FASL **26, ###-###**
Michigan Slavic Publications
2018

Definiteness of bare NPs as a function of clausal position: A corpus study of Czech[*]

Radek Šimík
Humboldt-Universität zu Berlin

Markéta Burianová
Praha (without affiliation)

[Preprint, 2018-02-08]

We provide novel corpus evidence that the definiteness of a bare (non-determined) noun phrase (NP) depends on the position of the NP in the clause, thus corroborating an intuition common among Slavic linguists since the 1970s. The most significant finding is that indefinite bare NPs are very unlikely to occur in clause-initial position, which is in line with Geist's (2010) predictions. A further notable result is that definiteness of a bare NP is affected by its absolute position in the clause (clause-initial vs. clause-final), but not its position relative to the verb (preverbal vs. postverbal). This has worrisome implications for theories according to which the verb partitions the clause into a presupposed and non-presupposed area (Kučerová 2007 and, with some reservations, Diesing 1992). Finally, we are able to tease apart the effect of clausal position from the effect of syntactic function, to the effect that being a subject or

[*] Besides FASL26 at Urbana-Champaign, the material was presented in various stages of development in Berlin, Köln, Tübingen, and Praha. We'd like to thank the audiences and particularly Petr Biskup, Jan Chromý, Berit Gehrke, Stephanie Harves, Tania Ionin, Roland Meyer, and Natalia Slioussar. We're also grateful to the two anonymous reviewers of this paper, whose comments led to various improvements. All remaining errors are solely ours. The work was partly supported by the German Research Foundation (DFG), via the project *Definiteness in articleless Slavic languages* granted to RŠ.

object (properties that strongly correlate with being clause-initial and clause-final, respectively) does not increase the likelihood of bare NPs to be interpreted as definite or indefinite, respectively.

The paper is organized as follows. Section 1 introduces the core empirical issue – what we call the definiteness–word order interaction. Theoretical approaches to this interaction and their predictions are discussed in section 2. Section 3 is the main contribution of this paper – a corpus study of the definiteness–word order interaction in Czech, designed to test for the validity of what we call the absolute position hypothesis (effect of clause-initiality/finality on (in)definiteness) and the relative position hypothesis (effect of pre-/postverbality on (in)definiteness). Section 4 discusses and rules out the potentially confounding factor of syntactic function (subject/object). Section 5 provides a discussion of the results and their theoretical implications.

1 Definiteness–word order interaction

It is a common and long-standing observation that the definiteness of bare NPs in articleless Slavic languages depends, at least in part, on word order. Descriptively speaking, a clause-final bare NP tends to be interpreted as indefinite and a clause-initial bare NP tends to be interpreted as definite. The observation has gradually been qualified (see e.g. Geist 2010): it is now often claimed that there is an effect of the initial but not the final position; the latter is believed to remain neutral with respect to (in)definiteness (consider also Chvany's and Krámský's intuition about *stole* 'table' in (1b)/(2b)).

(1) a. Na stole je **kniha**.
on table is book
'There is **a** book on the table.'

b. **Kniha** je na stole.
book is on table
'**The** book is on the table.'
[Cz; Krámský 1972:42]

(2) a. Na stole stojala **lampa**.
on table stood lamp
'There was **a lamp** on the desk.'

b. **Lampa** stojala na stole.
lamp stood on table
'**The** lamp was on a/the desk.'
[Ru; Chvany 1973:266]

(3) W pokoju siedziała dziewczyna. [Po; Szwedek1974:215]
in room sat girl
'There was a girl sitting in the room.'
a. Wszedł **chłopiec**.
entered boy
'**A boy** entered.'
b. **Chłopiec** wszedł.
boy entered
'**The boy** entered.'

Despite the fact that the definiteness–word order interaction has been well-known for half a century, there are important unresolved questions, the answers to which would be highly informative for the theories that aim to explain or model the interaction.

2 Approaches to the definiteness–word order interaction

Consider our examples (1)–(3) again. We stated, following a common opinion, that clause-initiality correlates with definiteness and, potentially, clause-finality with indefiniteness. But there are at least two other factors that could be held responsible for the effect: position with respect to the verb (preverbal → definite, postverbal → indefinite) and prosodic prominence (non-prominence → definite, prominence → indefinite). None of these three perspectives on the data pattern is a priori implausible, but each is a proxy for a potentially very different theory: clause-initiality is expected to correlate with topichood (and thereby definiteness), preverbality with presuppositionality (Diesing 1992, Kučerová 2007), and prosodic non-prominence with givenness, which in turn correlates with anaphoricity – one common kind of definiteness (Szwedek 2011). Yet another plausible analysis relies on the relative position of NPs: if indefinite NPs cannot precede definite NPs – as suggested for Russian double objects by Titov 2017 – then the definiteness of the subject NP in (1b/2b) follows from the definiteness of *stole* 'table'. Finally, one could expect there to be an effect of syntactic function (subject vs. object) or perhaps a devoted syntactic "subject position" (such as SpecTP). The idea that subjecthood is associated with definiteness goes back to Li & Thompson's (1976) work on Chinese.

