

## Editorial

The editors are pleased to welcome you to the double issue of the ninth volume of FULL, an open access international journal providing a platform for linguistic research on modern and older Finno-Ugric or other Uralic languages and dialects. FULL publishes comparative research as well as research on single languages, including comparison of just Uralic languages or comparison across family lines. We encourage both formal linguistic submissions and empirically oriented contributions.

The present issue contains two research articles and two papers that describe corpora.

This first research article, written by Katalin É. Kiss, is titled *Accusative or possessive? The suffix of pronominal objects in Ob-Ugric*. The paper seeks an answer to the question why pronominal objects in Mansi and Northern Khanty are personal pronouns bearing a possessive agreement morpheme encoding the person and number of the given pronoun, and why the possessive suffix of these pronouns is identified as an accusative case marker in Mansi and Northern Khanty grammars. It is argued that pronouns bearing a possessive agreement morpheme are formally reflexive pronouns functioning as referentially independent, emphatic, strong pronouns. In Ob-Ugric, 1st and 2nd person pronominal objects used to be – and in some dialects, still are – barred from topic position by the Inverse Topicality Constraint, and, as focal elements, they are represented by strong pronouns. In Northern Khanty and Northern Mansi, the consistent possessive marking of 1st and 2nd person object pronouns has been analogically extended to 3rd person pronouns, as well. Since only subjects and familiar objects can be topicalized, oblique pronouns have also been barred from topic position, and therefore they also appear in their strong forms. Since 1st and 2nd person (and in some languages, 3rd person) object pronouns have been consistently represented by the possessive-marked strong forms, the possessive morphemes of these forms have come to be interpreted as object markers.

The second article, *Focus in Udmurt: Positions, contrastivity and exhaustivity* by Erika Asztalos, presents the results of three surveys examining the positions and the interpretation of foci in Udmurt. While confirming earlier findings according to which the most acceptable focus position is the immediately preverbal one, and that sentence-final focusing is also grammatical for a part of the speakers, the results indicate that foci, with some limitations, can also occur in some preverbal but not verb-adjacent positions. From the perspective of interpretation, none of the focus positions turned out to be obligatorily contrastive or necessarily exhaustive. The sentence-final focusing strategy is interpreted as a phenomenon induced by Russian influence and as a sign of the ongoing SOV-to-SVO change of Udmurt. The results also reveal considerable inter-speaker variation in focus position preferences.

The third contribution, *Web Corpora of Volga-Kama Uralic Languages* by Timofey Arkhangelskiy, reports on a total of 11 electronic corpora of five minority Uralic languages that belong, or are adjacent to, the Volga-Kama area, which has been characterized as comprising a Sprachbund. The corpora, available at <http://volgakama.web-corpora.net>, contain written and, in one case, spoken texts in Udmurt, Komi, Meadow Mari, Erzya and Moksha languages. The described resources are “web corpora” both in terms of their accessibility through a web-based query interface, and, in most cases, in terms of the medium: almost all texts come from web resources, such as digital newspapers and social media. The paper describes the corpora from the user’s perspective. The main focus is on the search capabilities and on certain research questions that can be studied with the help of these corpora.

The fourth paper in the volume, *The INEL Dolgan Corpus: Insights into an endangered language of Northern Eurasia* by Chris Lasse Däbritz, presents a description of the INEL Dolgan Corpus, which has been created between 2016 and 2019 within the INEL project at the Institute for Finno-Ugric/Uralic Studies of the University of Hamburg. The corpus aims to provide a digital research infrastructure for Dolgan, an indigenous language of Northern Siberia. Though Dolgan is a Turkic language, the corpus is relevant for researchers of Uralic languages both due to the close areal connections of Uralic with Dolgan on the Taymyr peninsula and on account of the fact that it is an example of electronic research infrastructure developed for an endangered language. After introducing Dolgan and the INEL project, the paper describes the INEL Dolgan Corpus in detail, focusing on its linguistic content, annotation layers and search possibilities. Finally, the author provides an outlook on how the corpus contributes to furthering research on this endangered language.

We take this opportunity to thank the anonymous reviewers who generously lent their time and expertise to FULL. Our publications can be freely accessed and downloaded without any need for prior registration. At the same time, those who register, or have already registered, are provided with the benefit of getting notified of new issues, calls, etc. via email. FULL welcomes manuscripts from all the main branches of linguistics, including phonology, morphology, syntax, semantics and pragmatics, employing a diachronic or synchronic perspective, as well as from first language acquisition and psycholinguistics. Whatever the theoretical or empirical orientation of the contributions may be, our leading principle is to maintain the highest international standards.

The Editors

# Accusative or Possessive? The Suffix of Pronominal Objects in Ob-Ugric\*

É. Kiss Katalin

This paper seeks an answer to the question why pronominal objects in Mansi and Northern Khanty are personal pronouns bearing a possessive agreement morpheme encoding the person and number of the given pronoun, and why the possessive suffix of these pronouns is identified as an accusative case marker in Mansi and Northern Khanty grammars. The answer is derived from the morphosyntax of reflexive pronouns, and the morphosyntax of differential object marking in Ob-Ugric. It is argued that pronouns bearing a possessive agreement morpheme are formally reflexive pronouns functioning as referentially independent, emphatic, strong pronouns. In Ob-Ugric, 1st and 2nd person pronominal objects used to be – and in some dialects, still are – barred from topic position by the Inverse Topicality Constraint, and, as focal elements, they are represented by strong pronouns. In Northern Khanty and Northern Mansi, the consistent possessive marking of 1st and 2nd person object pronouns has been analogically extended to 3rd person pronouns, as well. Since only subjects and familiar objects can be topicalized, oblique pronouns have also been barred from topic position, and therefore they also appear in their strong forms. Subjects are topics in these languages, hence subject pronouns have been grammaticized in their weak forms. Since subject pronouns have been consistently represented by the weak (i.e., base) forms, and 1st and 2nd person (and in some languages, 3rd person) object pronouns have been consistently represented by the possessive-marked strong forms, the possessive morphemes of the latter have come to be interpreted as object markers.

Keywords: *accusative case, differential object marking (DOM), Inverse Agreement Constraint, possessive agreement, pronominal object*

## 1 The problem

In Mansi and Northern Khanty, pronominal objects bear suffixes encoding the person and number of the pronominal stem. These suffixes appear to be identical with the possessive agreement suffixes cross-referencing an overt or pro-dropped possessor on the possessums. A puzzle of Uralic morphosyntax is why pronominal objects bear a possessive agreement morpheme, and why the possessive suffix of these pronouns is identified as an accusative case marker in Mansi and Northern Khanty grammars. In these dialects, the possessive “accusative” suffix is also present on the pronominal stem when the pronoun is supplied with an oblique case marker. Pronominal subjects, on the contrary, never bear an agreement morpheme. In Hungarian, possessive agreement stands in, or can stand in, for accusative marking in the case of first and second person objects and objects with a first or second person possessor. So far, it has remained unexplained how possession is related to personal pronouns and to object function. After summarizing the relevant facts, this squib will attempt a hypothetical answer.

---

\* This research was carried out in the framework of NKFIH grant 129921. I owe thanks to Irina Burukina, Márta Csepregi, Katalin Gugán and two anonymous reviewers for their helpful comments and suggestions.

## 2 The facts

Observe the Northern Mansi pronominal paradigm, as described by Kálmán (1976). (The dual and plural 2nd and 3rd person forms not spelled out in the table display the same behavior as the 1st person forms.)

(1) Declension of personal pronouns in Northern Mansi (Kálmán 1976: 50)

|     | 1SG <sup>1</sup>           | 2SG             | 3SG            | 1DU               | 1PL              |
|-----|----------------------------|-----------------|----------------|-------------------|------------------|
| NOM | <i>am</i>                  | <i>naŋ</i>      | <i>taw</i>     | <i>me:n</i>       | <i>ma:n</i>      |
| ACC | <i>a:n<sup>u</sup>m</i>    | <i>naŋən</i>    | <i>tawe</i>    | <i>me:nmen</i>    | <i>ma:naw</i>    |
| DAT | <i>a:n<sup>u</sup>mn</i>   | <i>naŋənn</i>   | <i>tawen</i>   | <i>me:nmenn</i>   | <i>ma:nawn</i>   |
| ABL | <i>a:n<sup>u</sup>mnəl</i> | <i>naŋənnəl</i> | <i>tawenəl</i> | <i>me:nmennəl</i> | <i>ma:nawnəl</i> |
| COM | <i>a:n<sup>u</sup>mtəl</i> | <i>naŋəntəl</i> | <i>tawetəl</i> | <i>me:nmentəl</i> | <i>ma:nawtəl</i> |

The “accusative” suffixes are identical with the corresponding members of the paradigm of possessive agreement morphemes except for the epenthetic vowel connecting the suffix to the stem:

(2) Paradigm of possessive agreement in Northern Mansi (Kálmán 1976: 46)

| possessed | SG                    | SG         | SG        | DU          | PL         |
|-----------|-----------------------|------------|-----------|-------------|------------|
| possessor |                       |            |           |             |            |
| 1SG       | <i>-<sup>u</sup>m</i> |            |           |             |            |
| 2SG       |                       | <i>-ən</i> |           |             |            |
| 3SG       |                       |            | <i>-e</i> |             |            |
| 1DU       |                       |            |           | <i>-men</i> |            |
| 1PL       |                       |            |           |             | <i>-mw</i> |

Interestingly, the possessive suffix is also present in the oblique cases; it intervenes between the pronominal stem and the oblique case marker. (This is not unexpected – in fact, it is capitalized on – in the theories of Caha (2009) and Smith et al. (2019), assuming that morphological cases are internally complex with more complex cases containing less complex ones.)

Unlike Northern Mansi, Eastern Mansi has preserved the Proto-Ugric accusative suffix *-m*; still, 1st and 2nd person singular and plural pronominal objects, and a variant of the 3rd person singular pronominal object bear the corresponding possessive agreement morphemes instead. (In the case of the dual and 3rd person plural pronouns, the accusative form is the same as the nominative form (Virtanen 2015: 34).) Compare the nominative and accusative forms of these pronouns with the corresponding possessive agreement morphemes:

<sup>1</sup> The following abbreviations are used in the paper: 1 = first person, 2 = second person, 3 = third person, ABE = abessive case, ABL = ablative case, ACC = accusative case, APPR = approximative case, COM = comitative case, DAT = dative case, DEM = demonstrative, DU = dual, INSF = instructive-final case, LAT = lative case, LOC = locative case, NEG = negative particle, NOM = nominative case, PART = particle, PL = plural, POSS.AGR = possessive agreement, PST = past tense, SG = singular, TRA = translative case.

- (3) Nominative and accusative personal pronouns in Eastern Mansi (Kulonen 2007: 87)

|     |                          |             |             |               |             |
|-----|--------------------------|-------------|-------------|---------------|-------------|
|     | 1SG                      | 2SG         | 3SG         | 1PL           | 2PL         |
| NOM | <i>om</i>                | <i>näg</i>  | <i>täv</i>  | <i>möän</i>   | <i>nöän</i> |
| ACC | <i>oänəm<sup>2</sup></i> | <i>nä:n</i> | <i>tävə</i> | <i>möänəw</i> | <i>nöän</i> |

- (4) Partial paradigm of possessive agreement in Eastern Mansi (Kulonen 2007: 31)

|           |              |              |           |             |             |
|-----------|--------------|--------------|-----------|-------------|-------------|
| possessed | SG           | SG           | SG        | PL          | PL          |
| possessor |              |              |           |             |             |
| 1SG       | <i>-(ə)m</i> |              |           |             |             |
| 2SG       |              | <i>-(ə)n</i> |           |             |             |
| 3SG       |              |              | <i>-ə</i> |             |             |
| 1PL       |              |              |           | <i>-nəw</i> |             |
| 2PL       |              |              |           |             | <i>-ään</i> |

As pointed out by Virtanen (2014: 13), and illustrated by examples like (5a–b), the accusative morpheme is also absent on lexical objects that bear a 1st or 2nd person possessive morpheme cross-referencing a 1st or 2nd person possessor:

- (5) a. *Pim.syəsyk<sup>o</sup>-əm öät tə pümənt-əs-ləm.<sup>3</sup>*  
 son.dear-1SG NEG PART command-PST-SG<1SG<sup>4</sup>  
 ‘I have not commanded my dear son enough.’ (Virtanen 2015: 44)
- b. *Ääk-ən koməly woäxtl-əs-lən!*  
 uncle-2SG how leave-PST-SG<2SG  
 ‘How could you leave your uncle!’ (Virtanen 2014: 13)

A similar resemblance is attested between the “accusative” case endings of pronouns and the corresponding possessive agreement suffixes in Northern Khanty. The impoverished case system of Northern Khanty only includes a single oblique case. Notice that the possessive suffix is also present on the stem when it combines with the locative case suffix.

<sup>2</sup> So as to facilitate comparison, I have replaced Kulonen's (2007) *ø* character with *ə*.

<sup>3</sup> The suffix *-əm* cannot be interpreted as the combination of the *-ə* 3rd person possessive morpheme and the *-m* accusative morpheme because the 3rd person singular possessive accusative ending is represented by the portmanteu morpheme *-ääm/-ätääm*.

<sup>4</sup> The symbol < separates the object agreement morpheme, cross-referencing the number of the object, and the subject agreement morpheme, cross-referencing the number and person of the subject. (In the Ob-Ugric Database of the EuroBABEL project (<http://www.babel.gwi.uni-muenchen.de/>) the symbol > is used for this purpose. This paper adopts the convention of the Uralic databases of the Research Institute for Linguistics, Budapest (<http://www.nytud.hu/oszt/elmyelv/urali/adatbazisok.html>), where the direction of < corresponds to the relative prominence of object and subject.

## (6) Declension of personal pronouns in Northern Khanty (Nikolaeva 1999: 16)

|     | 1SG              | 3SG             | 1DU               | 1PL             |
|-----|------------------|-----------------|-------------------|-----------------|
| NOM | <i>ma</i>        | <i>luw</i>      | <i>min</i>        | <i>muŋ</i>      |
| ACC | <i>ma:ne:m</i>   | <i>luve:l</i>   | <i>mine:mən</i>   | <i>muŋe:w</i>   |
| LOC | <i>ma:ne:mna</i> | <i>luve:lna</i> | <i>mine:mənna</i> | <i>muŋe:wna</i> |

## (7) Partial paradigm of possessive agreement in Northern Khanty (Nikolaeva 1999: 14)

| possessed | SG          | SG                    | DU          | PL         |
|-----------|-------------|-----------------------|-------------|------------|
| possessor |             |                       |             |            |
| 1SG       | <i>-e:m</i> |                       |             |            |
| 3SG       |             | <i>-l<sup>5</sup></i> |             |            |
| 1DU       |             |                       | <i>-mən</i> |            |
| 1PL       |             |                       |             | <i>-uw</i> |

Eastern Khanty marks pronominal objects with a *-t* accusative suffix (the same suffix that functions as the general accusative morpheme in Hungarian).<sup>6</sup> In the Eastern Khanty pronominal paradigm, the possessive suffix appears on 1st person dative pronouns, following the dative morpheme. The rest of the case suffixes other than locative (lative, approximative, translative, instructive-final, comitative, and abessive) are attached to the pronoun+dative suffix+possessive suffix complex – systematically in 1st and 2nd person, and less systematically in 3rd person. That is, the dative form of the pronouns serves as their oblique stem, as opposed to Northern Mansi and Northern Khanty, where the accusative form performs this function. Only the singular pronominal paradigm is cited below, but the dual and plural forms, too, are constructed along parallel principles. The possessive suffixes *-əm*, *-ən* and *-əl*, cross-referencing a singular possessum, are underlined in the pronouns:

<sup>5</sup> Whereas the accusative suffix of the 3rd person singular pronoun (*-e:l*) contains the *-l* 3rd person singular possessive agreement suffix, the *-e:l* complex is formally identical with the 3rd person plural possessive agreement morpheme. I tentatively assume that the epenthetic vowel preceding *-l* has been replaced by *-e:-* analogically – since *-e:-* is present in the accusative forms of the other pronouns.

<sup>6</sup> Pronominal objects in the Baltic Finnic languages bear the same *-t* morpheme. According to Kulonen (1989), the suffix *-t* marked pronominal objects in Proto-Finno-Ugric.

- (8) Declension of personal pronouns in Eastern Khanty (Csepregi 2017: 105–106; forthcoming)

|      | 1SG                    | 2SG                         | 3SG                           |
|------|------------------------|-----------------------------|-------------------------------|
| NOM  | <i>ma</i>              | <i>nüŋ</i>                  | <i>üŋw</i>                    |
| ACC  | <i>mant</i>            | <i>nüŋat</i>                | <i>üŋwat</i>                  |
| DAT  | <i>mantem, manem</i>   | <i>nüŋati</i>               | <i>üŋwati</i>                 |
| LAT  | <i>mantema</i>         | <i>nüŋatena</i>             | <i>üŋwatila</i>               |
| LOC  | <i>manə</i>            | <i>nüŋnə</i>                | <i>üŋwnə</i>                  |
| ABL  | <i>mantemi, manemi</i> | <i>nüŋateni</i>             | <i>üŋwatili</i>               |
| APPR | <i>mantemnam</i>       | <i>nüŋatennam</i>           | <i>üŋwatilnam, üŋwatinnam</i> |
| TRA  | <i>mantemyə</i>        | <i>nüŋatiyə, nüŋatenyə</i>  | <i>üŋwatiyə, lükkə</i>        |
| INSF | <i>mantemat</i>        | <i>nüŋatinat, nüŋatiyat</i> | <i>üŋwatiyat</i>              |
| COM  | <i>mantemnat</i>       | <i>nüŋatenat</i>            | <i>üŋwatinat</i>              |
| ABE  | <i>mantemlay</i>       | <i>nüŋatilay</i>            | <i>üŋwatilay</i>              |

In Eastern Khanty, lexical objects bear no accusative suffix, which raises a further question: why are pronominal objects more likely targets of accusative morphology than lexical noun phrases in languages with differential accusative morphology?

Among the Ugric languages, Hungarian has removed farthest from Proto-Ugric and Proto-Uralic; nevertheless, it still has relics of a system of object marking resembling that surviving in Ob-Ugric, especially that preserved in Eastern Mansi. Namely, Hungarian 1st and 2nd person singular pronominal objects have a possessive ending instead of the accusative *-t*. The possessive ending is also present on the 1st and 2nd person plural pronominal objects, albeit it is followed by the accusative *-t* morpheme.

- (9) Nominative and accusative personal pronouns in Hungarian

|            | 1SG            | 2SG                        | 1PL             | 2PL              |
|------------|----------------|----------------------------|-----------------|------------------|
| NOM        | <i>én</i>      | <i>te</i>                  | <i>mi</i>       | <i>ti</i>        |
| ACC        | <i>en-g-em</i> | <i>té-g-ed<sup>7</sup></i> | <i>mi-nk-et</i> | <i>ti-tek-et</i> |
| POSS. AGR. | <i>-m</i>      | <i>-d</i>                  | <i>-nk</i>      | <i>-tEk</i>      |

The phenomenon observed in Eastern Mansi in connection with (5a–b), i.e., the lack of accusative case suffix on objects with a 1st or 2nd person possessor, has also survived in Hungarian as an option. The accusative marking of the object in Hungarian is optional if and only if the object has an overt or covert 1st or 2nd person possessor:

- (10) *Hova tetted*      *a kulcs-om(-at)* / *kulcs-od(-at)* / *kulcs-unke(-at)* /  
 where put.PST.2SG    the    key-1SG(-ACC)/    key-2SG(-ACC)/    key-1PL(-ACC)/  
*kulcso-tok(-at)?*  
 key-2PL(-ACC)  
 ‘Where have you put my key/your<sub>sg</sub> key/our key/your<sub>pl</sub> key?’

<sup>7</sup> In some dialects, the accusative *-t* has also appeared on *engem* and *téged*.

### 3 An explanation

The Ugric data surveyed above raise the following questions:

- i. Why do pronominal objects in Mansi and in Northern Khanty (and 1st and 2nd person pronominal objects in Hungarian) bear a possessive agreement morpheme agreeing with the person and number of the given pronoun?
- ii. Why is the possessive “accusative” suffix also present on the pronominal stem when the pronoun is supplied with an oblique case marker?
- iii. Why is it never present on subject pronouns?
- iv. Why is the possessive suffix of these pronouns identified as an accusative case marker in Mansi and Northern Khanty grammars?

The explanation to be proposed is derived from independently motivated analyses of two phenomena of Ugric grammar: reflexive pronouns, and differential object marking.

According to Volkova (2014), reflexive pronouns in Northern (Tegi) Khanty are represented by a possessive construction, where both the pro-dropped possessor and the possessum are personal pronouns of the same person and number, and the possessum bears an agreement suffix cross-referencing the possessor.<sup>8</sup> For example:

- (11) *Uttitexo<sub>i</sub>    luv-el<sub>i/j</sub>    isək-s-alle.*  
 teacher    he-3SG    praise-PST-SG<3SG  
 ‘The teacher praised himself/him.’

The assumption that (11) under the reflexive interpretation involves binding rather than coreference is confirmed by examples involving a quantified subject such as (12). If *luv-el* is understood as a reflexive, the sentence means ‘for no x, x a person, x praised x’.

- (12) *Nemxojat<sub>i</sub>    luv-el<sub>i/j</sub>    änt    isək-s-alle.*  
 nobody    he-3SG    NEG    praise-PST-SG<3SG  
 ‘Nobody praised him/himself’

In the Ob-Ugric languages, only contextually given objects elicit verbal agreement; the verb does not agree with objects introducing a new referent (Nikolaeva 2001; Dalrymple and Nikolaeva 2011). Accordingly, if the verb does not bear object agreement, as in (13), *luvel* cannot be bound by the subject; it only has a disjoint reading:

- (13) *Uttitexo<sub>i</sub>    luv-el<sub>\*i/j</sub>    isək-s.*  
 teacher    he-3SG    praise-PST.3SG  
 ‘The teacher praised him/\*himself.’

In the case of object–verb agreement, both the bound and the disjoint interpretations are possible. *Luvel* can be licensed as a referentially independent pronoun because reflexives also serve as intensifiers of a lexical NP or a pronominal across languages (Baker 1995). In a pro-drop language like Khanty, the pronominal associate of the intensifier may be silent, hence the reflexive itself is intuitively identified with the emphatic referent. In fact, a reflexive pronoun eliciting verbal agreement, e.g., that in (11), is ambiguous between the bound reflexive and the free emphatic interpretation because it

---

<sup>8</sup> This strategy is also employed in other Uralic languages. For an overview, see Burukina (2020).



is structurally ambiguous: it can represent the object, which yields the reflexive interpretation, or the modifier of a pro-dropped object, which yields the emphatic reading.

Reflexive pronouns are personal pronouns supplied with a possessive suffix corresponding in number and person to the stem in the Vasyugan dialect of Eastern Khanty, as well – but in this dialect, also an emphatic *-#-* morpheme intervenes between them (Filchenko 2007: 130–132), e.g.:

- (14) a. *män-t-im*  
I-*t*-1SG  
'myself'  
b. *nöŋ-t-in*  
you-*t*-2SG  
'yourself'  
c. *joy-t-il* / *loy-t-il*  
he-*t*-3SG / he-*t*-3SG  
'himself, herself'

As shown by Filchenko, these pronouns can function either as reflexives (15a) or as emphatic pronouns (15b) in Vasyugan Khanty, as well.

- (15) a. *Mä män-t-im sem-γəl-äm-nə t'i təγi əjnäm wu-γal-im.*  
I I-*t*-1SG eye-DU-1SG-COM DEM place all see-PST-1SG  
'I saw this all with my own eyes.'  
b. pro *joy-t-il küm lüyt-əs.*  
he-*t*-3SG out exit-PST.3SG  
'He himself went out.'

Reflexive pronouns are possessive constructions in Northern Mansi, too. Northern Mansi reflexive pronouns include a *-ki-* morpheme between the pronominal stem and the possessive suffix – see (16a). The personal pronoun that is modified by the emphatic pronoun can be spelled out, yielding a reduplicated structure (Riese 2001: 31) – see (16b).

- (16) a. *am-ki-na:m*  
I-KI-1SG  
'myself'  
b. *am am-ki-na:m*  
I I-KI-1SG  
'I myself'

The Mansi grammar of Riese (2001) calls *-ki* an emphatic clitic. Helimski (1982: 88–97) derived a similar *-ki* morpheme of the corresponding Selkup reflexive pronouns from a Samoyedic noun meaning 'shape, form, soul'. Helimski also related the *-g-* element intervening between the personal pronoun and the possessive suffix in the Hungarian 1st and 2nd person singular pronominal objects (see (9)) to this *-ki* morpheme. The assumption that Uralic reflexives with a possessive ending involve a lexical root that can be traced back to a proto-Uralic word meaning 'shadow, soul' goes back to Majtinskaja (1964). Den Dikken (2006) proposed a similar analysis for the Hungarian accusative pronouns *en-g-em* 'me' and *té-g-ed* 'you-ACC' based on synchronic considerations, claiming that they are

possessive constructions where *én* and *te* are the possessors, and *g* is the left-over of a possessum; possibly the left-over of *mag* ‘kernel’, the element corresponding to ‘self’ in Hungarian reflexive pronouns. These approaches are similar to that of Volkova in that they analyze the pronoun + person-number agreement complex as a (grammaticalized) possessive construction, but, whereas Volkova identifies the pronoun with the possessum, and assumes a pro-dropped possessor, Majtinskaja, Helimski and den Dikken identify the pronoun with the possessor, and assume an obsolete possessum.

An anonymous reviewer has suggested analyzing the person-number suffixes on personal pronouns simply as agreement morphemes independent of possession. It is, in principle, an appealing assumption that emphatic pronouns reduplicate their person and number feature in the form of a suffix, but Majtinskaja’s and Helimski’s proposals argue for preserving the traditional assumption that these pronominal suffixes are possessive agreement morphemes.

The Tegi and the Vasyugan data suggest that the (referentially independent) Ugric pronominal objects and oblique arguments that bear a possessive suffix are emphatically used reflexive pronouns. As argued by Cardinaletti and Starke (1994), pronouns tend to have a weak version and a morphologically more complex strong version, which have different distributions. Apparently, in some Ugric languages, the strong forms of personal pronouns are represented by the corresponding reflexives.

The use of reflexive forms as emphatic pronouns is attested cross-linguistically (Baker 1995). What needs to be explained is why the Ugric emphatic/strong object pronouns discussed above get no case suffixes, and why the emphatic object and oblique pronouns appear to have no weak equivalents without any possessive agreement.

