The order of subject, object, and verb in Tatyshly Udmurt

Daria Belova
Institute of Linguistics, RAS
HSE University
dd.belova@yandex.ru

Irina Khomchenkova
Institute of Linguistics, RAS
irina.khomchenkova@yandex.ru

Abstract

We conduct a preliminary study of the order of subject (S), object (O), and verb (V) in Tatyshly Udmurt (Finno-Ugric) on the basis of approximately 900 clauses from oral folklore and non-folklore narratives (including contemporary texts and texts recorded earlier) using a gradient approach. We show that the most frequent word orders are SOV, SV, and OV. In full clauses (with both S and O), in folklore texts SOV order (≈ 70%) is followed by OSV order (≈ 15%). In contemporary non-folklore texts, however, SOV order competes with SVO order (50% vs 30%), which may be explained by the influence of Russian. We note that full clauses may differ from clauses with only S or with only O: in contemporary folklore texts VS order is much more frequent in S-only clauses (≈ 23%) than in full ones (≈ 4%), and in contemporary non-folklore texts VO order is more frequent in full clauses (≈ 35%) than in O-only ones (≈ 12%). Moreover, we show that word order can depend on the type of clause. For example, in existential clauses the order is almost always SV, while clauses with verbs of speech are often VS.

1 Introduction

The study deals with word order in Tatyshly Udmurt, which is an understudied and low-resource subdialect of Udmurt (Permic < Finno-Ugric < Uralic). We will present quantitative data on the most important elements (S, O, V) of a clause.

Udmurt is mainly spoken in the Udmurt Republic (Russia); a significant number of speakers also live compactly in the Republics of Bashkortostan, Tatarstan and Mari El, in Perm Krai, and in Sverdlovsk and Kirov Oblasts (Russia), see, for example, (Edygarova, 2022, 507). The speakers of the Tatyshly subdialect live in the Tatyshly district of the Republic of Bashkortostan.

The vast majority of Udmurt speakers are bilingual in Udmurt and Russian (Salánki, 2007). Apart from this, Tatyshly Udmurt is significantly influenced by contact Turkic idioms (dialects of Tatar and Bashkir), see e.g. (Baidoullina, 2003, 5–7).

Udmurt is often considered a non-rigid SOV language: SOV order is the most common, while other orders are possible depending on the information structure (Bulyčov, 1947; Gavrilova, 1970; Timerkhanova, 2011; Karpova, 2015, *inter alia*). However, some researchers claim that VO order can be pragmatically unmarked as well as OV order. For example, Asztalos (2021) shows that Udmurt is undergoing a typological change from OV to VO based on her elicitation data: younger speakers tend to use VO order more often than older ones, which she explains by the influence of Russian basic SVO (Bailyn, 2012, 239–244). She also reports an areal difference: Udmurt speakers from Tatarstan used VO order less often than speakers from the Udmurt Republic. The author suggests that this may be explained by the additional influence of Tatar, which is also a (non-rigid) SOV language (Kashaeva, 2012, 77–78).

Tatyshly Udmurt may serve as another curious example of a "double-influenced" SOV language variety, since it is spoken in the area where both Russian and Turkic languages are spread. To show whether word order serves as a parameter of variation between the Udmurt dialects, one should analyze it in Tatyshly Udmurt and compare it with Tatar and Bashkir data on the one hand and the data of other Udmurt varieties on the other. However, existing research on Udmurt, Tatar, and Bashkir provides only qualitative analysis with detailed descriptions and illustrations of the possible arrangement of various sentence elements, and no statistical analysis is conducted. This hinders our ability to directly compare these data.

This paper focuses on Tatyshly Udmurt with the aim of getting reliable results and opening the Udmurt data to typological comparison by using statistical corpus methods.

Proceedings of the Fifth Workshop on NLP Applications to Field Linguistics, pages 60–65

2 Materials and methods

In linguistic typology, word order is an essential variation parameter. Some authors use a type-based approach (in terms of Levshina, 2019) and classify languages as SVO, SOV, etc based on the dominant order. Such word order types are listed for many languages in databases such as WALS (Dryer and Haspelmath, 2013) or Grambank (Skirgård et al., 2023).

There is another approach, the token-based (Levshina, 2019) or gradient (Levshina et al., 2023) approach. Within this framework, the labelling of languages as SOV, SVO etc. or as having fixed, flexible, or free word order is substituted for (or supplemented by) the proportion of different orders and the degree of word order variability (Levshina et al., 2023).