Table 1 summarizes the landscape of (i) plausible empirical generalizations that subsume the definiteness–word order interaction, (ii) the hypotheses that these could be a proxy for, and (iii) selected existing proposals that entail one of the hypotheses.

	GENERALIZATION	HYPOTHESIS		PROPOSALS
A	**Initial** **→ Definite**	Initial → Topic	Topic → Referential Referential ≈ Definite	Hlavsa 1975 Chvany 1983 King 1995 Geist 2010
B	**Subject** **→ Definite**	Subject → Topic		Li & Thompson 1976 Matthews & Yip 1994 Jenks to appear
C	**Preverbal** **→ Definite**	Preverbal → External to vP	Out of vP → Presuppositional Presuppositional ≈ Definite	Diesing 1992 Junghanns & Zybatow 1997 Brun 2001 Späth 2003 Biskup 2011 Mykhaylyk 2011
D		Preverbal → Pre-G-operator	Pre-G-operator → Presuppositional Presuppositional ≈ Definite	Kučerová 2007
E	**Precedes Referential** **→ Definite**	Precedes Referential → Referential (subcase of *Non-Prominent precedes Prominent) Referential ≈ Definite		Titov 2017 (extrapolation)
F	**Unstressed** **→ Definite**	Unstressed → Given	Given ≈ Anaphoric Anaphoric ≈ Definite	Szwedek 2011

Table 1: Approaches to the definiteness–word order interaction

A number of clarification remarks are due. To start with, the hypotheses listed in Table 1 entail a relation between some *formal property* (e.g. position) and some *semantic property* (e.g. referentiality), whereby the semantic property is not specifically definiteness. The notions of referentiality (A, B, E) and presuppositionality (C, D) are applicable to indefinites, too (so called specific indefinites). It has been argued, however, that bare NPs – if indefinite – cannot be specific. This

is a reasonable conjecture not just for Slavic (see Geist 2010 on Russian), but possibly for bare NPs in general (e.g. Dayal 2011). Thus, bare NPs are either definite or non-specific indefinite. If this assumption is correct, Hypotheses A through E establish a relatively safe connection between the respective formal property and definiteness.

The notion of anaphoricity, implicated under F, represents a more complicated case. In Szwedek's (2011) work, the lack of prosodic prominence is directly tied to anaphoricity. From a broader perspective, however, this is somewhat unorthodox. Most relevant literature postulates a connection between lack of stress and *givenness* (starting with Schmerling 1976; more recently Féry & Samek-Lodovici 2006) and while givenness is often defined in terms of discourse anaphoricity (e.g. Rochemont 1986, Schwarzschild 1999), it does not necessarily entail definiteness (Umbach 2001). I.e., indefinites (even non-specific ones) can also be given and subject to avoiding prosodic prominence (see Šimík & Wierzba 2015 for an experimental argument for Czech).

There is a caveat that concerns hypothesis C, which states that NPs located externally to vP are presuppositional. It is certainly a simplification to assume that whatever is preverbal is external to vP. While the vP-edge is not an unlikely position of the (finite) verb in Slavic languages (Bailyn 2004, Wiland 2009), it is by far not a settled matter (cf. Migdalski 2006). This problem is sidestepped in the approach of Kučerová (2007) (hypothesis D), which entails an intimate connection between overt verb position and the partition into the presuppositional and non-presuppositional area (mediated by the G-operator). Including preverbality as a factor will thus directly test a prediction of Kučerová's (2007) approach to the definiteness–word order interaction.

3 Corpus study

3.1 Motivation and aim

The above-mentioned approaches have rarely (if ever) been explicitly and systematically compared. Much of the existing work concentrates on proving a particular theory and centers around isolated and ad hoc observations. While we consider the development of theories about the definiteness–word order interaction important, we believe that a solid understanding of the empirical matter is equally important. In our view, there is plenty of work that needs to be done in order to establish even

the basic empirical generalization, namely which factor or factors are behind the pertinent interaction. Further unresolved questions are whether and how these factors interact and whether they are subject to cross-linguistic variation.