The answer can be derived from the system of differential object marking (DOM) reconstructed for Proto-Ugric. All the present-day Ugric languages and dialects display elements of DOM. As shown by Nikolaeva (1999; 2001) about Khanty, and by Skribnik (2001) about Mansi, the object in the Ob-Ugric languages elicits object–verb agreement if and only if it is a secondary topic, occupying a predicate–phrase–external position. Its topicalization is a resultant of its ‘referential’ and ‘contextually given’ features.<sup>9</sup> (The primary topic role is fused with the subject role in these languages, hence an object cannot be primary topic.) In Hungarian, the criterion of topic status has been replaced by definiteness: the object elicits verbal agreement if it is definite. In Eastern Khanty and in Hungarian, 1st and 2nd person objects elicit no agreement even though they refer to a given referent in most cases, which is derived by É. Kiss (2013; 2017) from the Inverse Agreement Constraint. The Inverse Agreement Constraint is a manifestation, or relic, of the Inverse Topicality Constraint, forbidding that the structurally less prominent secondary topic be more prominent than the primary topic in the topicality hierarchy ‘1st person/2nd person > 3rd person’. (In Hungarian, the hierarchy is more articulated; the 1st person is more prominent than the 2nd person, and singular pronouns are more prominent than plural pronouns of the same person.) If the object is of a higher person than the subject, it cannot be topicalized; it can only be formulated as a focus, eliciting no agreement. The topicality hierarchy is a hierarchy of referents based on how active a role they play in the discourse. Since possessive constructions with a 1st or 2nd person possessor denote a part

---

<sup>9</sup> In Northern Khanty, the object of a secundative construction and the causee of a causative construction have a grammaticalized [+topic] feature, which is independent of their referential and contextual status.

or a belonging of the 1st or 2nd person participant, some languages treat them similarly to 1st/2nd person nominals.<sup>10</sup>

In Eastern Mansi, topicalized objects are not only cross-referenced on the verb but are also marked by a *-m(ə)* accusative suffix (Kulonen 2007: 51; Virtanen 2014), whereas focal objects are caseless. At the same time, objects with a 1st or 2nd person referent, as well as objects with a 1st or 2nd person possessor (denoting a part or a belonging of a 1st or 2nd person referent) are never case-marked, even though they tend to refer to familiar referents, and tend to elicit verbal agreement. The lack of accusative marking on 1st and 2nd person referents used to be a manifestation, and is now a relic, of the same Inverse Topicality Constraint that blocks agreement with 1st and 2nd person objects in Eastern Khanty and in Hungarian: an object that was of a higher person than the subject could not figure as a secondary topic; it could only be formulated as a focus (unless the sentence was passivized and it was promoted to subject-topic.) The fact that Hungarian 1st and 2nd person singular objects bear no accusative suffix, and objects with a 1st or 2nd person possessor can also remain caseless is a consequence of the same type of DOM and the same Inverse Topicality Constraint that is attested in Eastern Mansi. In fact, it is a fossilized consequence both in Hungarian and in Eastern Mansi – because 1st and 2nd person objects are not barred from topic position any more in either language. For example, the Eastern Mansi caseless objects with 1st and 2nd person possessors in (5a–b) are both topics as is shown by the fact that they elicit verbal agreement.

Assuming that the elements of DOM that are shared by at least two Ugric languages represent Proto-Ugric heritage,<sup>11</sup> É. Kiss (2017) reconstructed for Proto-Ugric a system of object marking where the object bears accusative case and elicits verbal agreement if and only if it is topic, and where 1st and 2nd person objects are barred from topic position by the Inverse Topicality Constraint. At this stage of Proto-Ugric, 1st and 2nd person pronominal objects were always part of the predicate phrase, hence they never received accusative case. As they were necessarily focal, they could systematically be represented by strong pronouns. This was true – and still tends to be true in most varieties of Ob-Ugric – of pronominal oblique arguments, as well, as only subjects and objects can be topicalized. An underlying goal, locative, or other oblique argument can be topicalized via passivization (Kulonen 1989). The oblique argument is NP-moved into subject-topic position, where it receives nominative case, which overwrites its inherent case – see (17).

- (17) *Näy tak mujnēt-nə jɔχt-wə-n.*  
 you so guest.PL-LAT come-PASS-2SG  
 ‘Guests come to you.’ Lit.: ‘You are come by guests.’  
 (Eastern Mansi; Kulonen 1989: 158)

The Proto-Ugric system of DOM has been grammaticized to varying degrees in the different Ugric dialects. In Northern Khanty and Northern Mansi, the consistent possessive marking of object pronouns has been extended from 1st and 2nd person pronouns to 3rd person pronouns, as well; in Eastern Mansi, it has been extended to a variant of 3rd person singular pronouns. Since subject pronouns used to be (and still are) topics, whereas 1st and 2nd person object pronouns used to be foci, the possessive

<sup>10</sup> Pronominal imposters of this kind are attested across languages – see Collins (2014).

<sup>11</sup> A reviewer has called attention to the potential drawbacks of extrapolating conclusions based on pieces of evidence attested in different Ugric languages and dialects. Indeed, syntactic reconstruction is necessarily hypothetical.

marking of the focal pronouns has also served the purpose of distinguishing object pronouns from subject pronouns, which eventually led to the reinterpretation of pronominal possessive endings as accusative case suffixes.<sup>12</sup>

The fact that in Eastern Khanty, all lexical objects are caseless whereas pronominal objects can be case-marked can also be accounted for in this framework: lexical noun phrases tend to introduce new referents, therefore, they have grammaticized as foci in this language.

#### 4 Conclusion

By way of conclusion, let us give itemized answers to the questions raised at the beginning of section 2.

i. Pronouns bearing a possessive agreement morpheme agreeing with the person and number of the pronominal stem are reflexive pronouns functioning either as anaphors or as referentially independent strong pronouns. In Proto-Ugric, 1st and 2nd person pronominal objects were barred from topic position by the Inverse Topicality Constraint, i.e., they were necessarily focal, hence they were consistently represented by strong pronouns. In Northern Khanty and Northern Mansi, the consistent possessive marking of object pronouns has been extended to 3rd person pronouns analogically.

ii. Oblique pronouns could not be topicalized in Proto-Ugric, and still cannot be topicalized in various Ob-Ugric languages; therefore, they also appear in their strong forms.

iii. Subject pronouns are inherently topical, hence they always occur in their weak forms.

iv. Since subject pronouns, restricted to topic position, the domain of given information, have been represented by the weak (i.e., base) forms, and since 1st and 2nd person (and in some languages, 3rd person) object pronouns, restricted to the domain of new information, have been represented by the possessive-marked strong forms, the possessive morphemes of the latter have come to be interpreted as object markers.

#### References

- Baker, Carl L. 1995. Contrast, discourse prominence, and intensification, with special reference to locally-free reflexives in British English. *Language* 71(1). 63–101. <https://doi.org/10.2307/415963>
- Burukina, Irina. 2020. *Profile of reflexives in Hill Mari*. To appear in *Folia Linguistica*.
- Caha, Pavel. 2009. The nanosyntax of case. University of Tromsø, PhD dissertation.
- Cardinaletti, Anna & Michal Starke. 1994. The typology of structural deficiency. On the three grammatical classes. *University of Venice Working Papers in Linguistics* 4(2).
- Collins, Chris (ed.). 2014. *Cross-Linguistic Studies of Imposters and Pronominal Agreement*. Oxford: Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780199336852.001.0001>
- Csepregi, Márta. 2017. *Surgutskij Dialect Khantyjskogo Jazyka*. Khanty-Mansijsk: Obsko-ugorskij Institute.

---

<sup>12</sup> In the Samoyedic languages, e.g., in Tundra Nenets, the possessive agreement suffix also appears on some nominative pronouns, and this is the case also with the Hungarian dialectal nominative form *mi-nk* ‘we-1PL’. These must be analogical developments; *mi-nk* is obviously abstracted from *mi-nk-et* ‘we-1PL-ACC’.

- Csepregi, Márta. (forthcoming). Khanty. In Daniel Abondolo & Riitta-Liisa Valijärvi (eds.), *The Uralic Languages*. Abingdon: Routledge.
- Dalrymple, Mary & Irina Nikolaeva. 2011. *Objects and Information Structure*. Cambridge: Cambridge University Press. <https://doi.org/10.1353/lan.2013.0002>
- Dikken, Marcel den. 2006. Where Hungarians agree (to disagree): The fine structure of ‘phi’ and ‘art.’ CUNY Graduate Center. Manuscript.
- É. Kiss, Katalin. 2013. The Inverse Agreement Constraint in Uralic languages. *Finno-Ugric Languages and Linguistics* 2(1). 2–21.
- É. Kiss, Katalin. 2017. The Person–Case Constraint and the Inverse Agreement Constraint are manifestations of the same Inverse Topicality Constraint. *The Linguistic Review* 34(2). 365–396. <https://doi.org/10.1515/tlr-2017-0004>
- Filchenko, Andrey. 2007. A Grammar of Eastern Khanty. Houston: Rice University, PhD dissertation.
- Helimski, Evgenij A. 1982. *Drevnejšie vengerskogo-samodijskie jazыkovye paralleli*. Moskva: Nauka.
- Kálmán, Béla. 1976. *Chrestomathia Vogulica*. Budapest: Tankönyvkiadó.
- Kulonen, Ulla-Maija. 1989. *The Passive in Ob-Ugrian*. Mémoires de la Société Finno-Ougrienne 203. Helsinki: Finno-Ugrian Society.
- Kulonen, Ulla-Maija. 2007. *Itämansin kielioppi ja tekstejä*. Helsinki: Suomalais-Ugrilainen Seura.
- Majtinskaja, Klara E. 1964. *Mestoimenija v mordovskix i marijskix jazыkax*. Moskva: Nauka.
- Nikolaeva, Irina. 1999. *Ostyak*. Languages of the World/Materials 305. München: Lincom Europa.
- Nikolaeva, Irina. 2001. Secondary topic as a relation in information structure. *Linguistics* 39. 1–49. <https://doi.org/10.1515/ling.2001.006>
- Riese, Timothy. 2001. *Vogul*. Languages of the World/Materials 158. München: LINCOM Europa.
- Skribnik, Elena. 2001. Pragmatic structuring in Northern Mansi. In Tõnu Seilenthal (ed.), *Congressus Nonus Internationalis Fenno-ugristarum. Pars IV. Dissertationes sectionum: Linguistica III*. Tartu: Tartu University.
- Smith, Peter W., Moskal, Beata, Xu, Ting, Kang, Jungmin & Bobaljik, Jonathan D. 2019. Case and number suppletion in pronouns. *Natural Language and Linguistic Theory* 37(3). 1029–1101. <https://doi.org/10.1007/s11049-018-9425-0>
- Virtanen, Susanna. 2014. Pragmatic object marking in Eastern Mansi. *Linguistics* 52(2). 391–413. <https://doi.org/10.1515/ling-2013-0067>
- Virtanen, Susanna. 2015. Transitivity in Eastern Mansi. University of Helsinki, PhD dissertation.
- Volkova, Anna. 2014. *Licensing Reflexivity: Unity and variation among selected Uralic languages*. Utrecht: LOT.

Katalin É. Kiss  
 Research Institute for Linguistics, Budapest  
 ekiss@nytud.hu

# Focus in Udmurt: Positions, Contrastivity, and Exhaustivity<sup>1</sup>

Erika Asztalos

The paper presents the results of three surveys examining the positions and the interpretation of foci in Udmurt. While confirming Tánčzos's (2010) findings that the most acceptable focus position is the immediately preverbal one, and that sentence-final focusing is also grammatical for a part of the speakers, the results indicate that foci, with some limitations, can also occur in some preverbal but not verb-adjacent positions. Foci associated with the exhaustive particle *gine* 'only' were highly accepted in all tested positions. From the perspective of interpretation, none of the focus positions turned out to be obligatorily contrastive or necessarily exhaustive. Sentence-initial focusing is mostly available for subjects and for dative complements. As for direct object foci, preverbal but not verb-adjacent positions are mostly accessible for personal pronouns and, more broadly, for objects marked with the accusative case suffix. The more flexible distribution of personal pronoun objects as compared to morphologically unmarked objects is presumably related to the high degree of definiteness of the former. The sentence-final focusing strategy was interpreted as a phenomenon induced by Russian influence and as a sign of the ongoing SOV-to-SVO change of Udmurt. The results also show that speakers vary considerably in their focus position preferences.

Keywords: *focus positions, word order, contrastivity, exhaustivity, Udmurt*

## 1 Introduction

The information structure of the Udmurt sentence is a relatively unexplored area of research, where sometimes even basic questions remain poorly understood. The present paper, which has mainly descriptive aims, addresses two principal questions: i) whether the appearance of the focused constituent is restricted in Udmurt to the immediately preverbal and the sentence-final positions (as Tánčzos 2010 claims), and ii) whether any of the positions in which foci can occur is obligatorily exhaustive and/or contrastive.

The data presented in this paper may also be relevant from a typological point of view. Traditionally, Udmurt has been classified as an SOV language, but some recent works (e.g., Tánčzos 2013, Asztalos et al. 2017, Asztalos 2018) claim that it is undergoing an SOV-to-SVO change. Since SOV and SVO languages have different focus positioning tendencies (see Cypionka 2007), it is of interest to see how contemporary Udmurt behaves with regard to focus placement.

On the basis of the results of a fieldwork study carried out by means of three consecutive questionnaires filled out by native speakers of Udmurt, the paper argues that

---

<sup>1</sup> This work was financially supported by the following grants and projects: Erasmus Mundus "Aurora" program; project ERC\_15\_HU, OTKA-118079 "Languages under the Influence" of the National Science and Research Fund of Hungary; projects NKFI-125282 "Typological Database of the Volga Area Finno-Ugric Languages", NKFI-125206 "Nominal Structures in Uralic Languages", and NKFI-129921 "Implications of endangered Uralic languages for syntactic theory and the history of Hungarian" (National Research Innovation and Development Office of Hungary). I am grateful to two anonymous reviewers as well as to Balázs Surányi for their helpful comments and suggestions, which helped to improve this paper. Special thanks are due to my Udmurt consultants for filling in my questionnaires.

besides the immediately preverbal and the sentence-final positions (cf. Tánčzos 2010), foci can also occur preverbally but not adjacent to the verb. Namely, they can precede a preverbal adverbial and/or the subject, thus occurring sentence-medially or sentence-initially. Preverbal but not verb-adjacent placement of foci is, however, sensitive to the morphosyntactic properties of the focussed element. Sentence-initial focusing resulted to be mostly available for subjects and for dative complements. As for object foci, preverbal but not verb-adjacent positions were mostly accessible for personal pronoun objects and, more broadly, for objects marked with the accusative case suffix. The more flexible distribution of personal pronoun objects (and of accusative-marked objects in general) compared to morphologically unmarked objects is presumably related to the higher degree of definiteness of the former object types.

The results indicate that exhaustively and contrastively focused items can occur in all of the tested positions, however, none of these positions is *obligatorily* exhaustive or necessarily contrastive.

Speakers seem to vary extensively in their focus position preferences and flexibility with regard to focus placement. Certain speakers clearly preferred one focus position: most frequently, the immediately preverbal one, more rarely, the “pre-adverbial” or the sentence-final one. Other speakers were more permissive, as they consistently judged as grammatical more than one focus position.

From a typological point of view, Udmurt seems to behave like an SOV language which is undergoing a change towards the SVO type: while immediately preverbal focusing as a main focusing strategy is characteristic of SOV languages, sentence-final focusing is present in SVO languages but absent in SOV languages (see Czypionka 2007). The sentence-final focus position has presumably developed in Udmurt under the influence of Russian (see also Tánčzos 2010). It is interesting, however, that sentence-initial focusing, which is also available in Russian and is, actually, the most common focusing strategy in SVO languages and is also quite common in SOV languages (see Czypionka 2007), resulted to be more marked and is subject to restrictions in Udmurt.

The paper is structured as follows. Section 2 presents background information. After discussing neutral word order(s) in Udmurt, I outline the typological tendencies of focus placement in SOV and SVO languages. Afterwards, I offer an overview of previous works on Udmurt focus, then I introduce the notions of information structure the paper relies on and provide a short overview of the Russian focus positions. Section 3 introduces the research aims and the questionnaires by means of which the research was carried out. Section 4 presents and discusses the results. 4.1 is concerned with focus placement in relation to the morphosyntactic properties of the focused element. 4.2 addresses the question whether any of the Udmurt focus positions is necessarily contrastive and/or exhaustive. 4.3 provides a speaker-internal evaluation of the results. 4.4 discusses the results from a typological point of view and deals with the question to what extent Russian may have had an influence on focus placement in Udmurt. Section 5 draws the conclusion and points out some questions left for future research.

## 2 Background

### 2.1 Neutral order of sentence constituents in Udmurt

Udmurt has traditionally been claimed to be a non-rigid SOV (or head-final) language. Thus, the neutral order has been claimed to be SOV (or SXV) at the sentence-level (1) and

*modifier-head* at the phrasal level, while non-verb-final sentences and head-initial phrases have been considered to be pragmatically marked (cf., e.g., Bulyčov 1947; Gavrilova 1970; Csúcs 1990; Suihkonen 1990; Vilkuna 1998; Winkler 2001, 2011; Tánczos 2010; Timerxanova 2011).<sup>2</sup>

- (1) *Saša kníga-jez' hydž-i-ž.*<sup>3</sup>  
 Sasha book-ACC read-PST-3SG  
 'Sasha read the book.' (Tánczos 2010: 223)

Several recent studies (Tánczos 2013; Asztalos & Tánczos 2014; Asztalos 2016, 2018; Asztalos et al. 2017), however, claim that in contemporary Udmurt, both SOV and SVO orders can be neutral. By (discourse-)neutral sentences most of these papers mean to refer to *all-new* sentences, which include, for example, text-initial sentences and sentences answering the question 'What's new?'. The example in (2), e.g., is an *all-new* sentence with SVO order.

- (2) *Ogjaulonni-ys' starosta bića podpis-jos.*  
 dormitory-ELA head gather.3SG signature-PL  
 'The dormitory supervisor is gathering signatures.'  
 (Marajko, 25.08.2015, cited in Asztalos 2018: 79)

The authors of the cited papers assume (and their assumption will be adopted throughout the present study) that the contemporary Udmurt language is undergoing a typological change from the OV to the VO type under the influence of Russian. At the same time, it has to be noted that (S)VO order, and head-initial constituents both at the clausal and the phrasal level, are textually less frequent than (S)OV order and head-final constituents in general, and they are mainly produced and accepted by the younger generation (see Asztalos 2016, 2018).

---

<sup>2</sup> A typical example of pragmatically marked, non-verb-final sentences are emphatic sentences with discourse-old postverbal constituents, cf. (i) (cf. Ponarjadov 2010: 14, 23, 27):

- (i) *T'urma-yn šist-o mon ton-e!*  
 prison-INE putrify-FUT.1SG 1SG 2SG-ACC  
 'I will putrify you in the prison!' (Ponarjadov 2010: 27)

<sup>3</sup> The following abbreviations are used in the glosses and tables: 1 = first person, 2 = second person, 3 = third person, Acc, ACC = accusative case, CMPR = comparative, CNG = connegative form of the verb, CVB = converb, DAT = dative case, DET = determinative suffix, ELA = elative case, FUT = future tense, ILL = illative case, IMP = imperfect, INE = inessive case, INS = instrumental-comitative case, Nom = nominative case, NEG = negative auxiliary, PL = plural, PRF = perfect, PRS = present tense, PRT = perfectivizer, PST = past tense, PTCL = particle, PTCP = participle, Q = question particle, SG = singular. Other abbreviations used in the body text and the figures are the following: Adv = adverbial, Adv<sub>TEMP</sub> = temporal adverbial, Ins = noun phrase in the instrumental-comitative case, NP = noun phrase, O<sub>FOC</sub>/Ofoc = focused direct object, O<sub>PRON</sub> = personal pronoun object, S = subject, S<sub>PRON</sub> = personal pronoun subject, V = verb, w.o. = word order. Glosses, transcriptions and (in some cases) translations of cited examples are mine.



## 2.2 Focus positions in SOV and SVO languages

Examining how a language undergoing an SOV-to-SVO change, like Udmurt, behaves with regard to focus placement, is not of merely descriptive interest but also has broader typological relevance, since SOV and SVO languages have different focus positioning tendencies. Cypionka (2007), in a typological study examining 112 languages, finds a correlation between unmarked (neutral, or basic) word order and focus position, stating that SOV languages are more likely to encode focus preverbally than SVO languages. In her sample, 36% of SOV languages but only 7% of SVO languages, showed a preference for the immediately preverbal focus position.<sup>4</sup> On the other hand, none of the SOV languages had a sentence-final focus position, while 10% of SVO languages did have it. Postverbal focusing also resulted to be less common among SOV than among SVO languages (3% vs. 13%). Interestingly, sentence-initial focusing was available in roughly the same proportion of SOV and SVO languages (34% vs. 37%) (ibid.: 441–444).<sup>5</sup>

Many languages also allowed for other focus positions in addition to the most common one. Thus, for most of the languages, *in situ* focusing was also an option (ibid.: 441). Furthermore, for the majority of SOV languages with a preference for immediately preverbal focusing, the existence of a sentence-initial focus position is not explicitly excluded by the grammars consulted by the author. Similarly, the possibility of immediately preverbal focusing is not excluded for most SOV languages having a sentence-initial focus position (2007: 443). As for SVO languages, the postverbal focus position also often co-occurs with an alternative sentence-initial focus position (2007: 444).

Cypionka (2007) also deals with the question whether subject and non-subject foci show different positioning tendencies, and finds that when focus marking involves movement in a language (i.e., the placement of the focused item into a dedicated position as opposed to *in situ* focusing), subject and non-subject foci are moved to the same position (2007: 439, 443).

To sum up, Cypionka's (2007) data reveal that SOV and SVO languages show the following tendencies with regard to focus placement:

- Immediately preverbal focusing is more typical of SOV than of SVO languages.
- Sentence-final and postverbal focusing is more frequent in SVO than in SOV languages.
- Sentence-initial focusing is roughly as common in SOV as in SVO languages.
- Many languages have more than one focusing strategy.

---

<sup>4</sup> In fact, those 7% include only two languages, which, as Cypionka (ibid.: 5) points out, are not even entirely clear regarding this feature. In any case, immediately preverbal focusing does not seem to be a property of SVO languages specifically.

<sup>5</sup> Verb-initial (VSO, OVS) and object-initial (OSV, OVS) languages typically have a sentence-initial focus position in Cypionka's sample, but, as the number of these languages is much lower in the sample than the number of SOV and SVO languages, the author does not consider the results for the former languages as reliable as for the latter (Cypionka 2007: 445).

### 2.3 Previous works on Udmurt focus

Early grammars and works on Udmurt syntax contain some observations about the placement of so-called “logically stressed” constituents (in Russian, *logičeski udarjaemoe slovo*). Although the authors do not specify what they exactly mean by logically stressed constituents, on the basis of the usual interpretation of the term in the literature and the provided examples it is feasible that they refer by the term to constituents fulfilling a focus-like function.

The opinions concerning the placement of these items partly differ. Glezdenev (1921: 15, 45) and Baushev (1929: 10) claim that logically stressed elements immediately precede the predicate. Thus, in the sentence in (3), logical stress falls on the direct object *iz korka* ‘stone house’, which is in immediately preverbal position. For emphasizing another element of the sentence, e.g., the adverbial *tolon* ‘yesterday’, or the subject *vuž karis* ‘tradesman’, the order of the sentence has to be altered so that the emphasized element immediately precede the verb (Glezdenev 1921: 45).

- (3) *Tolon vuž kar-is kar-yn IZ KORKA bašt-i-ž.*<sup>6</sup>  
 yesterday product make-PTCP.IMP city-INE stone house buy-PST-3SG  
 ‘Yesterday the tradesman bought A STONE HOUSE in the city. / It was a stone house that the tradesman bought yesterday in the city.’ (Glezdenev 1921: 45)

Žujkov (1937: 18), however, provides examples in which logically stressed constituents are placed sentence-initially, without being immediately preverbal (4):

- (4) a. TUNNE *mon zavod-e myn-o.*  
 today 1SG factory-ILL go-FUT.1SG  
 ‘It is today that I will go to the factory.’  
 b. ZAVOD-E *tunne mon myn-o.*  
 factory-ILL today 1SG go-FUT.1SG  
 ‘It is to the factory that I will go today.’  
 c. MON *tunne zavod-e myn-o.*  
 1SG today factory-ILL go-FUT.1SG  
 ‘It is me who will go to the factory today.’ (Žujkov 1937: 18)

According to Bulyčov (1947: 77), logically stressed constituents can occur sentence-initially or stay in their “ordinary” position (1947: 78) (by which he probably means neutral or *in situ* position). Konjuxova (1964: 6) claims that logical stress can fall on any constituent of the sentence without entailing constituent reordering, which equals saying that constituents can be focused in their neutral position. Thus, the sentence in (5) may express different meanings depending on which constituent is logically stressed.

- (5) a. PINAL-JOS *kolhoz-yn už-a-žy.*  
 child-PL kolkhoz-INE work-PST.3PL  
 ‘It is the children who have worked in the kolkhoz.’

---

<sup>6</sup> Focused constituents are marked by small capitals throughout the whole study.

- b. *Pinaljos KOLHOZYN užazy.*  
‘It is in the kolkhoz that children have worked.’
- c. *Pinaljos kolbožyn UŽAZY.*  
‘Work was what children have done in the kolkhoz.’ (Konjuxova 1964: 6)

Summing up, early works mention three possible positions for logically stressed items: i) immediately preverbal, ii) sentence-initial and iii) neutral (*in situ*) position.

The first paper offering a thorough analysis of focus placement in Udmurt is written by Tánčzos (2010). According to her, topic and focus are structurally marked in the language. The topic position is sentence-initial and recursive (ibid.: 219). The focus position, which is not recursive, immediately precedes the predicate in the standard variety of Udmurt (6a), while it is sentence-final in a non-standard variety of the language (6b) (ibid.: 219). The author attributes the development of sentence-final foci in Udmurt to the influence of Russian (ibid.: 222), as in Russian, information foci are located sentence-finally (cf. Bailyn 2012: 275–278).

- (6) Context: ‘What did Sasha see in the cinema?’
- a. *Saša kinol'eatr-yn T'ERMINATOR-EZ ućk-i-ž.*  
Sasha cinema-INE Terminator-ACC watch-PST-3SG
- b. *Saša kinol'eatr-yn ućk-i-ž T'ERMINATOR-EZ.*  
Sasha cinema-INE watch-PST-3SG Terminator-ACC
- ‘It is the Terminator that Sasha saw in the cinema.’ (Tánčzos 2010: 225)

However, other papers (Vilkuna 1998; Timerxanova 2006, 2011; Asztalos 2012) suggest that the possibilities of focus placement are not limited to the immediately preverbal and the sentence-final positions. Vilkuna (1998: 195) claims that “focus does not appear to be positionally restricted” in Udmurt, and that the preverbal position is a frequent but not exclusive position for focused elements:

“The (...) Udmurt preverbal position seems to be a neutral and frequent focus and WH position, but this does not prohibit the placement of WH items and exhaustive foci elsewhere. (...) It seems that when the neutral position of a constituent is preverbal, it will remain there when focused, but, for example, a subject is not necessarily placed in this position for focusing purposes” (ibid.).

Timerxanova (2006), similarly to Žujkov (1937) and Bulyčov (1947), claims that logically stressed items are placed sentence-initially. In a later paper (Timerxanova 2011), however, she associates more than one order – namely, SVO (7a), OVS (7b) and OSV (7c) – with object focusing, which implies that besides the sentence-initial position, she also designates a sentence-final and an immediately preverbal focus position, at least for direct object foci:

- (7) a. *Mon adž-is'ko N'ULES-EZ.*  
1SG see-PRS.1SG forest-ACC
- b. *N'ULES-EZ adž-is'ko mon.*  
forest-ACC see-PRS.1SG 1SG
- c. *N'ULES-EZ mon adž-is'ko.*  
forest-ACC 1SG see-PRS.1SG
- ‘It is the forest that I see.’ (Timerxanova 2011: 183)

Asztalos (2012) presents the results of a small-scale experiment that tested the possible positions of direct object foci in two contexts, contrastive and non-contrastive (examples below are given in a non-contrastive context). Independently of whether the context was contrastive or not, the position accepted by most speakers was the immediately preverbal one (8a). However, sentence-final object foci (8b), as well as object foci preceding the verb *non-immediately* (8c) were also allowed by some speakers. Marginally, sentence-initial (8d) and postverbal but not sentence-final (8e) object foci were also accepted. No difference between the placement of contrastive and non-contrastive foci was found (Asztalos 2012: 10–11).

