Chapter 4

Bare nouns in Slavic and beyond

Olga Borik[a], Bert Le Bruyn[b], Jianan Liu[b] & Daria Seres[c]

[a]Universidad Nacional de Educacion a Distancia [b]Utrecht University
[c]University of Graz

The article presents a study of the distribution of singular bare nouns in three Slavic languages, Russian, Polish and Macedonian, based on parallel translation corpora. The distribution of bare singulars in Russian and Polish shows that they freely appear in definite and indefinite contexts, which makes it possible to classify these languages as truly articleless. Macedonian bare singulars frequently appear in indefinite contexts, alongside with nouns accompanied by an indefinite marker *one*, whose status require further scrutiny. The data reported in this study call for a theoretical account of bare nouns which allows for fine-grained variation in their distribution across domains and languages, taking into consideration a broader/narrower use of competing expressions.

1 Introduction

Referring is one of the main functions of natural language, and speakers of different languages use a variety of linguistic means and mechanisms to express different types of reference. In the empirical study that we present here, we focus on the syntax–semantics interface of bare nouns (BNs) and examine their distributional properties in Russian, Polish and Macedonian, languages that belong to the East Slavic, West Slavic and South Slavic subgroups, respectively. In particular, we address the issue of a comparative distribution of bare singular nouns (BSs) in the definite and indefinite domain across the three languages.

In terms of definiteness/indefiniteness marking, Russian and Polish are typically classified as articleless languages (Dryer & Haspelmath 2013), that is, having no dedicated morphosyntactic marker to express definiteness or indefiniteness.

Olga Borik, Bert Le Bruyn, Jianan Liu & Daria Seres. 2025. Bare nouns in Slavic and beyond. In Berit Gehrke, Denisa Lenertová, Roland Meyer, Daria Seres, Luka Szucsich & Joanna Zaleska (eds.), *Advances in formal Slavic linguistics 2022*, 107–128. Berlin: Language Science Press. DOI: 10.5281/zenodo.15394168

We thus expect nominals to appear in their bare form in all argument positions in both languages. This straightforward expectation is in line not only with the traditional descriptive grammars, such as Švedova (1980), but also with some formal semantic literature, such as Chierchia (1998), Geist (2010), among others. Other formal approaches, most notably Dayal (2004, 2018), argue that number plays a crucial role in the distribution of BNs in articleless languages, making different predictions for bare plurals (henceforth BPs) and BSs. In particular, Dayal (2004) argues that BSs do not get an indefinite interpretation in languages without articles, while BPs can get narrow scope indefinite readings. Therefore, BSs are predicted to be largely restricted to definite contexts. Our focus on BSs allows us to check the predictions made by Dayal's theory as opposed to more traditional approaches.

Macedonian is usually described as a language with a definite article (Friedman 1993, Tomić 2006, among many others). The definite article in Macedonian is postpositive and morphologically bound. It is typically added to the first element of a nominal phrase[1] (e.g., *kuče-to* 'the dog', *ubavo-to kuče* 'the beautiful dog'), and is inflected for number and gender (e.g., *maž-ot* M.SG 'the man', *žena-ta* F.SG 'the woman', *dete-to* N.SG 'the child', *maži-te/ženi-te* M/F.PL 'the men/the women', *deca-ta* N.PL 'the children'). BSs are also admissible in argument positions in Macedonian, while there is no agreement on their interpretation in the literature (Weiss 2004, Topolinjska 2009, among others). The most widely accepted assumption is that BSs in Macedonian appear in indefinite contexts, although it has also been noticed that a determiner *eden* 'one' is often used to mark indefiniteness in this language (Tomić 2006).

Our study aims at answering the following research questions:

1. What is the distribution of BSs in Russian and Polish as languages without articles? Do they appear in both definite and indefinite domains or do we observe significant differences in the distribution of BSs across domains?

2. What is the status of BSs in Macedonian in indefinite contexts as compared to Russian and Polish?

3. What is the status of BSs in Macedonian in definite contexts?

To address these questions, we ran a parallel corpus study to analyze nominal phrases that appear in both definite and indefinite contexts in the three languages, with a critical look at the distribution of BSs in each of the domains.

[1] We use the term *nominal phrase* to abstract away from the DP/NP debate, prominent mostly in the syntactic literature on Slavic. See, for instance, Bošković (2008).

For the definite domain, the expectations are rather straightforward: both traditional descriptive and formal literature seem to converge on the idea that BSs freely appear in definite contexts in Russian and Polish, whereas in Macedonian we expect the definite article to dominate. However, the status of the definite marker as an article in Macedonian is not uncontroversial: Rudin (2021: 313), for instance, suggests that it might be a type of demonstrative rather than an article. Semantic literature repeatedly stresses similarities between demonstrative NPs and definite descriptions (e.g., Roberts 2002, Elbourne 2008), as well as the need to differentiate between the two (Lyons 1999). We include demonstrative nominals in our empirical study and look at the relative distribution of NPs specified by demonstratives vs. definite nominal phrases in Macedonian or BSs in Russian and Polish in the definite domain.

For the indefinite domain, existing analyses diverge when it comes to predictions. Traditional descriptions do not report any irregularities or asymmetries in the distribution of BSs across definite vs. indefinite domains, so they seem to predict that BSs can freely appear in indefinite contexts. However, claims have been made that in Polish and Macedonian, the indefinite marker ONE[2] is acquiring (or has acquired) the status of an indefinite article (Hwaszcz & Kędzierska 2018, Molinari 2022 with reference to Polish; Tomić 2006 with reference to Macedonian). The prediction that these proposals make is that the marker ONE will frequently appear in the indefinite domain in these languages, competing with or prevailing over BSs. The same prediction is made by Dayal (2004, 2018), who takes Hindi as a representative example of an articleless language and argues that it typically resorts to a construction with ONE in those contexts where English uses the indefinite article. Applying Dayal's analysis to Russian and Polish,[3] we expect BSs in these languages to be severely restricted in the indefinite domain, as opposed to the construction with ONE, which should dominate. In other analyses, the marker ONE in Russian is assumed to mark specificity rather than function as an indefinite article (Ionin 2013), which predicts its appearance only in specific indefinite contexts, converging with the predictions of Geist (2010), who argues that BSs in Russian can only get a non-specific reading.[4]

To get a broader cross-linguistic perspective, we compare parallel-corpus data for Russian, Polish and Macedonian to corpus results for Mandarin and German,

[2] The English ONE is used as a cover term for language specific *odin* (Russian), *jeden* (Polish) and *eden* (Macedonian) and their respective forms.

