Left branch extraction and basic word order in Czech

Daniela Kořánová, Český jazyk a literatura – Obecná lingvistika, 3. ročník Bc.

1 Introduction

The present paper deals with the topic of left branch extraction in Czech. In the first part, the theoretical background of this phenomenon will be discussed. For the analysis will be the phase theory and the cyclic linearization introduced. To test the claims derived from these concepts, an experiment was conducted which will be the topic of the second part of this paper. The main hypotheses deal with the word order being based on the syntactic functions (subject – object) and based on the semantics of the constituents (humanness). Other than that, the definiteness of constituents in relation to the word order was tested as well as the possible effect on the word order of two types of adverbs. The results, which will be discussed lastly, bring some interesting insights into the phenomenon and the word order in general.

2 Left branch extraction

What is left branch extraction (LBE) and how does it connect to Czech? To discuss this, we firstly need to mention the proposal of left branch condition (LBC) by Ross (1967):

"No NP that is the leftmost constituent of a larger NP can be reordered out of this NP by a transformational rule."

This condition does not seem to apply to Czech, where a leftmost constituent of an NP can be moved. Examples in (1) show the difference between questions without LBE and with LBE.

(1)	a_1	[Která dívka] _{NP} četla?	without LBE
	a_2	Která ₁ četla [t ₁ dívka] _{NP} ?	with LBE
	b_1	[Kterou dívku] _{NP} jsi viděl?	without LBE
	b_2	Kterou ₂ jsi viděl [t ₂ dívku] _{NP} ?	with LBE

"Která dívka" and "kterou dívku" form NPs (as indicated in (1)). The Wh-constituent ("který" in the right gender and case in our examples) stands in the first position (before a clitic). Depending on the situation, "který" can either be moved to its place by itself (a_2 and b_2 with LBE) or as a whole NP (a_1 and b_1 without LBE). Both alternations of a and b share the same meaning.

Bošković (2008) argues that LBE is possible in languages that do not have articles, i.e., they have NP and not DP. Czech being an articleless language thus meets the requirement for LBE.

If looked upon from a different view, constructions with LBE are nonprojective constructions created by the division of NP. In a paper discussing the Czech nonprojective constructions in the Prague Dependency Treebank Hajičová, Havelka, Sgall, Veselá and Zeman (2004) state that the percentage of clauses with divided nominal groups is about 11 % and fronted wh-elements about 1.6 % (however it is not clear whether all the findings are strictly cases with LBE). As is apparent from these percentages, constructions with LBE are quite uncommon (but possible).

For our analysis, we will work with the phase theory and cyclic linearization.

2.1 Phase theory

To describe the structure of clauses, Chomsky (2000) suggests the phase theory. The clause is divided into phases. If some part of a phase is needed in another phase, the part is only accessible from the edge of the base phase (see 2).

(2) Phase-Impenetrability Condition

In phase a with head H, the domain of H is not accessible to operations outside a, only H and its edge are accessible to such operations.

In other words, to be able to work with a constituent from phase "x" in a higher phase "y" the constituent first needs to be moved to the edge of phase "x", from this edge the constituent can be then moved in phase "y". If a constituent stays in a phase, it undergoes the operation spellout, by which the phase, more precisely the complement of the phase head, is sent to logical and phonological form "and thus by hypothesis, out of the syntactic derivation" (Bondarenko & Davis, 2021). After this, the constituent is no longer able to move to other phases. This is called successive-cyclic movement.

Figure 1 shows a non-successive-cyclic phase exit (a) where V cannot be moved from the phase to T because V is not on the edge of vP. In the case of a successive-cyclic phase exit (b), V is first moved to v (which is possible because V is on the edge of its phase), from v it can then be moved further to T due to its position at the edge of vP.



Figure 1 – visualisation of cyclic movement, adapted and altered from Biskup (2017b)

2.2 Cyclic linearization

Cyclic linearization (Fox & Pesetsky, 2005) also deals with the operation of spell-out. Linearization which handles the word order at spell-out happens phase by phase. After spell-out follows a movement of a constituent. For a movement to be successive-cyclical, the word order has to be the same before and after the spell-out, otherwise, the constituent would be present twice in one clause, more precisely the linearization would cause a problem in the phonological form.