- (8) Context: ‘What did Vova drink yesterday?’
- a. *Vova tolon* SUR *ju-i-ŋ*.  
Vova yesterday beer drink-PST-3SG
  - b. %*Vova tolon ju-i-ŋ* SUR.  
Vova yesterday drink-PST-3SG beer
  - c. %*Vova* SUR *tolon ju-i-ŋ*.  
Vova beer yesterday drink-PST-3SG
  - d. %/?SUR *Vova tolon ju-i-ŋ*.  
beer Vova yesterday drink-PST-3SG
  - e. %/?*Vova ju-i-ŋ* SUR *tolon*.  
Vova drink-PST-3SG beer yesterday
- ‘It was beer that Vova drank yesterday.’ (on the basis of Asztalos 2012: 10)

In (8c), a temporal adverbial, whereas in (8d), the subject and a temporal adverbial stand between the focused object and the verb. As a matter of fact, Tánčzos (2010) also makes a brief observation (2010: 222), which implies that some of her respondents may have allowed the adverbial to appear between the focused element and the verb, but the author does not go into detail about this.<sup>7</sup>

To sum up, while the most comprehensive work on Udmurt focus (Tánčzos 2010) posits two focus positions (immediately preverbal in the standard variety and sentence-final in a non-standard variety of the language), other works (Žujkov 1937; Bulyčov 1947; Konjuxova 1964; Vilkuňa 1998; Timerxanova 2006, 2011 and Asztalos 2012) suggest that focus placement is not restricted to these two specific positions: instead focused phrases may occasionally occur sentence-initially, in a postverbal but not sentence-final position, or they may stay *in situ*, i.e. in their canonical position.

## 2.4 Terminology

This section introduces the key concepts that are relevant for the present study. Focus, along with its different subtypes, has been defined in a number of ways in linguistics. The present paper mainly relies on the definitions of É. Kiss (1998), who makes a distinction between two main focus types, *information focus* and *identificational focus*. Two semantic features, *exhaustivity* and *contrastivity*, that cross-linguistically may optionally or obligatorily

---

<sup>7</sup> ‘(...) in most cases, most of the speakers do not allow the adverbial to stand between the focused element and the verb’ (Tánčzos 2010: 222; translation mine).

be associated with foci, are also relevant for the purposes of this study. Additionally, the paper also refers to the notion of *corrective focus*.

Information focus, as defined by É. Kiss (1998), “conveys new, non-presupposed information [...] without expressing exhaustive identification” (É. Kiss 1998: 246). E.g., in the Hungarian sentence in (9), the constituent *egy kalapot* ‘a hat’ introduces new, non-presupposed information, and thus fulfils the role of information focus. The sentence does not imply that everything Mary picked for herself was a hat: the predicate can potentially hold for other elements, too.

- (9) Context: John and Mary are shopping.  
*Mari ki-néz-ett magá-nak EGY KALAP-OT.*  
 Mary out-watch-PST.3SG herself-DAT a hat-ACC  
 ‘Mary picked for herself a hat.’ (É. Kiss 1998: 249) (Hungarian)

Information foci typically appear *in situ* (or, in other words, in their base-generated position) (É. Kiss 1998: 249).

Identificational focus, on the other hand, identifies the exhaustive subset of “contextually or situationally given elements for which the predicate phrase [...] actually holds” (É. Kiss 1998: 245), and, according to É. Kiss’s (1998) analysis, it involves a specific structural position in a functional projection of the sentence. Thus, the English sentence in (10) and its Hungarian counterpart in (11) imply that from among various pieces of clothes, Mary picked for herself a hat, and she did not pick anything else (É. Kiss 1998: 249.) Exhaustivity is thus a semantic property of identificational focus in both languages. In English, identificational focus is realized via the cleft construction *It is...* (10), while in Hungarian identificational foci occupy the position immediately preceding the verb (11).

- (10) *It was a hat that Mary picked for herself.*  
 (11) *Mari EGY KALAP-OT néz-ett ki magá-nak.*  
 Mary a hat-ACC watch-PST.3SG out herself-DAT  
 ‘It was a hat that Mary picked for herself.’ (É. Kiss 1998: 249) (Hungarian)

Cross-linguistically, identificational focus can be obligatorily or optionally contrastive. A focus, according to É. Kiss (1998: 267), is contrastive if “it operates on a closed set of entities whose members are known to the participants of the discourse”. Thus, in the case of contrastive foci, “the identification of a subset of the given set also identifies the contrasting complementary subset” (ibid.). Identificational focus is obligatorily contrastive, for example, in Italian: the answer sentence in (12c) with sentence-initial identificational focus is only grammatical if it operates on a context with a closed set of possible entities known to the participants of the discourse (É. Kiss 1998: 269). Thus, the sentence in (12c) (which is equal to (13b)) is grammatical as an answer to the questions in (12a–b), but it is ungrammatical in the context of (13a), as the latter is a simple *wh*-question, which is a context with an open set of entities.

- (12) a. *Chi di voi due ha rotto il vaso?*  
 which of 2PL two have.3SG break.PTCP.PRF the vase  
 ‘Which one of you two broke the vase?’

- b. *L' ha rotto Giorgio, il vaso?*  
 it.ACC have.3SG break.PTCP.PRF George the vase  
 'Did George break the vase?'
- c. *MARIA ha rotto il vaso.*  
 Mary have.3SG break.PTCP.PRF the vase  
 'It is Mary who broke the vase.' (É. Kiss 1998: 269) (Italian)
- (13) a. *Chi ha rotto il vaso?*  
 who have.3SG break.PTCP.PRF the vase  
 'Who broke the vase?'
- b. \**MARIA ha rotto il vaso.*  
 Mary have.3SG break.PTCP.PRF the vase  
 'It is Maria who broke the vase.' (ibid.) (Italian)

In English and in Hungarian, the position reserved for identificational foci is not necessarily contrastive, which means that it can host both contrastive and non-contrastive items. The Hungarian example in (14) illustrates that the sentence in (14c) can be given as an answer both to a question with a closed set of entities known to the participants of the discourse (14a) (contrastive context), and to a simple *wh*-question, which operates on an open set of entities (14b) (non-contrastive context) (É. Kiss 1998: 267–268).

- (14) a. *Mari egy kalap-ot vagy egy sál-at néz-ett ki magá-nak?*  
 Mary a hat-ACC or a scarf-ACC watch-PST.3SG out herself-DAT  
 'Did Mary pick for herself a hat or a scarf?'
- b. *Mit néz-ett ki magá-nak Mari?*  
 what watch-PST.3SG out herself-DAT Mary  
 'What did Mary pick for herself?'
- c. *Mari EGY KALAP-OT néz-ett ki magá-nak.*  
 Mary a hat-ACC watch-PST.3SG out herself-DAT  
 'It was a hat that Mary picked for herself.' (Hungarian)

It is important to note that even if in a given language like Italian identificational focus is obligatorily contrastive, this does not imply that foci which occur in a contrastive context are obligatorily moved into the identificational focus position in that language. In fact, contrastively focused items in many languages can also stay *in situ*, and/or occur in the position where information foci are placed in the language. This is illustrated by the Italian example in (15c), which can also be given as a grammatical and congruent answer to the questions in (12a–b) (repeated here as (15a–b)).

- (15) a. *Chi di voi due ha rotto il vaso?*  
 which of 2PL two have.3SG break.PTCP.PRF the vase  
 'Which one of you two broke the vase?'
- b. *L' ha rotto Giorgio, il vaso?*  
 it.ACC have.3SG break.PTCP.PRF George the vase  
 'Did George break the vase?'
- c. *Il vaso, l' ha rotto MARIA.*  
 the vase it.ACC have.3SG break.PTCP.PRF Mary  
 'It is Maria who broke the vase.' (É. Kiss 1998: 269) (Italian)

To put it another way, information foci, in an appropriate context, can also be used contrastively, cf. (15c), but as opposed to identificational foci they are never associated with an *obligatorily* contrastive reading (recall that the main function of information foci is to introduce new, non-presupposed information). Surányi's (2011) study suggests that the situation is somewhat analogous to the exhaustivity of information foci in Hungarian. As stated at the beginning of this section, the Hungarian sentence in (9) (repeated here as (16)), with the constituent *egy kalapot* 'a hat' fulfilling the role of information focus, does not imply that Mary only picked a hat for herself. However, it does not explicitly *exclude* the possibility that Mary only picked for herself a hat: the sentence might well be continued, e.g., by a sentence which means 'She bought it immediately and then they left', which would in fact suggest that she didn't buy anything else.

- (16) Context: John and Mary are shopping.  
*Mari ki-néz-ett magá-nak EGY KALAP-OT.*  
 Mary out-watch-PST.3SG herself-DAT a hat-ACC  
 'Mary picked for herself a hat.' (É. Kiss 1998: 249) (Hungarian)

Thus, it might be appropriate to state that, as opposed to identificational focus, information focus *by itself* does not provide information about the exhaustivity of the focussed element (it does not encode exhaustivity semantically), but such information, in some cases, might be inferred pragmatically from the context. Thus, information foci can be associated with *pragmatic* exhaustivity (see Surányi 2011: 292–295). This is to be distinguished from the context-independent, semantically encoded type of exhaustivity presented above in relation to identificational foci. The present study is concerned with this latter type of exhaustivity in Udmurt.

It has to be noted that the context in (12b–c), which is considered by É. Kiss (1998) a contrastive one, is, in fact, a so-called *correction*. Foci used in corrections are often regarded in the literature as instances of a distinct (sub)type of focus, *corrective focus*. However, as there is also a long-standing tradition of using corrections as a means for the elicitation of contrastive foci (see Repp 2016: 280–281, 283), in this paper I will consider corrective focus as a subtype of contrastive focus.

## 2.5 Focus positions in Russian

Udmurt is subject to strong Russian influence. According to Salánki's (2007) sociolinguistic study, 98% of Udmurt speakers are bilingual and speak both Udmurt and Russian (Salánki 2007: 81). However, generations differ concerning their competence in Udmurt and Russian (*ibid.*: 89, 205): while older Udmurts are usually Udmurt-dominant speakers and middle-aged speakers typically have an equal command of Udmurt and Russian (*ibid.*: 82), the young generation frequently has higher proficiency in Russian than in Udmurt (that is, they are either balanced or Russian-dominant bilinguals) (*ibid.*: 82, 85).

Russian influence can be detected at every linguistic level in Udmurt (Csúcs 1990: 21). Morphosyntactic phenomena induced by Russian influence include, among others, the usage of plural forms after numerals, number agreement on attributive adjectives, the usage of Russian conjunctions and complementizers, the spreading of finite subordination to the detriment of non-finite subordination, etc. (see Salánki 2007: 158–185). The ongoing SOV-to-SVO change of Udmurt has also been attributed (at least partly) to the influence of Russian (see Asztalos et al. 2017; Asztalos 2018). From this general perspective, it may be of interest to examine whether Russian may have had an impact on the focusing

strategies of Udmurt. Thus, in what follows I will give an overview of the Russian focus positions and their interpretation on the basis of the related literature.

Foci in Russian may occur sentence-finally or preverbally. Sentence-final foci (17) have been analysed as information foci by King (1995), Neeleman & Titov (2009), Dyakonova (2009), Titov (2012), and Bailyn (2012).

- (17) Context: ‘Who is reading the book?’  
*Knigu čita-jet IVAN.*  
 book.ACC read-3SG Ivan  
 ‘It is Ivan who is reading the book.’ (Bailyn 2012: 276) (Russian)

As introduced in the previous subsection, cross-linguistically information foci are not associated with an *obligatory* contrastive or exhaustive reading, but optionally, in an appropriate context, they may have such readings. This is also true for Russian sentence-final information foci, as Dyakonova (2009: 67–68) shows.

As for Russian preverbal foci, Dyakonova (2009: 64) points out that they can occur in three distinct positions (at least in colloquial Russian): they can precede the verb immediately (18a), occur in the middle-field but not adjacent to the verb (18b), or appear sentence-initially (18c):

- (18) a. *Oni emu ŠČENKA podarili.*  
 3PL 3SG.DAT puppy.ACC give.PST.3PL  
 b. *Oni ŠČENKA emu podarili.*  
 3PL puppy.ACC 3SG.DAT give.PST.3PL  
 c. *ŠČENKA oni emu podarili.*  
 puppy.ACC 3PL 3SG.DAT give.PST.3PL  
 ‘They gave him a PUPPY.’ (Dyakonova 2009: 64) (Russian)

Whether preverbal foci in Russian are necessarily contrastive and/or exhaustive is a matter of some dispute. King (1995) and Titov (2012: 272–282) claim that they are necessarily contrastive. Neeleman & Titov (2009) discuss sentence-initial foci and regard them as contrastive. However, Dyakonova (2009) and Bailyn (2012) argue that preverbal foci are not necessarily contrastive, nor are they obligatorily exhaustive, as they may also occur in non-contrastive contexts, e.g., as answers to *wh*-questions (Dyakonova 2009: 71–73; Bailyn 2012: 281–282).

Summing up, foci can occur sentence-finally or preverbally in Russian. Preverbal foci can be left-adjacent to the verb, sentence-initial, or occur in the middle-field but not adjacent to the verb. Sentence-final foci are instances of information focus. All positions can host contrastive foci and none of them is necessarily exhaustive. There is no consensus on whether preverbal foci are necessarily contrastive, but the fact that they can also answer *wh*-questions suggests that they are not associated with an obligatorily contrastive reading.

### 3 Research aims and the questionnaires

The primary goal of the fieldwork study presented in this paper was to test to what extent native speakers of Udmurt accept sentence-initial, non-immediately preverbal and postverbal (but not sentence-final) foci compared to immediately preverbal and sentence-final ones (identified by Tánzos 2010), and to reveal whether focus placement is



influenced by the syntactic function and, in case of direct object foci, the morphological marking and the lexical subcategory (noun/personal pronoun) of the focused item. Second, the investigations aimed at examining whether any of the focus positions is associated in Udmurt with an obligatorily contrastive or exhaustive reading. The third aim was to compare the revealed properties of Udmurt foci with those of the Russian preverbal and sentence-final focus positions, and to check to what extent focus placement and focus interpretation in Udmurt may be influenced by Russian.

The research was carried out by means of three consecutive questionnaires (hereinafter: Questionnaire 1, 2 and 3) that were compiled and filled out, respectively, in 2013, 2014 and 2016. Questionnaire 1 and 2 were filled out each by 12 native speakers of Udmurt, who were mainly employees and students of the Udmurt State University. Questionnaire 3, which was designed together with Katalin É. Kiss (and first reported in Asztalos & É. Kiss 2016), was an online survey sent out through the social networking sites *Facebook* and *Vkontakte*. In the latter survey, 36 complete and 24 incomplete responses were collected.<sup>8</sup>

Questionnaire 1 concentrated exclusively on direct object foci. Udmurt has differential object marking: non-specific direct objects are morphologically unmarked (formally identical to the nominative), whereas specific objects (including personal pronouns) are accusative-marked (see É. Kiss & Tánzos 2018: 738–739, 752–753). Questionnaire 1 aimed at examining whether the placement of object foci is influenced by their morphological marking and/or lexical subcategory (proper noun vs. personal pronoun). This question may be legitimate because Vilkuna's (1998) results point to a possible relationship between the morphological marking and the position of direct objects (for more on this, see Section 4.1.3 below). The related questionnaire items consisted of *wh*-questions and a set of possible answer sentences associated to each question, as illustrated by the examples in Appendix A and their glossed and translated version in (19)–(20). For each *wh*-question, the respondents had to choose from the related list all those sentences that, in their opinion, can figure as grammatical and congruent answers to the question. The *wh*-questions contained (besides the *wh*-element) a subject (S), a locative adverbial (Adv), and a verb (V). The answer sentences contained the same elements as the *wh*-questions, except for the object, which was realized in the answers by a noun phrase or a personal pronoun (which was interpreted as a focus, labelled O<sub>FOC</sub>). The only difference between the possible answer sentences belonging to one question consisted in the order of the constituents, and especially in the position of the focused object.

In order to help the respondents to keep in mind that it is the direct object that has to be elicited by the questions, the object was written with capital letters and a photo illustrating it was attached to the answer sentences (see Appendix A). The answer sentences appeared in randomized order within each item.

- (19) *Mar Lera magažin-yś bašt-i-ž?*  
 what Lera grocery-ELA buy-PST-3SG  
 ‘What did Lera buy at the grocery?’

---

<sup>8</sup> The sets of respondents of Questionnaire 1 and Questionnaire 2 partly overlapped. None of the questionnaires contained filler items, and respondents were not compensated for their participation in the survey(s).

- (20) a. *Lera* KUREG *magaşin-yş* *başt-i-z*. (SO<sub>FOC</sub>AdvV)  
 Lera chicken grocery-ELA buy-PST-3SG
- b. *Lera magaşin-yş* *başt-i-z* KUREG. (SAdvVO<sub>FOC</sub>)  
 Lera grocery-ELA buy-PST-3SG chicken
- c. *Lera magaşin-yş* KUREG *başt-i-z*. (SAdvO<sub>FOC</sub>V)  
 Lera grocery-ELA chicken buy-PST-3SG
- d. KUREG *Lera magaşin-yş* *başt-i-z*. (O<sub>FOC</sub>SAdvV)  
 chicken Lera grocery-ELA buy-PST-3SG
- e. *Lera başt-i-z* *magaşin-yş* KUREG. (SVAdvO<sub>FOC</sub>)  
 Lera buy-PST-3SG grocery-ELA chicken
- f. *Lera başt-i-z* KUREG *magaşin-yş*. (SVO<sub>FOC</sub>Adv)  
 Lera buy-PST-3SG chicken grocery-ELA
- Intended meaning: ‘It is chicken that Lera bought at the grocery.’

The placement of contrastive foci was tested with alternative *wh*-questions of the type *What did Lera buy at the grocery, chicken or duck?* This type of question is called “interrogative discourse with alternative question”, and it is identified by Repp (2016: 281) as one of the tests commonly used for the elicitation of contrastive foci. The related answer sentences were completed by a clause negating one of the objects, and the negated object was illustrated by a photo that was crossed out. This is illustrated by the examples in Appendix B and their glossed version in (21)–(22).

- (21) *Mar Lera magaşin-yş* *başt-i-z*, *kureg jake* *çöş?*  
 what Lera grocery-ELA buy-PST-3SG chicken or duck  
 ‘What did Lera buy at the grocery, chicken or duck?’
- (22) a. *Lera magaşin-yş* KUREG *başt-i-z*, *çöş* *öş* *baştı*.  
 Lera grocery-ELA chicken buy-PST-3SG duck NEG.PST.3 buy.CNG.SG  
 (SAdvO<sub>FOC</sub>V)
- b. *Lera magaşin-yş* *başt-i-z* KUREG, *çöş* *öş* *baştı*.  
 Lera grocery-ELA buy-PST-3SG chicken duck NEG.PST.3 buy.CNG.SG  
 (SAdvVO<sub>FOC</sub>)
- c. KUREG *Lera magaşin-yş* *başt-i-z*, *çöş* *öş* *baştı*.  
 chicken Lera grocery-ELA buy-PST-3SG duck NEG.PST.3 buy.CNG.SG  
 (O<sub>FOC</sub>SAdvV)
- d. *Lera başt-i-z* KUREG *magaşin-yş*, *çöş* *öş* *baştı*.  
 Lera buy-PST-3SG chicken grocery-ELA duck NEG.PST.3 buy.CNG.SG  
 (SVO<sub>FOC</sub>Adv)
- e. *Lera* KUREG *magaşin-yş* *başt-i-z*, *çöş* *öş* *baştı*.  
 Lera chicken grocery-ELA buy-PST-3SG duck NEG.PST.3 buy.CNG.SG  
 (SO<sub>FOC</sub>AdvV)

- f. *Lera bašt-i-ž magažin-yś KUREG, čöž öž bašty.*  
 Lera buy-PST-3SG grocery-ELA chicken duck NEG.PST.3 buy.CNG  
 (SAdvO<sub>FOC</sub>)

Intended meaning: ‘It is chicken that Lera bought at the grocery, not duck.’<sup>9</sup>

Table 1 summarizes the different levels of the two main factors (context and object type) tested in Questionnaire 1. The context was either non-contrastive or contrastive, while the object was either a common noun, or a proper noun, or a pronoun. Common nouns appeared either in the nominative or in the accusative. Proper nouns and pronouns uniformly appeared in the accusative. Each ‘Nom’ or ‘Acc’ value in the table below corresponds to exactly one item in the questionnaire.

| Context →          | Non-contrastive |     | Contrastive |     |
|--------------------|-----------------|-----|-------------|-----|
| Lexical subclass ↓ |                 |     |             |     |
| <b>Common noun</b> | Nom             | Acc | Nom         | Acc |
| <b>Proper noun</b> |                 | Acc |             | Acc |
| <b>Pronoun</b>     |                 | Acc |             | Acc |

Table 1: *Object types and contexts tested in Questionnaire 1*

In each questionnaire item, the following focus positions and word orders were tested:<sup>10</sup>

<sup>9</sup> It has to be noted, however, that (partly due to the presence of the second clause, which negates the other possible alternative) the answer sentences in (22a–f) allow for more than one interpretation (thanks to Balázs Surányi for drawing my attention to this). In the one given in (22), the object of both the first and second clause are focused. This interpretation implies that the speaker who answers the question presupposes that the other speaker expects ‘duck’ to be the correct answer, and the first clause corrects this information. In this case, the focused object in the first sentence is a corrective focus. Another possible interpretation is ‘Lera bought CHICKEN at the grocery, duck she did not buy’, in which case the object of the first clause is a proper contrastive focus, whereas the object of the second clause is a contrastive topic. A third theoretically possible interpretation is ‘Chicken, Lera did buy at the grocery, duck, she did not buy’, in which case the object is a contrastive topic in both clauses. However, as contrastive topics appear in Udmurt at the left periphery of the sentence structure (Surányi et al., to appear), for sentences with *kureg* ‘chicken’ in postverbal position such an interpretation can be excluded. The reason why the object in the second clause can be interpreted both as a focus and as a contrastive topic is that standard Udmurt lacks an element used only for constituent negation, thus, constituent negation is not distinguishable from predicate negation (see Edygarova 2015: 284–285).

<sup>10</sup> The relative order of the subject and the adverbial was not examined here, the subject preceded the adverbial in each case, although the reverse order is also grammatical.

- (23) Focus positions tested in Questionnaire 1:
- a. immediately preverbal (SAdvO<sub>FOC</sub>V order)
  - b. non-immediately preverbal:
    - i. preceding a locative adverbial (SO<sub>FOC</sub>AdvV)
    - ii. sentence-initial, preceding the subject and the locative adverbial (O<sub>FOC</sub>SAdvV)
  - c. sentence-final (SVAdvO<sub>FOC</sub> and SAdvVO<sub>FOC</sub>)<sup>11</sup>
  - d. postverbal but not sentence-final (SVO<sub>FOC</sub>Adv)

Questionnaire 2 was also mainly concerned with direct object foci. The main aim of this survey was to test whether any of the positions is associated with an obligatorily contrastive and/or exhaustive reading. The following focus positions and permutations of S, O<sub>FOC</sub> and V were examined:

- (24) Focus positions tested in Questionnaire 2:
- a. immediately preverbal (SO<sub>FOC</sub>V)
  - b. non-immediately preverbal:
    - i. preceding a locative adverbial (SO<sub>FOC</sub>AdvV)
    - ii. sentence-initial, preceding the subject (O<sub>FOC</sub>SV)
  - c. sentence-final (SVO<sub>FOC</sub>)

The respondents had to evaluate on a rating scale (*good/odd/incorrect*) the grammaticality of sentences constituting short dialogues, and they had to correct the sentences that they found odd or unacceptable. Both the focus-eliciting sentences and the sentences containing the focused item itself had to be evaluated (and corrected in case they were found odd or ungrammatical), but for the purposes of the present study only judgements on the latter will be taken into consideration (even if the focus eliciting context also contained a focused element).

The contrastive test contexts were corrections like the dialogue presented in (25) (the focused element is immediately preverbal in the example, but all of the positions listed in (24) were tested):

- (25) – *Nadja Saša-jez=a byrj-i-ž?*  
 Nadja Sasha-ACC=Q choose-PST-3SG  
 ‘Did Nadja choose Sasha?’  
 – *Öž so VOLOD’A-JEZ byrj-i-ž.*  
 NEG.PST.3 3SG Volodja-ACC choose-PST-3SG  
 Intended meaning: ‘No, it was Volodja whom she chose.’

Exhaustivity was tested by means of the exhaustive identification test applied by É. Kiss (1998) to Hungarian, cf. (26)–(27). According to É. Kiss, the dialogue is felicitous only if negation in sentence (b) can be interpreted as the negation of the exhaustivity of the

---

<sup>11</sup> Thus, sentence-final foci were tested in two contexts, with the adverbial either preceding or following the verb. The purpose of this was to lower the possibility that speakers reject a variant with sentence-final focus only because of the position of the adverbial. The two word order variants were then collapsed into a single option of “sentence-final focus” at the speaker-internal evaluation of the results, see Section 4.3.

focused element of the sentence in (a) (É. Kiss 1998: 251). Thus, according to É. Kiss (1998), (26) is a felicitous dialogue while (27) is not, and *egy kalapot* ‘a hat’ fulfils the role of exhaustive identificational focus in (26b) (which occupies the immediately preverbal position in Hungarian), whereas it is a non-exhaustive information focus in (27b) (which is postverbal in Hungarian).

- (26) a. *Mari* EGY KALAP-OT *néz-ett* *ki* *magá-nak*.  
 Mary a hat-ACC watch-PST.3SG out herself-DAT  
 ‘It was a hat that Mary picked for herself.’  
 b. *Nem, egy kabát-ot is ki-néz-ett*.  
 no, a coat-ACC too out-look-PST.3SG  
 ‘No, she picked a coat, too.’ (É. Kiss 1998: 251) (Hungarian)
- (27) a. *Mari ki-néz-ett magá-nak* EGY KALAP-OT.  
 Mary out-watch-PST.3SG herself-DAT a hat-ACC  
 ‘Mary picked for herself a hat.’ (É. Kiss 1998: 249)  
 b. #*Nem, egy kabát-ot is ki-néz-ett*.  
 no a coat-ACC too out-look-PST.3SG  
 ‘No, she picked a coat, too.’ (É. Kiss 1998: 251) (Hungarian)

At this point it has to be noted that the above exhaustivity test is not entirely reliable: not every speaker of Hungarian agrees that (26) is a felicitous dialogue (see also Onea & Beaver 2011).<sup>12</sup>

The dialogue in (28) illustrates the test for Udmurt as in the questionnaire (the focused element is sentence-final in the example, but again all of the positions listed in (24) were tested):

- (28) – *Ljuba jarat-e* ARTUR-EZ.  
 Ljuba love-3SG Arthur-ACC  
 Intended meaning: ‘Ljuba loves ARTHUR.’/‘It is Arthur whom Ljuba loves.’  
 – *Ug, so jarat-e Artjom-ez no*.  
 NEG.3SG 3SG love-3SG Artjom-ACC too  
 Intended meaning: ‘No, she loves Artjom, too.’

Further questionnaire items consisted of dialogues that were similar to the above one with the exception that they also contained the focus particle *gine* ‘only’ (which follows the focused element). Thus, while in (28) the exhaustive interpretation was meant to arise solely from the context, in (29a), exhaustivity was lexically marked, as well. Again, all of the positions mentioned in (24) were tested.

- (29) – *Ljuba jarat-e* ARTUR-EZ GINE.  
 Ljuba love-3SG Arthur-ACC only  
 Intended meaning: ‘It is only Arthur whom Ljuba loves.’