[3] Dayal does discuss Russian, and we assume that the proposal extends to other languages without articles like Polish, as it is based on general, language-independent semantic principles and mechanisms.

[4] Geist's (2010) predictions should be relativized to the information structure since she argues that indefinite BSs cannot serve as aboutness topics.

two non-Slavic languages. Mandarin functions as a control language for Russian and Polish, as it is usually assumed to be an articleless language (Li 2021), whereas German functions as a control language for Macedonian, as both have a definite article.

In order to investigate the distribution of BSs in definite and indefinite contexts in the three Slavic languages we ran a parallel-corpus study, described in detail in §2. We present the results of our study in §3, followed by a general discussion in §4. §5 concludes the paper.

2 Data and methodology

We use parallel corpora to study the distribution of grammatical items in different languages in parallel, an approach that has recently gained traction in the formal literature for the study of a variety of empirical domains, for example, tense and aspect (see – among others – Fuchs & González 2022; Gehrke 2022; Mo 2022; Mulder et al. 2022), negation (de Swart 2020) and reference (Bremmers et al. 2021). Parallel corpus research builds on the assumption that the meanings of the original and the translations are as closely related to each other as the grammars of the respective languages allow them to be. Another important assumption is that translations are representative of their target languages (*the target language representativeness hypothesis*). For a more detailed discussion of the methodology and its caveats, see Le Bruyn et al. (2022), Le Bruyn & de Swart (2022).

This study uses a translation corpus built on the first chapter of J. K. Rowling's *Harry Potter and the Philosopher's Stone*, a novel written in English and translated into many typologically diverse languages. English grammatically marks the distinction between definiteness and indefiniteness, which allows us to easily detect all definite and indefinite referential expressions in the source text. We selected all (in)definite referential expressions (*a N, the N, N-s, the N-s*) with their aligned translations in Russian, Polish, Macedonian, Mandarin and German (n=284) and manually annotated the corresponding NP forms in all the target languages.[5] At this point, it is important to emphasize that our methodology involves the annotation of forms (but not meanings) in the same contexts across the languages under study.

[5] Because some referential expressions are not translated and because of issues of automatic alignment, some data are literally lost in translation. Our dataset for this study includes referential expressions that have translations in all five languages under scrutiny. These numbers are expected not to be identical to the ones in Liu et al. (2023), a study that we conducted for a wider set of languages using the same methodology.

Since this paper focuses on the singular domain, we limit our quantitative analysis to the singular paradigm only.[6] Apart from theoretical reasons discussed in §1, plurals were excluded due to their relatively low frequency in our dataset and the interaction of plural definites with proper names (e.g., *The Potters, The Dursleys*). Thus, our final dataset includes the translations of *a N* (n=82) and *the* N_{sing} (n=124) constructions into Russian, Polish, Macedonian, as well as Mandarin and German, which are used as control languages in this study.

An example of an English source *the N* expression (1a) and its translations from the parallel corpus are shown below.

(1) a. Mr Dursley might have been drifting into an uneasy sleep, but *the cat* on the wall outside was showing no sign of sleepiness.

b. Dolgoždannyj i nespokojnyj son uže prinjal v svoi ob"jatija mistera Darsli, a sidevšaja na ego zabore *koška* spat' soveršenno ne sobiralas'. Russian [N]

c. Pan Dursley zapadł w niezbyt zresztą spokojny sen, ale *kot* na murku nie okazywał najmniejszych oznak senności. Polish [N]

d. Gospodinot Darsli možebi potona vo nemiren son, no *mačkata* na dzidot nadvor ne pokažuvaše ni troška sonlivost. Macedonian [N+the]

e. Mr Dursley mochte in einen unruhigen Schlaf hinübergeglitten sein, doch *die Katze* draußen auf der Mauer zeigte keine Spur von Müdigkeit. German [the N]

f. Désīlǐ xiānshēng mímíhúhú, běnlái kěnéng húluàn shuì-shàng yí jiào, kě huāyuán qiángtóu shàng *nà zhī māo* què méiyǒu sīháo shuìyì. Mandarin [demonstrative+classifier+N]

In the definite domain, we examine the forms that Russian, Polish and Macedonian use for the translation of the English *the N*. In particular, we check whether and to what extent BSs that we expect to find in Russian and Polish, and singular definites that we expect to find in Macedonian, interact with demonstratives in the definite domain. We then contrast the results obtained for the three Slavic languages with Mandarin (as a control for Russian and Polish) and German (as a control for Macedonian).

In the indefinite domain, we look to determine which forms are used for the translations of the English *a N* in all three languages. We evaluate to which extent BSs are used in singular indefinite contexts in Russian vs. Polish vs. Macedonian and check for the interactions with the forms using the marker ONE. Once

[6] Although there will be a short discussion of plurals in §4.1.2.

again, we compare the results obtained for Slavic languages with the results for Mandarin in the indefinite domain.

3 Results

3.1 Singular definite contexts

As far as definite contexts are concerned, there are no major surprises found in our data. The overall results are presented in Figure 1, which reflects absolute frequencies and includes all translations in the target languages. The category *Rest* contains all those translations that do not present any immediate interest for us (e.g., pronouns, possessives, etc.).

As we can observe, BSs are, indeed, the default option for rendering English *the N* both in Russian and in Polish, as shown in Figure 1. The differences in the occurrence of bare nominals in definite contexts are not significant for these two languages ($p = 0.37$, Fisher's Exact Test (FET)). Regarding Macedonian, the most prominent form in singular definite contexts is the one with a definite article, a result which is also fully in accordance with our initial expectations. In all three languages, there are practically no demonstratives used in singular definite contexts.