The degree of variability in word order is evaluated using Shannon entropy (Levshina et al., 2023, 850). The Shannon entropy is computed as follows (Shannon, 1948):[1]

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log_2(p(x_i))$$

Shannon entropy quantifies the average level of uncertainty associated with a set of possible outcomes. Low entropy indicates high predictability. For example, the entropy is zero if there is only one outcome. The entropy is maximal if there is a uniform distribution; in this case, the outcome is the hardest to predict. Thus, the higher the entropy, the more variable the word order.

In our study, we use this gradient approach and count the proportion and degree of variability of different word orders in Tatyshly Udmurt using corpus data. This approach is chosen because Udmurt in general is considered as a SOV language, but there is clear interdialectal variability, and it is important to have quantitative data.

The corpus of Tatyshly Udmurt[2] contains approximately 69.5 thousand tokens (as of November 2025). The corpus consists of two subcollections. The first one is sound-aligned transcriptions of recordings done in 2019, 2021–2025 during fieldwork. The second one is texts from previously published literature that were collected in 1963–2003. Further in the article we will call them the *newer* and the *older* texts. Note that both subcorpora are oral and not written texts. It is also important to note that since we do not have access to source materials, we do not know whether the published texts have undergone any editorial work. Nevertheless, this data is important to assess language change.

The main genres in the corpus are folklore narratives, non-folklore narratives, and non-folklore dialogues. For our preliminary analysis, we used the folklore narratives: 13 newer texts with 3338 tokens (from 9 speakers) and 12 older texts (which were mainly recorded in 1970) with 2899 tokens. The folklore sample was then compared with non-folklore narratives: 24 newer texts with 5288 tokens from 13 speakers and 11 older texts with 1387 tokens. In total, we analyzed texts from 15 speakers (9 women, 6 men) from 23 to 79 years old (the median age is 60).

From the sample, we manually selected and annotated 928 declarative clauses (both affirmative and negative) with noun phrases (including pronouns) as S and O, see Table 1. We did not include clauses without overt S and O and several other types of clauses (e.g., existential, with verbs of speech, with non-verbal predicates, imperative and interrogative sentences), since word order proportions in such clauses may differ (see some observations in Section 4).

	Newer		**Older**	
	Folklore	Other	Folklore	Other
S and O	70	48	44	11
Only S	179	155	122	55
Only O	61	107	29	47
Total	310	310	195	113

Table 1: Clause types under consideration.

Note that folklore and non-folklore subcorpora differ in the number of clauses with both S and O, which we will call "full clauses". It may be related to the fact that there are more nominal subjects in folklore texts than in other types of narratives. It also makes it difficult to balance the sample, especially with the older texts. In this article, we examined all the older texts and all the newer folklore texts available in the corpus to maximize the sample size. The number of newer non-folklore monologues was chosen so that it roughly corresponded to the newer folklore in number of clauses.

[1] The Shannon entropy can be calculated, e.g., here: https://www.mytimecalculators.com/tools/entropy-calculator/.

[2] The corpus is available at the following link: https://udmurt.web-corpora.net/tatyshly/index_en.html

Let us examine the types of clauses attested in our sample. The first group consists of full clauses (with both S and O), as in (1).

(1) *äd'žämi vedra kut-em*
person bucket grasp-PST2
'The man took a bucket.' (older)

Clauses with only S are mainly intransitive (2). Transitive clauses without an overt O, as in (3), were also classified as S-only clauses.

(2) *äǯ'ämi lə̑kt-e*
person come-PRS.3SG
'The man comes.' (newer)

(3) *tolez' žal'a-∅*
moon pity-PRS.3SG
'The moon pities [her].' (newer)

Finally, clauses with only O are transitive clauses without an overt subject, as in (4).

(4) *tolez'-ez a$^{?}$ǯ'-i-z*
moon-ACC see-PST-3SG
'[She] saw the moon.' (newer)

3 Results

3.1 SOV vs other word orders

The most frequent word order in both subcorpora is SOV, see Tables 2–3. In folklore texts, it is attested in about 70% of all clauses. The second most common word order is OSV, which was attested in about 15% of all folklore examples. Other orders are much less common. The entropy of the newer and the older folklore subcorpora is quite similar: $H_{newer} = 1.27$ and $H_{older} = 1.24$.

	Folklore	Other
SOV	**51 (72.85%)**	**24 (50%)**
OSV	10 (14.3%)	4 (8.3%)
SVO	6 (8.6%)	**15 (31.25%)**
VSO	2 (2.85%)	1 (2.1%)
OVS	1 (1.4%)	3 (6.25%)
VOS	0	1 (2.1%)

Table 2: Word order: newer subcorpus.