---

<sup>12</sup> As an anonymous reviewer points out, this is likely to be due to the fact that exhaustivity is not asserted but presupposed content in these dialogues, and presuppositions cannot be negated directly, as they need a move like “Hey, wait a minute” (see von Stechow 2004).

- *Ug, so jarat-e Art'om-ez no.*  
 NEG.3SG 3SG love-3SG Artjom-ACC too  
 Intended meaning: ‘No, she loves Artjom, too.’

The third and most comprehensive questionnaire (Questionnaire 3) (cf. Asztalos & É. Kiss 2016) was concerned with the focus positions which are most often made reference to in the literature, i.e., the immediately preverbal, sentence-final and sentence-initial positions (cf. Section 2.3). The aim of the questionnaire was to test, on the one hand, whether focus placement is influenced by the syntactic function of the focused element. For that, subject, direct object, dative, instrumental-comitative and temporal adverbial foci were tested. The respondents had to give their grammaticality judgements of the test sentences on a 5-point Likert scale (where 5 meant ‘perfectly acceptable’ and 1 stood for ‘unacceptable’).

Contexts eliciting non-contrastive foci were *wh*-questions and sentences containing a superlative adjunct construed with one of the constituents of the sentence, see e.g. (30). Superlative adjuncts, in fact, entail the presence of a focused item in the sentence (see F. Farkas & É. Kiss 2000).

- (30) Context: ‘Yesterday a beauty contest was organized at the Philharmonia Concert Hall.’  
 (VIKTORIJA PUŠINA-LY) *žuri* (VIKTORIJA PUŠINA-LY) *tuž-ges no*  
 Victoria Pushina-DAT jury V.P.-DAT very-CMPR PTCL  
*tros ball sot-i-z* (VIKTORIJA PUŠINA-LY).<sup>13</sup>  
 many score give-PST-3SG V.P.-DAT  
 Intended meaning: ‘The jury gave the highest score TO VICTORIA PUSHINA.’

Questionnaire 3 was also concerned with exhaustive and contrastive foci. Contrastive contexts included alternative questions like the one in (31), and corrections similar to (25) and (32).

- (31) – *Ku ton Votkinsk-e košk-o-d, čukaže=a jake*  
 when 2SG Votkinsk-ILL leave-FUT-2SG tomorrow=Q or  
*čukaže uly-sa=a?*  
 tomorrow be-CVB=Q  
 ‘When are you leaving for Votkinsk, tomorrow or the day after?’  
 – (ČUKAŽE) *mon Votkinsk-e* (ČUKAŽE) *košk-o* (ČUKAŽE).  
 tomorrow 1SG Votkinsk-ILL tomorrow leave-FUT.1SG tomorrow  
 Intended meaning: ‘I will leave for Votkinsk TOMORROW.’ / ‘It is tomorrow that I will leave for Votkinsk.’
- (32) – *Tunne mi'emby kyrža-lo-z Anna.*  
 today 1PL.DAT sing-FUT-3SG Anne  
 ‘Today ANNE will sing for us.’

---

<sup>13</sup> Here and henceforth, examples in which the same element occurs in brackets in different positions illustrate the distribution of a *single* occurrence of that element.

- $U_{\text{NEG.FUT.3SG}}$  (D'IANA) *tunne* (D'IANA) *mil'emly* *kyrʒa-lo-ʒ* (D'IANA).  
 NEG.FUT.3SG Diana today D. 1PL.DAT sing-FUT-3SG D.  
 Intended meaning: ‘No, today DIANA will sing for us.’ / ‘No, it is Diana who will sing for us today.’

Exhaustivity was tested by checking the meaning of numerically modified noun phrases. According to É. Kiss (2006), numerals in natural languages have an ‘at least *n*’ meaning unless they are “associated with a particular structural position with an encoded [+exhaustive] feature”, in which case they have an ‘exactly *n*’ reading, as illustrated by the Hungarian examples in (33)–(34). (In (34), the postverbal position of the verbal prefix indicates that the numerically modified phrase occupies the immediately preverbal focus position.)

- (33) *János 15 palacsintá-t meg-esz-ik.*  
 John 15 pancake-ACC PRT-eat-3SG  
 ‘John eats (at least) 15 pancakes.’ (É. Kiss 2006: 447) (Hungarian)
- (34) *János 15 palacsintá-t esz-ik meg.*  
 John 15 pancake-ACC eat-3SG PRT  
 ‘John eats (exactly) 15 pancakes.’ (ibid.) (Hungarian)

The meaning of numerically modified items was also tested in each of the above mentioned positions (sentence-initial, immediately preverbal, and sentence-final). Respondents had to answer questions like the one presented in (35):

- (35) A professor says: “Who scores 91 points at the exam is going to receive a present.” Now, Kostja had 100 points. Is he going to get a present?

Every “no” answer was interpreted as an ‘exactly *n*’ interpretation of the numeral (by virtue of  $100 \neq 91$ ), while “yes” answers were taken to be ‘at least *n*’ interpretations (by virtue of  $100 > 91$ ).<sup>14</sup>

It has to be noted that a shortcoming of all three questionnaires is that they only contained non-neutral sentences, that is, they did not test the word orders under discussion in neutral baseline sentences. As a reviewer points out, the results presented in Section 4 would be better interpretable when compared to results received for neutral sentences.

In the next section, I am going to present the results of the questionnaires following a thematic classification (i.e., not the chronology of the tests). In 4.1.1, I will discuss to what extent focus placement is determined by the syntactic function of the focused constituent. In 4.1.2–4.1.4, I will turn to direct object foci and to the question whether two factors, namely, morphological marking and the lexical subcategory of the focused object plays any role in focus placement. In 4.2, I will deal with the semantic features of exhaustivity and contrastivity. In 4.3, I will provide a speaker-internal evaluation of the results.

---

<sup>14</sup> However, it has to be noted that extralinguistic factors (general knowledge about the world) may have had an impact on speakers’ answers: in fact, the typical situation is that when a smaller achievement is being rewarded a bigger one is also rewarded.

## 4 Results and discussion

### 4.1 Focus placement and morphosyntactic properties of the focused element

#### 4.1.1 Syntactic function

As mentioned in Section 3, Questionnaire 3 (cf. Asztalos & É. Kiss 2016) tested the grammaticality of the immediately preverbal, sentence-final and sentence-initial focus positions in relation to the syntactic function and certain morphosyntactic properties of the focused element. Proper noun subject foci, definite (morphologically marked) and non-specific indefinite (morphologically unmarked) direct object foci, as well as proper noun dative, instrumental-comitative, and temporal adverbial foci were examined by means of different questionnaire items. The test sentences belonging to one item differed only in the position of the focused element. For each test sentence (containing the focused element in a given position) the average rating given by the speakers on the 5-point Likert scale was calculated. Table 2 shows the lowest and the highest *average* ratings belonging to a given focus position in a range. The table also indicates what syntactic functions turned out to be less acceptable in a given position.

|                              | Lowest and highest<br>average rating | Less accepted<br>syntactic functions                          |
|------------------------------|--------------------------------------|---|
| <b>Immediately preverbal</b> | 4,37–4,86                            | –   |
| <b>Sentence-final</b>        | 3,81–4,57                            | Adv <sub>TEMP</sub> (3,81–4,03)                               |
| <b>Sentence-initial</b>      | 3,03–4,45                            | Adv <sub>TEMP</sub> (3,74–3,88), Ins (3,32),<br>O (3,03–3,43) |

Table 2: *Lowest and highest average ratings of the test sentences/focus positions on a 5-point Likert scale*

Sentences that were given a score equivalent to or higher than 4 on average were considered as grammatical, while those with an average between 3 and 4 were regarded as degraded in grammaticality (but not ungrammatical). It is important to note that none of the test sentences was given an average score below 3, thus, none of them turned out to be completely ungrammatical.

The immediately preverbal focus position turned out to be grammatical independently of the syntactic function of the focused element, cf. (36)–(41). The sentence-final focus position resulted to be almost as acceptable as the immediately preverbal one, cf. (36)–(40), but (temporal) adverbials were slightly less accepted sentence-finally (41). The sentence-initial position turned out to be grammatical with subject (36) and with dative foci (39), and somewhat degraded in acceptability with temporal adverbial (41), instrumental-comitative (40) and direct object foci (37)–(38), especially with non-specific, unmarked direct objects (38).

- (36) Subject focus  
 (KAT<sup>PA</sup>) *tuž-ges no čeber kart'ina-jež* (KAT<sup>PA</sup>) *daša-ž* (KAT<sup>PA</sup>).  
 Kate very-CMPR PTCL nice picture-ACC K. make-PST.3SG K.  
 'It was Kate who made the nicest picture.'



- (37) Object focus (morphologically marked object)  
 Context: ‘Whom did Peter beat?’  
 (?ART’OM-EZ) *Petyr* (ART’OM-EZ) *žug-i-ž* (ART’OM-EZ).  
 Artjom-ACC Peter Artjom-ACC beat-PST-3SG A.-ACC  
 ‘It was Artjom whom Peter beat.’
- (38) Object focus (unmarked object)  
 – *Lera perepeč* *ši-je*.  
 Lera perepechi[Udmurt national dish] eat-3SG  
 ‘Lera is eating perepechi.’  
 – *Ug*, (??PEIŃAŃ) *Lera* (PEIŃAŃ) *šij-e* (PEIŃAŃ).  
 NEG.3SG pelmeni Lera pelmeni eat-3SG pelmeni  
 ‘No, Lera is eating PELMENI.’ / ‘No, it is pelmeni that Lera is eating.’
- (39) Focus = NP in the dative case  
 (VIKTORIJA PUŠINA-LY) *žuri* (VIKTORIJA PUŠINA-LY) *tuž-ges* *no*  
 Victoria Pushina-DAT jury V.P.-DAT very-CMPR PTCL  
*tros ball šot-i-ž* (VIKTORIJA PUŠINA-LY).  
 many score give-PST-3SG V.P.-DAT  
 Intended meaning: ‘The jury gave the highest score TO VICTORIA PUSHINA.’
- (40) Focus = NP in the instrumental-comitative case  
 – *Vadim Vera-jen=a ekt-i-ž?*  
 Vadim Vera-INS=Q dance-PST-3SG  
 ‘Did Vadim dance with Vera?’  
 – *Öž*, (EUBA-JEN) *Vadim* (EUBA-JEN) *ekt-i-ž* (EUBA-JEN).  
 NEG.PST.3 Ljuba-INS Vadim L.-INS dance-PST-3SG L.-INS  
 ‘No, Vadim danced WITH LJUBA.’ / ‘No, it was Ljuba whom Vadim danced with.’
- (41) Temporal adverbial focus  
 – *Ku pešataj-ed-ly žingyrt-o-d?*  
 when grandfather-2SG-DAT telephone-FUT-2SG  
 ‘When are you going to telephone your grandfather?’  
 – (ČUKAŽE) *pešataj-e-ly* (ČUKAŽE) *žingyrt-o* (ČUKAŽE).  
 tomorrow grandfather-1SG-DAT tomorrow telephone-FUT.1SG tomorrow  
 ‘I’m going to telephone my grandfather TOMORROW.’ / ‘It is tomorrow that I’m going to telephone my grandfather.’

In what follows, I will concentrate on the placement of direct object foci in relation to their morphological marking and lexical subcategory (proper noun/personal pronoun).

#### 4.1.2 Direct object foci: overall results of Questionnaire 1

Figure 1 illustrates the overall results of Questionnaire 1. For each questionnaire item the percentage of speakers who accepted a given permutation of S, Adv, O<sub>FOC</sub> and V as a grammatical and congruent answer to the related *wh*-question was calculated. Then, the results received for all questionnaire items were aggregated and the average percentage of speakers accepting a given word order (independently of the tested factors) was calculated.

On the whole, word orders and focus positions which were accepted by at least 50% of the respondents were considered as grammatical, while those that were chosen by less than 50% but at least 30% of the respondents, as marginally acceptable.

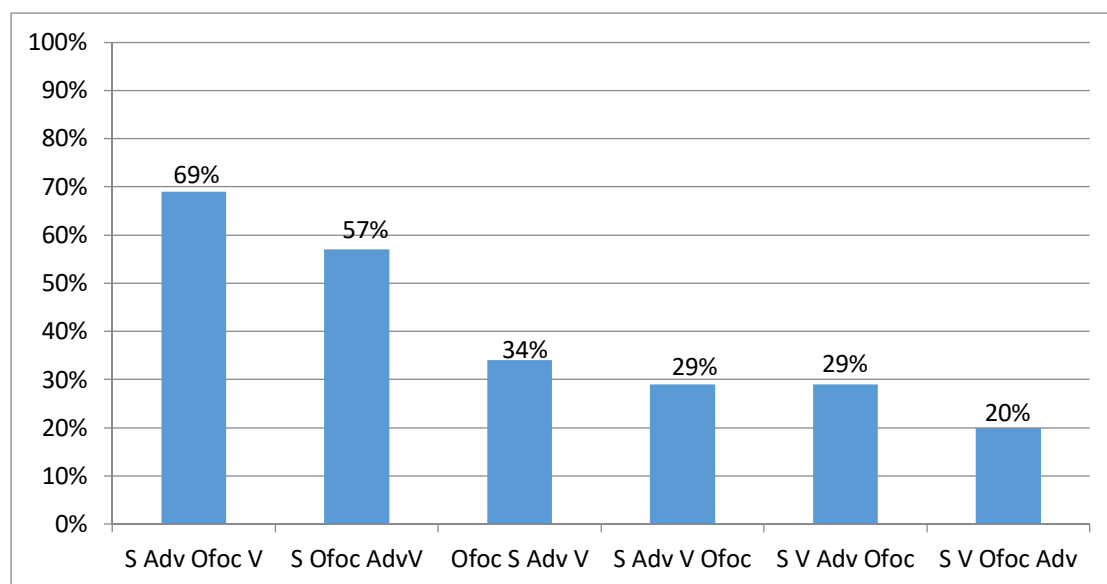


Figure 1: Average percentages of speakers accepting the tested word orders in Questionnaire 1 (all items included)<sup>15</sup>

Each tested word order variant was considered as a grammatical answer by at least one respondent to at least one question, but, as expected, the individual word orders did not turn out to be equally acceptable. Overall, the following tendencies were observed:

- The most accepted focus position resulted to be the immediately preverbal one (SAdvO<sub>FOC</sub>V order).
- Preverbal foci were given more favourable judgements than postverbal ones.
- Besides the immediately preverbal focus position, the “pre-adverbial” one (SO<sub>FOC</sub>AdvV order) also turned out to be grammatical.
- Sentence-initial foci preceding the subject and the locative adverbial (O<sub>FOC</sub>SAdvV order) resulted to be marginally acceptable.
- Sentence-final foci (SVAdvO<sub>FOC</sub> and SAdvVO<sub>FOC</sub> orders) were judged ungrammatical. (This contradicts the results of Questionnaire 3 (cf. Section 4.1.1), see Section 4.1.5 for a more detailed discussion of this problem.)
- Postverbal but not sentence-final foci (SVO<sub>FOC</sub>Adv order) also resulted to be ungrammatical.

However, the grammaticality of certain focus positions varies to some extent in relation to the morphosyntactic properties of the focused object. This will be discussed in the following subsections.

<sup>15</sup> 100% refers to the total number of questionnaire items (8) multiplied by the number of respondents (12) = 96.

### 4.1.3 *Direct object foci: morphological marking*

Four questionnaire items in Questionnaire 1 aimed at examining whether morphological marking plays a role in the placement of object foci. As anticipated in Section 4.1.1, Udmurt has differential object marking: direct objects can either be morphologically unmarked (formally identical to the nominative) (42), or case-marked (accusative) (43)–(44). Object marking is related to definiteness and specificity: non-specific indefinite objects are morphologically unmarked (42), whereas specific indefinites (43) and definites (44) are marked with the accusative case suffix (É. Kiss & Tánčzos 2018: 738–739, 752–753).

- (42) *Mon kníga lydz'-i.*  
 1SG book read-PST.1SG  
 'I read a book.' (É. Kiss & Tánčzos 2018: 738)
- (43) *Mon odíg puny-jez utća-ško.*  
 1SG one dog-ACC search-PRS.1SG  
 'I am searching for a (specific) dog.' (É. Kiss & Tánčzos 2018: 753)
- (44) *Mon Saša-jez magažin-ys adž'-i.*  
 1SG Sasha-ACC grocery-ELA see-PST.1SG  
 'I saw Sasha at the grocery.' (É. Kiss & Tánčzos 2018: 752)

Vilkuna (1998: 188) observes a relationship between the position and the morphological marking of direct objects: in the corpus she studied (compiled mainly of texts of 20th century prose (1998: 227)), the vast majority (88%) of unmarked objects immediately preceded the verb, while only less than half (42,8%) of accusative objects did so. There thus seems to be a tendency for unmarked objects to immediately precede the verb. This tendency has sometimes been described in the literature as a sort of incorporation of the object into the verb, as the unmarked object in such cases often forms a prosodic and morphosyntactic unit with the verb (Alatyrev et al. 1970: 169). Thus, the percentage of preverbal but not verb-adjacent objects was much higher in Vilkuna's corpus among accusative objects (42,1%) than among nominative ones (8,6%), and postverbal positioning was also more typical of marked objects than of unmarked ones (15,1% vs. 3,4%).

However, in contemporary blog texts, as Asztalos (2018)'s investigations indicate, the difference in the ability of unmarked and marked direct objects to occur postverbally seems to attenuate. This is accompanied by a strong increase of the proportion of postverbal direct objects, be they marked or unmarked: in Asztalos (2018)'s corpus, 35,5% of accusative-marked and 33% of unmarked object NPs appeared postverbally (2018: 78). (The calculations in both Vilkuna's (1998) and Asztalos's (2018) paper are made independently of the discourse function of the objects, that is, the counts of the authors are not limited to objects with focus function only.)

It may thus be of interest to see whether morphologically marked and unmarked focused objects show different tendencies with regard to their placement in the sentence.

Questionnaire 1 contained four related questionnaire items: two with a morphologically unmarked common noun object, and two with a marked common noun object. Figure 2 illustrates the average results:

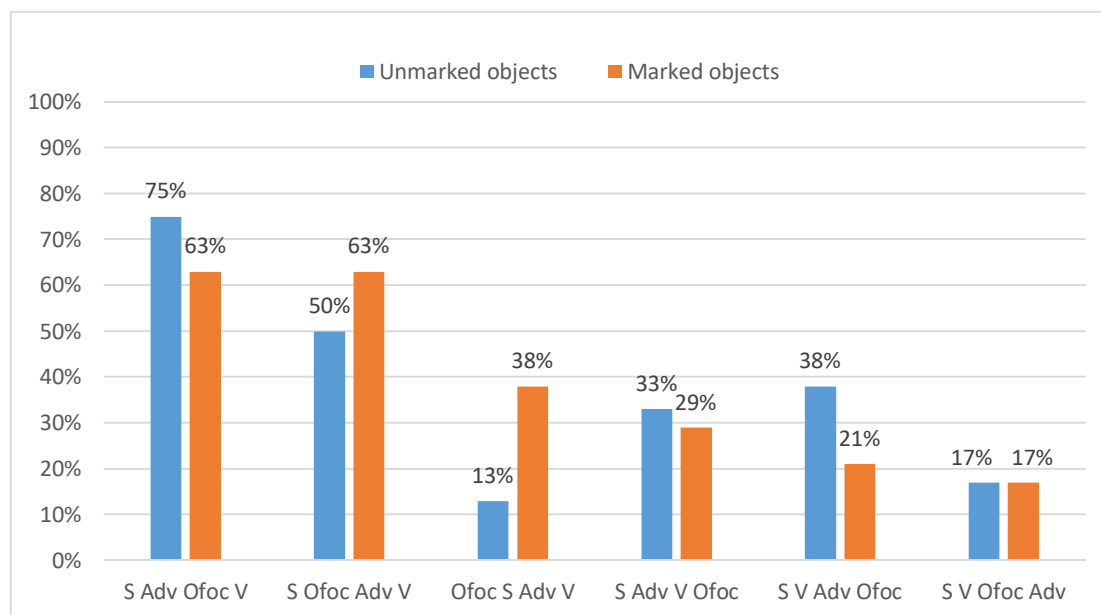


Figure 2: Percentages of speakers accepting the tested word orders with morphologically unmarked and marked focused objects (Questionnaire 1)<sup>16</sup>

Immediately preverbal focusing turned out to be grammatical with both object types (45)–(46), though it gave slightly better results with unmarked objects (45) than with marked ones (46).

(45) *Magazsin-yś Lera KUREG baśt-i-ż.*<sup>17</sup>  
 grocery-ELA Lera chicken buy-PST-3SG  
 ‘It is chicken that Lera bought at the grocery.’

(46) *Prazdńik-e Lera TA KUREG-EZ vaj-i-ż.*<sup>18</sup>  
 celebration-ILL Valerie this chicken-ACC bring-PST-3SG  
 ‘It is this chicken that Lera brought to the party.’

<sup>16</sup> 100% refers to the number of related questionnaire items (2) multiplied by the number of respondents (12) = 24.

<sup>17</sup> The focus-eliciting contexts for all sentences meaning ‘It is chicken that Lera bought at the grocery’ are given in (19) and (21).

<sup>18</sup> The focus-eliciting contexts for all sentences meaning ‘It was this chicken that Lera brought to the party’ are given in (i) and (ii):

(i) *Ma-je Lera prazdńik-e vaj-i-ż?*  
 what-ACC Lera celebration-ILL bring-PST-3SG  
 ‘What did Lera bring to the party?’

(ii) *Ma-je Lera prazdńik-e vaj-i-ż; ta kureg-eż=a jake so-że?*  
 what-ACC Lera celebration-ILL bring-PST-3SG this chicken-ACC=Q or that-DET.ACC  
 ‘What did Lera bring to the party: this chicken or that one?’

Pre-adverbial focusing (SO<sub>FOC</sub>AdvV) turned out to be grammatical with both object types, but it turned out to be more acceptable with objects in the accusative (47), while with objects in the nominative (48) it just reached the margin of grammaticality.

- (47) *Lera* TA KUREG-EZ *praždńik-e* *vaj-i-ž*.  
 Lera this chicken-ACC celebration-ILL bring-PST-3SG  
 ‘It is this chicken that Lera brought to the party.’
- (48) *Lera* KUREG *magažin-ys* *bašt-i-ž*.  
 Lera chicken grocery-ELA buy-PST-3SG  
 ‘It is chicken that Lera bought at the grocery.’

Sentence-initial focusing was marginally accepted with marked objects (49), while it turned out to be ungrammatical with unmarked ones (50) (note that unmarked, non-specific objects received less favourable judgements than marked ones in sentence-initial position in Questionnaire 3 as well, see Section 4.1.1):

- (49) ?TA KUREG-EZ *Lera praždńik-e* *vaj-i-ž*.  
 this chicken-ACC Lera celebration-ILL bring-PST-3SG  
 ‘It is this chicken that Lera brought to the party.’
- (50) \*KUREG *Lera magažin-ys* *bašt-i-ž*.  
 chicken Lera grocery-ELA buy-PST-3SG  
 Intended meaning: ‘It is chicken that Lera bought at the grocery.’

The above tendencies are in line with Vilkuna’s results (1998: 185–189) that non-verb-adjacent positions in the preverbal field are preferred in Udmurt with morphologically marked objects, and unmarked objects have a tendency to occur in the immediately preverbal position. Besides the above mentioned point that unmarked objects sometimes show incorporated object-like properties (Alatyrev et al. 1970: 169), a further reason for the dispreference for OS(Adv)V sentences with unmarked objects may lie in processing difficulties related to case-ambiguity. Studies on German (Gorrell 2000; Hemforth & Konieczny 2000; Schlesewsky & Bornkessel 2004) point to a processing difficulty of OS structures with case-ambiguous objects, and Levshina’s (2019) study reveals that cross-linguistically, formally overlapping subjects and objects tend to have rigid word order relative to each other. In the case of Udmurt, this may imply a difficulty to obtain an OSV reading for sentences which contain two morphologically unmarked nouns, given that the basic word order is SOV.<sup>19</sup>

Interestingly, sentence-final foci resulted to be marginally acceptable with objects in the nominative, while ungrammatical with objects in the accusative (51).

- (51) a. ?*Lera magažin-ys* *bašt-i-ž* KUREG / \*TA KUREG-EZ.  
 Lera grocery-ELA buy-PST-3SG chicken / this chicken-ACC

---

<sup>19</sup> However, as a reviewer points out, the animacy difference between the two morphologically unmarked nouns is sharp enough in (50) to ease the identification of the syntactic functions of the two nouns.

- b. ?Lera *bašt-i-ž* *magažin-ys* KUREG / \*TA KUREG-EZ.  
 Lera buy-PST-3SG grocery-ELA chicken / this chicken-ACC  
 'It is chicken/\*this chicken that Lera bought at the grocery.'

Postverbal but not sentence-final focusing resulted to be ungrammatical with both object types:

- (52) \*Lera *bašt-i-ž* KUREG *magažin-ys*.  
 Lera buy-PST-3SG chicken grocery-ELA  
 'It is chicken that Lera bought at the grocery.'
- (53) \*Lera *vaj-i-ž* TA KUREG-EZ *praždnik-e*.  
 Lera bring-PST-3SG this chicken-ACC celebration-ILL  
 'It is this chicken that Lera brought to the party.'

#### 4.1.4 Direct object foci: lexical subcategory (proper nouns vs. personal pronouns)

In Questionnaire 1, four items (two with a proper noun direct object and two with a personal pronoun direct object) were concerned with the question whether proper noun and pronominal object foci tend to be placed into different positions.<sup>20</sup> The results are summarized in Figure 3.

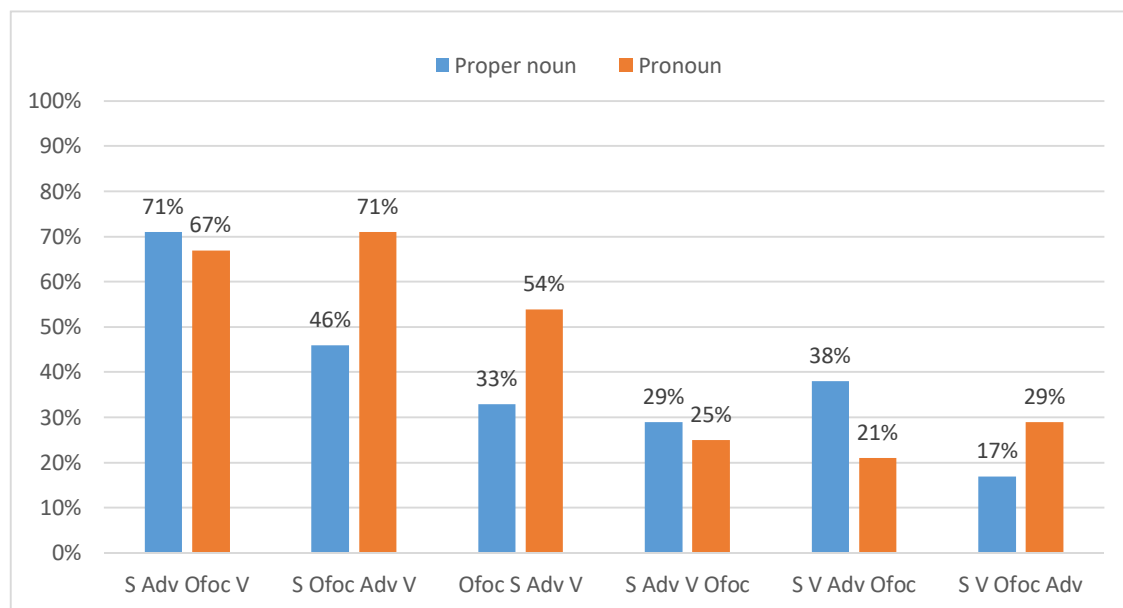


Figure 3: Percentages of speakers accepting the tested word orders with proper noun and pronominal focused objects (Questionnaire 1)<sup>21</sup>

<sup>20</sup> Both object types are morphologically marked: proper noun objects as specific and definite nouns are marked by the accusative case suffix by rule, whereas personal pronouns always have different forms in the subject and in the object function (nominative vs. accusative).