Bondarenko & Davis (2021) point out that the contradiction between linearization before and after spell-out can be solved without successive-cyclical movement. If constituent "x" is to be moved from a place that is not an edge of a phase to a higher phase, it is needed to also move the constituent "y" which is between the higher phase and the base position of constituent "x". By doing so, the word order will be the same before and after the spell-out.

To demonstrate this, we use examples (figures 2 and 3) from our experiment which is further discussed in chapter 3.



Který političku komentoval článek? *Figure 2 – example of a linearization problem*

NP₁ "političku" is firstly moved from a non-edge base position into SpecvP, then to SpecTP. In NP₂ left branch is extracted and moved to SpecCP. The head of NP₂ remains in its base position. The word order before spell-out is "političku" – "který" – "článek", after the operation "který" – "političku" – "článek", and thus a linearization problem occurs. The clause "Který političku komentoval článek?" would be acceptable by this theory if the whole NP₂ (since it is in between the SpecvP and the base position of NP₁ moved without successive-cyclical pass of the edge) would move to SpecCP. In this case, the clause would be "Který článek političku komentoval?" (and without LBE).



Který politička komentovala článek?

Figure 3 – example of a coherent linearization

This clause, on the other hand, shows consistency in linearization. LBE happens in NP₁, the pronoun "který" moves to SpecvP, from there to SpecCP, therefore avoiding the movement through the edge of the phase. NP₂ is moved from its base position to SpecTP. The word order before spell-out is "který" – "politička" – "článek", which is the same as the word order after spell-out in the final clause.

Following this analysis, LBE should only be possible if it is an extraction from an object over a subject (figure 3). An extraction from a subject over an object (figure 2) causes problems with linearization.

Titov (2012) argues that the semantic nature of constituents, rather than the syntactic function, has precedence in constructing the word order. After being asked the question "Co se stalo?" the ordering in (3) a is a more natural response and has a neutral word order. The humanness of the constituent is dominating the syntactic function.

(3) a Běžkyni srazilo auto. OVS, human – nonhuman

b Auto srazilo běžkyni. SVO, nonhuman – human

Deriving from this, the extraction from a nonhuman over a human should be possible.

3 Experiment

3.1 Design

The experiment consists of 1 main experiment and 5 fillers. All items are in the form of an indirect question, thus include a main clause followed by a subordinate clause. Our focus is on the latter one which contains LBE (except for F5, see 3.1.6) and is introduced by the pronoun "který" in the right gender and case. LBE is in all cases either from an object over a subject or from a subject over an object. The main clause was added for context and to make some of the clauses more plausible, each item has one main clause for all conditions (exceptions discussed in 3.1.2). We tried to use different noun phrases and verbs in all the items to reduce priming, though some of the matrix verbs appear more than once due to the fact that only a section of verbs can be used as a matrix verb in indirect questions. On the other hand, this should not be a problem since these verbs are generally quite common (e.g., say, ask, know), especially in the context of an indirect question, and should not interfere with the acceptability of the clauses in general. The nouns figuring in LBE (either as the noun in NP from which is extracted or as the intervening constituent) are mostly female, alternatively other gender with different forms of NOM and ACC, because it is clearer for the recipient to see which parts belong together (and avoiding a garden path).

3.1.1 Main experiment

The design of the main experiment is $2 \ge 2$, variables being the constituent order (i.e. the type of extraction) and the order of humanness of nouns, in this case either human – nonhuman or nonhuman – human. It consists of 24 items. The nouns of the subordinate clauses were chosen so that they can function as a subject as well as an object with the same verb, thus all four conditions in one item use the same words.

	subject – object	object – subject
human – nonhuman	с	b
nonhuman – human	a	d
Table 1 – Main experiment – design		

Example of an item:

- 14 a Detektiv na začátku vyšetřování tušil, kterou omáčka otrávila kuchařku.
 - b Detektiv na začátku vyšetřování tušil, která kuchařku otrávila omáčka.
 - c Detektiv na začátku vyšetřování tušil, kterou kuchařka otrávila omáčku.
 - d Detektiv na začátku vyšetřování tušil, která omáčku otrávila kuchařka.