Comparing the results of Russian and Polish with their control language, Mandarin, we see that BNs in Mandarin are the most frequent form in the definite context as well. However, we also observe an important difference in the relative distribution of BNs vs. NPs specified by demonstratives in the definite domain: in Mandarin, the tendency to resort to demonstratives is higher. The differences are significant for the comparison of Mandarin and Polish ($p < 0.001$, FET), and for Mandarin and Russian ($p = 0.016$, FET).

As for Macedonian and its control language German, the two languages are quite uniform in the distribution of nominal forms in the definite domain. BSs and demonstrative NPs are either absent or clearly outnumbered in singular definite contexts in both Macedonian and German.[7] In §4.1.1, we will come back to the issue of definiteness marking in Macedonian and discuss some of the examples with BSs.

[7] More specifically, there is only one BS found in German and three in Macedonian.

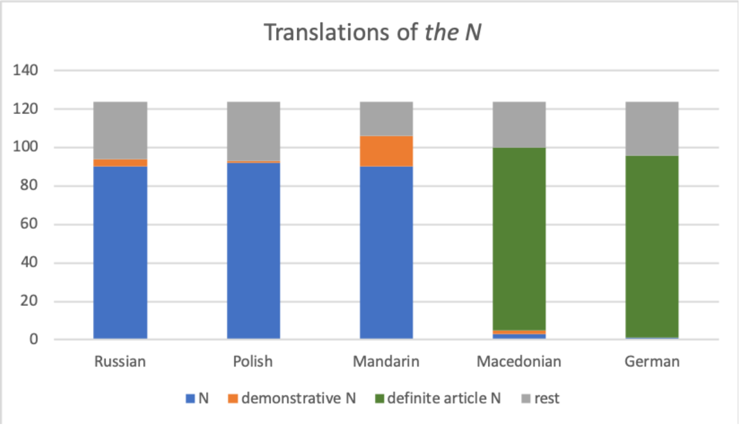

Figure 1: Russian, Polish and Mandarin BNs vs. demonstrative-N; Macedonian and German the-N vs. demonstrative-N

3.2 Singular indefinite contexts

The parallel-corpus data showed that a bare noun is the default option for rendering singular indefinite nominals in both Russian and Polish (see Figure 2). These two languages do not use the ONE+N construction in indefinite contexts in a statistically relevant way. The differences in distribution of bare nominals and nominals preceded by ONE are not significant for Russian and Polish ($p = 0.5$ FET).

In Macedonian, however, while a BS is still the most frequent form in the indefinite domain, the English *a N* construction is more often translated with the numeral ONE than in Russian or Polish. The differences are significant for the comparison of both Macedonian and Russian ($p < 0.001$, FET), and Macedonian and Polish ($p < 0.001$, FET).

As for the control language, Mandarin, where the numeral ONE precedes the nominal in a large number of cases in indefinite contexts, it shows a sharp contrast with Russian and Polish, which hardly ever use this structure. Moreover, Mandarin also shows contrast with Macedonian, where the use of ONE is not as frequent. The differences are significant for Mandarin and Russian ($p < 0.001$, FET), and Mandarin and Polish ($p < 0.001$, FET), as well as for Macedonian and Mandarin ($p < 0.001$, FET).

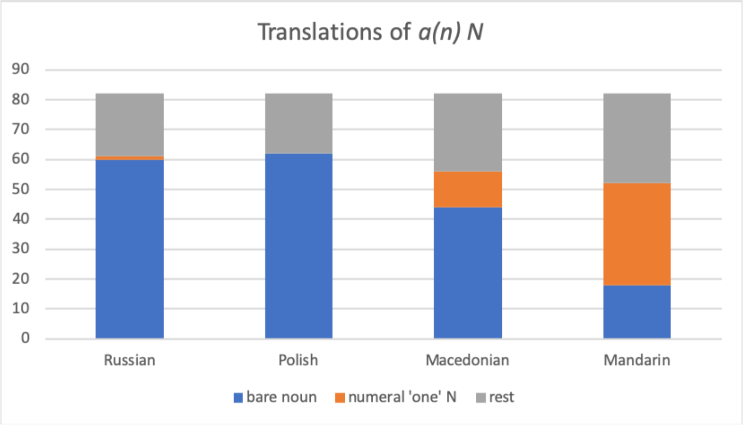

Figure 2: Russian, Polish, Macedonian and Mandarin bare nominals vs. ONE+N

3.3 Recap

Summing up the results of our parallel corpus study, it can be said that Russian and Polish freely use bare nouns in both singular indefinite and singular definite contexts, in accordance with the Slavic descriptive literature. They are, however, in sharp contrast with Mandarin, where the numeral ONE seems to be the default option in the indefinite domain and the demonstrative is competing with bare NPs in the definite domain.

As for Macedonian, in the indefinite domain it seems to occupy an intermediate position between Russian and Polish, on the one hand, and Mandarin on the other: The ONE+N construction appears in the translations of *a N* quite frequently, but not as often as in Mandarin. In the definite domain, Macedonian uses NPs with a definite marker in *the* N_{sing} contexts as often as German.

4 Discussion

We structure our main discussion points in the same way we presented the results of the study, that is, according to the distribution of various forms in a specific context. We begin by evaluating the parallel corpus results obtained for the definite domain. In the discussion of the indefinite domain, we reflect not only on the distribution of BSs, but also on the role of ONE+N construction in the grammar of all three target languages.

4.1 Definite contexts in Russian, Polish and Macedonian

Both traditional descriptions and formal semantic analyses seem to be in full agreement on attributing a possible definite reading to BSs in languages without articles. Our data cast no doubt on this claim for Russian or Polish: BSs prevail in definite contexts in both languages and the distributional behaviour of BSs is therefore in full accordance with their standard semantic descriptions and/or analyses. In Macedonian, the prevailing form is the definite singular, i.e., our data also confirm the status of Macedonian as a language with a definite article. Even though the interpretation of our main results seems to be rather straightforward, there are two points of interest that we would like to discuss.