<sup>21</sup> 100% refers to the number of related questionnaire items (2) multiplied by the number of respondents (12) = 24.

Immediately preverbal foci were considered as grammatical independently of the lexical subcategory of the object:

- (54) *Žeńa bazar-ys* AEONA-JEZ / TON-E *adž'-i-ž.*<sup>22</sup>  
 Zhenja market-ELA Aljona-ACC / 2SG-ACC see-PST-3SG  
 'It was Aljona/you whom Zhenja saw at the market.'

Pre-verbal but not verb-adjacent focus positions (SO<sub>FOC</sub>AdvV and O<sub>FOC</sub>SAdvV orders, cf. (55)–(56)) turned out to be grammatical with personal pronoun objects, and marginally acceptable with proper nouns. More precisely, SO<sub>FOC</sub>AdvV order was highly acceptable with personal pronouns, and sentence-initial object focusing resulted to be clearly grammatical, among all examined object types (nominative/accusative, proper noun/personal pronoun), with personal pronouns only. This is, in fact, also in line with Vilkuna's results: personal pronoun objects (along with demonstrative pronoun objects) turned out to be the most "movable" object type in her corpus as well, which means that pronominal objects occurred more frequently in preverbal but not verb-adjacent and in postverbal positions than other object types (1998: 188).

- (55) (TON-E) *Žeńa* (TON-E) *bazar-ys* *adž'-i-ž.*  
 2SG-ACC Zhenja 2SG-ACC market-ELA see-PST-3SG  
 'It was you whom Zhenja saw at the market.'

- (56) (?AEONA-JEZ) *Žeńa* (AEONA-JEZ) *bazar-ys* *adž'-i-ž.*  
 Aljona-ACC Zhenja A.-ACC market-ELA see-PST-3SG  
 'It was Aljona whom Zhenja saw at the market.'

The accessibility of preverbal but not verb-adjacent focus positions for personal pronoun objects may be related to the high degree of definiteness of personal pronouns. Personal pronouns are located on top of the so-called *definiteness scale* (cf. Aissen 2003), cf.

---

<sup>22</sup> The focus-eliciting questions of all sentences meaning 'It was Aljona whom Zhenja saw at the market' are given in (i) and (ii), while those of the sentences meaning 'It was you whom Zhenja saw at the market', in (i) and (iii).

- (i) *Kin-e Žeńa bazar-ys adž'-i-ž?*  
 who-ACC Zhenja market-ELA see-PST-3SG  
 'Whom did Zhenja see at the market?'

- (ii) *Kin-e Žeńa bazar-ys adž'-i-ž, Aljona-jež jake Aloša-jež?*  
 who-ACC Zhenja market-ELA see-PST-3SG Aljona-ACC or Aljoshka-ACC  
 'Whom did Zhenja see at the market, Aljona or Aljoshka?'

- (iii) *Kin-e Žeńa bazar-ys adž'-i-ž, mon-e=a jake Aloša-jež?*  
 who-ACC Zhenja market-ELA see-PST-3SG me-ACC=Q or Aljoshka-ACC  
 'Whom did Zhenja see at the market, me or Aljoshka?'

(57). The more to the left a grammatical entity is placed on the scale, the more it counts as definite:

- (57) *Definiteness scale* (Aissen 2003)  
 Personal pronoun > Proper name > Definite NP > Indefinite specific NP >  
 Non-specific NP

Cross-linguistically, categories located at the top of the hierarchy can behave differently from those at the bottom of the scale. This may imply for Udmurt, in this case, that personal pronouns have a freer distribution (at least in the preverbal field) than categories lower on the hierarchy: thus, even when they have a special discourse role (i.e., that of focus), they can occupy positions which are less accessible for categories lower on the scale. As we have seen in Section 4.1.3, preverbal but not verb-adjacent focus positions are more available for accusative objects (which are definite) than for morphologically unmarked objects (which are indefinite and non-specific). Overall, it seems that personal pronoun objects have the most flexible distribution, and morphologically unmarked, non-specific objects the least flexible distribution in the preverbal field in Udmurt, while accusative-marked definite NP objects are located between the two extremities, which fits what one could expect on the basis of the definiteness scale.<sup>23</sup>

Postverbal object foci (independently of whether they were proper nouns or personal pronouns) were in most cases accepted only by a small fraction of speakers, the average judgment not reaching the margin of grammaticality. The only exception was the SVAdvO<sub>FOC</sub> order, which resulted to be marginally acceptable with proper noun objects.

#### 4.1.5 *Interim summary*

Let us sum up what has been presented so far in this section.

The immediately preverbal focus position turned out to be grammatical independently of the syntactic function of the focused element, and, in the case of direct object foci, independently of their morphological marking and lexical subcategory.

The sentence-initial position, according to the results of Questionnaire 3, is more readily available for subject and dative foci than for direct object foci.

Preverbal but not verb-adjacent positions (i.e., the sentence-initial one and the one with an adverbial standing in between the focused object and the verb) seem to be sensitive to the morphological marking and to the lexical subcategory of the object. While morphologically unmarked object foci cannot occur sentence-initially, morphologically marked focused object nouns turned out to be marginally acceptable, and personal pronoun focused objects resulted to be grammatical in the sentence-initial position. The “pre-adverbial” position was more easily available for morphologically marked objects than for unmarked ones, and more easily available for personal pronouns than for proper nouns. The fact that the sentence-initial position is not available for unmarked direct objects may be explained, at least partly, by processing reasons: given the SOV character of Udmurt, obtaining an OSV reading for sentences that display two morphologically unmarked noun phrases in preverbal position may result in processing difficulties (similarly to German, see Gorrell 2000; Hemforth & Konieczny 2000; Schlesewsky & Bornkessel 2004). On the

---

<sup>23</sup> Nevertheless, the question remains why accusative-marked proper nouns were less accepted in preverbal but not verb-adjacent positions than accusative-marked, definite common nouns (cf. Figure 2 and 3).



other hand, the different degree of definiteness of the tested object types may also play a role. Personal pronouns, which are highly definite, seem to have the most flexible distribution, whereas unmarked objects, which are non-specific and sometimes behave similarly to incorporated objects (Alatyrev et al. 1970: 169), the least flexible distribution, at least in the preverbal field.

Postverbal but not sentence-final object foci were acceptable only for a small part of the speakers, thus, overall, they resulted to be ungrammatical in Questionnaire 1.

Sentence-final placement of foci also turned out to be on the whole ungrammatical in Questionnaire 1, but marginally acceptable with unmarked common nouns and with personal pronouns. However, in Questionnaire 3, sentence-final foci did turn out to be grammatical; what is more, they were evaluated as being almost as good as immediately preverbal foci.

The low acceptability of sentence-final foci in Questionnaire 1 is presumably due to normative reasons. In fact, all respondents of Questionnaire 1 were either students or employees of the Faculty of Udmurt Philology of the Udmurt State University. In Udmurt prescriptive linguistics, there exists a general normative restraint according to which non-verb-final sentences are to be avoided, and this may have had a considerable impact on the choices of the respondents of Questionnaire 1 because of respondents' education in Udmurt philology. In contrast with this, Questionnaire 3 was distributed via the social networking sites *Facebook* and *Vkontakte*, thus, the respondents were drawn from a more heterogeneous group.

#### 4.2 Focus interpretation: contrastivity and exhaustivity

As mentioned in Section 3, in Questionnaire 1, all sentences were tested both in non-contrastive contexts (as answers to *wh*-questions), and in contrastive contexts (as answers to alternative *wh*-questions). None of the tested focus positions resulted to be obligatorily contrastive: no focus position turned out to be grammatical with contrastive foci and at the same time ungrammatical with non-contrastive foci. This is illustrated in Figure 4.

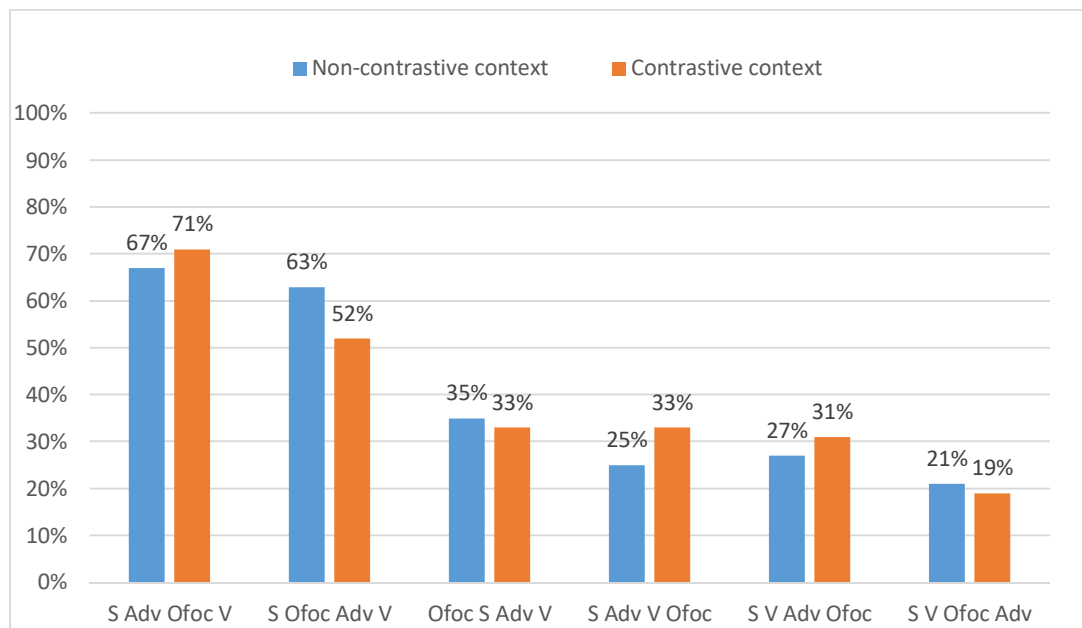


Figure 4: Percentages of speakers accepting the tested word orders in non-contrastive and contrastive contexts (Questionnaire 1)<sup>24</sup>

Thus, immediately preverbal foci and pre-adverbial foci resulted to be grammatical both in contrastive and non-contrastive contexts, see (57)–(58):

- (57) Context<sub>1</sub>: ‘What did Lera buy at the grocery?’  
 Context<sub>2</sub>: ‘What did Lera buy at the grocery, chicken or duck?’  
*Magažin-yś Lera KUREG bašt-i-ž.*  
 grocery-ELA Lera chicken buy-PST-3SG  
 ‘It is chicken that Lera bought at the grocery.’
- (58) Context<sub>1</sub>: ‘What did Lera buy at the grocery?’  
 Context<sub>2</sub>: ‘What did Lera buy at the grocery, chicken or duck?’  
*Lera TA KUREG-EZ prazdnik-e vaj-i-ž.*  
 Lera this chicken-ACC celebration-ILL bring-PST-3SG  
 ‘It is this chicken that Lera brought to the party.’

Sentence-initial foci were also judged similarly in the two different contexts. As presented in 4.1.4, sentence-initial object foci turned out to be grammatical with personal pronouns only, cf. (59), but here again, the fact whether the context was contrastive or not did not play a role:

<sup>24</sup> 100% refers to the number of related questionnaire items (4) multiplied by the number of respondents (12) = 48.

- (59) Context<sub>1</sub>: ‘Whom did Zhenja see at the market?’  
 Context<sub>2</sub>: ‘Whom did Zhenja see at the market, me or Aliosha?’  
 TON-E      *Žeńa    bazar-yś      adž-i-ž:*  
 2SG-ACC    Zhenja   market-ELA    see-PST-3SG  
 ‘It was you whom Zhenja saw at the market.’

The acceptability of postverbal (including sentence-final) foci was below 50% independently of the contrastivity of the context.

The results of Questionnaire 2 also suggest that none of the tested focus positions is associated with an obligatorily contrastive reading. As mentioned in Section 3, contrastive focus was tested in Questionnaire 2 by means of corrections. As opposed to them, non-contrastive exhaustive foci were examined. The latter were tested by two means: exhaustivity was either meant to arise exclusively from the context, or it was also lexically marked by the particle *gine* ‘only’.

It has to be noted that, since in Questionnaire 1 sentence-initial and pre-adverbial object foci were judged more favourably with personal pronouns than with non-pronominal elements (cf. Section 4.1.4), SO<sub>FOC</sub>AdvV and O<sub>FOC</sub>SV orders in Questionnaire 2 were only tested with pronominal objects. (Moreover, the subject was also pronominal in these test sentences.)

Figure 5 shows the percentage of speakers who considered the tested word orders as grammatical:

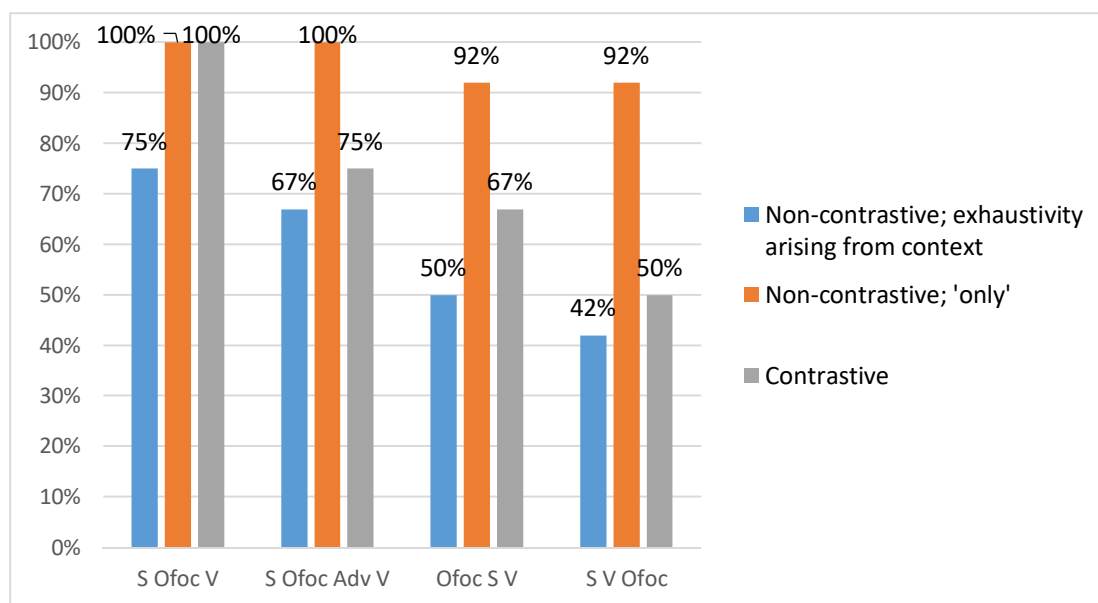


Figure 5: Percentages of speakers accepting the tested word orders in non-contrastive and contrastive contexts (Questionnaire 2)<sup>25</sup>

The results indicate that contrastive foci can occur in all of the tested positions (immediately preverbal (60a), sentence-final (60a), sentence-initial (60b), pre-adverbial (61)), though, sentence-final contrastive foci barely reached the margin of grammaticality.

<sup>25</sup> 100% refers to the number of related questionnaire items (1) multiplied by the number of respondents (12) = 12.

- (60) Context: ‘Did Nadja choose Sasha?’
- a.  $\ddot{O}\ddot{z}$ , *so* (VOLOD’A-JEZ) *byrj-i-ž* (VOLOD’A-JEZ).  
 NEG.PST.3 3SG Volodja-ACC choose-PST-3SG V.-ACC  
 ‘No, it was Volodja whom she chose.’
- b.  $\ddot{O}\ddot{z}$  TON-E *so* *byrjiž*.  
 NEG.PST.3 2SG -ACC 3SG choose-PST-3SG  
 ‘No, it was you whom she chose.’
- (61) Context: ‘Did Nastja choose Cyril among the boys?’
- $\ddot{O}\ddot{z}$  *so* MON-E *pi-os* *pöl-yś* *byrj-i-ž*.<sup>26</sup>  
 NEG.PST.3 3SG 1SG-ACC boy-PL among-ELA choose-PST-3SG  
 ‘No, it was me whom she chose among the boys.’

Similarly to the results of Questionnaire 1, no focus position turned out to be clearly grammatical with contrastive foci and at the same time clearly ungrammatical with non-contrastive foci. Thus, none of the focus positions resulted to be obligatorily contrastive.

As for exhaustive foci, the results indicate that those marked with the particle *gine* ‘only’ can grammatically appear in all tested positions (immediately preverbal (62), pre-adverbial (63), sentence-initial (64), and sentence-final (65)), which confirms Vilkuina’s claim that phrases with *gine* are freely placed in the sentence (1998: 196). However, when exhaustivity was meant to arise solely from the context, all word orders were much less accepted than in the case of *gine*-marked foci (and also less accepted than with contrastive foci) – though they all resulted to be grammatical with the exception of SVO<sub>FOC</sub>, which was somewhat below the margin of grammaticality. The lower acceptability of sentence-final foci is probably due to the same reason as in the case of Questionnaire 1 (see Section 4.1.5), i.e., the respondents of Questionnaire 2 were also students or teachers of the Faculty of Udmurt Philology of the Udmurt State University and thus, the normative restraint according to which they should avoid non-verb-final sentences may have had an impact on their choices.

The lower acceptability of all word orders in the case of lexically non-marked exhaustive foci, however, is likely to be due to the relative oddity (mentioned in Section 3) of the test dialogue itself.

- (62) – *D’ima* JULIJA-JEZ (*gine*) *jarat-e*.  
 Dima Julia-ACC only love-3SG  
 ‘It is Julia whom Dima (only) loves.’
- *Ug*, *so* *Annajež* *no* *jarat-e*.  
 NEG.3SG 3SG Anne-ACC also love-3SG  
 ‘No, he also loves Anne.’
- (63) – *Oleg* TON-E (*gine*) *klub-yś* *adž-i-ž*.  
 Oleg 2SG-ACC only disco-ELA see-PST-3SG  
 ‘It was (only) you whom Oleg saw at the disco.’

---

<sup>26</sup> The object occupied the same positions in the first and second sentences of the dialogues. If a respondent left the position of the object unchanged in the *test sentence* and changed it only in the *context sentence* of the dialogue, the related word order/focus position was regarded as accepted by that speaker.

- $\ddot{O}\ddot{z}$             *so*    *ton-e*            *no*            *ot-ys'*            *adž'i-ž*.  
 NEG.PST.3    3SG    2SG-ACC    also            there-ELA            see-PST-3SG  
 'No, he also saw you there.'
- (64) – MON-E    (*gine*) *so*    *jarat-e*.  
 2SG-ACC only 3SG    love-3SG  
 'It is (only) me whom (s)he loves.'
- *Ug*,            *mon-e*            *no*    *so*    *jarat-e*.  
 NEG.3SG 1SG-ACC    also    3SG    love-3SG  
 'No, (s)he also loves me.'
- (65) – *Ljuba* *jarat-e*            ARTUR-EZ    (*gine*).  
 Ljuba    love-3SG    Arthur-ACC    only  
 'It is (only) Arthur whom Ljuba loves.'
- *Ug*,            *so*    *jarat-e*            *Art'ome-ž*            *no*.  
 NEG.3SG 3SG    love-3SG    Artjom-ACC    also  
 'No, she also loves Artjom.'

As mentioned in Section 3, Questionnaire 3 concentrated on immediately preverbal, sentence-initial and sentence-final foci. Table 3 illustrates that the focus positions under discussion were given similar scores on average in non-contrastive and contrastive contexts, which again confirms the claim that their grammaticality does not depend on contrastivity, cf. (66)–(67), and that none of the positions is associated with an obligatorily contrastive reading.

|                              | Non-contrastive | Contrastive |
|------------------------------|-----------------|-------------|
| <b>Immediately preverbal</b> | 4,64            | 4,79        |
| <b>Sentence-final</b>        | 4,36            | 4,35        |
| <b>Sentence-initial</b>      | 3,74            | 3,47        |

Table 3: *Acceptability of focus positions in non-contrastive and contrastive contexts (average ratings on a 5-point Likert scale)*

- (66) Context: 'Who telephoned yesterday?'  
 (?L'UDMILA) *tolon*            (L'UDMILA) *žingyrt-i-ž*            (L'UDMILA).  
 Ludmila    yesterday    L.            telephone-PST-3SG    L.  
 'It is Ludmila who telephoned yesterday.'
- (67) Context: 'Today Anne will sing for us.'  
 $U\ddot{z}$             (?D'IANA) *tunne* (D'IANA) *mil'emly*    *kyrž'a-lo-ž*            (D'IANA).  
 NEG.FUT.3SG    Diana    today D.            1PL.DAT    sing-FUT-3SG    D.  
 'No, it is Diana who will sing for us today.'

The results of the test with numerical modifiers of Questionnaire 3 (see Section 3) suggest that none of the examined focus positions is necessarily exhaustive, either: independently of the position of the numerically modified phrase, around 80% of the

respondents preferred the ‘at least *n*’ interpretation over the ‘exactly *n*’ one for the sentences in (68)–(70).<sup>27</sup>

- (68) *Kin ekzamen-yn 91 ball l'uka-z, kuźym bašt-o-z.*  
 who exam-INE 91 score gather-PST.3SG present receive-FUT-3SG  
 ‘Who gets 91 points at the exam is going to receive a present.’
- (69) *Ađami-os-ly, kud-jos-yz 3 kn'iga magaźin-yśty-my bašt-o, duntek*  
 people-PL-DAT which-PL-DET 3 book shop-ELA-1PL buy-3PL free  
*d'isk' šot-o-m.*  
 disc give-FUT-1PL  
 ‘To those people who buy 3 books in our shop, we will give a free disk.’
- (70) *Kin-len vań kyk nylpi-jez, so-ly kun-my kvart'ira šot-e.*  
 who-GEN be two child-3SG 3SG-DAT state-1PL flat give-3SG  
 ‘To those who have two children, our state will give a flat.’

Overall, the results of Questionnaire 2 and 3 suggest that exhaustive interpretation is available in each tested focus position, but none of these positions is *obligatorily* exhaustive.

### 4.3 Variation across speakers

The results of Questionnaire 1 and 3 were evaluated speaker-internally, as well. In order to see how flexible speakers are with regard to object focus placement, in Questionnaire 1, the average number of speakers’ word order choices per item was calculated: the number of total word order choices was counted per speaker (the maximal number of possible choices, as presented in Section 3, was six for each questionnaire item), then the amount received was divided by the number of questionnaire items (= 8). Table 4 summarizes the average numbers, as well as the maximal and minimal numbers of word orders accepted by the speakers. To put it another way, the table illustrates speakers’ degree of flexibility with regard to object focus placement:

---

<sup>27</sup> However, as noted in Section 3, extralinguistic factors such as a general knowledge about the world may also have had an impact on respondents’ answers.

| Speaker    | Average nr. of<br>w.o. choices (max. value = 6) | Range of<br>w.o. choices |
|------------|---|--------------------------|
| Speaker 1  | 1   | 1–1                      |
| Speaker 2  | 1   | 1–1                      |
| Speaker 3  | 1   | 1–1                      |
| Speaker 4  | 1,8   | 1–3                      |
| Speaker 5  | 1,8   | 1–3                      |
| Speaker 6  | 1,9   | 1–3                      |
| Speaker 7  | 2,1   | 2–3                      |
| Speaker 8  | 2,3   | 2–3                      |
| Speaker 9  | 3   | 2–4                      |
| Speaker 10 | 3,3   | 3–4                      |
| Speaker 11 | 3,8   | 2–6                      |
| Speaker 12 | 5,9   | 5–6                      |

Table 4: *Average number and range of speakers' word order choices per item in Questionnaire 1*  
(Max. value = 6)

As Table 4 illustrates, speakers' flexibility varies considerably. 25% of respondents (Speaker 1, 2, and 3) considered as grammatical only one (though, not in every case the same) word order variant throughout the whole questionnaire. More than half of the respondents (Speaker 4, 5, 6, 7, 8, 9 and 10) marked most frequently 2 or 3 word order variants as correct. Finally, some respondents considered all variants in certain items as grammatical (Speaker 11), or throughout almost the whole questionnaire (Speaker 12).

Speakers seem to vary greatly in relation to their focus position preferences, as well. In the case of Questionnaire 1, it was counted, speaker by speaker, how many times they accepted a given word order variant throughout the whole questionnaire. SAdvVO<sub>FOC</sub> and SVAdvO<sub>FOC</sub> orders were both counted as instances of sentence-final foci, and therefore, no matter whether a respondent marked only one or both of them as grammatical in a questionnaire item, they were only counted once. Afterwards, the percentages in which each focus position was chosen were calculated speaker by speaker. The results are presented in Figure 6.

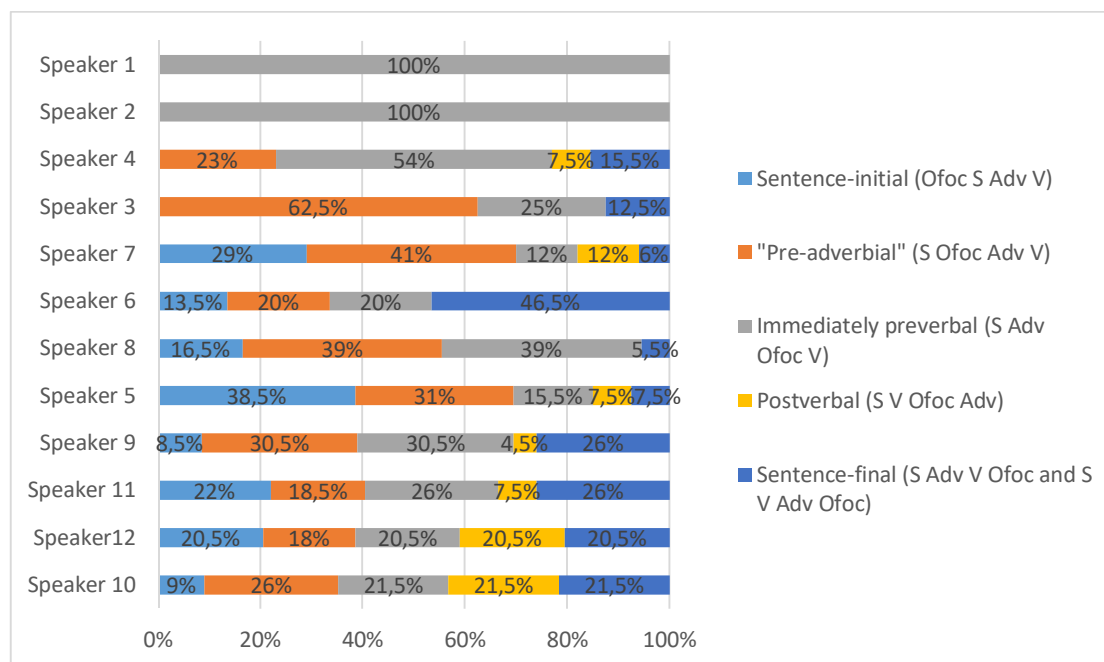


Figure 6: *Speakers' overall focus position choices in Questionnaire 1*<sup>28</sup>

Two respondents (Speaker 1 and Speaker 2) opted consistently, and one (Speaker 4) in more than 50% of the cases for the immediately preverbal focus position. Speaker 3 and Speaker 7 chose most frequently the “pre-adverbial” focus position, while Speaker 6 opted most frequently for sentence-final foci. Speaker 8 had an equal preference for pre-adverbial and immediately preverbal foci, and Speaker 5, a roughly equal preference for the sentence-initial and the pre-adverbial focus position. Speaker 9 chose sentence-final foci almost as frequently as pre-adverbial or immediately preverbal ones. The rest of the respondents did not show any obvious preference for any of the focus positions, or considered all options to be equally or almost equally good. No speaker had a preference for postverbal but not sentence-final foci.