In non-folklore newer texts, the entropy is higher: $H = 1.8$, which means that word order is more variable. SOV order is still dominant (50%), but the second most common order is SVO (≈ 30%). In comparison, in newer folklore texts SVO order is rather rare (≈ 9%). The more common use of SVO order in non-folklore newer texts may be due to the influence of Russian, and its rarity in newer folklore texts may be explained by the genre: folklore texts tend to conserve more archaic features, including fossilized verbal constructions.

	Folklore	Other
SOV	**31 (70.4%)**	9 (81.8%)
OSV	8 (18.2%)	0
SVO	4 (9.1%)	1 (9.1%)
VSO	1 (2.3%)	0
OVS	0	1 (9.1%)
VOS	0	0

Table 3: Word order: older subcorpus.

The number of clauses with both S and O in non-folklore older texts is rather low, so it is difficult to draw any reliable conclusions based on these data.

3.2 SV vs VS

To analyze order of S and V, apart from the examples with full clauses (with both S and O) discussed above, we also used S-only clauses. We found that the difference between full clauses and S-clauses in the newer folklore subcorpus is statistically significant according to the Fisher exact test ($p = 0.0003$), see Table 4. The entropy differs as well: $H_{S,O} = 0.26$ and $H_S = 0.78$.

	S and O	Only S
SV	67 (95.7%)	138 (77.1%)
VS	3 (4.3%)	41 (22.9%)

Table 4: The order of S and V: newer folklore texts.

However, in newer non-folklore texts, there was no such correlation ($p = 0.6317$), see Table 5. The total entropy is $H_{S,O} = 0.57$.

	S and O	Only S	Total
SV	43	133	176 (86.3%)
VS	5	23	28 (13.7%)

Table 5: The order of S and V: newer non-folklore texts.

As for the older subcorpus (see Table 6), the difference between these two types of clauses is not statistically significant ($p = 0.2911$ and $p = 1$). Word order in the non-folklore subcorpus is a little more variable, cf. the entropy scores: $H_{folklore} = 0.35$ and $H_{other} = 0.55$.

	S and O	Only S	Total
Folklore			
SV	43	111	154 (93.3%)
VS	1	10	11 (6.7%)
Other			
SV	9	45	54 (87.1%)
VS	1	7	8 (12.9%)

Table 6: The order of S and V: older texts.

To sum up, the most frequent order is SV. In the newer folklore subcorpus, VS order in S-clauses is more common than in the other subcorpora.

It seems curious that one subcorpus out of four stands out in such a manner. We suppose that in this case we observe the influence of a specific speaker. Table 7 presents the distribution of SV and VS order in S-only clauses from two individual speakers with the most data in the folklore subcorpus (other 7 speakers in our dataset have less than 20 S-only clauses each). Both speakers under consideration are women, Speaker A was born in 1951, and Speaker B was born in 1978. The speakers also differ in profession and place of residence. Speaker B uses VS clauses much more frequently than Speaker A, and this difference is statistically significant ($p = 0.008702$).

	Speaker A	Speaker B
SV	41 (87.2%)	41 (65.1%)
VS	6 (12.8%)	22 (34.9%)

Table 7: Word order in S-only clauses in newer folklore texts of two speakers

Thus, the explanation through individual preferences seems plausible, but the existing data are insufficient to draw any strong conclusions based on sociolinguistic factors.

Apart from this, different clause types are worthy of separate consideration when analyzing SV/VS order. As an example, we will discuss clauses with existential verbs (Subsection 4.1) and with verbs of speech (Subsection 4.2).

3.3 OV vs VO

In addition to full clauses, we analyzed examples without an overt subject. We found that the difference between full clauses and O-clauses is statistically significant in the newer non-folklore subcorpus ($p = 0.0016$), but not in the newer folklore subcorpus ($p = 0.1017$), see Tables 8–9. The more frequent occurrence of VO order in newer non-folklore texts in full clauses with both S and O can also probably be explained by the Russian influence (cf. Table 2: we have shown that SVO order is frequent in newer non-folklore texts).[3]

	S and O	Only O
OV	31 (64.6%)	94 (87.9%)
VO	17 (35.4%)	13 (12.1%)

Table 8: The order of O and V: newer non-folklore texts.

	S and O	Only O	Total
OV	62	47	109 (83.2%)
VO	8	14	22 (16.8%)

Table 9: The order of O and V: newer folklore texts.

The difference between these two clause types is absent in both folklore and non-folklore texts from the older subcorpus, see Table 10.