The first observation concerns the distribution of BSs vs. NPs specified by demonstratives in the definite domain. In the previous section, we pointed out that demonstratives do not seem to occupy a prominent place in either of the three Slavic languages in *the* N_{sing} contexts. Russian and Polish as languages without articles can be contrasted to Mandarin in this respect, one of the two control languages used in this study, where the higher rate of demonstratives in the definite domain suggests that the demonstrative plays a much more significant role in definiteness marking in Mandarin. In fact, Liu et al. (2023) hypothesise that Mandarin is developing a definite article (and an indefinite one), but in this paper, we limit ourselves to empirical statements with respect to Mandarin. In Macedonian, there were only three contexts where a demonstrative was used, all corresponding to anaphoric uses of *the N* in the source text.

The second point that we would like to discuss is the status of the definite article in Macedonian, which will be examined in the next subsections.

4.1.1 The Macedonian definite article in the singular domain

The status of the Macedonian definite article has been subject to some debate in the semantic literature, as pointed out in §1. No consensus emerges from the literature concerning the semantic contribution of this marker. One of the features that it exhibits (and that distinguishes it from a typical definite article) is that it morphologically marks a proximal–neutral–distal distinction, just like demonstratives in many languages do (Lyons 1999). It should be noted that the question about the status of a definite marker in any language is essentially semantic and cannot be definitively resolved without looking into the meaning of this expression, but the distribution of any definite marker/article also plays a significant role in a potential analysis and this is what our study can inform about.

We looked into the properties of the definite article in Macedonian by running a comparative study of the distribution of the definite article in Macedonian and

German. In particular, we measured their co-occurrence in the same contexts by calculating Normalized Pointwise Mutual Information (NMPI, Bouma 2009), which is a bidirectional measure for parallel data (Le Bruyn et al. 2022). The result shows that the NPMI of the two articles reaches 0.48 (with a maximum of 1). That means that the likelihood of the articles in the two languages occurring in the same contexts is higher than chance but not at ceiling. In other words, the bidirectional mapping pattern suggests that the distribution of definite articles in German and Macedonian across the definite contexts is not completely identical.

Table 1: Bi-directional mapping patterns between the German and Macedonian (singular) definite article

		Macedonian		
		definite article N	rest	
German	definite article N	88	24	112
	rest	20	74	94
		108	98	206

Looking into the contexts where Macedonian and German did not coincide in the use of the definite article, we find some interesting examples of BSs. For instance, in (2), Macedonian uses a BS, while German opts for a definite article, just like the English source:

(2) a. At half past eight, Mr Dursley picked up his briefcase, pecked Mrs Dursley on *the cheek* and tried to kiss Dudley goodbye but missed… English (source): [the N]

b. Vo osum i pol gospodinot Darsli ja zede svojata aktenčanta, ja kolvna gospoǵáta Darsli vo *obraz* i se obide da go bakne Dadli za razdelba, no ne uspea… Macedonian: [N]

c. Um halb neun griff Mr Dursley nach der Aktentasche, gab seiner Frau einen Schmatz auf *die Wange* und versuchte es auch bei Dudley mit einem Abschiedskuss… German: [the N]

Another example of the same type of article mismatch is presented in (3):

(3) a. A man appeared on the corner the cat had been watching, appeared so suddenly and silently you’d have thought he’d just popped out of *the ground.* English: [the N]

b. Na agolot što go nabljuduvaše mačkata se pojavi čovek, tolku nenadejno i tivko, kako da izniknal od *zemja*. Macedonian [N]

c. An der Ecke, die sie beobachtet hatte, erschien ein Mann, so jäh und lautlos, als wäre er geradewegs aus *dem Boden* gewachsen. German: [the N]

Although we cannot reach any firm conclusions on the basis of only few examples, we can hypothesise that they both present cases of weak definites (possessive weak definites, Barker 2005, in the case of (2)), so that the contexts where the uniqueness of a definite description is questioned are potentially very good candidates for the absence of the definite article in Macedonian. Needless to say, additional empirical investigation is needed to check this hypothesis.

4.1.2 Some remarks on the plural definite domain

Even though we did not run statistical analyses for the plural domain and there were not too many data points in our dataset, we would like to draw attention to some observations concerning the use of the definite article with plurals in Macedonian that appear important. For instance, Macedonian seems to use definite articles in plural generic contexts, while English resorts to bare plurals and German presents variation.

(4) a. *Cats* couldn't read maps or signs. English (source): [Ns]

b. *Mačkite* ne možat da čitaat ni mapi ni oznaki. Macedonian: [Ns+the]

c. *Katzen* konnten weder Karten noch Schilder lesen. German: [Ns]

The use of the definite article with generic plurals as illustrated in (4) may suggest that Macedonian – at least in some aspects – is rather comparable to Romance languages in its use of definite plurals than to Germanic languages.

Existential contexts in the plural definite domain require further scrutiny. We detected several examples in our dataset where both Macedonian and German use a definite article whereas English uses a bare plural in the same context. One of those examples is (5).

(5) a. And finally, *bird-watchers* everywhere have reported that the nation's owls have been behaving very unusually today. English: [Ns]

b. I konečno *nabljuduvačite* na ptici od site strani javija deka buvo - vite vo našata zemja deneska se ondesuvale mnogu neobično. Macedonian: [Ns + the]

c. Und hier noch eine Meldung. Wie *die Vogelkundler* im ganzen Land berichten, haben sich unsere Eulen heute sehr ungewöhnlich verhalten. German: [the Ns]

In this particular case, the presence of the article in Macedonian could be due to a specific syntactic construction used in the example (prepositional phrase *watchers of birds* instead of the nominal compound *bird-watchers* in the source text). This, however, would not explain the presence of the definite article in German. It might also be the case of a so-called FUNCTIONAL reading of BPs in English, discussed at length for English by Condoravdi (1994). The availability of this reading for BPs is language-specific, so we conclude that our data demonstrate some cross-linguistic variation worth a more systematic investigation. It is not surprising to see this variation in the distributional patterns, as cross-linguistic differences in the use of the definite article are very well documented and widely discussed in the literature. The corpus data of the current study is not sufficient to arrive at any firm conclusions, but it may be reasonably suggested that German and Macedonian, just like many other languages with grammatical marking of definiteness, do not fully coincide in definiteness marking patterns: the overlap in the use of the definite article is only partial, not absolute.