The present work supplies corpus evidence from Czech, which sheds new light on generalizations/hypotheses A through D. More particularly, our study is designed to directly assess the adequacy of generalization A (Initial → Definite), as compared to generalization C/D (Preverbal → Definite). An additional post-hoc analysis also tests for the adequacy of generalization B (Subject → Definite) and compares it with generalization A, for which it constitutes a potential confound.

While we find a strong dependency of definiteness on clause-initiality (and finality), our data support neither the view that definiteness depends on pre-/postverbality, nor that it depends on subject/objecthood. We interpret these results as a step towards reducing the hypothesis space. We will further show that the most clearly pronounced restriction is one on clause-initial indefinites, in line with Geist (2010).

3.2 Hypotheses

The two hypotheses that we aim to compare are in (4) and (5).[1]

(4) ABSOLUTE POSITION HYPOTHESIS: The absolute clausal position of bare NPs (initial/final) has an impact on their (in)definiteness.
   a. Clause-initial bare NPs are more likely to be definite.
   b. Clause-initial bare NPs are less likely to be indefinite.
   c. Clause-final bare NPs are more likely to be indefinite.
   d. Clause-final bare NPs are less likely to be definite.

(5) RELATIVE POSITION HYPOTHESIS: The position of bare NPs relative to the verb (pre-/postverbal) has an impact on their (in)definiteness.
   a. Preverbal bare NPs are more likely to be definite.
   b. Preverbal bare NPs are less likely to be indefinite.
   c. Postverbal bare NPs are more likely to be indefinite.
   d. Postverbal bare NPs are less likely to be definite.

[1] For presentational and rhetoric purposes, we treat (in)definiteness as the dependent variable, such that position is assumed to have a (causal) impact on (in)definiteness. Technically, however, we can only measure a correlation between (in)definiteness and position. It cannot be ruled out that it is (in)definiteness that affects position.

The one-tailed directional sub-hypotheses in (a) through (d) are expected manifestations of the respective "matrix" hypotheses. They need not all be true in order for the matrix hypothesis to hold. As discussed above, the intuitions expressed in the literature give us a reason to believe that (4a/b) are more likely to hold than (4c/d). A comparable expectation holds for (5). Biskup (2011), for instance, claims that bare NPs in the preverbal position (in the CP phase) are obligatorily specific or definite, but the postverbal position (the vP phase) has no effect on NP interpretation. This is inherited from the classical works on semantic effects of scrambling, particularly Diesing (1992) and de Hoop (1992).

3.3 Method, material, annotation

Our basic method is very simple: we annotated bare NPs for (in)definiteness and looked whether their (in)definiteness correlates with (i) the absolute position in the clause and (ii) the relative position to the verb.[2] Our sample was drawn from the Czech National Corpus and particularly from the SYN2010 subcorpus – a representative corpus of synchronic written Czech (at the time when the research was carried out). In order to ensure a certain stylistic homogeneity and at the same time an affinity to colloquial Czech, we concentrated on fiction only. As argued in Berger (1993), style and register are factors relevant for the formal expression of definiteness in Czech. However, we had no intention and capacity to include genre as a factor into the analysis. We further excluded translations, in order to avoid interference from other languages. The resulting subcorpus of SYN2010 had about 15 million tokens.

We proceeded by a search for nouns, followed by an automatic removal of proper names and nouns with determiners. Out of the resulting 2.37 million tokens (0.16 i.p.m.), we drew a random sample of 800 noun (phrase) occurrences. These underwent further manual filtering, whereby the following NPs were removed from the sample:

[2] The corpus research originated as Burianová (2016), which was carried out under the supervision of RŠ. As presented here, the corpus study consists in a re-annotation of the original sample by RŠ; the raw results remain largely unaffected, but the present work departs from Burianová (2016) significantly in its theoretical anchoring. The raw data, annotations, analyses, as well as selected glossed corpus examples are made available at https://osf.io/jauhw.

- the remaining determined NPs,
- NPs that were parts of idioms or collocations (motivated by the assumption that these cannot be meaningfully (in)definite),
- NP fragments or appositions (no clear clausal position),
- attributive NPs (significantly reduced freedom of position),
- predicative NPs (no referential properties, hence no clear definiteness),
- kind-denoting NPs (inherently hard to judge for definiteness), and
- cases where definiteness was simply too hard to decide on.

We ended up with a final sample of 315 bare NP occurrences, which then entered an annotation for (i) DEFINITENESS (definite, indefinite), (ii) ABSOLUTE POSITION (initial, medial, final), and (iii) RELATIVE POSITION (preverbal, postverbal). For each occurrence we included an auxiliary annotation for SYNTACTIC FUNCTION (subject, object, adverbial), DEFINITENESS TYPE (unique, anaphoric, plus a number of subtypes of each), INDEFINITENESS TYPE (presentational, quantified-over), REFERENT TYPE (entity, event, temporal interval, …), GRAMMATICAL NUMBER (singular, plural), MODIFICATION (none, premodified, postmodified, both), GIVENNESS (given, new), and FOCUS (narrow focus, part of focus, part of background).[3] For an analysis of some of these auxiliary factors (e.g. modification), see Burianová (2016). In this paper, we will only concentrate on syntactic function (see section 4).