	S and O	Only O	Total
Folklore			
OV	39	26	65 (89%)
VO	5	3	8 (11%)
Other			
OV	10	47	57 (98.3%)
VO	1	0	1 (1.7%)

Table 10: The order of O and V: older texts.

Overall, OV order is the most preferred in all of the corpora; however, it is the most common in older non-folklore texts. The entropy scores are as follows: $H_{newerfolklore} = 0.65$, $H_{olderfolklore} = 0.5$, $H_{olderother} = 0.13$.

4 Clause types

4.1 Existential clauses

Asztalos (2021) in her elicitation tasks analyzed SV/VS order in existential sentences separately. (In general Asztalos (2021) considered only existential and predicative possessive sentences when discussing SV vs VS order.) She showed that the speakers of the older generation demonstrated unanimity in their choice of the SV order, while the younger participants showed greater variability in their preferences: the majority of respondents

[3] Unfortunately, we are not able to check whether there is any individual influence, because we have even fewer data on each speaker than in SV/VS comparison (see Subsection 3.2).

from the Udmurt Republic and a third of respondents from the Republic of Tatarstan judged VS order at least as good as the SV order.

We annotated 29 existential clauses, such as (5), with the markers *van'* 'EXST', *jevâl* 'NEG' *val* 'be.PST', *vâlem* 'be.PST2', and the verb *luəńə* 'be'.

(5) *mânam tod-em-e van'*
1SG.GEN know-NMLZ-POSS.1SG EXST
'I have knowledge.' (newer)

The difference between subcorpora has not been attested, see Table 11. As we have shown in Subsection 3.2, in newer folklore texts S-only clauses have VS order more often than in clauses with S and O. Thus the distribution of SV/VS order in existential clauses and S-only clauses differs. However, this difference is not statistically significant.

	Newer	Older	Total
Folklore			
SV	17	9	26 (89.7%)
VS	3	0	3 (10.3%)
Other			
SV	28	19	47 (95.9%)
VS	1	1	2 (4.1%)

Table 11: The order of S and V: existential sentences.

4.2 Verbs of speech

In some cases, on the contrary, VS order is more widespread. This is typical for verbs of speech, see Table 12. We annotated these clauses in the folklore subcorpus, since they were almost absent in the non-folklore texts.

	Newer	Older	Total
SV	8	12	20 (27.4%)
VS	22	31	53 (72.6%)

Table 12: The order of S and V: verbs of speech.

VS order is attested when the reported speech precedes the verb (6) while SV order is used when it follows the verb (7). (This is also noted in (Vakhrushev, 1974, 130).)

(6) *d'źeź gəne iź-i-∅ no, šü-e*
good only sleep-PST-1SG ADD say-PRS.3SG
äd'źämi
person
'I slept pretty well, the man says.' (older)

(7) *pijosmurt šü-em: danag ȯj uža*
man say-PST2 plenty NEG.PST.1SG work
'The man said: "I haven't done much work."' (older)

Thus, if one uses automatically annotated texts to study the order of S and V and does not remove speech predicates from consideration, the overall proportion of word orders may change significantly.

5 Conclusion

We presented a preliminary study of the order of S, O and V in the corpus of Tatyshly Udmurt. We analyzed subcorpora that differ in genre (folklore, non-folklore) and in time period (2019–2025, 1963–2003). The most frequent orders are SOV, SV, and OV. In full sentences (with both S and O), SOV word order is the most common in all subcorpora. In the newer non-folklore subcorpus, the second most frequent word order is SVO. This fact distinguishes this subcorpus from both newer folklore and older non-folklore texts. Hence, it may be taken as a sign of Russian influence on contemporary Tatyshly Udmurt speech. Our findings show that folklore texts should be treated separately from non-folklore ones. An approach that combines these two text classes as a single sample (such as in (Karpova, 2015)) may be too simplistic. Moreover, word order proportions depend on argument structure and clause type. They may be different in full, S-only and O-only clauses. Clauses with verbs of speech demonstrate "deviating" word order distributions. Unlike other S-only clauses in all four subcorpora, in clauses with verbs of speech VS order is the most common.

Limitations

Although comparing contemporary speech with texts from 30 to 60 years ago gives us important insights into contact-induced language change, consideration of sociolinguistic parameters such as age and language skills would be beneficial for future research. These limitations are closely tied to the Tatyshly subdialect being a low resource language variety: the volume of the corpus, especially the older subcorpus, and different proportions of texts from individual speakers in different subcorpora impede our ability to properly assess whether and how sociolinguistic parameters influence our results. The use of other methods such as elicitation is needed to verify corpus findings, as well.

