpus research and argued that our Hindi data are overall compatible with them but that the same does not hold for our Mandarin and Russian data, leading us to explore a number of extensions and modifications of Dayal’s analysis. For Mandarin, our data led us to hypothesize a role for the numeral as an indefinite article and for demonstratives as definite articles. For Hindi and Russian, we argued that the only way to account for the two languages was to re-open at least one of the two paths to indefinite readings Dayal closed off and to hypothesize that Hindi – unlike Russian – is developing an indefinite article. The overall conclusion we arrive at is that so-called ‘articleless’ languages are not created equal.

Along the way, we pointed out that there remains quite some theoretical and empirical work, in particular to properly pin down what it means for a language to be developing a definite and an indefinite article (see Liu et al. 2022 and Bremmers et al. 2022 for some first steps). Relevant follow-up empirical work also includes replication and triangulation of our results as well as extensions to a broader set of languages (see Borik et al. 2022). At a more general theoretical level, it is important to assess the impact of our data on the neo-Carlsonian framework, paying special attention to the desirability of reversing Dayal’s extensions and how these would fare with later updates of the framework (see Liu et al. 2023 for discussion).

References

Bogaards, M. (2022). The Discovery of Aspect: A heuristic parallel corpus study of ingressive, continuative and resumptive viewpoint aspect. *Languages 7(3)*, 158.

Borik, O., B. Le Bruyn, J. Liu, and D. Seres (2022). Bare nouns in Slavic and beyond. Talk presented at Formal description of Slavic languages 15, Berlin, October 5.

Bremmers, D, J. Liu, M. van der Klis, and B. Le Bruyn (2022). Translation Mining: Definiteness across Languages (A Reply to Jenks 2018). *Linguistic Inquiry 53(4)*, pp. 735-752.

Bronnikov, G. (2006). A critique of Dayal (2004). *Term Paper,* University of Texas at Austin.

Carlson, G. N.(1977). A unified analysis of the English bare plural. *Linguistics and philosophy 1(3)*, pp. 413-457.

Chierchia, G. (1998). Reference to kinds across language. *Natural language semantics 6(4)*, pp. 339-405.

Dayal, V. (2004). Number marking and (in) definiteness in kind terms. *Linguistics and philosophy 27*, pp. 393-450.

Dayal, V, and L. J. Jiang (2021). The Puzzle of Anaphoric Bare Nouns in Mandarin: A Counterpoint to Index!. *Linguistic inquiry*, pp. 1-20.

Dayal, V. (2011). Hindi pseudo-incorporation. *Natural Language & Linguistic Theory*, pp. 123-167.

de Swart, H., C. Grisot, B. Le Bruyn, and T. M. Xiqués (2022a). Perfect variations in Romance. *Isogloss. Open Journal of Romance Linguistics 8(5)*, pp. 1-31.

de Swart, H., J. Tellings, and B. Wälchli (2022b). Not… until across European languages: A parallel corpus study. *Languages 7(1)*, p. 56.

Fuchs, M., and P González (2022). Perfect-Perfective Variation across Spanish Dialects: A Parallel-Corpus Study. *Languages 7(3)*, p. 166.

Gehrke, B. (2022). Differences between Russian and Czech in the use of aspect in narrative discourse and factual contexts. *Languages 7(2)*, p. 155.

Jenks, P. (2018). Articulated definiteness without articles. *Linguistic Inquiry 49(3)*, pp. 501-536.

Jiang, L. J. (2017). Mandarin associative plural-men and NPs with-men. *International Journal of Chinese Linguistics 4(2)*, pp. 191-256.

Le Bruyn, B., H. De Swart, and Joost Zwarts (2016). From HAVE to HAVE-verbs: Relations and incorporation. *Lingua 182*, pp. 49-68.

Le Bruyn, B., M. Fuchs, M. van der Klis, J. Liu, C. Mo, J. Tellings, and H. De Swart (2022). Parallel corpus research and target language representativeness: The contrastive, typological, and translation mining traditions. *Languages 7(3)*, p. 176.

Le Bruyn, B., and H. de Swart (2023). Introduction: Tense and Aspect across Languages. *Languages 8(1)*, p. 33.

Le Bruyn, B., and H. de Swart (submitted). *Cross-linguistic research, parallel corpora, and replication in the* Translation Mining *tradition*.

Liu, J., X. Dong, and B. Le Bruyn (2022). Mandarin bare indefinites. In *Proceedings of Sinn und Bedeutung* (Vol. 26, pp. 575-591).

Liu, J., S. Patil, H. Schurr, D. Seres, O. Borik, and B. Le Bruyn (2023). The theory of argument formation: between kinds and properties. Poster accepted for presentation at Semantics and Linguistic Theory 33.

Mo, C. (2022). *The Compositionality of Mandarin Aspect: A Parallel Corpus Study*. PhD dissertation, University Utrecht.

Mulder, G., G.-J. Schoenmakers, O. Hoenselaar, and H. de Hoop (2022). Tense and aspect in a Spanish literary work and its translations. *Languages 7(3)*, p. 217.

Pustejovsky, J, and P. Bouillon (1995). Aspectual coercion and logical polysemy. *Journal of semantics 12(2)*, pp. 133-162.

Seres, D., and O. Borik (2021). Definiteness in the absence of uniqueness: The case of Russian. *Advances in formal Slavic linguistics*, pp. 339-363.

Šimík, R., and C. Demian (2020). Definiteness, uniqueness, and maximality in languages with and without articles. *Journal of Semantics 37(3)*, pp. 311-366.

Simpson, A., and Z. Wu (2022). Constraints on the representation of anaphoric definiteness in Mandarin Chinese. *New Explorations in Chinese Theoretical Syntax: Studies in honor of Yen-Hui Audrey Li* 272, p. 301.

Tellings, J., and M. Fuchs (2021). *Sluicing and Temporal Definiteness*. Manuscript. Utrecht University.

Van der Klis, M., B. Le Bruyn, and H. De Swart (2022). A multilingual corpus study of the competition between past and perfect in narrative discourse. *Journal of Linguistics 58(2)*, pp. 423-457.