The annotation of the two position factors was not particularly complicated and included only a number of relatively uncontroversial assumptions, namely: (i) position of the whole NP was considered (not just the N from the concordance), (ii) clause-initial function words, such as conjunctions or complementizers, were ignored (i.e., in sequences like 'although new car…', 'new car' would count as an initial NP), and (iii) the position of the lexical verb was considered (not, e.g., of an auxiliary).

The annotation of definiteness was, expectedly, less trivial. For each NP occurrence, we inspected the preceding context (up to where it felt necessary, often the whole paragraph) and considered (i) whether adding an overt indefinite (e.g. *nějaký* 'some') or definite (demonstrative *ten*)

[3] The annotation of the third core information structural category, namely topic (vs. comment), was not performed (despite its relevance), because it has proved to be particularly difficult (see Cook & Bildhauer 2013), and would require an extra study.

determiner to the NP is possible without a meaning change, (ii) whether uniqueness of the referent is satisfied – by means of contextual bridging, binding, etc., (iii) whether the translation to English yielded a definite or indefinite NP (a method used for some cases by RŠ). Our annotation methodology was, of course, not without shortcomings. The annotation was performed by the authors of the study and was sequential – first done by MB (Burianová 2016) and later revised by RŠ. Because the two annotations were not mutually independent, there was no way to meaningfully measure the interannotator agreement (Cohen 1960). The annotation procedure was relatively informal: there was no decision tree and the three above-mentioned criteria were used in a case-by-case fashion – depending on which one(s) suited best the occurrence at hand. Despite this, the annotation was done with great care and in an unbiased manner, so we are confident that it represents a robust and useful approximation of the facts.

3.4 Results

Table 2 presents the results qua the absolute position hypothesis. The numbers in boldface represent the attested frequencies; for instance, of all the 315 occurrences, there were 61 definite bare NPs in clause-initial position. The bracketed numbers indicate the frequencies expected under the null hypothesis; for instance, had there been no effect of position on definiteness (or of definiteness on position), we would have found about 43 definite bare NPs in the initial position.[4]

	INITIAL		FINAL		MEDIAL		TOTAL
DEF	**61**	(43.4)	**88**	(113.0)	**58**	(50.6)	207
INDEF	**5**	(22.6)	**84**	(59.0)	**19**	(26.4)	108
TOTAL	66		172		77		315

Table 2: Results qua the absolute position hypothesis

Overall, there were more definite than indefinite NPs (207:108). Higher frequency of definites should not come as a surprise, however, as an auxiliary search of the German corpus (using articles as a proxy for

[4] Expected frequencies can be intuitively grasped if one realizes (by inspecting the table) that their ratio matches the ratio of attested total frequencies (e.g. 43.4 : 113.0 : 50.6 (DEF) ~ 66 : 172 : 77 (TOTAL) or 43.4 : 22.6 (INITIAL) ~ 207 : 108 (TOTAL)).

definiteness) yields a 4:1 ratio in favor of definites. If anything, we should therefore be surprised to have found so few definites.[5] But let us leave the issue at that and move on to our main interest: the definiteness–word order interaction. We find that the absolute position hypothesis is confirmed: the position of the NP has an effect on its definiteness ($\chi^2(2) = 40.22$, $p < .001$, $n = 315$) – with numbers clearly departing from the null hypothesis in initial and final position. We find more initial definites & fewer initial indefinites than expected ($\chi^2(1) = 20.90$, $p < .001$, $n = 66$) and fewer final definites & more final indefinites than expected ($\chi^2(1) = 16.16$, $p < .001$, $n = 172$). Medial position has no or only marginal effect on definiteness ($\chi^2(1) = 3.16$, $p = .08$, $n = 77$).[6]

Let us now turn to the relative position hypothesis. In order to assess this hypothesis properly, we need to focus our attention on the 77 medial NPs, i.e. NPs that are neither initial, nor final, as represented in Table 3. The reason for that is that if we included initial and final NPs into the dataset of pre- and postverbal NPs, respectively, we would not be able to tear apart the effect of pre- vs. postverbality from the one of initality vs. finality. In fact, because the frequency of initial/final NPs is higher than the one of medial NPs, we would see *mainly* the effect of initiality vs. finality. This is a trap that Czardybon, Hellwig & Petersen (2014) fell into when they concluded – based on a Polish corpus study, similar to the present one – that preverbality increases the likelihood of definiteness and postverbality of indefiniteness: they included initial and final NPs into their dataset and it is thus possible that what they observed is an effect of absolute rather than relative position.