4.2 Indefinite contexts in Russian, Polish and Macedonian

Indefinite contexts constitute the most interesting case in our study, as they convincingly illustrate several theoretically relevant points. First, there is variation both within and outside the Slavic family in the distribution of BSs in the indefinite domain, which has direct repercussions for existing theoretical analyses of BSs. Second, intricate interactions of ONE+N with BSs in the indefinite domain can elucidate the grammatical status of ONE in a given language. Third, our data pose some very specific constraints and requirements for an accurate and empirically adequate theoretical analysis of BSs in languages without articles. We discuss each of these points in the three subsections that follow.

4.2.1 BSs in Russian, Polish and Macedonian

One of the main results of our study concerns the distributional pattern of BSs in Russian and Polish. In particular, the data from the parallel corpus show that in Russian and Polish BSs freely appear in indefinite singular contexts as counterparts of *a N* in the source text. One rather typical example of an indefinite in an existential context is given below:

(6) a. English (source): [a N]
There was *a tabby cat* standing on the corner of Privet Drive, but there wasn't a map in sight.

b. Russian: [N]
Na uglu Praivet Draiv dejstvitel'no sidela *polosataja koška*, no nikakoj karty vidno ne bylo.

c. Polish: [N]
Na rogu Privet Drive rzeczywiście stał *bury kot*, ale nie studiował żadnej mapy.

Our Russian and Polish data directly support traditional descriptive approaches to BSs in Slavic languages without articles and those formal approaches which do not rule out an indefinite interpretation for BSs, e.g., Chierchia (1998), Krifka (2003). The results of our study are also compatible with the proposal that bare NPs in Russian are essentially indefinite and a definite reading is achieved through pragmatic strengthening (Seres & Borik 2021).

On the other hand, our empirical findings are in conflict with Dayal's (2004) proposal, whose prediction – as we mentioned in §1 – is that BSs should never give rise to indefinite readings in regular argument position in languages without articles. Dayal examines the behavior of BSs in Hindi, Russian and Mandarin, and argues that an overt indefiniteness marker has to appear in those contexts where an indefinite reading has to be expressed. This prediction holds for Hindi, where ONE functions as such a marker,[8] but it is very clear that Russian and Polish behave differently. In fact, in our data ONE is only used twice in Russian in the indefinite domain, whereas the Polish data do not contain a single occurrence of this item. Thus, our data allow us to conclude that both Russian and Polish are truly articleless languages where BSs dominate in both definite and indefinite contexts. No competing forms emerge in our study in either of the two contexts in either of the two languages.

In contrast to Russian and Polish, Macedonian uses both BSs and ONE+N constructions. Our data show that Macedonian differs from truly articleless languages, and the construction ONE+N competes with BSs in the indefinite domain in Macedonian. This difference can be illustrated with the translation of example (6a) above into Macedonian: where Russian and Polish use a BS, Macedonian uses ONE+N.

(7) Na agolot na Šimširovata uliča stoeše *edna neobična šarena mačka*, no nikade nemaše mapa.

[8] This result has been confirmed by a parallel corpus study reported in Liu et al. (2023).

If we look outside the Slavic family, our control language, Mandarin, shows a strong tendency for the ONE+N construction to appear in singular indefinite contexts (see Figure 2). Macedonian clearly occupies an intermediate position between Mandarin (relatively low percentage of BSs) and Russian/Polish (predominantly BSs) with respect to the use of BSs in the singular indefinite domain.

Note that this kind of variation in the use of BSs comes out unexpected on most analyses. In general, articleless languages are perceived as a homogeneous group that either do or do not use BSs in a certain domain, but the kind of variation that we see in our data is rather challenging for theoretical approaches. We will come back to this point at the end of this section, but first we will take a better look at the closest competitor of a BS in the indefinite domain, the indefinite marker ONE.

4.2.2 The status of ONE in the indefinite domain

It is well known that the numeral ONE is a predecessor of the indefinite article in many languages (Heine 1997, van Gelderen 2011, among many others). Looking once again at the distribution of nominal forms in the indefinite domain in Figure 2, we observe a clear interaction between BSs and ONE+N constructions: the frequency of ONE+N in our data goes from being at floor in Russian and Polish to a significant percentage in Macedonian and to predominance in Mandarin. This raises a question about the grammatical status of the marker ONE in different languages.

The differences in the use of the ONE+N construction across languages may be accounted for by different stages of its grammaticalisation as an article. Typically, the stages of grammaticalisation of the indefinite article are defined in the following order: 1. the numeral, 2. the presentative marker, 3. the specificity marker, 4. the non-specific marker, 5. the generalised article (Givón 1981, Heine 1997, among others).[9] Even though defining the exact stage of grammaticalisation of ONE in the languages under study is out of the scope of this paper, our data offer several discussion points relevant for the issue.

Our empirical findings for Russian and Polish, where BSs overwhelmingly dominate in the indefinite domain, seem to be in conflict with the proposal of Hwaszcz & Kędzierska (2018), who claim that in Russian ONE is grammaticalised as a presentative marker, that is, it marks a newly introduced referent, which is intended to be used in the subsequent discourse and is usually specific and topical. The authors also claim that in Polish ONE is grammaticalised even further, being

[9] These stages are coarsely defined and may have substages.

used as a specific and sometimes as a non-specific marker. Neither of the two claims is confirmed by our data, as some representative examples can illustrate:

(8) a. English (source): [a N]
The Dursleys had *a small son* called Dudley and in their opinion there was no finer boy anywhere.

b. Russian: [N]
U mistera i missis Darsli byl *malen'kij syn* po imeni Dadli, i, po ix mneniju, ėto byl samyj čudesnyj rebenok na svete.

c. Polish: [N]
Syn Dursleyów miał na imię Dudley, a rodzice uważali go za najwspanialszego chłopca na świecie.

(9) a. English (source): [a N]
He was sure there were lots of people called Potter who had *a son* called Harry.

b. Russian: [N]
Mister Darsli legko ubedil sebja v tom, čto v Anglii živet množestvo semej, nosjaščix familiju Potter i imejuščix *syna* po imeni Garri.

c. Polish: [N]
Mnóstwo ludzi może się nazywać Potter i mieć *syna* Harry'ego.