3.1.2 Filler 1 – pronouns

In the first filler, our goal was to look at how the overtness of the intervening constituent interacts with the naturalness of the whole clause. The filler includes 8 items with the design 2 x 2 x 2. The variables are the same as in the main experiment (see 3.1.1) plus the category of the intervening constituent. The constituent is either a noun (more precisely NP) or a pronoun (in the case of a covert subject, the pronoun is omitted as it is common to do so in Czech). The subject and object of one item are again interchangeable. After creating the first version of

items, we decided to mention the intervening constituent (in its overt form) in the main clause, so that the covert constituent would more likely be interpreted as a human or a nonhuman depending on the noun. Without this, we found the covert form to be naturally interpreted as a human which we do not want. This leads to one item having two versions of the main clause to capture both NPs (each clause for four conditions).

	nonhuman – human	human – nonhuman
subject – object	a	e
object – subject	g	c
pro – object	b	f
pro - subject	h	d
Table 2 – Filler 1 – design		

Example of an item:

- 8 a Z článku o novém projektu jsem bohužel nezjistil, kterou ten projekt shání investorku.
 - b Z článku o novém projektu jsem bohužel nezjistil, kterou shání investorku.
 - c Z článku o zdatné investorce jsem bohužel nezjistil, který tu investorku shání projekt.
 - d Z článku o zdatné investorce jsem bohužel nezjistil, který ji shání projekt.
 - e Z článku o zdatné investorce jsem bohužel nezjistil, který ta investorka shání projekt.
 - f Z článku o zdatné investorce jsem bohužel nezjistil, který shání projekt.
 - g Z článku o novém projektu jsem bohužel nezjistil, která ten projekt shání investorka.
 - h Z článku o novém projektu jsem bohužel nezjistil, která ho shání investorka.

3.1.3 Filler 2 – humanness balanced

The second filler consists of 8 2 x 2 items. The variables are the constituent order and the humanness of both constituents so that the NPs have balanced humanness. Because of this, two conditions of one item share the same human subject and human object, the other two conditions have the same nonhuman subject and nonhuman object. The main clause remains identical for all four conditions.

	subject – object	object – subject
nonhuman – nonhuman	a	b
human – human	с	d
Table 3 – Filler 2 – design		

Example of an item:

- 4 a Na výstavě o vesmíru jsme zjistili, kterou satelit fotografoval planetu.
 - b Na výstavě o vesmíru jsme zjistili, který planetu fotografoval satelit.
 - c Na výstavě o vesmíru jsme zjistili, které umělkyně fotografovala astronauty.
 - d Na výstavě o vesmíru jsme zjistili, která astronauty fotografovala umělkyně.

3.1.4 Filler 3 – definiteness of the intervening constituent

With the design of 2 x 2, the third filler focuses on the nature of the intervening NP. The first variable is once more the order of constituents (all the NPs are human). For the second variable, the category of definiteness of the intervening constituent was chosen. We came up with contexts where in a group of people is one person unique (e.g., the context of an orchestra, where there are many players but only one conductor). The unique person should be semantically more definite (since we can be quite sure about whom we hear when the person is

just one) than someone from the group (using our previous example, a conductor is more definite than a violin player in an orchestra with twenty other violin players).

	subject – object	object – subject
definite intervening const.	a	b
indefinite intervening const.	с	d
Table 4 – Filler 3 – design	•	

Example of an item:

- 2 a Mezi hráči v orchestru se řešilo, kterou dirigent nenávidí trumpetistku.
 - b Mezi hráči v orchestru se řešilo, která dirigenta nenávidí trumpetistka.
 - c Mezi hráči v orchestru se řešilo, kterou houslistka nenávidí trumpetistku.
 - d Mezi hráči v orchestru se řešilo, která houslistku nenávidí trumpetistka.