	PREVERBAL		POSTVERBAL		TOTAL
DEF	**28**	(28.6)	**30**	(33.9)	58
INDEF	**10**	(9.4)	**9**	(11.1)	19
TOTAL	38		39		77

Table 3: Results qua the relative position hypothesis

[5] The relatively high frequency of indefinites – even singulars, where the ratio is 157:81 – could be interpreted as worrisome for Dayal's (2004) proposal that singular bare NPs in articleless languages are never genuinely indefinite.

[6] Expected values used for pairwise comparisons are the same as in the full contingency table (Table 2 for the case at hand). Bonferroni-adjusted p is assumed for pairwise comparisons throughout the paper.

Table 3 shows that definite and indefinite NPs are distributed around the verb in full accordance with the null hypothesis ($\chi^2(1) = .11$, $p = .74$, $n = 77$). There are neither more preverbal definites / fewer preverbal indefinites than expected ($\chi^2(1) = .06$, $p = .82$, $n = 38$), nor fewer postverbal definites / more postverbal indefinites than expected ($\chi^2(1) = .05$, $p = .82$, $n = 39$). We found no evidence for the relative position hypothesis.

A preliminary conclusion is that the definiteness of bare NPs depends on the absolute position in the clause but not on the position relative to the verb. We postpone further discussion until after we discuss the apparent effect of syntactic function on definiteness, which turns out to be a potential confound for the absolute position hypothesis.

4 Syntactic function and definiteness

4.1 Basic observations

A naked-eye observation of Table 4 makes it clear that there is an effect of syntactic function on definiteness ($\chi^2(2) = 19.22$, $p < .001$, $n = 315$). More particulary, there are more definite & fewer indefinite subjects than expected ($\chi^2(1) = 10.75$, $p = .001$, $n = 78$) and more indefinite & fewer definite objects than expected ($\chi^2(1) = 8.35$, $p = .004$, $n = 127$). Being an adverbial has no effect on definiteness ($\chi^2(1) = .12$, $p = .73$, $n = 110$).[7]

	SUBJECT		OBJECT		ADVERBIAL		TOTAL
DEF	**65**	(51.3)	**68**	(83.5)	**74**	(72.3)	207
INDEF	**13**	(26.7)	**59**	(43.5)	**36**	(37.7)	108
TOTAL	78		127		110		315

Table 4: Effect of syntactic function on definiteness

The effect of being a subject vs. being an object is thus qualitatively similar – although not so statistically robust – to being in the initial vs. in the final clausal position. It further turns out (see Table 5) that there is a strong correlation between being a subject and being initial on the one

[7] Nominative-marking functioned as the proxy for subjecthood in the annotation. What could be of relevance is that 65 out of the 78 subjects were agents. An NP was annotated as an object when it was an obligatory internal argument (including 17 PPs). The majority of objects were accusative-marked direct objects (90 out of the 127).

hand and being an object and being final on the other ($\chi^2(2) = 74.21$, $p < .001$, $n = 315$). More particularly, there are more initial & fewer final subjects than expected ($\chi^2(2) = 48.65$, $p < .001$, $n = 78$) and more final & fewer initial objects than expected ($\chi^2(2) = 20.50$, $p < .001$, $n = 127$). There is no statistically significant tendency for adverbials to be in any particular position ($\chi^2(2) = 5.03$, $p = .08$, $n = 110$).[8]

	SUBJECT		OBJECT		ADVERBIAL		TOTAL
INITIAL	**41**	(16.3)	**9**	(26.6)	**16**	(23.0)	66
FINAL	**20**	(41.1)	**90**	(66.9)	**56**	(58.0)	166
MEDIAL	**17**	(20.6)	**28**	(33.5)	**38**	(29.0)	108
TOTAL	78		127		110		315

Table 5: Interaction between syntactic function and position

Given this state of affairs, syntactic function could be a confounding factor for the absolute position hypothesis – at present we cannot rule out the possibility that the in/decreased likelihood of (in)definiteness reported in section 3.3, is caused by syntactic function rather than clausal position. That syntactic function (esp. being a subject) can have an effect on definiteness is a well-known hypothesis, as discussed in section 2, so the confound needs to be addressed properly.