Example (8) is a typical context where a new specific referent is introduced by a modified indefinite in the source text, which is then rendered by a BS both in Russian and in Polish, just like the non-specific indefinite *a son* in (9). At least in Russian, ONE+N cannot be used instead of N in (8) and (9), unless ONE is interpreted as a numeral.[10] Our data show no sign of any significant difference between Russian and Polish with respect to the grammatical status of ONE: this marker does not show up regularly or systematically in either a presentative, specific or any other type of context.

Macedonian ONE, on the other hand, is more frequent. We have not conducted any specific study of the contexts where ONE appears in Macedonian, as our dataset is too small to yield sensible results, but we can provide some indicative examples here that can help us map out a path for future research. For instance, Macedonian uses a BS in translations of both example (8) and (9) above, but there are other specific and non-specific contexts where ONE+N construction appears:

[10] We thank an anonymous reviewer for stressing this point.

(10) a. English (source): [a N] (non-specific)
My dear Professor, surely *a sensible person* like yourself can call him by his name?

b. Macedonian [one N]
Draga moja profesorke, ne misliš li deka *edna tolku razumna ličnost'* kako što si ti slobodno može da go narekuva po ime?

(11) a. English (source): [a N] (specific)
Professor McGonagall pulled out *a lace handkerchief* and dabbed at her eyes beneath her spectacles.

b. Macedonian [one N]
Profesorkata Mekgonagl izvadi *edno tanteleno maramče* i gi protri očite pod očilata.

The mixed data across specific and non-specific contexts indicate that the ONE+N construction is not really established in these types of contexts. The data obtained in our study are, in principle, in line with Hwaszcz & Kędzierska (2018), who claim that ONE in Macedonian is used with both specific and non-specific indefinite NPs. Our Macedonian data show that both specific and non-specific indefinite NPs may also appear as bare, as illustrated in the above examples, which may indicate a certain degree of optionality in the use of ONE for marking specific and non-specific nominals.[11] This flexibility (possibly translated as optionality) provides a contrast with English and German, languages where an indefinite article is obligatorily used in all the examples discussed in this subsection. Thus, Macedonian does differ from languages with established indefinite articles, and we therefore conclude this discussion by saying that the status of ONE cannot be unequivocally defined as an indefinite article in Macedonian, contra, e.g., Tomić (2006).[12] Rather, ONE is an indefinite marker that might evolve into an article, but further research is needed to substantiate this claim.

4.2.3 Theoretical implications

As the discussion in the previous sections indicates, the main challenge that our data pose for theoretical approaches striving for empirical adequacy is the problem of language variation. The variation in the definite domain, especially in the

[11] One of the limitations of corpus studies is that it is impossible to determine the optionality of an element. In order to research the (non-)obligatoriness of ONE in certain linguistic environments, linguistic experiments with native speakers need to be carried out.

[12] In this respect, the Macedonian data resemble the situation in Bulgarian, as reported in Geist (2013).

distribution of the definite article across languages, is relatively well known and discussed in the semantic literature (e.g., Dryer 2005). Our analysis of the definite article in Macedonian vs. German adds one more study case to this discussion.

In the indefinite domain, however, variation in the distribution of BSs in languages without articles (or without an indefinite article) is less expected. For instance, the approach to BNs in general and BSs in particular developed in Dayal (2004), Dayal (2018), and Dayal & Sağ (2020) is based on the claim that BSs do not allow for indefinite readings in articleless languages. The formal machinery of this approach does not leave much room for variation: the denotation of a noun in regular argument positions is derived by type-shifting operators and, crucially, Dayal's analysis cuts off the possibility of an existential type-shift for BSs. The logic behind this move, we believe, applies universally. Our data for Russian and Polish, though, strongly suggest that there should be an easy way to allow for a BS to appear in the singular indefinite domain, which may be achieved via standard type-shifting operations, like an existential type shift. However, allowing for this type shift to be subject to parametric variation will considerably weaken Dayal's formal theory, at least in the absence of any independent principle underlying such variation.

The Macedonian data, where we see a competition between BSs and the ONE+N construction, suggest that there should be a way to allow for BSs in those contexts where the other construction does not appear on a regular basis. In other words, there should be an account of an interaction between nominal forms that coexist in the indefinite domain. Dayal's approach cannot easily accommodate such interaction either, because ONE+N is predicted to be the only option in the indefinite domain in the absence of an indefinite article. Thus, we conclude that the semantic theory of bare nominals advocated in Dayal (2004), Dayal (2018), and Dayal & Sağ (2020) has considerable difficulties accounting for an overall empirical picture that emerges from our data.[13]

Mandarin, our control language, clearly prefers the ONE+N construction to BNs in the indefinite domain. As Liu et al. (2023) argue, this fact does not really follow from Dayal's analysis either, since in Mandarin, which lacks grammatical number, BNs are expected to easily get an indefinite reading, just like BPs in other languages do. If Mandarin BNs behave like BPs rather than BSs, they are predicted to get a narrow scope indefinite reading and hence, they should be visibly prominent in indefinite (singular and plural) contexts. In our data, however, the ONE+N construction wins over BNs in the singular indefinite domain. In fact, it looks like what Dayal (2004, 2018) predicts for Mandarin occurs in Russian and

[13] See also Liu et al. (2023).

Polish, with a proviso for number marking, and what her analysis predicts for Russian and Polish seems to hold for Mandarin.

An analysis that our data calls for should allow for a formal way to derive an existential interpretation of a BS via type-shifting, but only if there is no competing form with an overt marker that would block this shift. Chierchia's (1998) or Krifka's (2003) classical analyses, for instance, state that while in some languages type shifts are indicated by overt determiners, in languages that lack them, type shifts apply covertly whenever the linguistic context requires it. Covert type-shifting is restricted by the Blocking Principle, which roughly states that if a language has an overt means to express a type shift, then it must be used. This analysis seems to be much better equipped to handle our data. For instance, we have seen no evidence that ONE+N in Russian (*odin N*) and Polish (*jeden N*) function as an article-like expression. Thus, the covert application of the existential type-shift is not blocked, which allows for BSs to be freely used in indefinite contexts. For Macedonian, a language with an emerging indefinite marker ONE, the existential type shift would be blocked for a BN only in those contexts where *eden* appears. Our cross-linguistic data provide a serious argument in favour of a classical blocking semantic analysis of bare nominals, in which fine-grained variation in the distribution of bare nominals follows from the broader/narrower use of article-like expressions.

