Bare Nouns in Slavic and beyond

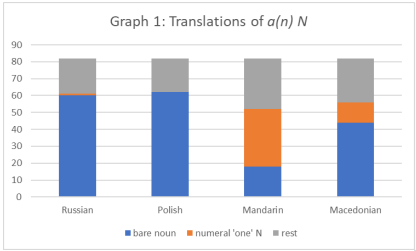
Olga Borik[1], Bert Le Bruyn[2], Jianan Liu[2], Daria Seres[3]
[1]*UNED,* [2]*Utrecht University,* [3]*UAB/Humboldt-Universität Berlin*

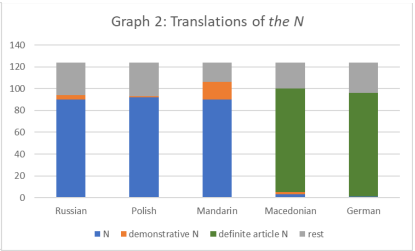
1. Introduction | We study the syntax-semantics interface of reference in Russian, Polish and Macedonian. In terms of (in)definiteness marking, the literature typically divides these languages into: (i) no articles (Russian/Polish), (ii) definite article only (Macedonian). For Russian and Polish, we ask whether (singular) bare nouns (BNs) take on both definite and indefinite readings – in line with the view of traditional Slavic approaches – or only definite readings – in line with Dayal (2004). To complement our Russian and Polish data, we introduce Mandarin as an articleless control language. For Macedonian, we evaluate its status as a language without an indefinite article by comparing its use of BNs in indefinite contexts to that of Russian, Polish and Mandarin. To evaluate its status as a language with a definite article, we add German as a control language.

2. Methodology | We adopt a *Translation Mining* parallel corpus approach to cross-linguistic semantics (Bremmers et al. 2021). We selected all (in)definite referential expressions (*a(n) N*, *the N*, *N-s*, *the N-s*) from the first chapter of *Harry Potter and the Philosopher's Stone* with their aligned translations in Russian, Polish, Macedonian, Mandarin and German (N=284). The *Translation Mining* approach builds on the assumption that the meaning of a source text is kept constant in its translations, so we assume that the meanings of the translations of the different referential expressions are as closely related to each other as the grammars of the respective languages allow them to be. Plural nominals are a relatively infrequent category in our dataset (*the N-s* ($n = 34$) and *N-s* ($n = 44$)) and interact with proper names (e.g., 'The Potters', 'The Dursleys'). We consequently decided to run our quantitative analyses for the singular paradigm and to only include the plural paradigm in our qualitative interpretation of the data.

To evaluate whether Russian and Polish BNs take on definite and indefinite readings, we compare Russian, Polish and Mandarin translations of *a(n) N* ($n = 82$) and *the N* ($n = 124$) and check whether the distribution of BNs interacts with that of the numeral *one* (for *a(n) N*) or demonstratives (for *the N*). To evaluate whether Macedonian BNs take on indefinite readings, we look into translations of *a(n) N* and compare the use of Macedonian BNs to that of Russian, Polish and Mandarin BNs. To evaluate the status of Macedonian as a language with a definite article, we compare the German and Macedonian translations of *the N* and the bi-directional mapping patterns between their respective definite articles.

3. Results | Graphs 1 and 2 present the translation data of *a(n) N* and *the N*. Table 1 presents the bi-directional mapping data between the German and the Macedonian (singular) definite articles.





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

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

Observation 1 | Graph 1 shows that the BN is the default option for rendering singular indefinites in Russian and Polish whereas Mandarin also relies on the numeral *one*. The differences in distributions of BNs and the numeral are not significant for Russian and Polish ($p = 0.5$, Fisher's