4.2 Ruling out the syntactic function confound

In order to separate the correlating factors position and syntactic function from one another, we need to look at four data subsets. These are suitable for testing the effect of the two pertinent factors in isolation, as summarized in (6). The rationale behind this is simple: if, e.g., the effect of position on definiteness is real, we should find it even by looking at subjects only (comparing initial and final subjects) or at objects only (comparing initial and final objects).

(6) a. **Subjects only & Objects only**
→ Testing for the effect of position on definiteness (without the interference of syntactic function).

[8] We are coding position as the dependent variable for presentational purposes. The results are comparable if syntactic function is coded as the dependent variable.

b. **Initial NPs only & Final NPs only**
→ Testing for the effect of syntactic function on definiteness (without the interference of clausal position).

Table 6 demonstrates that the effect of position on definiteness is preserved even without the interference of syntactic function, esp. for the subset of subjects ($p < .001$, $n = 61$) and, less clearly but significantly so, for the subset of objects ($p = .017$, $n = 103$). More particularly, we find more initial definite & fewer initial indefinite subjects than expected ($p = .003$, $n = 41$) and more final indefinite & fewer final definite objects than expected ($p = .001$, $n = 20$). The position effect in the subset of objects is caused by the effect of the initial position, where there are more initial definite & fewer initial indefinite objects than expected ($p = .021$, $n = 9$). We find no effect of object finality on definiteness ($p = .27$, $n = 94$).[9]

	SUBJECTS ONLY				OBJECTS ONLY			
	INITIAL		FINAL		INITIAL		FINAL	
DEF	**40**	(33.6)	**10**	(16.4)	**8**	(4.5)	**44**	(47.5)
INDEF	**1**	(7.4)	**10**	(3.6)	**1**	(4.5)	**50**	(46.5)

Table 6: Effect of position on definiteness of subjects & objects only

The pattern revealed by Table 7 is strikingly different: when considering initial NPs only and final NPs only – in order to test for the effect of syntactic function on definiteness – we find no departure from the null hypothesis (initial NPs: $p = .33$, $n = 50$; final NPs: $p = .49$, $n = 114$).

	INITIAL NPS ONLY				FINAL NPS ONLY			
	SUBJECT		OBJECT		SUBJECT		OBJECT	
DEF	**40**	(39.4)	**8**	(8.6)	**10**	(9.5)	**44**	(44.5)
INDEF	**1**	(1.6)	**1**	(0.4)	**10**	(10.5)	**50**	(49.5)

Table 7: Effect of syntactic function on definiteness of initial & final NPs only

Based on this post-hoc analysis, we can conclude that the effect of position (initial vs. final) on definiteness is real, while the effect of syntactic function (subject vs. object) on definiteness is a mere illusion,

[9] Due to low expected frequencies (below 5), one-tailed Fisher exact test (rather than Pearson chi-square) is used for Tables 6 and 7.

caused by the fact that subjects are typically initial and objects are typically final.[10]

5 Discussion and outlook

We found strong support for the absolute position hypothesis, repeated for clarity in (7).

(7) ABSOLUTE POSITION HYPOTHESIS: The absolute clausal position of bare NPs (initial/final) has an impact on their (in)definiteness.
   a. Clause-initial bare NPs are more likely to be definite.
   b. Clause-initial bare NPs are less likely to be indefinite.
   c. Clause-final bare NPs are more likely to be indefinite.
   d. Clause-final bare NPs are less likely to be definite.

Of the sub-hypotheses (a)–(d), (b) turned out to be the most strongly supported one: there are **4.5-times fewer initial indefinites** than what is expected under the null hypothesis. A post-hoc analysis confirmed this strong trend for both subjects and objects individually (although the numbers are very low and so is the level of confidence). This finding lends support to the specific proposal of Geist (2010), who takes the effect of initial position to be a "restriction on indefiniteness" (rather than a requirement to be definite). In her proposal, indefinite bare NPs are ruled out in the initial position by the conjunction of the following three assumptions: (i) initial bare NPs are topics (exception: thetic sentences in the sense of Sasse 1987), (ii) topics are referential (Reinhart 1981), and (iii) indefinite bare NPs cannot be referential.