5 Conclusions

In this paper, we have reported the results of a parallel translation corpus study on the distribution of BSs in three Slavic languages, Russian, Polish and Macedonian. We built our corpus on the text of the first chapter of *Harry Potter and the Philosopher's Stone* and complemented the results obtained for Slavic languages with the results for Mandarin as a control language for Russian and Polish, and German as a control language for Macedonian.

In view of the empirical data presented here, it can be concluded that Russian and Polish are truly articleless languages and freely allow their BSs to take on definite and indefinite readings across domains. In Macedonian, BSs are restricted to the indefinite domain where they compete with the indefinite marker ONE, whereas in the definite domain, Macedonian uses the definite article, just as expected. Therefore, we conclude that Macedonian is a language with a definite article and with an emerging indefinite marker whose exact grammatical status requires further empirical investigation.

Slavic languages present challenging theoretically relevant contrasts with their control languages. In case of Macedonian, we have stressed the need to further scrutinize the conditions and the contexts where the definite article is used because we have shown that the overlap between the definite articles in Macedonian and German is partial. We also see the need to extend the investigation to the plural domain to get a full picture of the distribution of the definite article in Macedonian. As for Russian and Polish, they present a striking contrast with Mandarin in the indefinite singular domain, where the two Slavic languages show a clear preference for BSs and Mandarin opts for the one+N construction as a counterpart of the English *a N*. Macedonian occupies an intermediate position: one+N is used rather frequently in Macedonian, but not as often as in Mandarin singular indefinite contexts.

We have argued that these contrasts call for a theoretical approach where the observed variation in the distribution of BSs and competing forms can be naturally accounted for. We suggest that the Blocking Principle as formulated in Chierchia (1998) can serve as a foundation for such an approach.

Abbreviations

f	feminine
m	masculine
n	neuter
pl	plural
sg	singular

Acknowledgments

The first and the last authors acknowledge financial support by Spanish MINECO (PID2020–112801GB–100). The last author also acknowledges support by Generalitat de Catalunya (2021-SGR-00787). The second author acknowledges support by the Dutch Research Council (NWO) (#360-80-070). The third author acknowledges support under her CSC grant. At the time of writing the article, the last author was a Margarita Salas fellow, funded by the European Union – Next Generation EU.

References

Barker, Christian. 2005. Possessive weak definites. English (US). In Kim, Ji-yung, Lander, Yury, Partee & H. Barbara (eds.), *Possessives and beyond: Semantics and syntax*. 89–113. Amherst, MA: GLSA Publications.

Bošković, Željko. 2008. What will you have, DP or NP? In Emily Elfner & Martin Walkow (eds.), *Proceedings of NELS 37*, 101–114. University of Illinois Urbana-Champaign.

Bouma, Gerlof J. 2009. Normalized (pointwise) mutual information in collocation extraction. In Christian Chiarcos, Richard Eckart de Castilho & Manfred Stede (eds.), *Proceedings of the Biennial International Conference of the German Society for CLLT*, 31–40. Tübingen: Gunter Narr Verlag.

Bremmers, David, Jianan Liu, Martijn van der Klis & Bert Le Bruyn. 2021. Translation mining: Definiteness across languages (A reply to Jenks 2018). *Linguistic Inquiry* 53(4). 735–752. DOI: 10.1162/ling_a_00423.

Chierchia, Gennaro. 1998. Reference to kinds across languages. *Natural Language Semantics* 6(4). 339–405. DOI: 10.1023/a:1008324218506.

Condoravdi, Cleo. 1994. *Descriptions in context*. New Haven: Yale University. (Doctoral dissertation).

Dayal, Veneeta. 2004. Number marking and (in)definiteness in kind terms. *Linguistics and Philosophy* 27(4). 393–450. DOI: 10.1023/b:ling.0000024420.80324.67.

Dayal, Veneeta. 2018. (In)definiteness without articles: Diagnosis, analysis, implications. In Ghanshyam Sharma & Rajesh Bhatt (eds.), *Trends in Hindi linguistics*, 1–26. Berlin: De Gruyter Mouton. DOI: 10.1515/9783110610796-001.

Dayal, Veneeta & Yağmur Sağ. 2020. Determiners and bare nouns. *Annual Review of Linguistics* 6(1). 173–194. DOI: 10.1146/annurev-linguistics-011718-011958.

de Swart, Henriëtte. 2020. Double negation readings. In Viviane Déprez & M. Teresa Espinal (eds.), *The Oxford handbook of negation*, 479–496. Oxford: Oxford University Press. DOI: https://doi.org/10.1093/oxfordhb/9780198830528.013.26.

Dryer, Matthew S. 2005. Definite articles. In Martin Haspelmath, Matthew S. Dryer, David Gil & Bernard Comrie (eds.), *The world atlas of language structures*, 154–157. Oxford: Oxford University Press.

Dryer, Matthew S. & Martin Haspelmath (eds.). 2013. *WALS online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. https://wals.info/.

Elbourne, Paul D. 2008. Demonstratives as individual concepts. *Linguistics and Philosophy* 31(4). 409–466. DOI: 10.1007/s10988-008-9043-0.

Friedman, Victor A. 1993. Macedonian. In Bernard Comrie & Greville G. Corbett (eds.), *The Slavonic Languages*, 249–305. London: Routledge.

Fuchs, Martín & Paz González. 2022. Perfect-perfective variation across Spanish dialects: A parallel-corpus study. *Languages* 7(3). 166. DOI: 10.3390/languages7030166.

Gehrke, Berit. 2022. Differences between Russian and Czech in the use of Aspect in narrative discourse and factual contexts. *Languages* 7(2). 155. DOI: 10.3390/languages7020155.