3.1.5 Filler 4 – adverbs

All of the above-mentioned fillers (including the main experiment) have a subordinate clause constituted of a subject, an object and a verb. In the fourth filler, an adverb was added to the LBE clause. It consists of 8 items, four of which include the extraction from an object, the other four extraction from a subject. Each item has 2 conditions depending on the adverb used in the clause, either an adverb of manner or time. In the items, where the extraction from an object is present, is the subject omitted. Clauses with the extraction from a subject have an intransitive verb (in all the other cases only transitive verbs were used) since we do not change the type of extraction in these cases and our focus is on the type of the adverb. Each item has, again, one main clause for both conditions.

	object – subject (items 1–4)	subject – object (items 5–8)
adverb of manner	a	a
adverb of time	b	b
Table 5 – Filler 4 – design		

Example of items:

2	a	Studenti se dohadovali, který nudně mluvil přednášející.
	b	Studenti se dohadovali, který včera mluvil přednášející.

- 7 a Knihkupec se zákaznice přeptával, kterou těžce sháněla knihu.
 - b Knihkupec se zákaznice přeptával, kterou předevčírem sháněla knihu.

3.1.6 Filler 5 – control

The last filler consists of 24 items. These clauses do not contain LBE. Half of the items were created to correspond with the standard word order, the other half are disturbed by a nonstandard word order (in most cases a wrongly situated clitic, i.e., the clitic is not on the Wackernagel's position). This filler was added so that the participants would not be overloaded with clauses with LBE and to filter the participants. The items were created to have a similar form as the items with LBE from the main experiment and the other filler.

3.2 Predictions

As we are dealing with multiple hypotheses, this section will be divided into parts, each focusing on a different one. Filler 4 will be discussed separately.

3.2.1 Word order based on syntactic function

For this hypothesis, we consider the word order where an object follows a subject as conventional, Czech is categorized as SVO language (Sgall et al., 1980; Uhlířová, Kučerová, 2017 and others) although it has quite a flexible word order (more precisely, free word order). The order subject – object should be more acceptable than a reversed word order. This means, that the extraction from an object should be generally more acceptable than the extraction from a subject. As mentioned in 2.2, the analysis by Bondarenko & Davis (2021) suggests that only the extraction from an object over a subject should be possible.

For the main experiment, conditions a and c should be rated more natural than b and d. The same goes for fillers 2 and 3. In the first filler conditions a, b, e, and f are supposed to be perceived better than the rest (i.e., conditions c, d, g, h).

3.2.2 Word order based on semantics

The first hypothesis in this category is based on the observations found in Titov (2012) and Jasinskaja & Šimík (in print). It is based on the animacy (humanness) of the constituents. The word order is ought to have the human constituent as a preceding to the nonhuman constituent (i.e., NP human – NP nonhuman). In our case, nonhuman is always a thing, animals were excluded.

In the main experiment, conditions b and c are predicted to be more acceptable than a and d. For filler 1, conditions c, d, e, and f are ought to be more natural than a, b, g, and h. In the other fillers, the humanness is either balanced (F2, F3) or is not in focus (F4).

Filler 3 works with the definiteness of NP. The second hypothesis based on the semantics of the constituents is that a definite NP should precede the less definite (nondefinite) NP. Conditions a and b based on this are predicted to be more natural than c and d.

3.2.3 Type of adverb

One of the variables in the fourth filler is the type of adverb in the subordinate clause. Our hypothesis stands on 2 assumptions: 1) that the adverb of time is based higher in a clause than a subject, 2) that the adverb of manner is based lower than a subject. The second assumption leads to a problem with linearization and antilocality when LBE happens from a subject.¹ There should not be an issue when an adverb of manner is present in LBE from an object, for adverbs of time (due to their higher base position) both types of extractions should be allowed and possible.

Condition a from items 1–4 is therefore predicted to be rated as less natural than the other three conditions.