The effect of clause-initiality on definiteness – sub-hypothesis (a) – is less pronounced: there are **1.3x more initial definites** than what is expected under the null hypothesis. This effect is stronger for objects (1.8x) than for subjects (1.2x), which correlates with the fact that subjects are initial by default. Despite the common assumption that clause-final position has no impact on bare NPs' (in)definiteness, we did

[10] As noted by an anonymous reviewer, a comparable refutation might not be applicable to languages like Mandarin Chinese, where the initial (or rather preverbal) position of subjects is basically obligatory. For a related corpus-based discussion of pre/postverbal subjects in Russian, see Slioussar (2011).

find a trend in the expected direction: there are **1.4x more final indefinites** (sub-hypothesis (c)) and **1.3x fewer final definites** (sub-hypothesis (d)) than what is expected under the null hypothesis. Our post-hoc analysis reveals that this trend is clearly visible for subjects (1.6x fewer definites and 2.8x more indefinites), but virtually non-existent for objects, whose (in)definiteness remains unaffected by being placed in final position. A plausible explanation of this subject–object asymmetry builds on the notion of focus: clause-final objects correlate with focus-size neutrality (availability of "focus projection"), whereas clause-final subjects strongly correlate with narrow subject-focus. If, in turn, focus correlates with novelty and novelty with indefiniteness (Heim 1982), the observed subject-specific effect follows (and particularly the strong tendency towards indefiniteness).[11]

Our findings fail to support the relative position hypothesis – the idea that the position of bare NPs relative to the verb (pre-/postverbal) has an impact on their (in)definiteness. This sheds doubt on the traditional concept of verb as a "transition" between a contextually dependent and a contextually independent area of the sentence (Firbas 1965), as well as on what could be considered its generative incarnation – Kučerová's (2007) G-operator-based approach, which establishes an intimate connection between overt verb position and the presupposed–non-presupposed partition. The consequences for Diesingian (1992) approaches are pending a precise (and perhaps case-by-case) analysis of the syntactic position of the main verb.

Last but not least, our findings fail to support the idea that syntactic function (being a subject or object) has an effect on bare NP definiteness. As revealed by our post-hoc analysis (despite the relatively low numbers), any effect on (in)definiteness that could apparently be attributed to syntactic function is directly derivative of the effect of clausal position. This is because subjects are likely to be initial and objects are likely to be final.

We hope that this work has proved the usefulness of applying corpus methodology to test the existing generalizations and hypotheses about the definiteness–word order interaction. Hopefully it also demonstrates

[11] Unfortunately, this explanation finds no support in our annotation, as all of the clause-final indefinite subjects are found to be parts of focus, not narrow foci.

the need to systematically control for closely related factors such as absolute vs. relative position or position vs. syntactic function.

References

Bailyn, John. 2004. Generalized inversion. *Natural Language & Linguistic Theory* 22(1): 1–49. https://doi.org/10.1023/B:NALA.0000005556.40898.a5

Berger, Tilman. 1993. Das System der tschechischen Demonstrativpronomina. Habilitation thesis, Ludwigs-Maxmilians-Universität München.

Biskup, Petr. 2011. *Adverbials and the phase model.* Amsterdam: John Benjamins.

Brun, Dina. 2001. Information structure and the status of NP in Russian. *Theoretical Linguistics* 27(2–3): 109–135. https://doi.org/10.1515/thli.2001.27.2-3.109

Burianová, Markéta. 2016. Referenční vlastnosti holé jmenné fráze v češtině. Masters thesis, Charles University in Prague.

Chvany, Catherine V. 1973. Notes on root and structure-preserving in Russian. In *You take the high node and I'll take the low node,* ed. Claudia W. Corum and Thomas Cedric Smith-Stark and Ann Weiser, 52–290. Chicago, IL: Chicago Linguistic Society.

Chvany, Catherine V. 1983. On definiteness in Bulgarian, English, and Russian. In *American Contributions to the Ninth International Congress of Slavists, Vol. 1: Linguistics,* ed. Michael S. Flier, 71–92. Kiev.

Cohen, Jacob. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 20(1): 37–46. https://doi.org/10.1177/001316446002000104

Cook, Philippa and Felix Bildhauer. 2013. Identifying "aboutness topics": Two annotation experiments. *Dialogue and Discourse* 4(2): 118–141. https://doi.org/10.5087/dad.2013.206

Czardybon, Adrian, Oliver Hellwig, and Wiebke Petersen. 2014. Statistical analysis of the role of word order in Polish determination types. In *Proceedings of PolTAL2014,* ed. Adam Przepiórkowski and Maciej Ogrodniczuk, 144–150. Springer International Publishing.

Czech National Corpus – SYN2010. Institute of the Czech National Corpus, Praha 2010. http://www.korpus.cz

Dayal, Veneeta. 2004. Number marking and (in)definiteness in kind terms. *Linguistics and Philosophy* 27(4): 393–450. https://doi.org/10.1023/B:LING.0000024420.80324.67

Dayal, Veneeta. 2011. Bare noun phrases. In *Semantics: An international handbook of natural language meaning, Vol. 2,* ed. Klaus von Heusinger and Claudia Maienborn and Paul Portner, 1088–1109. Berlin: de Gruyter. https://doi.org/10.1515/9783110255072.1088

Diesing, Molly. 1992. *Indefinites.* Cambridge, MA: MIT Press.