In the case of Questionnaire 3, speaker-internal evaluation of the results consisted in checking, speaker by speaker, how they evaluated, throughout the whole questionnaire, the three tested focus positions compared to each other. As Table 5 illustrates, 38% of the respondents gave consistently better judgements to the immediately preverbal focus position than to the other options. Almost half (48%) of the respondents considered the sentence-final position to be as good, or almost as good, as the immediately preverbal one. Thus, sentence-final foci were given much more favourable judgements in Questionnaire 3 than in Questionnaire 2. However, only a negligible proportion (3%) of speakers preferred the sentence-final position over all other options. A small portion (11%) of respondents judged all focus positions to be equally good. Finally, no speaker had a preference for sentence-initial foci.

<sup>28</sup> 100% refers to the total number of questionnaire items in Questionnaire 1 (8) multiplied by the number of possible answer sentences per item (6) = 48.



| Preferred position(s)                    | % of respondents |
|--|------------------|
| Immediately preverbal                    | 38%              |
| Immediately preverbal + sentence-final   | 48%              |
| Sentence-final                           | 3%               |
| No preference (all options equally good) | 11%              |

Table 5: *Speakers' focus position preference in Questionnaire 3*

#### 4.4 Typological implications and the influence of Russian

Let us now consider the Udmurt data from a typological perspective. As presented in 2.2, Czypionka (2007), in a typological study carried out on 112 languages, shows that the most common syntactic focus positions in SOV languages are the immediately preverbal one and the sentence-initial one. On the other hand, postverbal focusing resulted to be really rare in SOV languages, and none of the SOV languages examined in her study had sentence-final focusing as its main focusing strategy (Czypionka 2007: 441–443). As for SVO languages, they rarely showed a preference for immediately preverbal focusing, while postverbal and sentence-final focusing was more common in them than in SOV languages. Interestingly, the main focusing strategy in SVO languages resulted to be the sentence-initial one, which was slightly more frequent in SVO than in SOV languages.

The fact that the immediately preverbal position resulted to be the most commonly accepted focus position in Udmurt corresponds to what one may expect on the basis of the traditional classification of Udmurt as an SOV language. However, according to Questionnaire 3, the sentence-final focus position is almost as acceptable in contemporary Udmurt as the immediately preverbal one (see also Tánčzos 2010). As sentence-final focusing is more typical of SVO than of SOV languages, this finding may further confirm the claim that contemporary Udmurt is undergoing an SOV-to-SVO change (cf. Tánčzos 2013; Asztalos 2016, 2018; Asztalos et al. 2017). Since information foci in Russian are sentence-final, and Udmurt is subject to strong Russian influence (see Section 2.5), there is also good reason to attribute the development of the sentence-final focus position in Udmurt to the influence of Russian (see also Tánčzos 2010).

Sentence-initial (and, more generally, preverbal but not verb-adjacent) appearance of foci seems to be subject to restrictions in Udmurt, and understanding the exact conditions of sentence-initial focusing needs further investigation (e.g, it is a possibility that sentence-initial subject foci in Udmurt are in fact instances of *in situ* focusing). Given the fact that sentence-initial foci are approximately as common in SOV as in SVO languages, one could argue that the possibility of sentence-initial focusing does not necessarily have to be interpreted as a phenomenon induced by the influence of Russian: it could also arise from the SOV nature of Udmurt. However, speaker-internal evaluation of the results suggests that this is not necessarily the case. If the possibility of sentence-initial focusing were stemming from the SOV character of Udmurt, one would expect respondents with a preference for immediately preverbal focusing to have judged sentence-initial foci more favourably than sentence-final ones. As Figure 6 in Section 4.3 illustrates, this was not a typical pattern in Questionnaire 1. As for Questionnaire 3, the respondents either had a preference for the immediately preverbal position, or a roughly equal preference for the immediately preverbal and the sentence-final one, but no speaker showed a preference for the immediately preverbal and the sentence-initial positions. Even the respondents with a clear preference for immediately preverbal foci gave consistently better judgements for sentence-final foci than for sentence-initial ones. All in all, there do not seem to be strong

reasons to assume that the possibility of sentence-initial focusing originates from the SOV grammar of Udmurt.

The question whether sentence-initial focusing is then induced by Russian influence could be addressed within the frame of the present study by comparing the interpretation of sentence-initial foci in the two languages. As presented in Section 4.2, none of the focus positions resulted to be obligatorily exhaustive or contrastive in Udmurt. In Russian, preverbal (including sentence-initial) foci have also been claimed not to be necessarily exhaustive, but there is no consensus in the literature whether they are obligatorily contrastive or not (see Section 2.5). However, as Dyakonova (2009) and Bailyn (2012) present examples with preverbal foci in non-contrastive contexts, a non-obligatorily contrastive analysis seems to be more plausible. In this latter case, the focus positions may not differ too much in terms of contrastivity and exhaustivity in the two languages, and the possibility of having Russian influence behind sentence-initial focusing cannot to be excluded.

## 5 Summary

While Tánzos (2010) identified an immediately preverbal and a sentence-final focus position in the Udmurt sentence structure, the investigations presented in this paper confirm the claims and sporadic observations made in the literature (cf. Vilkuna 1998; Timerxanova 2011; Asztalos 2012) that the possibilities of focus placement are not limited in Udmurt to the aforementioned two positions. While confirming the findings that the most acceptable focus position is the immediately preverbal one and that sentence-final placement of foci is also grammatical for a part of the speakers, the results of this paper indicate that focused items can also appear in certain preverbal but not verb-adjacent positions. Namely, they can precede a preverbal adverbial and/or the subject. The occurrence of foci in these positions is, however, subject to limitations. Sentence-initial focusing resulted to be mostly available for subjects, for dative complements and for personal pronoun direct objects. The pre-adverbial position proved to be accessible mainly for personal pronoun objects and, in a wider sense, for objects marked with the accusative case suffix. The more flexible distribution of personal pronoun objects and of morphologically marked objects (as compared to morphologically unmarked ones) is presumably related to the different degree of definiteness of the different object types, personal pronouns being at the top of the definiteness scale and non-specific (unmarked) objects at the bottom of it. In addition, the dispreference for  $O_{\text{FOC}}SV$  order with morphologically unmarked objects may also arise from processing difficulties: given the SOV nature of Udmurt, obtaining an OSV reading for sentences that contain two noun phrases without overt case-marking may require an extra processing cost (cf. Gorrell 2000; Hemforth & Konieczny 2000; Schlesewsky & Bornkessel 2004), thus, the order of unmarked objects relative to the subject may tend to be rigid in Udmurt (cf. Levshina 2019).

From an interpretive perspective, none of the focus positions turned out to be obligatorily contrastive or necessarily exhaustive. Thus, the acceptability of the tested focus positions does not depend on the contrastivity or on the exhaustivity of the focused item. However, when exhaustivity is lexically marked with the particle *gine* ‘only’, all of the tested focus positions (immediately preverbal, pre-adverbial, sentence-initial, sentence-final) are accepted to a much higher degree than when exhaustivity has to be retrieved solely on the basis of the test context.

Speakers vary notably in relation to their focus position preference and flexibility with regard to focus placement. Certain respondents considered as grammatical only one focus position throughout the whole questionnaire, or (in Questionnaire 3) had a clear preference for a certain focus position: in most cases this was the immediately preverbal position, in some (more rare) cases the pre-adverbial or the sentence-final one. Other speakers allowed more or all of the given possibilities. In Questionnaire 3, sentence-final foci were considered as grammatical only by respondents who also judged immediately preverbal foci to be grammatical. Finally, there were also speakers with no clear preference for any of the tested focus positions.

The Udmurt data presented in this paper may also be interesting from a typological point of view. According to Czypionka (2007), immediately preverbal focusing is much more typical of SOV than of SVO languages, while sentence-final focusing occurs in the latter but is not typical of the former. Thus, the fact that besides the most common strategy – i.e., immediately preverbal focusing – sentence-final focusing is also available for a part of the speakers, is itself a further argument for the claim that contemporary Udmurt is undergoing an SOV-to-SVO change (cf. Tánčzos 2013; Asztalos 2016, 2018; Asztalos et al. 2017). Since Russian has a sentence-final information focus position (cf. Section 2.5), and Udmurt is subject to strong Russian influence, it is feasible that the development of the sentence-final focus position in Udmurt is induced by Russian influence (see also Tánčzos 2010; Asztalos et al. 2017; Asztalos 2018). However, interestingly, sentence-initial focusing, which is actually the main focusing strategy in SVO languages and is also common in SOV languages, did not result to be widely accepted in Udmurt. This is somewhat surprising also when taking into consideration that Russian (besides its sentence-final position for information foci) has a sentence-initial focus position, as well. In any case, the exact conditions of sentence-initial focusing need to be further studied.

This paper had mainly descriptive aims and was principally concerned with the linear positions and the interpretation of foci in those positions. Several questions regarding focus in Udmurt remain to be answered by future work. *In situ* focussing, for instance, was not examined in detail here, nor was the interaction of word order with prosody studied in focus marking. The question whether any of the linearly determined focus positions is to be explained in terms of a position in hierarchical constituent structure (in other words, whether Udmurt is discourse-configurational with regard to any of its linearly identified focus positions), as well as the task of offering a possible syntactic analysis of focus positioning have also been left for future research.

## Appendix A

A questionnaire item eliciting non-contrastive focus in Questionnaire 1

7. юан: Мар Лера магазиньсь басьтїз?

↓ (курег)



1. *ответ:* Лера КУРЕГ магазиньсь басьтїз.
2. *ответ:* Лера магазиньсь басьтїз КУРЕГ.
3. *ответ:* Лера магазиньсь КУРЕГ басьтїз.
4. *ответ:* КУРЕГ Лера магазиньсь басьтїз.
5. *ответ:* Лера басьтїз магазиньсь КУРЕГ.
6. *ответ:* Лера басьтїз КУРЕГ магазиньсь.

Question 7:

*Mar Lera magažin-ys bašt-i-ž?*  
 what Lera grocery-ELA buy-PST-3SG  
 ‘What did Lera buy at the grocery?’

(*kureg* ‘chicken’)

- |                |                   |                   |                    |                          |
|----------------|-------------------|-------------------|--------------------|--------------------------|
| 1. <i>Lera</i> | KUREG             | <i>magažin-ys</i> | <i>bašt-i-ž.</i>   | (SO <sub>FOC</sub> AdvV) |
| Lera           | chicken           | grocery-ELA       | buy-PST-3SG        |                          |
| 2. <i>Lera</i> | <i>magažin-ys</i> | <i>bašt-i-ž</i>   | KUREG.             | (SAdvVO <sub>FOC</sub> ) |
| Lera           | grocery-ELA       | buy-PST-3SG       | chicken            |                          |
| 3. <i>Lera</i> | <i>magažin-ys</i> | KUREG             | <i>bašt-i-ž.</i>   | (SAdvO <sub>FOC</sub> V) |
| Lera           | grocery-ELA       | chicken           | buy-PST-3SG        |                          |
| 4. KUREG       | <i>Lera</i>       | <i>magažin-ys</i> | <i>bašt-i-ž.</i>   | (O <sub>FOC</sub> SAdvV) |
| chicken        | Lera              | grocery-ELA       | buy-PST-3SG        |                          |
| 5. <i>Lera</i> | <i>bašt-i-ž</i>   | <i>magažin-ys</i> | KUREG.             | (SVAdvO <sub>FOC</sub> ) |
| Lera           | buy-PST-3SG       | grocery-ELA       | chicken            |                          |
| 6. <i>Lera</i> | <i>bašt-i-ž</i>   | KUREG             | <i>magažin-ys.</i> | (SVAdvO <sub>FOC</sub> ) |
| Lera           | buy-PST-3SG       | chicken           | grocery-ELA        |                          |
- Intended meaning: ‘It is chicken that Lera bought at the grocery.’

## Appendix B

A questionnaire item eliciting contrastive focus in Questionnaire 1

**4. юан: Мар Лера магазиньсь басьтїз, курег яке чөж ?**

↓ (Курег. Чөж өз)




1. *ответ:* Лера магазиньсь КУРЕГ басьтїз, чөж өз басьты.
2. *ответ:* Лера магазиньсь басьтїз КУРЕГ, чөж өз басьты.
3. *ответ:* КУРЕГ Лера магазиньсь басьтїз, чөж өз басьты.
4. *ответ:* Лера басьтїз КУРЕГ магазиньсь, чөж өз басьты.
5. *ответ:* Лера КУРЕГ магазиньсь басьтїз, чөж өз басьты.
6. *ответ:* Лера басьтїз магазиньсь КУРЕГ, чөж өз басьты.

Question 4:

*Mar Lera magažin-yś bašt-i-ž, kureg jake čöž?*  
 what Lera grocery-ELA buy-PST-3SG chicken or duck  
 ‘What did Lera buy at the grocery, chicken or duck?’

*(Kureg. Čöž öz)*

chicken duck NEG.PST.3

‘Chicken, not duck’ (lit. ‘Chicken. Duck she didn’t’)

1. *Lera magažin-yś KUREG bašt-i-ž čöž öz bašty.*  
 Lera grocery-ELA chicken buy-PST-3SG duck NEG.PST.3 buy.CNG.SG  
 (SAdvO<sub>FOC</sub>V)
2. *Lera magažin-yś bašt-i-ž KUREG, čöž öz bašty.*  
 Lera grocery-ELA buy-PST-3SG chicken duck NEG.PST.3 buy.CNG.SG  
 (SAdvVO<sub>FOC</sub>)
3. *KUREG Lera magažin-yś bašt-i-ž čöž öz bašty.*  
 chicken Lera grocery-ELA buy-PST-3SG duck NEG.PST.3 buy.CNG.SG  
 (O<sub>FOC</sub>SAdvV)
4. *Lera bašt-i-ž KUREG magažin-yś, čöž öz bašty.*  
 Lera buy-PST-3SG chicken grocery-ELA duck NEG.PST.3 buy.CNG.SG  
 (SVO<sub>FOC</sub>Adv)

5. Lera KUREG *magažin-yś* *bašt-i-ž* *čöž* *öž* *bašty*.  
Lera chicken grocery-ELA buy-PST-3SG duck NEG.PST.3 buy.CNG.SG  
(SO<sub>FOC</sub>AdvV)
6. Lera *bašt-i-ž* *magažin-yś* KUREG, *čöž* *öž* *bašty*.  
Lera buy-PST-3SG grocery-ELA chicken duck NEG.PST.3 buy.CNG.SG  
(SVAdvO<sub>FOC</sub>)

Intended meaning: ‘It is chicken that Lera bought at the grocery, not duck.’

## References

- Aissen, Judith. 2003. Differential Object Marking: Iconicity vs. Economy. *Natural Language & Linguistic Theory* 21(3). 435–483. <https://doi.org/10.1023/a:1024109008573>
- Alatyrev, V. I., Vaxrušev, V. M., Zaxarov, V. N. & Kalinina, L. I. 1970. *Grammatika sovremennogo udmurtskogo jazyka. Sintaksis prostogo predloženiya* [Grammar of the contemporary Udmurt language. The syntax of simple clauses]. Izhevsk: Izdatel'stvo „Udmurtija”.
- Asztalos, Erika É. 2012. Pozicij sfokuszovannogo objekta v udmurtskom jazyke [Positions of the focused object in Udmurt]. *Ezšegodnik finno-ugorskikh issledovanij “Yearbook of Finno-Ugric Studies”* 4. 7–12. <https://doi.org/10.35634/efui>
- Asztalos, Erika. 2016. A fejevű grammatikától a fejkezdettű felé: generációs különbségek a mai udmurt beszélőközösségben a szórendhasználat és -megítélés terén [From head-final towards head-initial grammar: generational differences concerning word order usage and judgement in present-day Udmurt speech community]. In É. Kiss Katalin, Hegedűs Attila & Pintér Lilla (eds.), *Nyelvelmélet és kontaktológia 3* [Linguistic theory and contactology 3], 126–156. Budapest – Piliscsaba: Szent István Társulat.
- Asztalos, Erika. 2018. *Szórendi típusváltás az udmurt nyelvben* [Word order type change in Udmurt]. Eötvös Loránd University, PhD dissertation. Manuscript.
- Asztalos, Erika & É. Kiss, Katalin. 2016. *Discourse-Motivated Word Order Variation in Udmurt*. Conference presentation. Olomouc Linguistics Colloquium, Palacky University, Olomouc.
- Asztalos, Erika, Gugán, Katalin & Mus, Nikolett. 2017. Uráli VX szórend: nyenyec, hanti és udmurt mondatszerkezeti változatok [VX order in Uralic: sentence structures in Nenets, Khanty and Udmurt]. In É. Kiss Katalin, Hegedűs Attila & Pintér Lilla (eds.), *Nyelvelmélet és diakrónia 3* [Linguistic theory and diachrony 3], 30–62. Budapest – Piliscsaba: PPKE BTK Elméleti Nyelvészeti Tanszék – Magyar Nyelvészeti Tanszék.
- Asztalos, Erika & Tánczos, Orsolya. 2014. *Competing Grammars in contemporary Udmurt*. Conference presentation. 7th Budapest Uralic Workshop, Research Institute for Linguistics of the Hungarian Academy of Sciences, Budapest.
- Bailyn, John F. 2012. *The Syntax of Russian*. New York: Cambridge University Press.
- Baušev, K. M. 1929. *Sintaksičeskij stroj votskoj reči i genezis častic sojužnogo porjadka* [Syntactic structure of the Votyak language and the genesis of conjunction-like particles]. Gosudarstvennoe izdatel'stvo Moskva–Leningrad.
- Bulyčov, M. N. 1947. *Porjadok slov v udmurtskom prostom predloženiü* [Word order in Udmurt simple clauses]. Izhevsk: Udmurtgosizdat.
- Csúcs, Sándor. 1990. *Chrestomathia Votiacica* [Votyak Chrestomathy]. Budapest: Tankönyvkiadó.

- Czypionka, Anna. 2007. Word Order and Focus Position in the World's Languages. *Linguistische Berichte* 212. 439–454.
- Dyakonova, Marina. 2009. *A Phase-based Approach to Russian Free Word Order*. Utrecht: LOT.
- Edygarova, Svetlana. 2015. Negation in Udmurt. In Matti Miestamo, Anne Tamm & Beáta Wagner-Nagy (eds.), *Negation in Uralic Languages* [Typological Studies in Language 108], 265–291. Amsterdam/Philadelphia: John Benjamins Publishing Company. <https://doi.org/10.1075/tsl.108.10edy>
- É. Kiss, Katalin. 1998. Identificational Focus Versus Information Focus. *Language* 74. 245–273. <https://doi.org/10.2307/417867>
- É. Kiss, Katalin. 2006. Jól megoldottuk? Rosszul oldottuk meg? Az összefoglaló és a kirekesztő kifejezést tartalmazó mondatok szórendjének magyarázata [Have we solved it well? Have we solved it badly? An explanation for the word order in clauses containing summarizing and exhaustive expressions]. *Magyar Nyelv* 102(4). 442–459.
- É. Kiss, Katalin & Tánzos, Orsolya. 2018. From possessor agreement to object marking in the evolution of the Udmurt *-jez* suffix: A grammaticalization approach to morpheme syncretism. *Language* 94 (4). 733–757. <https://doi.org/10.1353/lan.2018.0052>
- F. Farkas, Donka & É. Kiss, Katalin. 2000. On the comparative and absolute readings of superlatives. *Natural Language and Linguistic Theory* 18. 417–455. <https://doi.org/10.1023/a:1006431429816>
- Gavrilova, T. G. 1970. *Porjadok slov v udmurtskom prostom povestvovatel'nom predloženii* [Word order in Udmurt simple declarative sentences]. Izhevsk: Zapiski Udmurtskogo NII istorii, ekonomiki, literatury i jazyka pri Sovete Ministrov Udmurtskoj ASSR.
- Glezdenev, P. B. 1921. *Kratkaja grammatika jazyka naroda udmurt* [A short grammar of the language of the Udmurt nation]. Vjatka: Izdanie Vjatskogo Gubernskogo Otdelenija Gosizdata.
- Gorrell, Paul. 2000. The subject-before-object preference in German clauses. In Barbara Hemforth & Lars Konieczny (eds.), *German sentence processing*, 25–63. Dordrecht: Kluwer Academics Publishers. [https://doi.org/10.1007/978-94-015-9618\\_2](https://doi.org/10.1007/978-94-015-9618_2)
- Hemforth, Barbara & Konieczny, Lars. 2000. Cognitive parsing in German e an introduction. In Barbara Hemforth & Lars Konieczny (eds.), *German sentence processing*, 1–24. Dordrecht: Kluwer Academics Publishers. [https://doi.org/10.1007/978-94-015-9618-3\\_1](https://doi.org/10.1007/978-94-015-9618-3_1)
- King, Tracy H. 1995. *Configuring Topic and Focus in Russian*. Stanford: CSLI Publications.
- Konjuxova, A. V. 1964. *Udmurt kyl. Grammatika. Kyketi ljuketez. Sintaksis* [The Udmurt language. A grammar. Second part. Syntax]. Izhevsk: Udmurt knižnoj izdatel'stvo.
- Levshina, Natalia. 2019. Token-based typology and word order entropy: A study based on Universal Dependencies. *Linguistic Typology* 23(3). 533–572. <https://doi.org/10.1515/lingty-2019-0025>
- Marajko (2015) Blogpost retrieved from [https://marjamoll.blogspot.com/2015/08/blog-post\\_25.html](https://marjamoll.blogspot.com/2015/08/blog-post_25.html)
- Neeleman, Ad & Titov, Elena. 2009. Focus, contrast, and stress in Russian. *Linguistic Inquiry* 40(3). 514–524. <https://doi.org/10.1162/ling.2009.40.3.514>
- Onea, Edgar & Beaver, David. 2011. Hungarian focus is not exhausted. In Ed Cormany, Satoshi Ito & David Lutz (eds.), *Proceedings of the 19th Semantics and Linguistic Theory Conference 19*, 342–359. <https://doi.org/10.3765/salt.v0i0.2524>
- Ponarjadov, Vadim V. 2010. *Porjadok slov v permskix jazykax v sravnitel'no-tipologičeskom osveščanii (prostoe predloženie)* [Word order in the Permian languages in a comparative-typological approach (simple clauses)]. Syktyvkar: Komi nauchnyj centr UrO.

- Repp, Sophie. 2016. Contrast: Dissecting an elusive information-structural notion and its role in grammar. In Caroline Féry & Shinichiro Ishihara (eds.), *The Oxford Handbook of Information Structure*, 270–289. Oxford: Oxford University Press. <https://doi.org/10.1093/oxfordhb/9780199642670.013.006>
- Salánki, Zsuzsanna. 2007. Az udmurt nyelv mai helyzete [The present-day situation of the Udmurt language]. Eötvös Loránd University, PhD dissertation. Manuscript.
- Schlesewsky, Matthias & Bornkessel, Ina. 2004. On incremental interpretation: degrees of meaning accessed during sentence comprehension. *Lingua* 114. 1213–1234. <https://doi.org/10.1016/j.lingua.2003.07.006>
- Suihkonen, Pirkko. 1990. Korpustutkimus kielitypologiassa sovellettuna udmurttiin [Computer corpus analysis in language typology applied to Udmurt]. *Mémoires de la Société Finno-ougrienne* 207.
- Surányi, Balázs, Asztalos, Erika, Brattico, Pauli & Sakhai, Heete. To appear. In Anne Tamm & Anne Vainikka (eds.), *Uralic Syntax*. Cambridge: Cambridge University Press.
- Surányi, Balázs. 2011. A szintaktikailag jelöletlen fókusz pragmatikája [On the pragmatics of syntactically unmarked focus]. In Bartos Huba (ed.), *Általános Nyelvészeti Tanulmányok XXIII. Új irányok és eredmények a Magyar mondattani kutatásban*, 281–313. Budapest: Akadémiai Kiadó.
- Szabolcsi, Anna. 1981. The Semantics of Topic-Focus Articulation. In Jeroen Groenendijk, Theo Janssen & Martin Stokhof (eds.), *Formal Methods in the Study of Language*, 513–540. Amsterdam: Mathematisch Centrum.
- Tánczos, Orsolya. 2010. Szórendi variációk és lehetséges okaik az udmurtban [Word order variation and its possible causes in Udmurt]. *Nyelvtudományi Közlemények* 107. 218–228.
- Tánczos, Orsolya. 2013. Hogy... hogy? Kettős kötőszók az udmurt mondatban [That... that – double complementizers in Udmurt]. In Agyagási Klára, É. Kiss & Hegedűs Attila (eds.), *Nyelvelmélet és kontaktológia 2* [Language theory and contactology 2], 95–112. Pilisscaba: PPKE BTK Elméleti Nyelvészeti Tanszék – Magyar Nyelvészeti Tanszék.
- Timerxanova, Nadezhda N. 2006. *Udmurt kyl. Nyrýs kutskúsjosly dyšetskon kníga* [Udmurt language. A course book for beginners]. Ižkar – Pilisscaba: Udmurt universitet.
- Timerxanova, Nadezhda N. 2011. Osobennost' porjadka slov v prozaicheskikh proizvedenijakh G. E. Vereshchagina i v sovremennom udmurtskom jazyke [Word order in the prosaic works of G. E. Vereščagin and in the contemporary Udmurt language]. In T. A. Krasnova (ed.), *Tipologičeskie aspekty mnogojazyčija v sovremennom obrazovatel'nom prostranstve* [Typological aspects of multilingualism in the modern educational space], 180–185. Izhevsk: Izdatel'stvo "Udmurtskij Universitet".
- Titov, Elena. 2012. Information Structure of Argument Order Alternations. University College London, PhD dissertation. Manuscript retrieved from [https://discovery.ucl.ac.uk/id/eprint/1370567/1/PhDThesis\\_TITOV.pdf](https://discovery.ucl.ac.uk/id/eprint/1370567/1/PhDThesis_TITOV.pdf)
- Vilkuna, Maria. 1998. *Word Order in European Uralic*. In Anna Siewierska (ed.), *Constituent Order in the Languages of Europe* (Empirical approaches to language typology 20 (1)), 173–233. Berlin–New York: Mouton de Gruyter. <https://doi.org/10.1515/9783110812206.173>
- Žujkov, S. P. 1937. *Osnovy grammatiki udmurtskogo jazyka: tezisy k pervoj respublikanskoj jazykovej konferencii* [Bases of the grammar of the Udmurt language: thesis for the first republican linguistics conference]. Izhevsk: Udmurtgosizdat.



- von Fintel, Kai. 2004. Would you believe it? The king of France is back! (Presuppositions and Truth-Value Intuitions). In Marga Reimer & Anne Bezuidenhout (eds.), *Descriptions and beyond*, 315–341. Oxford: Oxford University Press.
- Winkler, Eberhard. 2001. Udmurt. *Languages of the World* 212. München: Lincom Europa.
- Winkler, Eberhard. 2011. Udmurtische Grammatik. *Veröffentlichungen der Societas Uralo-Altaica* 81. Wiesbaden: Harrassowitz.