3.3 Participants

The participants were obtained via the seminar "Participation in linguistic and psychological experiments in LABELS Lab" with the help of doc. Jan Chromý. The university students (from different faculties) filled out the experiment in return for credits. After being sent the link to the experiment, the participants had one week to fill it in. Altogether, 142 participants completed the experiment.

¹ The adverb would have to move from SpecvP to SpecvP higher than the subject, which is not possible.

3.3.1 Procedure

The whole experiment was conducted on the L-Rex platform. The participants were instructed to rate the naturalness of sentences (one at a time) on a scale from 1 to 7, where 1 was "not natural" and 7 was "natural". Each participant was shown 80 items (for every item just one condition). The items were pseudo-randomized.

After collecting all the data, the participants were filtered based on time spent filling in the whole experiment and on ratings from filler 5 (as mentioned in 3.1.6).

In the end, we were left with 70 participants whose data were then analysed.

3.4 Results

The first thing apparent from the results is that the clauses with LBE are generally perceived as unnatural, throughout the experiment the majority of participants rated the constructions on the lower portion of the scale.

3.4.1 Main experiment

As is evident from figures 4 and 5, only one condition shows a difference from the others. It is condition c where a human subject intervenes the extraction from a nonhuman object. Conditions a, b and d have close similarities as for their rating.

The constituent order in the main experiment does not influence the acceptability as well as the order of humanness. Both are statistically significant.



Figure 4 – Main experiment: boxplot



Figure 5 – Main experiment: z-scores with 95% CI

3.4.2 Filler 1

The items in filler 1 were rated as one of the most natural clauses in the experiment. The conditions with constituent order subject – object (a, b, e, and f) have higher ratings than the order object – subject. The nature of the intervening constituent reflects on the naturalness of the clause in a way that the overt NP is worse than the covert (or omitted) constituents. In the case of the extraction from an object (constituent order subject – object), the order human – nonhuman was rated as slightly better than the opposite order. Interestingly, the participants stated that the extraction from a subject is more acceptable if the intervening constituent in an overt form is nonhuman. Condition c has the lowest rating from this filler.



Figure 6 – Filler 1: boxplot



Figure 7 – Filler 1: z-scores with 95% CI

3.4.3 Filler 2

In the second filler, the type of extraction reflects strongly on the naturalness of the clauses. Both conditions with the subject – object order are rated higher than the ones with the opposite constituent order. As for humanness, there is only a slight difference in the acceptability, the clauses with nonhumans have lower ratings than the ones with only humans.



Figure 8 – *Filler* 2: *boxplot*



Figure 9 – Filler 2: z-scores with 95% CI

3.4.4 Filler 3

The results of the third filler again show the same higher rating of clauses with the constituent order subject – object. In both types of extraction, the intervening constituent being nondefinite leads to higher ratings (in the extraction from a subject only slightly).



Figure 10 – Filler 3: boxplot



Figure 11 – Filler 3: z-scores with 95% CI

3.4.5 Filler 4

The adverb type of time in clauses is rated higher than the items with an adverb of manner. The best condition is then the one with the extraction from an object paired with an adverb of time. Altogether, both variables were significant.



Figure 12 – Filler 4: boxplot



Figure 13 – Filler 4: z-scores with 95% CI

3.5 Discussion

Let's first discuss the role of the constituent order on the rating of naturalness in our experiment. In the main experiment, the constituent order alone does not show any significance and thus does not match our predictions. In all four fillers, nevertheless, the clauses with the extraction from an object were rated higher than the ones with the extraction from a subject (the humanness was balanced in these cases).

Secondly, the order of humanness of the NPs was looked at. The ratings of the main experiment do not correspond with our predictions (similarly to the hypothesis of the constituent order). In the case of filler 1, human – nonhuman is rated according to the hypothesis only when the extraction is from an object. Slightly smaller is the relationship between the case of a human and a nonhuman object pronoun with the extraction from a subject, but still in agreement with the prediction. The ratings of the extraction from a subject when the object is overt are contrary to our hypothesis.