Féry, Caroline and Vieri Samek-Lodovici. 2006. Focus projection and prosodic prominence in nested foci. *Language* 82(1): 131–150. https://doi.org/10.1353/lan.2006.0031

Firbas, Jan. 1965. A note on transition proper in functional sentence analysis. *Philologica Pragensia* 8: 170–176.

Geist, Ljudmila. 2010. Bare singular NPs in argument positions: Restrictions on indefiniteness. *International Review of Pragmatics* 2(2): 191–227. https://doi.org/10.1163/187731010X528340

Heim, Irene. 1982. The semantics of definite and indefinite noun phrases. PhD dissertation, University of Massachusetts at Amherst.

Hlavsa, Zdeněk. 1975. *Denotace objektu a její prostředky v současné češtině.* Praha.

de Hoop, Helen. 1992. Case configuration and noun phrase interpretation. PhD dissertation, University of Groningen.

Jenks, Peter. To appear. Articulated definiteness without articles. To appear in *Linguistic Inquiry.*

Junghanns, Uwe and Gerhild Zybatow. 1997. Syntax and information structure of Russian clauses. In *Formal Approaches to Slavic Linguistics (FASL) 4: The Cornell Meeting 1995,* ed. Wayles Browne, 289–319. Ann Arbor, MI: Michigan Slavic Publications.

King, Tracy Holloway. 1995. *Configuring topic and focus in Russian.* Stanford, CA: CSLI Publications.

Krámský, Jiří. 1972. *The article and the concept of definiteness in language.* The Hague: Mouton.

Kučerová, Ivona. 2007. The syntax of givenness. PhD dissertation, MIT, Cambridge, MA.

Li, Charles N. and Sandra A. Thompson. 1976. Subject and topic: A new typology of language. In *Subject and topic,* ed. Charles N. Li, 457–489. New York: Academic Press.

Matthews, Stephen and Virginia Yip. 1994. *Cantonese: A comprehensive grammar*. London: Routledge.
Migdalski, Krzysztof. 2006. The syntax of compound tenses in Slavic. PhD dissertation, Tilburg University.
Mykhaylyk, Roksolana. 2011. Middle object scrambling. *Journal of Slavic Linguistics* 19(2): 231–272. https://doi.org/10.1353/jsl.2011.0012
Reinhart, Tanya. 1981. Pragmatics and linguistics: An analysis of sentence topics. *Philosophica* 27(1): 53–94. http://logica.ugent.be/philosophica/fulltexts/27-4.pdf
Rochemont, Michael. 1986. *Focus in generative grammar*. Philadelphia: John Benjamins.
Sasse, Hans-Jürgen. 1987. The thetic/categorical distinction revisited. *Linguistics* 25(3): 511–580. https://doi.org/10.1515/ling.1987.25.3.511
Schmerling, Susan. 1976. *Aspects of English sentence stress*. Austin, TX: University of Texas Press.
Schwarzschild, Roger. 1999. Givenness, AvoidF and other constraints on the placement of accent. *Natural Language Semantics* 7(2): 141–177. https://doi.org/10.1023/A:1008370902407
Šimík, Radek and Marta Wierzba. 2015. The role of givenness, presupposition, and prosody in Czech word order: An experimental study. *Semantics & Pragmatics* 8(3): 1–103. https://doi.org/10.3765/sp.8.3
Slioussar, Natalia. 2011. Russian and the EPP requirement in the Tense domain. Lingua 121(4): 2048–2068. https://doi.org/10.1016/j.lingua.2011.07.009
Späth, Andreas. 2003. The linearization of argument DPs and its semantic reflection. In *Mediating between concepts and grammar,* ed. Holden Härtl and Heike Tappe, 165–195. Berlin: de Gruyter.
Szwedek, Aleksander. 1974. A note on the relation between the article in English and word order in Polish (Part 1 and 2). *Papers and Studies in Contrastive Linguistics* 2: 213–225.
Szwedek, Aleksander. 2011. *The thematic structure of the sentence in English and Polish: Sentence stress and word order*. Frankfurt am Main: Peter Lang.
Titov, Elena. 2017. The canonical order of Russian objects. *Linguistic Inquiry* 48(3): 427–457. https://doi.org/10.1162/ling_a_00249

Umbach, Carla. 2001. (De)accenting definite descriptions. *Theoretical Linguistics* 27(2–3): 251–280.
Wiland, Bartosz. 2009. Aspects of order preservation in Polish and English. PhD dissertation, University of Poznań.

radek.simik@hu-berlin.de
burianova.marketa@gmail.com