Geist, Ljudmila. 2010. Bare singular NPs in argument positions: Restrictions on indefiniteness. *International Review of Pragmatics* 2(2). 191–227. DOI: 10.1163/187731010x528340.

Geist, Ljudmila. 2013. Bulgarian *edin*: The rise of an indefinite article. In Uwe Junghanns, Dorothee Fehrmann, Denisa Lenertová & Hagen Pitsch (eds.), *Formal Description of Slavic Languages: The Ninth Conference. Proceedings of FDSL 9, Göttingen 2011*, 125–148. Frankfurt am Main: Peter Lang.

Givón, Talmy. 1981. On the development of the numeral 'one' as an indefinite marker. *Folia Linguistica Historica* 15(2-1). 35–54. DOI: doi:10.1515/flih.1981.2.1.35.

Heine, Bernd. 1997. *Cognitive foundations of grammar*. Oxford: Oxford University Press. DOI: 10.1093/oso/9780195102512.001.0001.

Hwaszcz, Krzysztof & Hanna Kędzierska. 2018. The rise of an indefinite article in Polish: An appraisal of its grammaticalisation stage (Part 1). *Studies in Polish Linguistics* 13. 93–121. DOI: 10.4467/23005920SPL.18.005.8744.

Ionin, Tania. 2013. Pragmatic variation among specificity markers. In Cornelia Ebert & Stefan Hinterwimmer (eds.), *Different kinds of specificity across languages*, 75–103. Dordrecht: Springer. DOI: 10.1007/978-94-007-5310-5_4.

Krifka, Manfred. 2003. Bare NPs: Kind-referring, indefinites, both, or neither? In Robert B. Young & Yuping Zhou (eds.), *Proceedings of SALT 13*, 180–203. DOI: https://doi.org/10.3765/salt.v13i0.2880.

Le Bruyn, Bert & Henriëtte de Swart. 2022. *Cross-linguistic semantics: Methodological advances*. Course materials, ESSLLI. shorturl.at/doqV6.

Le Bruyn, Bert, Martín Fuchs, Martijn van der Klis, Jianan Liu, Chou Mo, Jos Tellings & Henriëtte de Swart. 2022. Parallel corpus research and target language representativeness: The contrastive, typological, and translation mining traditions. *Languages* (3). 176. DOI: 10.3390/languages7030176.

Li, Xuping. 2021. The Semantics of Chinese Noun Phrases. In *Oxford Research Encyclopedias: Linguistics*. Oxford: Oxford University Press. https://oxfordre.com/linguistics/view/10.1093/acrefore/9780199384655.001.0001/acrefore-9780199384655-e-886.

Liu, Jianan, Shravani Patil, Daria Seres, Olga Borik & Bert Le Bruyn. 2023. 'Articleless' languages are not created equal. In Maria Onoeva, Anna Staňková & Radek Šimík (eds.), *Proceedings of Sinn und Bedeutung 27*, 381–398. DOI: 10.18148/sub/2023.v27.1076.

Lyons, Christopher. 1999. *Definiteness*. Cambridge: Cambridge University Press. DOI: https://doi.org/10.1017/CBO9780511605789.

Mo, Chou. 2022. *The compositionality of Mandarin Aspect: A parallel corpus study*. Utrecht: University Utrecht. (Doctoral dissertation). https://www.lotpublications.nl/Documents/628_fulltext.pdf.

Molinari, Luca. 2022. The syntax of Polish *jeden* 'one' as an indefinite determiner. *Annali di Ca' Foscari. Serie occidentale* 56. 63–84. DOI: http://doi.org/10.30687/AnnOc/2499-1562/2022/10/004.

Mulder, Gijs, Gert-Jan Schoenmakers, Olaf Hoenselaar & Helen de Hoop. 2022. Tense and aspect in a Spanish literary work and its translations. *Languages* 7(3). 217. DOI: https://doi.org/10.3390/languages7030217.

Roberts, Craige. 2002. Demonstratives as definites. In Kees van Deemter & Roger Kibble (eds.), *Information sharing: Reference and presupposition in language generation and interpretation*, 89–136. Stanford: CSLI Press.

Rudin, Catherine. 2021. Demonstratives and definiteness: Multiple determination in Balkan Slavic. In Andreas Blümel, Jovana Gajić, Ljudmila Geist, Uwe Junghanns & Hagen Pitsch (eds.), *Advances in formal Slavic linguistics 2018* (Open Slavic Linguistics 4), 305–338. Berlin: Language Science Press. DOI: 10.5281/zenodo.5483114.

Seres, Daria & Olga Borik. 2021. Definiteness in the absence of uniqueness: the case of Russian. In Andreas Blümel, Jovana Gajić, Ljudmila Geist, Uwe Junghanns & Hagen Pitsch (eds.), *Advances in formal Slavic linguistics 2018* (Open Slavic Linguistics 4), 339–363. Berlin: Language Science Press. DOI: 10.5281/zenodo.5155544.

Švedova, Natalija J. 1980. *Russkaja grammatika*. Moskva: Nauka.

Tomić, Olga Mišeska. 2006. *Balkan Sprachbund morpho-syntactic features*. Dordrecht: Springer. DOI: 10.1007/1-4020-4488-7.

Topolinjska, Zuzanna. 2009. Definiteness (synchrony). In Sebastian Kempgen, Peter Kosta, Tilman Berger & Karl Gutschmidt (eds.), *Die slavischen Sprachen / the Slavic languages*, 176–188. Berlin: De Gruyter Mouton. DOI: doi:10.1515/9783110214475.1.3.176.

van Gelderen, Elly. 2011. *The linguistic cycle: Language change and the language faculty*. New York: Oxford University Press. DOI: 10.1093/acprof:oso/9780199756056.001.0001.

Weiss, Daniel. 2004. The rise of an indefinite article: The case of Macedonian *eden*. In Walter Bisang, Nikolaus P. Himmelmann & Björn Wiemer (eds.), *What makes grammaticalization? A look from its fringes and its components*, 139–168. Berlin: De Gruyter Mouton. DOI: doi:10.1515/9783110197440.2.139.