Erika Asztalos

Research Institute for Linguistics, Budapest and Eötvös Loránd University

aszterik@nytud.hu

# Web Corpora of Volga-Kama Uralic Languages<sup>1</sup>

Timofey Arkhangelskiy

This paper presents corpora of five minority Uralic languages that belong or are adjacent to the Volga-Kama area, which has been characterized as a Sprachbund (Bereczki 1983, Helinski 2003). A total of 11 corpora contain written and, in one case, spoken texts in Udmurt, Komi, Meadow Mari, Erzya and Moksha languages. The described resources are “web corpora” both in terms of their accessibility (all of them are accessible through a web-based query interface) and, in most cases, in terms of the medium (almost all texts come from web resources, such as digital newspapers and social media). The paper describes the corpora from the user perspective. The main focus is on the search capabilities and on certain research questions that can be studied with the help of these corpora. All corpora are available at <http://volgakama.web-corpora.net/>.

## 1 Introduction

Linguistic corpora as research tools and corpus linguistics as a methodology have experienced exponential growth since the 1990s. Multiple general-use reference corpora, as well as thousands smaller research-specific corpora, have been developed for major languages of the world. The Uralic family is no exception. For example, already in early 2000s there existed a number of large annotated corpora for Hungarian, such as the Hungarian National Corpus (Váradi 2002); somewhat smaller, but syntactically annotated Szeged corpus (Csendes et al. 2004); vast Hungarian web corpus (Halácsy et al. 2004); historical corpus (Pajzs 2000), etc. However, the minority Uralic languages spoken in Russia, even the largest and most vital ones, had a different fate. Until mid-2010s, only digital text collections of a limited size were created for some of them, e.g. by Suihkonen (1998), or small spoken corpora recorded by researchers in the field. First reasonably large publicly available written corpora for these languages only started appearing in 2014-2015, when the first versions of the literary Komi corpora (by the Syktyvkar-based FU-Lab team headed by Marina Fedina), the Udmurt corpus (by Maria Medvedeva and Timofey Arkhangelskiy) and Mari corpora (Bradley 2015) were created.

The corpora described in this paper were mostly developed in 2017-2019 by Timofey Arkhangelskiy with the purpose of filling this gap. The two exceptions are the “main” Udmurt corpus, which was started earlier in collaboration with Maria Medvedeva, and the spoken Udmurt corpus, which contains the data collected by Ekaterina Georgieva (see below). All corpora are available at <http://volgakama.web-corpora.net/>.

Since the languages in question share many properties such as some grammatical features or Cyrillic-based orthography, and have comparable level of digital presence, or digital vitality (Kornai 2016), similar methods and tools were used for developing the corpora. The vast majority of texts in all written corpora come from the web; my goal was to collect all or most texts written on the internet in the relevant languages. For each language, a rule-based morphological analyzer was developed; all of them are open source and can be found through the links in the respective corpus pages. Each analyzer contains a grammatical dictionary and a formalized description of the inflectional (as well as some

---

<sup>1</sup> The work is supported by RFBR grant 20-512-14003 ASCF\_a “Linguistic diversity in the Volga-Kama region. Typology and language documentation between Volga and Urals”.

productive derivational) morphology. Since the analyzers are dictionary-based, not all words in the corpora will have a morphological analysis. Words which are not covered in the dictionary or that contain spelling mistakes or non-standard/dialectal affixes do not receive analyses. The proportion of analyzed words is different for different corpora and varies between 80% and 96%. Also, most analyzers do not take word's context into account. This leads to ambiguity, whereby each word receives all potentially possible analyses, even though only one of them is correct in the given context. For instance, an Erzya token *valdo* can in principle be analyzed either as the base form of the adjective *valdo* 'bright', or as the ablative of the word *val* 'word' (*val-do* word-ABL).<sup>2</sup> Without disambiguation, both analyses will be assigned to each *valdo* token in the entire corpus.

More detailed technical information about the corpus development process can be found in (Arkhangelskiy 2019).

## 2 Sources

For each language, two written corpora were created: a "main" corpus and a social media corpus. The latter contains texts from social media (*vkontakte*, which is the most popular social media platform in Russia, and, in some cases, forums), while the former contains all other digital texts. Other social media, such as Facebook, Twitter or Odnoklassniki, presumably contain far fewer posts in minority Uralic languages than *vkontakte*, and were not included at this stage.

The reason for this dichotomy is that linguistic properties of these two types of texts are so different that different processing pipelines and different metadata are required for them. One significant difference is code switching, which is ubiquitous on social media, but rather limited or nonexistent in other texts (even in blogs). As a consequence, the social media corpora contain sentence-level language tagging and offer an option of searching in Russian sentences written on pages that also contain Uralic posts. The number of misspellings and dialectal material is also higher in social media, which is why a slightly different approach was taken for tagging them. The social media corpora are generally smaller than their "main" counterparts and contain between 0.014 and 3.59 million words in the target languages (as well as several times more words in Russian). Their sizes are summarized in Table 2.

The "main" corpora mainly consist of contemporary digital press but include other digital texts as well. Table 1 presents the genre distribution in the five "main" corpora and their total sizes. The "other" column subsumes fiction, scientific papers, Bible translations, Wikipedia articles (filtered by quality), official texts and some other genres. Most texts in the corpora were written between 2010 and 2019, but there are some earlier texts as well.

Metadata for both kinds of corpora include year of creation (exact date in the case of newspaper articles), title and author (when known). The main corpora also contain genre metadata. The social media corpora contain information about relevant distinctions, e.g. whether the text was taken from a post or a comment, or whether it appeared on a group page or a personal page. Additionally, it includes sociolinguistic data about the

---

<sup>2</sup> The following abbreviations are used in the paper: 1 = first person, 2 = second person, ABL = ablative case, FUT = future tense, ILL = illative case, M = million, NOM = nominative case, NP = noun phrase, P = possessive suffix, PL = plural, SG = singular.

authors (in aggregated, non-identifying form) whenever the authors indicated them in their profile.

| Language    | size in words | press (%) | blogs (%) | other |
|-------------|---------------|-----------|-----------|-------|
| Udmurt      | 9.57M         | 91.3%     | 5.1%      | 3.6%  |
| Komi-Zyrian | 1.75M         | 100%      | 0%        | 0%    |
| Meadow Mari | 2.63M         | 84%       | 0%        | 16%   |
| Erzya       | 2.3M          | 67.4%     | 6%        | 26.6% |
| Moksha      | 1.74M         | 86.4%     | 0.7%      | 12.9% |

Table 1: *Size and composition of the “main” corpora*

| Language    | size in words<br>(Uralic part) | size in words<br>(Russian part) |
|-------------|--------------------------------|---------------------------------|
| Udmurt      | 2.66M                          | 9.83M                           |
| Komi-Zyrian | 2.14M                          | 16.12M                          |
| Meadow Mari | 3.59M                          | 15.1M                           |
| Erzya       | 0.83M                          | 5.23M                           |
| Moksha      | 0.014M                         | 0.17M                           |

Table 2. *Size of the social media corpora*

Although the sizes of these corpora are several orders of magnitude smaller than those of e.g. contemporary Hungarian corpora, it is likely that the majority of digital texts available in these languages on the web has been included. A significant expansion of these corpora would necessarily require adding digitized texts from traditional media (books and newspapers), which requires a much higher level of time and resources.

The only spoken corpus so far contains transcribed Udmurt recordings made by Ekaterina Georgieva in several Udmurt dialects (Arkhangelskiy and Georgieva 2018). Although very different in its size and composition from the rest, it was processed using approximately the same pipeline and published through the same search interface as the other corpora.

### 3 Search capabilities

For the linguistic data to be reusable, it is crucial that they come with a tool that allows for complex search queries. As an example, the literary Komi corpus by FU-Lab, which is amazing in terms of its contents (over 50 million words of texts in a variety of genres, spanning almost a century), only allows very basic search requests, and therefore is difficult to use in some kinds of research.

All corpora described in this paper are published through the *tsakorpus* search platform that I started developing in 2017 and maintain now.<sup>3</sup> When developing it, I had several primary objectives:

- Provide an intuitive user interface that would allow complex linguistic queries without the need to learn a full-fledged query language such as CQP, used in Corpus Workbench (Evert and Hardie 2011), or AQL, used in ANNIS (Krause 2019).

<sup>3</sup> <https://bitbucket.org/tsakorpus/tsakorpus>

- Treat various corpus types (written, sound-aligned, parallel etc.) in a uniform way.
- Make sure the platform is fast enough to enable even sophisticated queries on mid-sized corpora (1–100 million words) with heavy annotation.
- Make the platform ambiguity-friendly. When it comes to POS tagging, it is assumed in most corpora of major languages that each analyzed word can have exactly one analysis. It might indeed be possible to choose one analysis out of several theoretically correct ones based on the context with very high precision, e.g. using neural networks trained on large manually tagged datasets, for major languages. However, for under-resourced languages this is usually not the case. Since there are no such datasets for them, any kind of statistical analyzer that only leaves one analysis for each word will make too many mistakes. Even with a 5% error rate the linguist risks not being able to find many relevant, but incorrectly tagged examples. Keeping ambiguous analyses makes the linguist's work more time-consuming, but reduces the chances of missing something important in the data.

The tsakorpus platform is open-source and language-independent. Since its creation, it has been used in a number of projects other than the one described here, e.g. INEL Selkup corpus (Brykina et al. 2020; <https://inel.corpora.uni-hamburg.de/SelkupCorpus/search>), Spoken corpus of Khakas (Maltseva and Sokur 2020, [https://linghub.ru/oral\\_khakas\\_corpus/](https://linghub.ru/oral_khakas_corpus/)), or Bashkir National Corpus (<http://bashcorpus.ru/>). The search interface is available in English and Russian.

There is a concise description of the search functionality in the Help window in each corpus (orange question mark at the top of the page). Instead of listing individual features, I will now describe a single research question that requires building a rather complex query, to demonstrate the capabilities of the platform. Udmurt Social media corpus will be taken as an example; the same search functionality is available in all other corpora (although the grammatical tags are language-specific).

Just as in other Volga-Kama languages, most spatial relations in Udmurt are expressed by inflected postpositions, or relational nouns, which have a nominal or pronominal dependent. In Standard Udmurt, the only available construction of this kind requires the dependent to be in the nominative and not cross-referenced on the head, as in Example 1. This is prescribed in most grammars and textbooks. However, there are other options available in the dialects. In one of them, 1st and 2nd person pronominal dependents are still in the nominative, but trigger appropriate possessive marking on the head, as in Example 2 (which is highly unusual for an Udmurt NP). This option has been mentioned in the grammar by Winkler (2011) without any remarks about its dialectal nature; other than that, it is unknown where exactly and why this construction exists.

(1) *mon dor-i*  
 I.NOM at-ILL  
 ‘towards me / to my place’

(2) *mon dor-a-m*  
 I.NOM at-ILL-P.1SG  
 ‘towards me / to my place’

Since the social media corpus contains geographical metadata (place of birth and current location) for some authors, it would make sense to search the second construction and see whether its approximate areal distribution can be established.

Here is how an appropriate search request can be built in the web interface:

- By default, tsakorpus shows one block of search fields that corresponds to one search term. Since the construction in question involves two words, a second block should be added by clicking the plus sign (“add word”) in the right-side pane of the first block.

- If your search includes multiple words, the default behavior is to find all sentences that include all of them regardless of their mutual order or distance. Since we want the first word to be located immediately to the left of the second, a distance requirement has to be added. This is done by clicking the “add distance” button (two arrows pointing in opposite directions) in the second block. The default values (distance of at least 1 word and at most 1 word from the word #1) describe exactly the scenario that we need.

- The first word, i.e. the dependent, has to be a personal pronoun of first or second person. The easiest way to specify this constraint is to list all four possible variants in the Lemma field or in the Word field.<sup>4</sup> The expression that has to be put there is *мон|мон|му|мӱ*. The pipe symbol stands for logical OR in the Word, Lemma and Grammar fields; the words separated by it are the lemmata of the Udmurt 1SG, 2SG, 1PL and 2PL pronouns, respectively. Putting this string in the Word field means that the first word in the construction must coincide exactly with one of these four options. Since in the case of pronouns, the lemma coincides with the nominative form, this will be sufficient for our purposes. If, instead of that, this expression is pasted in the Lemma field, by default it means that all forms of these four pronouns must be found. In our case, we would have to additionally specify that only the nominative has to be found by typing *nom* in the Grammar field. The *nom* tag stands for the nominative (or, in the case of nouns, unmarked accusative); the entire tagset, i.e. the list of grammatical tags used in the corpus, can be found at the start page of each corpus. Instead of typing, the values can also be selected from a pop-up window that appears after clicking the button at the right end of the Grammar field. The two methods (putting the pronouns in the Word field or putting them in the Lemma field while specifying their case) may look the same; nevertheless, the latter yields more precise results. The reason for that is that some frequent misspellings, such as missing diacritics in *мӱ* you.PL.NOM, are handled correctly by the analyzer. Since the misspelled word *mu* will be found by the lemma+case query, but missed by the word query, the lemma+case query is preferable in the case of noisy texts.

- The second word can be any relational noun with a 1st or 2nd person possessive suffix. Additionally, we will limit the search to the three most frequent spatial cases that relational nouns combine with: locative (inessive), illative and elative. This constraint can be set by putting the following expression in the Grammar field of the second block: *rel\_n,(1sg|2sg|1pl|2pl),(loc|el|ill)*. Again, the pipe symbol stands for the logical OR; comma stands for AND, and parentheses are used for grouping.



- Finally, a metadata constraint has to be added to narrow down the search. In tsakorpus, two kinds of metadata are distinguished. The first kind is text-level metadata, such as title, author, or creation year of the text. Their values can be used for limiting the search to a subset of corpus texts, e.g. all texts written by a certain author, by clicking the “Select subcorpus” button. The second kind is the sentence-level metadata, which pertain to individual sentences. In the case of social media corpora, sentence-level metadata contain the information about the author of each particular sentence or post, while text-level metadata refer to the owner of the page where that post was written. Since we are

---

<sup>4</sup> I am omitting the 1pl inclusive pronoun (Maksimov and Panina 2018), which coincides with a possessive form of the reflexive pronoun, because it behaves differently in this respect.

interested in the areal distribution of the phenomenon in question, only those sentences are relevant for which the author’s place of birth (which is an approximation of their dialect) is known. Since only a minority of users indicate their birth place in their profile, the “non-empty birth place” requirement will cut off many irrelevant search hits and thus save the researcher’s time.


Sentence-level metadata requirements can be set by clicking a downwards arrow in any of the two blocks. In our case, the “Account type (post-level)” field should be set to *user*, so that posts authored by groups are excluded. The “Birth place (post-level)” has to be set to  $\sim(\textit{unknown}|\textit{other})$ , where  $\sim$  stands for negation. This expression will cut off sentences written by users whose birth place is either not indicated (which is expressed by the value of *unknown* in the corpora), or indicated, but not recognized by the geographical classifier at annotation time (the value of *other*).


Udmurt social media corpus RU | EN |  

**Word #1**

Word:

Lemma:

Grammar:  

Gloss:  

---

Translation (ru):

2nd lemma:

2nd transl. (ru):

---

Metafield\_author\_id\_post:

Year (post-level):

Sex (post-level):


Current place (post-level):

Birth place (post-level):

Birth year (post-level):

Account type (post-level):

Post type:

Analyses:  


Position in sentence:


Language/tier:

**Word #2**

Word:

Lemma:

Grammar:  

Gloss:  

---

Language/tier:

Distance to word #

from

to

Full-text search:   Precise match






Figure 1: Search query in *Tsakorpus* interface of the Udmurt social media corpus

Clicking “Search sentences” will yield a number of search hits (21 as of May 2020), where the construction in question is highlighted. The examples are sorted randomly. First, this prevents the user from reconstructing the entire text, which would be a copyright violation. Second, in the case of a large number of results, the user can easily see how the construction in question behaves on average by looking at the first 100 or 200 sentences, for which it is crucial to have an unbiased sample.

The final step is going through the sentences found and assessing them manually. As it almost always happens, only a part of the search hits contain the construction that is

being looked for. For instance, the sentence in (3) technically conforms to the query. However, the pronoun there is the subject rather than the dependent of the relational noun, which has no overt dependent:

- (3) *Berjtsk-o-d*            *ton*            *dor-a-m.*  
 return-FUT-2SG    you.SG.NOM    at-ILL-P.1SG  
 ‘You will return to me.’

After sifting through the hits, we find that only 5 sentences make it to the final list of genuine examples. Sentence-level metadata for each of them can be seen in the upper right corner when hovering the mouse pointer over the sentence.

#### 4 Social media corpora and dialectology

The corpora presented here can be used for researching a number of topics in the areas of lexicography, morphology and syntax. However, the metadata in the social media corpora make it possible to conduct research on sociolinguistics and dialectology. This prospect seems especially important to me, since these disciplines have not benefited from corpora as much as other areas of linguistics. Besides, dialectological research with its fieldwork in multiple locations is a very expensive and time-consuming undertaking. Therefore, it is important to know to which extent social media data can be used to learn about areal distributions of words and grammatical phenomena.

As I have demonstrated elsewhere (Arkhangelskiy 2019), the social media corpora can be used in studies of dialectal vocabulary. By comparing the data extracted from social media corpora with the results of traditional dialectological surveys, I showed that although corpus data does not provide enough information on some varieties, the information it does provide does not contradict the facts established by traditional dialectology. Therefore, social media corpora can be used as incomplete, but relatively reliable sources of dialectological data. As such, they can be used in preliminary studies, e.g. when planning dialectological fieldwork.

Since Uralic dialectology has paid much more attention to phonology and vocabulary than to morphosyntax, relatively little is known about dialectal distribution of syntactic constructions such as the one described in Section 4. Social media corpora could prove a great help here. The examples of the non-standard construction found in the corpus belong to the authors born in Igra and Sharkan districts, which allows us to very roughly outline the area where this phenomenon exists. My preliminary fieldwork shows that it indeed exists there, while being either infrequent or altogether nonexistent elsewhere.

#### 5 Future work

The corpora described in this paper were last updated in 2018–2019. In order to keep them up to date, I am working on a semi-automatic pipeline that would make it easy to add new texts from social media, blogs and newspapers each 6 months. Geographical metadata has to be added to the social media corpora to enable the dialectological research described above; right now, it is only available in Udmurt and Meadow Mari (to a certain extent) corpora. Another direction of improvement is the functionality of the search platform; I



expect the next major release to be ready in late 2020. Finally, I am collaborating with other teams who have spoken corpora of Volga-Kama Uralic languages in order to make them available through tsakorpus and provide the functionality necessary for searching them. At the moment, this includes a spoken Meadow Mari corpus (Anna Volkova, Aigul Zakirova, Linguistic Convergence Laboratory at Higher School of Economics); I will be happy to collaborate with other researchers and teams as well.

## 6 Conclusion

I have presented 11 corpora of five Uralic languages of the Volga-Kama area. All of them have morphological annotation and are publicly available through a web interface. These corpora can be used in various kinds of linguistic research, such as lexicography, morphology and syntax. Additionally, the social media corpora may be used in studies of sociolinguistics and dialectology. I hope that these corpora will help linguists who specialize in these under-resourced Uralic languages and boost the research on them.

## References

- Arkhangelskiy, Timofey. 2019. Corpora of social media in minority Uralic languages. In Tommi A. Pirinen, Heiki-Jaan Kaalep & Francis M. Tyers (eds.), *Proceedings of the Fifth International Workshop on Computational Linguistics for Uralic Languages*, 125–140. Tartu: Association for Computational Linguistics. <https://doi.org/10.18653/v1/w19-0311>
- Arkhangelskiy, Timofey & Georgieva, Ekaterina. 2018. Sound-aligned corpus of Udmurt dialectal texts. In Tommi A. Pirinen, Michael Riebler, Jack Rueter, Trond Trosterud & Francis M. Tyers (eds.), *Proceedings of the Fourth International Workshop on Computational Linguistics of Uralic Languages*, 26–38. Helsinki: Association for Computational Linguistics. <https://doi.org/10.18653/v1/w18-0203>
- Berezki, Gábor. 1983. A Volga-Káma vidék nyelveinek areális kapcsolatai. In: Balázs János (ed.), *Areális nyelvészeti tanulmányok*, 207–236. Budapest: Tankönyvkiadó.
- Bradley, Jeremy. 2015. Corpus.mari-language.com: A Rudimentary Corpus Searchable by Syntactic and Morphological Patterns. In *Proceedings of the First international workshop on computational linguistics for Uralic languages*. Septentrio Conference Series. Tromsø: Septentrio Academic Publishing. <https://doi.org/10.7557/5.3468>
- Brykina, Maria, Orlova, Svetlana & Wagner-Nagy, Beáta. 2020. INEL Selkup Corpus. Version 1.0. Publication date 2020-06-16. Archived in Hamburger Zentrum für Sprachkorpora. <http://hdl.handle.net/11022/0000-0007-CAE5-3>. In Wagner-Nagy Beáta; Alexandre Arkhipov, Anne Ferger; Daniel Jettka & Timm Lehmberg (eds.), *The INEL corpora of indigenous Northern Eurasian languages*.
- Csendes, Dóra, Csirik, János & Gyimóthy, Tibor. 2004. The Szeged Corpus: A POS Tagged and Syntactically Annotated Hungarian Natural Language Corpus. In P. Sojka, I. Kopeček & K. Pala (eds.), *Text, Speech and Dialogue*, 41–47. Berlin/Heidelberg: Springer. [https://doi.org/10.1007/978-3-540-30120-2\\_6](https://doi.org/10.1007/978-3-540-30120-2_6)
- Evert, Stephan & Hardie, Andrew. 2011. Twenty-first century Corpus Workbench: Updating a query architecture for the new millennium. In *Proceedings of the Corpus Linguistics 2011 Conference*. Birmingham, UK.

- Halácsy, Péter, Kornai, András, Németh, László, Rung, András, Szakadát, István & Trón, Viktor. 2004. Creating open language resources for Hungarian. In Maria Teresa Lino, Maria Francisca Xavier, Fátima Ferreira, Rute Costa & Raquel Silva (eds.), *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, 203–210. European Language Resources Association. Lisbon, Portugal.
- Helinski, Eugene. 2003. Areal groupings (Sprachbünde) within and across the borders of the Uralic language family: A survey. *Nyelvtudományi Közlemények* 100. 156–167.
- Kornai, András. 2016. Computational linguistics of borderline vital languages in the Uralic family. In Tommi A. Pirinen, Simon Eszter, Francis M. Tyers & Vincze Veronika (eds.), *Proceedings of the Second International Workshop on Computational Linguistics for Uralic Languages*. Szeged: Szegedi Tudományegyetem. (Available online at <http://kornai.com/Drafts/iwclul.pdf>, accessed on 05.11.2018.)
- Krause, Thomas, 2019. ANNIS: A graph-based query system for deeply annotated text corpora. Humboldt-Universität zu Berlin, PhD thesis.
- Maksimov, Sergey & Panina, Tatjana. 2018. On the category of clusivity in the Udmurt language. *Linguistica Uralica* 54(3), 213–224. <https://doi.org/10.3176/lu.2018.3.05>
- Maltseva, Vera & Sokur, Elena. *Spoken corpus of the dialects of Khakas*. Moscow: Institute of Linguistics; Moscow: Linguistic Convergence Laboratory, NRU HSE. (Available online at [https://linghub.ru/oral\\_khakas\\_corpus/](https://linghub.ru/oral_khakas_corpus/), accessed on 04.08.2020.)
- Pajzs, Júlia. 2000. Making Historical Dictionaries with the Computer. In Ulrich Heid, Stefan Evert, Egbert Lehmann & Christian Rohrer (eds.), *Proceedings of EURALEX 2000*, 249–259. Stuttgart: Universität Stuttgart.
- Suihkonen, Pirkko Marjatta. 1998. *Documentation of the Computer Corpora of Uralic Languages at the University of Helsinki*. Helsinki: Department of General Linguistics, University of Helsinki. Technical paper.
- Váradi, Tamás. 2002. The Hungarian National Corpus. In Manuel González Rodríguez, Carmen Paz Suarez Araujo (eds.), *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*, 385–389. Las Palmas, Canary Islands, Spain.
- Winkler, Eberhard. 2011. *Udmurtische Grammatik* (Veröffentlichungen der Societas Uralo-Altaica 81). Wiesbaden: Harrassowitz Verlag.

Timofey Arkhangelskiy  
Universität Hamburg  
timarkh@gmail.com

# The INEL Dolgan corpus: Insights into an endangered language of Northern Eurasia<sup>1</sup>

Chris Lasse Däbritz

The paper at hand presents a description of the INEL Dolgan Corpus that has been created from 2016 to 2019 within the INEL project, located at the Institute for Finno-Ugric/Uralic Studies of the University of Hamburg. The corpus aims to provide a digital research infrastructure for Dolgan, an indigenous language of Northern Siberia. Though Dolgan is a Turkic language, the corpus is relevant for researchers of Uralic languages both due to the close areal connections of Uralic with Dolgan on the Taymyr peninsula and on account of the fact that it is an example of electronic research infrastructure developed for an endangered language. After introducing Dolgan and the INEL project, the paper describes the INEL Dolgan Corpus in detail, focusing on its linguistic content, annotation layers and search possibilities. Finally, the paper provides an outlook on how the corpus contributes to furthering research on this endangered language.

Keywords: *corpus, INEL project, Dolgan, languages of Northern Siberia, endangered languages*

## 1 Introduction

Dolgan is a Turkic language that is spoken by 1,054 people (VPN 2010) on the Taymyr peninsula and in adjacent areas in the extreme north of the Russian Federation. Several features call for the documentation and investigation of this indigenous language of Northern Siberia. First, Dolgan has been regarded a dialect of Sakha (Yakut) for a long time. As recently as in the 1980s, Ubrjatova (1985) pointed out that Dolgan is a language on its own that arose from Sakha (Yakut) under heavy Evenki (< Tungusic) substrate. Until today this has led to many accounts to Dolgan that are biased by Sakha (Yakut). Second, Dolgan was and is in contact with many surrounding languages (Sakha (Yakut), Evenki, to a lesser extent Nganasan and Enets, as well as Standard Russian, local Russian varieties and Taymyr Pidgin Russian). Especially the contact scenario, out of which Dolgan arose, is not fully understood yet, neither is the intensity of possible Samoyedic–Dolgan contacts. Therefore, the investigation of Dolgan has a particular relevance for Samoyedic studies, too. Finally – like many other indigenous languages of Siberia – Dolgan faces extinction, which is a sufficient reason on its own for conducting documentation work, collecting language material and compiling a linguistic corpus.

The INEL Dolgan Corpus<sup>2</sup> aims at founding the empirical base for the investigation of the language, which is the main goal of all INEL corpora (see section 2). In order to reach this goal, material from as many sources as possible is collected, digitized and linguistically annotated; moreover, some linguistic research already has been carried out on the basis of the INEL Dolgan Corpus (see section 3). Finally, the INEL Dolgan Corpus may, thus, contribute to an up-to-date documentation of Siberian languages, being useful for a wide range of both linguistic and even non-linguistic research (see section 4).

---

<sup>1</sup> This publication has been produced in the context of the joint research funding of the German Federal Government and Federal States in the Academies' Programme, with funding from the Federal Ministry of Education and Research and the Free and Hanseatic City of Hamburg. The Academies' Programme is coordinated by the Union of the German Academies of Sciences and Humanities.