Abbreviations

1, 3 — 1st, 3rd person; ACC — accusative; ADD — additive particle; EXST — existential marker; GEN — genitive; NMLZ — nominalization; NEG — negation; POSS — possessive suffix; PRS — present tense; PST, PST2 — past tense; SG — singular.

Acknowledgments

This research is supported by Russian Science Foundation, RSF project No. 24-18-00199 "Clause structure and positional phenomena in SOV languages" carried out at the Institute of Linguistics, Russian Academy of Sciences.

References

Erika Asztalos. 2021. From head-final towards head-initial grammar. generational and areal differences concerning word order usage and judgement among Udmurt speakers. In *Language contact in the territory of the former Soviet Union*, pages 143–182, Amsterdam/Philadelphia. John Benjamins Publishing Company.

Anna Baidoullina. 2003. Tatyshlinskii govor udmurtskogo yazyka: fonetika i morfologiya [Tatyshly subdialect of Udmurt: phonetics and morphology]. Master's thesis, University of Tartu.

John Bailyn. 2012. *The syntax of Russian*. Cambridge University Press, Cambridge.

Mikhail Nikandrovich Bulyčov. 1947. *Porjadok slov v udmurtskom prostom predlozhenii [Word order in Udmurt simple clauses]*. Udmurtgosizdat, Izhevsk.

Matthew Dryer and Martin Haspelmath. 2013. Wals online (v2020.4). Dataset.

Svetlana Edygarova. 2022. Udmurt. In Marianne Bakró-Nagy, Johanna Laakso, and Elena Skribnik, editors, *The Oxford Guide to the Uralic Languages*, pages 507–522. Oxford University Press, Oxford.

Tatyana Gennadievna Gavrilova. 1970. Porjadok slov v udmurtskom prostom povestvovatel'nom predlozhenii [Word order in Udmurt simple declarative sentences]. In *Zapiski. Issue 21: Philology*, pages 107–118, Izhevsk. Udmurtskij NII istorii, ekonomiki, literatury i jazyka pri Sovete Ministrov Udmurtskoj ASSR.

Lyudmila Leonidovna Karpova. 2015. Osobennosti poryadka slov v udmurtskoi narodno-razgovornoi rechi (na materiale severnykh dialektov) [Peculiarities of word order in the Udmurt local colloquial speech (on the material of northern dialects)]. *Bulletin of Udmurt University. History and Philology Series*, 25(1):85–91.

Goljihan Kashaeva. 2012. The Tatar IP-field. In *Generative Grammar in Geneva*, volume 8, pages 77–94, Geneva. University of Geneva.

Natalia Levshina. 2019. Token-based typology and word order entropy. *Linguistic Typology*, 23(3):533–572.

Natalia Levshina, Savithry Namboodiripad, Marc Allassonnière-Tang, Mathew Kramer, Luigi Talamo, Annemarie Verkerk, Sasha Wilmoth, Gabriela Garrido Rodriguez, Timothy Michael Gupton, Evan Kidd, Zoey Liu, Chiara Naccarato, Rachel Nordlinger, Anastasia Panova, and Natalia Stoynova. 2023. Why we need a gradient approach to word order. *Linguistics*, 61(4):825–883.

Zsuzsanna Salánki. 2007. *The present-day situation of the Udmurt language*. Ph.D. thesis, Eötvös Loránd University.

Claude Shannon. 1948. A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423.

Hedvig Skirgård, Hannah J. Haynie, Harald Hammarström, Damián E. Blasi, Jeremy Collins, Jay Latarche, Jakob Lesage, Tobias Weber, Alena Witzlack-Makarevich, Michael Dunn, Ger Reesink, Ruth Singer, Claire Bowern, Patience Epps, Jane Hill, Outi Vesakoski, Noor Karolin Abbas, Sunny Ananth, Daniel Auer, and 100 others. 2023. Grambank v1.0. Dataset.

Nadezhda Nikolaevna Timerkhanova. 2011. Osobennost' porjadka slov v prozaičeskix proizvedenijax G. E. Vereshchagina i v sovremennom udmurtskom jazyke [Word order in the prosaic works of G. E. vereshchagin and in contemporary Udmurt]. In *Tipologicheskie aspekty mnogojazychija v sovremennom obrazovatel'nom prostranstve [Typological aspects of multilingualism in the modern educational space]*, pages 180–185, Izhevsk. Udmurtskij Universitet.

Vasily Maksimovich Vakhrushev. 1974. *Grammatika sovremennogo udmurtskogo yazyka: sintaksis slozhnogo predlozheniya [The grammar of contemporary Udmurt: compound sentence syntax]*. Udmurtiya, Izhevsk.