The definiteness of the intervening constituent in filler 4 plays with the naturalness of the clauses in the opposite way than we predicted.

As mentioned in section 3.4.2, the items in F1 were the most natural from the entire experiment. This is most certainly due to the fact that the intervening constituent in each clause was also included in the main clause. The contextual involvement thus leads to a better acceptance of clauses.

Lastly, the results of the fourth filler are quite puzzling. Our prediction was that only the adverb of manner as an intervening constituent of the extraction from a subject would be rated worse than the others. The clauses with an adverb of manner were however rated poorly in both types of extractions, meaning that the base position is unknown.

All this said the suggested theory cannot explain the results as one, or rather the results are inconsistent with the theory.

According to the results, we propose that the canonical word order in Czech is the following: a subject that is a human - an object that is a nonhuman. If humanness is balanced, a subject has a prior position to an object. When these conditions are matched, LBE is possible.

4 Conclusions

This paper gives an insight into the phenomenon of LBE in Czech. As the results suggest, LBE is a cognitively challenging operation to process. The proposed analysis is not successful in explaining the tendencies found in the ratings in the experiment. The question of the syntactic analysis of LBE in Czech is open for future studies.

Note: This paper is partly adapted from my bachelor's thesis in progress supervised by doc. Mgr. Radek Šimík, Ph.D.

5 References

Biskup, P. (2017a): Left branch condition. In: Karlík, P.; Nekula, M; Pleskalová, J. (eds.), CzechEncy – Nový encyklopedický slovník češtiny [online].

Biskup, P. (2017b): Podmínka fázové impenetrability. In: Karlík, P.; Nekula, M; Pleskalová, J. (eds.), CzechEncy – Nový encyklopedický slovník češtiny [online].

Bondarenko, T. – Davis, C. (2021): Sub-extraction asymmetries and linearization in Russian. *Journal of Slavic Linguistics*. 29 (FASL 28 extra issue).

Boškovič, Ž. (2008): What will you have, DP or NP? In Elfner, E.; Walkow, M. (eds.), *Proceedings of NELS*, 37, s. 101–115.

Fansellow, G. – Lenertová, D. (2011): Left peripheral focus: mismatches between syntax and information structure. *Natural Language & Linguistic Theory*, 29(1), s. 169–209.

Fox, D. – Pesetsky, D. (2005): Cyclic linearization of syntactic structure. *Theoretical Linguistics*, 31 (1–2), s. 1–45.

Hajičová, E. – Havelka, J. – Sgall, P. – Veselá, K. – Zeman, D. (2004): Issues of projectivity in the Prague Dependency Treebank. *The Prague Bulletin of Mathematical Linguistics*, 81, s. 5–22.

Chaves, R. P. (2021): Island phenomena and related matters. In: S. Müller – A. Abeillé – R. D. Borlsey – J. P. Koenig (eds.), *Head-Driven Phrase Structure Grammar: The handbook*. Berlin: Language Science Press, s. 665–725.

Chomsky, N. (2000): Minimalist inquiries: The framework. In R. Martin – D. Michaels – J. Uriagereka (eds.), *Step by step: Essays on minimalism in honor of Howard Lasnik*. Cambridge, MA: MIT Press.

Jasinskaja, K. – Šimík, R. (in print): Slavonic free word order. In J. Fellerer – N. Bermel (eds.), The Oxford guide to the Slavonic languages. Oxford: Oxford University Press.

Ross, J. R. (1967): Constraints on Variables in Syntax. Dissertation. MIT.

Sgall, P. – Hajičová, E. – Buráňová, E. (1980): Aktuální členění věty v češtině. Praha: Academia.

Titov, E. (2012): Information Structure of Argument Order Alternations. Dissertation. University College London.

Uhlířová, L. – Kučerová, I. (2017): Slovosled. In: Karlík, P.; Nekula, M; Pleskalová, J. (eds.), CzechEncy – Nový encyklopedický slovník češtiny [online].

Veselovská, L. (2021): *Wh-questions: a case study in czech*. Olomouc: Palacký University Olomouc.