Exact Test (FET)) but they are for Mandarin/Russian ($p < 0.001$, FET) and Mandarin/Polish ($p < 0.001$, FET). **Observation 2 |** Graph 2 shows that the BN is the default option for rendering singular definites in Russian and Polish whereas Mandarin has a slightly higher tendency to resort to demonstratives. The differences in distribution are not significant for Russian and Polish ($p = 0.37$, FET) but they are for Mandarin/Polish ($p < 0.001$, FET) and for Mandarin/Russian ($p = 0.016$, FET). **Observation 3 |** Graph 1 shows that *a(n) N* is more often translated with the numeral *one* in Macedonian than in Russian and Polish but not as often as in Mandarin. The differences are significant for Macedonian/Russian ($p < 0.001$, FET), Macedonian/Polish ($p < 0.001$, FET) and Macedonian/Mandarin ($p < 0.001$, FET). **Observation 4 |** Graph 2 shows that the distributions of the Macedonian and German definite articles are close to identical in their translations of *the N*. However, Table 1 shows that the two articles only overlap in part. Adopting Normalized Pointwise Mutual Information (NPMI, Bouma (2009)) as a bidirectional measure for parallel data, we find that the two articles' NPMI reaches 0.48 (with a maximum of 1). The likelihood of the two articles occurring in the same contexts is higher than chance but not at ceiling.

4. Discussion | Our data convincingly show that **Russian and Polish** freely use BNs in singular indefinite and definite contexts, in accordance with the Slavic descriptive literature. This free use of BNs is in sharp contrast with Mandarin, where the numeral *one* seems to be the default option in indefinite contexts and the demonstrative is competing with BNs in the definite domain (Obs 1 & Obs 2). Russian and Polish thus turn out to be truly articleless languages (*contra*, i.a., Hwascz & Kedzierska 2018), unlike Mandarin. For a formal semantic account of BNs in articleless languages, the contrasts between Russian/Polish and Mandarin favor a classical blocking analysis (e.g., Krifka 2004, de Swart & Zwarts 2010) and argue against an analysis of BNs as uniformly conveying definiteness (Dayal 2004). For **Macedonian**, our data show that it differs from truly articleless languages like Russian and Polish and suggest that it has both a definite article (Obs.4), and an emerging indefinite marker (Obs. 3). As for Macedonian *one*, its distribution is different from that of its Mandarin counterpart, and its status cannot be unequivocally defined as indefinite article (e.g., Weiss 2004). In this respect, Macedonian data resemble the situation in Bulgarian, as reported in, for instance, Geist (2013). Our data also suggest caution for the analysis of the Macedonian definite article: despite the fact that in definite contexts it occurs equally frequently as its German counterpart, their low NPMI value suggests that they are not identical. This is a more general issue with definite articles across the languages in our data which is also reflected in the plural domain: translations of English bare plurals often lead to the use of definite articles. This happens in Macedonian in generic contexts as in (1), and in both Macedonian and German in indefinite/existential contexts as in (2).

(1) **Cats** couldn't read maps or signs. (Eng) | **Mački<u>te</u>** ne možat da čitaat ni mapi ni oznaki. (Mac) | **Katzen** konnten weder Karten noch Schilder lesen. (Ger)

(2) And finally, **bird-watchers** everywhere have reported that the nation's owls have been behaving very unusually today. (Eng) | I konečno **nabljuduvačite** na ptici od site strani javija deka buvo – vite […] (Mac) | Wie **die Vogelkundler** im ganzen Land berichten [...] (Ger)

The question that data like (2) raise is how strongly presuppositional definite articles really are across languages (cf. Šimík & Demian 2020).

5. Conclusion | We conclude (i) that Russian and Polish are truly articleless languages and freely allow their BNs to take on definite and indefinite readings, (ii) that Macedonian has a definite article and an emerging indefinite marker whose status requires further scrutiny, (iii) that our cross-linguistic data argue against a uniform definiteness semantics of singular BNs (Dayal 2004) and

can best be accounted for with a classical blocking analysis in which fine-grained variation in the distribution of BNs follows from the broader/narrower use of article-like expressions.

References (selected) | **Bouma** (2009). NPMI in collocation extraction. *Proceedings of the Biennial International Conference of the German Society for CLLT*. **Bremmers** et al. (2021). Translation Mining: Definiteness across Languages—A Reply to Jenks (2018). *Linguistic Inquiry*, 1-30. **Hwaszcz & Kędzierska** (2018). *Studies in Polish Linguistics*, 13 (1). 93–112. **de Swart** & **Zwarts** (2010). Optimization principles in the typology of number and articles. *The Oxford Handbook of Linguistic Analysis*. **Geist** (2013). Bulgarian edin: The rise of an indefinite article. *Proceedings of FDSL 9*. **Šimík** & **Demian** (2020). *Journal of Semantics* 37(3). 311–366. **Weiss** (2004). The rise of an indefinite article: The case of Macedonian eden. *What makes Grammaticalization? A Look from its Fringes and its Components*.