<sup>2</sup> PID: <http://hdl.handle.net/11022/0000-0007-CAE7-1>

## 2 INEL and the INEL corpora

The acronym “INEL” stands for *Grammatical Descriptions, Corpora and Language Technology for Indigenous Northern Eurasian Languages*, and refers to a long-term research project, being carried out at the Institute for Finno-Ugric/Uralic Studies of the University of Hamburg.<sup>3</sup> Its major aim is to create digital linguistic corpora as well as research infrastructure for several lesser-described Northern Eurasian languages and varieties. It is scheduled for 18 years (2016–2033), allowing three years for each language/variety dealt with. Table 1 shows the finalized and ongoing subprojects. In the future, further languages such as Ket and Nenets (Taymyr and Kanin variety) are planned to be included.

| Language                                 | Period            |
|--|-------------------|
| Selkup (all varieties)                   | 01/2016 – 12/2021 |
| Kamas                                    | 01/2016 – 12/2018 |
| Dolgan                                   | 09/2016 – 08/2019 |
| Evenki (Northern and Southern varieties) | 01/2019 – 12/2021 |

Table 1: *Languages dealt with in the INEL project*

As can be seen from the table above, the languages dealt with in the INEL project come mostly from Western Siberia, being under-resourced and exhibiting clear areal connections. Although the INEL project contributes to the documentation of these languages, it differs from many language documentation projects in an important way: The material that is processed often comes from existing archives and collections, rather than being collected within the project itself. This leads to a broad variety of material included, which will be described in detail for Dolgan in section 3. This language material is digitized and, thus, made accessible for linguistic annotation and the compilation of linguistic corpora. Up to now, the INEL project published three open-access corpora, namely the *INEL Selkup Corpus* (Brykina et al. 2020), the *INEL Kamas Corpus* (Gusev et al. 2019), and the *INEL Dolgan Corpus* (Däbritz et al. 2019).<sup>4</sup> The following Table 2 sums up basic statistical information on those corpora.

<sup>3</sup> The principal investigator is Prof. Beáta Wagner-Nagy, and the funding was applied for by Prof. Beáta Wagner-Nagy, Dr. Michael Rießler, Hanna Hedeland and Timm Lehmborg. The current project members are Prof. Dr. Beáta Wagner-Nagy, Dr. Alexandre Arkhipov (research coordinator), Timm Lehmborg (technical coordinator), Dr. Maria Brykina, Chris Lasse Däbritz (linguistic team), Anne Ferger, Daniel Jettka (technical team). The project website is available at <https://www.slm.uni-hamburg.de/inel/>.

<sup>4</sup> The corpora are available under the terms and conditions of Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International Public License (CC BY-NC-SA 4.0), cf. <https://creativecommons.org/licenses/by-nc-sa/4.0/legalcode>, last access: 22/04/2020.

| Corpus      | Transcripts <sup>5</sup> | Tokens | Speakers        | Genres   |
|-------------|--------------------------|--------|-----------------|--|
| INEL Selkup | 264                      | 42,466 | 74              | folklore, narrative, translations, songs, conversations  |
| INEL Kamas  | 158                      | 63,824 | 4 (+ 2 unknown) | folklore, narrative, songs, miscellaneous (e.g. riddles) |
| INEL Dolgan | 116                      | 77,636 | 61              | folklore, narrative, translations, songs, conversations  |

Table 2: *INEL corpora – statistics*

All INEL corpora are compiled following similar principles and guidelines. However, each corpus certainly has its peculiarities and special characteristics. The INEL Selkup Corpus is composed of the personal archive of Angelina Ivanova Kuzmina (1924–2002). It explicitly aims at covering all dialects of Selkup, which makes possible comparative studies of Northern, Central and Southern dialects. The INEL Kamas Corpus – as can be seen from the table – has a much smaller amount of speakers included, which is of course to be explained by the fact that Kamas is extinct, and there is simply no more material available. Nevertheless, the corpus contains transcripts from a relatively wide range of time, including both old texts from the 1910s and newer texts of Klavdiya Plotnikova from the 1960s and 1970s. The INEL Dolgan Corpus, finally, is the first corpus that covers a language, which is to some extent spoken in everyday life. Therefore, it was possible to include a higher amount of free conversations (radio interviews) into the corpus than in the cases of Selkup (especially Central and Southern dialects) and Kamas.

Thus, the INEL project provides an infrastructure for the compilation of structurally similar corpora of diverse languages, including diverse language material. For a concise description of the INEL project in general as well as those corpora, see also Arkhipov & Däbritz (2018).

### 3 The INEL Dolgan Corpus

As was mentioned already in the introduction, the INEL Dolgan Corpus aims at enabling the investigation of this rarely studied indigenous language of Northern Siberia on an empirically solid base. Given this, the content of the INEL Dolgan Corpus has to fulfil several criteria: as balanced a provenance as possible, as transparent a linguistic representation as possible and as accessible a technical representation as possible. The following paragraphs describe how the INEL Dolgan Corpus seeks to fulfil these criteria. The material included into the INEL Dolgan corpus comes from four very different sources:

- 1) texts from the published volume *Fol'klor Dolgan* [FD 2000] (Efremov et al. 2000),
- 2) audio material obtained from the *Taymyr House of National Arts* (TDNT),
- 3) audio material obtained from the collection of Eugénie Stapert, and
- 4) audio material collected during fieldwork in Dudinka in 2017.

---

<sup>5</sup> The term “transcript” is used here as a cover term for all items (texts, conversations or the like) included into the corpus.

Overall, the INEL Dolgan Corpus contains 116 transcripts (16 conversations, 50 folklore texts, 44 narratives, 2 songs, 4 translations from Russian) of 61 speakers (33 female, 28 male) with 11,329 utterances and 77,636 tokens. 81 communications can be linked to a corresponding audio file, making up a total of 10:42:14 hours of audio material. The following Figure 1 shows the number of tokens (green bars) and communications (blue bars) of each genre.

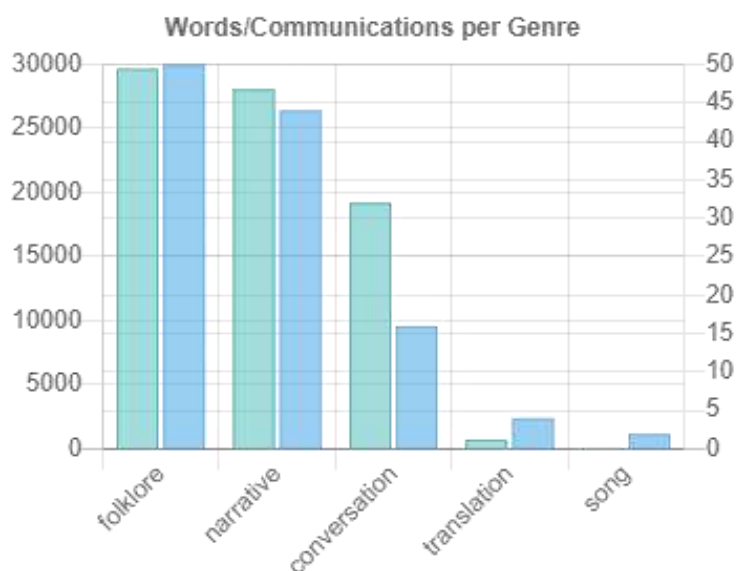


Figure 1: *Words/communications per genre in the INEL Dolgan Corpus*

The INEL Dolgan Corpus is published through the INEL infrastructure, the latter being partly based on existing infrastructure of the Hamburg Center for Language Corpora (Hamburger Zentrum für Sprachkorpora, HZSK).<sup>6</sup> The data is stored in XML-based format provided by the EXMARALDA program package.<sup>7</sup> To be able to browse the corpus and use the data locally, the relevant software tools (Partitur Editor<sup>8</sup>, Corpus Manager<sup>9</sup>, EXAKT<sup>10</sup>) have to be installed. In addition, the corpus – like the other INEL corpora, too – can be searched online using the Tsakonian Corpus Platform<sup>11</sup> (see Arkhangel'skiy, Ferger & Hedeland 2019 for technical details).

As for the content of the communications, there is always a phonological transcription of the Dolgan speech, morphological glossing as well as further annotations and translations into various languages. The principles of transcribing, glossing, annotating and translating are summarized in a user documentation file that is provided with the corpus data<sup>12</sup>, and is additionally published (Däbritz 2020).

The phonological transcription is based on principles used in all INEL corpora, which include elements from both IPA and FUT, the morphological glossing follows the

<sup>6</sup> <http://hdl.handle.net/11022/0000-0007-CAE7-1>, last access: 27/04/2020

<sup>7</sup> <https://exmaralda.org/en/>, last access: 27/04/2020

<sup>8</sup> <https://exmaralda.org/en/partitur-editor-en/>, last access: 27/04/2020

<sup>9</sup> <https://exmaralda.org/en/corpus-manager-en/>, last access: 27/04/2020

<sup>10</sup> <https://exmaralda.org/en/exakt-en/>, last access: 27/04/2020

<sup>11</sup> <https://bitbucket.org/tsakorpus/>, last access: 27/04/2020. Search can be performed through the following link: <https://inel.corpora.uni-hamburg.de/DolganCorpus/search>

<sup>12</sup> <http://hdl.handle.net/11022/0000-0007-CAE7-1>, last access: 28/04/2020.

principles of the Leipzig Glossing Rules (2015).<sup>13</sup> Lexical glosses are provided in English, German and Russian; grammatical glosses do not differ between the languages of analysis. Further annotation tiers contain the annotation of Semantic Roles (SeR), Syntactic Functions (SyF), Information Status (IST), Information Structure (Top and Foc), Borrowing (BOR) and Code-switching (CS). The annotations of SeR, SyF and IST are based on the principles developed for the *Nganasan Spoken Language Corpus* (NSLC; Brykina et al. 2018), described by Wagner-Nagy et al. (2018). The annotations of Top, Foc, BOR and CS were developed within the INEL project in close cooperation with the compilers of NSLC, see also Arkhipov (2020) for details of the latter two. Finally, free translations into English, German and Russian are provided. If the transcript was already published (transcripts from FD 2000) or had been translated by our native speaker assistants (transcripts from TDNT), this literal translation is given, too.

The deep annotation of the corpus data enables the user to conduct varied and complex searches. The grammatical glossing is form-oriented, i.e. grammatical forms are analyzed with respect to their components. As an example, Figure 2 contains the item *babuska-ŋ* ‘midwife-2SG’, which would be found via a search of *midwife* or the possessive suffix of the 2<sup>nd</sup> person singular. The further annotations, however, are function-oriented. Therefore, one would find the same item *babuska-ŋ* ‘midwife-2SG’ when searching for an agent (Semantic Roles), a subject (Syntactic Functions), a given referent (Information Status), a topic (Information Structure), or a cultural Russian borrowing (Borrowing). The function-oriented annotation tiers particularly contribute to the wide applicability of the corpus, since they enable the user to search specifically for these functional categories, even without having deep knowledge of the Dolgan language. This is relevant for typologists and/or theoretical linguists working with many languages and seeking for specific empirical data for their work. In order to illustrate this, Figure 2 shows the various annotations in a narrative text.

---

<sup>13</sup> The Leipzig Glossing Rules were developed and are regularly updated by the Max Planck Institute for Evolutionary Anthropology. The current version is available online at <https://www.eva.mpg.de/lingua/resources/glossing-rules.php> (last access: 27/04/2020).

|           | 242 [02:17.0]   | 243 [02:17.5] | 244 [02:18.0] | 245 [02:18.4] | 246 [02:19.2]        | 247 [02:19.8]          | 248 [02:20.4] | 249 [02:21.0] |
|-----------|---|---------------|---------------|---------------|----------------------|------------------------|---------------|---------------|
| ref       | SuAA_20XX_Birth_nar.029 (001.029)   |               |               |               |                      |                        |               |               |
| st        | Оччого буоллагына дуо бу баабускаҥ кэлэр ураһа дьиэгэ, төрүүр дьиэгэ, дьукаага. |               |               |               |                      |                        |               |               |
| ts        | Oččogo buollagina duo bu ba:buskaŋ keler uraha d'iege, törür d'iege, d'uka:ga.  |               |               |               |                      |                        |               |               |
| tx        | Oččogo  | buollagina    | duo           | bu            | ba:buskaŋ            | keler                  | uraha         | d'iege,       |
| mb        | oččogo  | buollagina    | duo           | bu            | ba:buska-ŋ           | kel-er                 | uraha         | d'ie-ge       |
| mp        | oččogo  | buollagina    | duo           | bu            | ba:biska-ŋ           | kel-Ar                 | uraha         | d'ie-GA       |
| ge        | then  | though        | MOD           | this          | midwife-2SG. [NOM]   | come-PRS. [3SG]        | pole. [NOM]   | tent-DAT/LOC  |
| gg        | dann  | aber          | MOD           | dieses        | Hebamme-2SG. [NOM]   | kommen-PRS. [3SG]      | Stange. [NOM] | Zelt-DAT/LOC  |
| gr        | тогда   | однако        | MOD           | этот          | повитуха-2SG. [NOM]  | приходить-PRS. [3SG]   | шест. [NOM]   | чум-DAT/LOC   |
| mc        | adv   | ptcl          | ptcl          | dempro        | n-n:(poss). [n:case] | v-v:tense. [v:pred.pn] | n. [n:case]   | n-n:case      |
| ps        | adv   | ptcl          | ptcl          | dempro        | n                    | v                      | n             | n             |
| SeR       | adv:Time  |               |               |               | np.h:A               |                        |               | np:G          |
| SyF       |   |               |               |               | np.h:S               | v:pred                 |               |               |
| IST       |   |               |               |               | giv-active           |                        |               | giv-active    |
| Top       |   |               |               |               | top.int. coner       |                        |               |               |
| Foc       |   |               |               |               |                      | foc.int                |               |               |
| BOR       |   |               |               |               | RUS:cult             |                        |               |               |
| BOR-Phon  |   |               |               |               | Vsub Csub            |                        |               |               |
| BOR-Morph |   |               |               |               | dir:infl             |                        |               |               |
| CS        |   |               |               |               |                      |                        |               |               |

Figure 2: Deep annotation in the INEL Dolgan Corpus

The metadata of the corpus is stored in the *Corpus Manager (Coma)* component of the EXMARaLDA system. The metadata of transcripts (called “communications” in EXMARaLDA) contains information about the place and date of recording or the genre of the transcript, as well as information on who did what in the transcription, glossing and annotation. The metadata of speakers contains the basic biographical data of the relevant speaker, i.e., place and date of birth, education, language competence, ethnic composition of the family, place(s) of living, etc. Figure 3 shows an example of speaker metadata in the INEL Dolgan Corpus.



| <b>Speaker: SuAA (Antonina Alekseevna Suzdalova, Sex: female)</b> |   |
|---|---|
| <b>Description (Speaker)</b>                                      |   |
| 1a Family name  | Suzdalova   |
| 1b Family name (RU)   | Суздадова   |
| 2a Given name   | Antonina  |
| 2b Given name (RU)  | Антонина  |
| 3a Patronymic   | Alekseevna  |
| 3b Patronymic (RU)  | Алексеевна  |
| 4 Locations   |   |
| <b>Basic biogr. data (Location)</b>                               |   |
| <b>Description (Location)</b>                                     |   |
| 1a Place of birth   | Novo-Letov`ye (Zhdanixa)  |
| 1b Place of birth (RU)  | Ново-Летовье (Жданиха)  |
| 2 Region  | Таумыр (Dolgano-Nenets) Autonomous Okrug                        |
| 3 Country   | Russia  |
| 4 Date of birth   | 1940.06.05.   |
| 5 Date of death   | 2015  |
| 6a Former residences  | Novo-Letov`ye (Zhdanixa), Хатанга, Красноярск, Хета, Сындасско, |
| 6b Former residences (RU)   | Ново-Летовье (Жданиха), Хатанга, Красноярск, Хета, Сындасско,   |
| 7a Domicile   | ...   |
| 7b Domicile (RU)  | ...   |
| <b>Education (Location)</b>                                       |   |
| <b>Description (Location)</b>                                     |   |
| 1a Education  | school (10 classes)   |
| 1b Education (RU)   | школа (10 классов)  |
| 2a Higher education   | pedagogical high school for kindergarden                        |
| 2b Higher education (RU)  | педагогическое училище (дошкольное)                             |
| 3a Occupation   | educator, folklore specialist                                   |
| 3b Occupation (RU)  | воспитатель, методист по фольклору                              |
| <b>Ethnicity (Location)</b>                                       |   |
| <b>Description (Location)</b>                                     |   |
| 1 Ethnicity   | Dolgan  |

Figure 3: *Speaker metadata in the INEL Dolgan Corpus*

As was mentioned above, the INEL Dolgan Corpus can be searched using either the EXAKT tool form the EXMARaLDA program package or the web-based search via the Tsakonian Corpus Platform. Each tool has respective strengths. In EXAKT (Figure 4), concordance searches can easily be combined with metadata automatically extracted from COMA (see above). In Figure 4, a test-search for the partitive case in Dolgan is presented. As can be seen, the respective token (marked red) is shown within its context. Additionally, further columns with annotations and/or metadata can be included. Here, the annotation of syntactic functions (mostly NP objects) and the dialect of the given text

(Upper vs. Lower) was chosen. The concordance could be filtered for any value within these annotations, e.g., one could display only those tokens that come from the Upper Dolgan dialect.

| #   | S    | Communication | Speaker | Left Context                | Match        | Right Context           | ge   | SyF   | 3 Dialect[C] |
|-----|------|---------------|---------|-----------------------------|--------------|-------------------------|------|-------|--------------|
| 1   | BaA  | 193...        | BaA     | ari ihikker holu.rgar ...   | u.ta         | bahan tolör, belemn...  | PART | s:adv | ...          |
| 2   | AKEE | 19...         | AKEE    | maččittar, masta:ŋ, ...     | uotta        | ottuŋ!"                 | PART | np:O  | Upper        |
| 3   | AKEE | 19...         | AKEE    | küörte:, iald'it kelle, ... | uotta        | ep!"                    | PART | np:O  | Upper        |
| 4   | AsKS | 19...         | AsKS    | "Tinna:k                    | goronuokta   | egeliem, kü:ten olör."  | PART | np:O  | Upper        |
| 5   | BaA  | 193...        | BaA     | ikum d'ogus, ulakan ...     | pabara:ŋkita | du:, komuosta du: ege   | PART | np:O  | ...          |
| 6   | BaA  | 193...        | BaA     | ulakan pabara:ŋkita ...     | komuosta     | du: ege", diebit.       | PART | np:O  | ...          |
| 7   | BaRD | 19...         | Ba...   | "Ha:tar                     | beliete      | bier", dien.            | PART | np:O  | ...          |
| 8   | BaRD | 19...         | Ba...   | "Haŋa ira:s če:lke:         | öldü:mne     | ani tis konugunan hars  | PART | np:O  | ...          |
| 9   | BaRD | 19...         | Ba...   | "                           | Oldo:nno     | ", diebit.              | PART | np:O  | ...          |
| ... | BeES | 19...         | BeES    | "Oŋoruŋ taŋara              | d'iete       | ".                      | PART | np:O  | Upper        |
| ... | ChGS | U...          | Ch...   | "                           | Nöŋüöte      | egeliŋ, nöŋüöte."       | PART | np:O  | Lower        |
| ... | ChGS | U...          | Ch...   | "Nöŋüöte egeliŋ,            | nöŋüöte      | "                       | PART | np:O  | Lower        |
| ... | ChGS | U...          | UoPP    | "Aha,                       | nöŋüöte      | egeliŋ.                 | PART | np:O  | Lower        |
| ... | ChPK | 19...         | Ch...   | Oŋoruŋ kömüs                | ilimne       | , kömüs ti:ta, kömüs... | PART | np:O  | Upper (?)    |

"Tinna:k **goronuokta** egeliem, kü:ten olör."

|              |       |
|--------------|-------|
| ge           | PART  |
| SyF          | np:O  |
| 3 Dialect[C] | Upper |

Figure 4: Concordance search in EXAKT

The Tsakonian Corpus Platform, in turn, has the advantage that it is web-based and does not require the whole corpus to be downloaded and stored locally. Additionally, it directly links the given token with its sound. By placing the cursor over the token, further information and annotations are given, if available in the respective transcript. In Figure 5 below, the same test-search for partitive singular is shown using the Tsakonian Corpus Platform.

Finally, it should be mentioned here that native speakers of Dolgan were involved in the work as much as possible, as it is the case for other languages, too. Here, it is especially noteworthy that Nina Kudryakova (the person responsible for Dolgan culture and folklore in TDNT), together with her relatives, transcribed and translated large parts of the TDNT material into Russian very reliably and quickly, using the intuitive and user-friendly software SayMore.<sup>14</sup> Without this collaboration, the amount of material included in the corpus would probably have been smaller. Additionally, Chris Lasse Däbritz and Eugénie Stapert (as a research fellow) conducted four weeks of fieldwork in Dudinka in summer 2017. Working up to eight hours with Dolgan informants per day, this fieldwork brought the project significantly forward, especially when it comes to clarifying uncertainties in texts and grammar; furthermore, they transcribed a great deal of material from Eugénie Stapert's collection.

<sup>14</sup> <https://software.sil.org/saymore/>, last access: 27/04/2020.

The screenshot displays the INEL Dolgan Corpus search interface. At the top, the URL is <https://inel.corpora.uni-hamburg.de/DolganCorpus/search>. The interface includes a search bar with the text "Word #1" and fields for "Word:", "Lemma:", "Grammar:", and "Gloss: PART". A "Language/tier:" dropdown is set to "all". Below the search bar, there are buttons for "Search sentences", "Search words / lemmata", and "Select subcorpus".

The search results show 21 occurrences of the word "goronuokta" found in approximately 13 documents. The results are organized into sections: "The amulet" and "History of the Settle".

**The amulet**

- "Тi:нна:k **goronuokta** enelliem. kii:ten olor."
- "Тыыннаак горо **goronuokta** тэн олоҕор."
- "Живого гороно **goronuok** n goronuok-ta ermine-PART gr: part in, warte."
- "I will bring a livi **goronuok** n goronuok-ta ermine-PART gr: part in, warte."
- "Ich bringe einen **goronuok** n goronuok-ta ermine-PART gr: part in, warte."

**History of the Settle**

- "Оһорун таһара **d'iete**".
- "Онүөрүн таһара дыиэтэ".
- "Постройте церковь".
- "Build a church."
- "Baut eine Kirche."

A detailed tooltip for the word "goronuokta" is visible, showing its morphological and semantic information: "goronuok n goronuok-ta ermine-PART gr: part trans\_en: ermine trans\_de: Hermelin trans\_ru: горноста́й SyF: np:O SeR: np:Th Foc: foc.nar IST: new-Q".

Figure 5: Concordance search using the Tsakonian Corpus Platform

#### 4 Conclusion

The publication of the INEL Dolgan Corpus fills a considerable gap in the documentation and investigation of this under-studied language. It is now possible to conduct empirically based research on Dolgan, irrespective of the object of interest and/or the theoretical approach applied. Several studies (e.g., Däbritz 2018, Däbritz 2019) have already made use of this methodological advantage. We hope that the INEL Dolgan Corpus will encourage the linguistic community to conduct similar studies and to contribute as much as possible to the investigation of the Dolgan language.

Finally, the INEL Dolgan Corpus – as well as the other INEL corpora – may hopefully show that language documentation and corpus building projects do not necessarily depend on gathering new linguistic material. In many cases, especially when it comes to the indigenous languages of the Russian Federation, there is already very valuable material that “waits” to be located and worked upon – the INEL project may be a kick-off and an inspiration for projects having comparable agendas in the field of Uralic languages, and beyond.

## References

- Arkhangelskiy, Timofej, Anne Ferger & Hanna Hedeland. 2019. Uralic multimedia corpora: ISO/TEI corpus data in the project INEL. In: *Proceedings of the Fifth Workshop on Computational Linguistics for Uralic Languages*, 115–124. Available online: <https://www.aclweb.org/anthology/W19-0310.pdf>
- Arkhipov, Alexandre. 2020. *INEL Corpora General Transcription and Annotation Guidelines*. In: *Working Papers in Corpus Linguistics and Digital Technologies: Analyses and Methodology 5*. Szeged & Hamburg: University of Szeged, Department of Finno-Ugric Studies & Universität Hamburg, Zentrum für Sprachkorpora.
- Arkhipov, Alexandre & Chris Lasse Däbritz. 2018. Hamburg Corpora for Indigenous Northern Eurasian Languages. *Tomsk Journal of Linguistics and Anthropology* 3 (21). 9–18.
- Brykina, Maria, Svetlana Orlova & Beáta Wagner-Nagy. 2020. INEL Selkup Corpus. Version 1.0. In: Wagner-Nagy, Beáta, Alexandre Arkhipov, Anne Ferger, Daniel Jettka & Timm Lehmborg (eds.). *The INEL corpora of indigenous Northern Eurasian languages*. Publication date 30/06/2020. Archived in Hamburger Zentrum für Sprachkorpora. <http://hdl.handle.net/11022/0000-0007-E1D5-A>
- Däbritz, Chris Lasse. 2018. Predicative possession in Dolgan. *Tomsk Journal of Linguistics of Anthropology* 2 (20). 29–38.
- Däbritz, Chris Lasse. 2019. First person imperative in Dolgan – Clusivity or number distinction? *Finnisch-Ugrische Mitteilungen* 43. 1–12.
- Däbritz, Chris Lasse. 2020. *User's Guide to INEL Dolgan Corpus*. Working Papers in Corpus Linguistics and Digital Technologies: Analyses and Methodology 4. Szeged & Hamburg: Department of Finno-Ugric Studies of the University of Szeged & Universität Hamburg, Zentrum für Sprachkorpora. <https://doi.org/10.14232/wpcl.2020.4>.
- Däbritz, Chris Lasse, Nina Kudryakova & Eugénie Stapert. 2019. INEL Dolgan Corpus. Version 1.0. In: Wagner-Nagy, Beáta, Alexandre Arkhipov, Anne Ferger, Daniel Jettka & Timm Lehmborg (eds.). *The INEL corpora of indigenous Northern Eurasian languages*. Publication date 31/08/2019. <http://hdl.handle.net/11022/0000-0007-CAE7-1>. Archived in Hamburger Zentrum für Sprachkorpora.
- Efremov, Prokopij E. et al. (eds.) 2000. *Fol'klor Dolgan*. Pamyatniki fol'klora narodov Sibiri i Dal'nego Vostoka 19. Novosibirsk: Izdatel'stvo Instituta Arkheologii i Etnografii Sibirskogo Otdelenija Rossijskoj Akademii Nauk.
- Gusev, Valentin, Tiina Klooster & Beáta Wagner-Nagy. 2019. INEL Kamas Corpus. Version 1.0. In: Wagner-Nagy, Beáta, Alexandre Arkhipov, Anne Ferger, Daniel Jettka & Timm Lehmborg (eds.). *The INEL corpora of indigenous Northern Eurasian languages*. Publication date 15/12/2019. <http://hdl.handle.net/11022/0000-0007-DA6E-9>. Archived in Hamburger Zentrum für Sprachkorpora.
- Ubrjatova, Elizaveta I. 1985. *Jazyk Noril'skich Dolgan*. Novosibirsk: Nauka.
- VPN 2010 = *Vserossijskaja perepis' naselenija 2010*. 4. Nacional'nyj sostav i vladenie jazykami [All-Russian census 2010. 4. National composition and command of languages]. [http://www.gks.ru/free\\_doc/new\\_site/perepis2010/croc/Documents/Vol4/pub-04-05.pdf](http://www.gks.ru/free_doc/new_site/perepis2010/croc/Documents/Vol4/pub-04-05.pdf), last access: 22/04/2020.

Wagner-Nagy, Beáta, Sándor Szeverényi & Valentin Gusev. 2018. *User's Guide to Nganasan Spoken Language Corpus*. Working Papers in Corpus Linguistics and Digital Technologies: Analyses and Methodology 1. Szeged & Hamburg: Department of Finno-Ugric Studies of the University of Szeged & Hamburger Zentrum für Sprachkorpora der Universität Hamburg. <https://doi.org/10.14232/wpcl.2018.1>

Chris Lasse Däbritz

Institute for Finno-Ugric/Uralic Studies, University of Hamburg  
chris.lasse.daebritz@uni-hamburg.de