Language documentation in repeatedobservations design: directionality of syntactic projections in Urum

Stavros Skopeteas

Proceedings of Conference on

Language Documentation & Linguistic Theory 3

Edited by Peter K. Austin, Oliver Bond, Lutz Marten & David Nathan

19-20 November 2011 School of Oriental and African Studies, University of London

Hans Rausing Endangered Languages Project
Department of Linguistics
School of Oriental and African Studies
Thornhaugh Street, Russell Square
London WC1H 0XG
United Kingdom

Department of Linguistics:
Tel: +44-20-7898-4640
Fax: +44-20-7898-4679
linguistics@soas.ac.uk
http://www.soas.ac.uk/academics/departments/linguistics

Hans Rausing Endangered Languages Project:
Tel: +44-20-7898-4640
Fax. +44-20-7898-4349
elap@soas.ac.uk
http://www.hrelp.org

ISBN: 978-0-7286-0398-1

This publication can be cited as:

Stavros Skopeteas. 2011. Language documentation in repeatedobservations design: directionality of syntactic projections in Urum. In Peter K. Austin, Oliver Bond, Lutz Marten & David Nathan (eds) *Proceedings of Conference on Language Documentation and Linguistic Theory 3*, 257-266. London: SOAS.

or:

Stavros Skopeteas. 2011. Language documentation in repeatedobservations design: directionality of syntactic projections in Urum. In Peter K. Austin, Oliver Bond, Lutz Marten & David Nathan (eds) *Proceedings of Conference on Language Documentation and Linguistic Theory 3*. London: SOAS. www.hrelp.org/eprints/ldlt3_27.pdf

Language documentation in repeated-observations design: directionality of syntactic projections in Urum

STAVROS SKOPETEAS
University of Bielefeld

1. PRELIMINARIES

Urum is a poorly documented and highly endangered Turkic language spoken by ethnic Greek speakers in the Small Caucasus, Georgia (see Podolsky 1986). This population was originally located in Kars and spoke an Anatolian dialect of Turkish. There has been intensive language contact with Russian, Armenian and Georgian in the last 200 years (since the speakers moved to the Caucasian territory) with the result that the currently spoken language shows substantial influence from these languages, both in lexicon and in grammar. The most important source of influence is Russian (all speakers are bilingual) and Russian loanwords account for 23.7% of the lexicon based on the *World Loanword Database* word list (Haspelmath & Tadmor 2009).

This paper deals with the linearization of syntactic projections in Urum. The empirical basis is a text collection compiled from a repeated-observations documentation design which is presented in Section 2. The empirical findings are summarized in Section 3. Section 4 contains a summary of the findings and the conclusions of the article.

2. REPEATED-OBSERVATIONS IN LANGUAGE DOCUMENTATION

The data collection discussed here was created by the *Urum documentation project* (University of Athens, University of Bielefeld, University of Bremen, and University of Potsdam, see http://urum.lili.uni-bielefeld.de/) and is based on a repeated-observations design. The background assumption of this data collection is that linguistic phenomena vary in two dimensions: (a) across individuals and (b) across linguistic instantiations (see Clark 1973). Hence, the grammar has to be examined on these two dimensions of variation. In order to meet this requirement in a dataset that is compiled prior to the formulation of grammatical statements, we need data collection that is balanced with respect to these sources of variation. The data collection examined here is part of the text collection of the *Urum documentation project* and has exactly these properties: sixteen Urum native speakers produced four narratives following exactly the same instructions. The outcome is a collection of 4*16 = 64 short narrative texts (totalling 9,031 words). The four narratives are: (a) the ancestor story (1,878 words), (b) a traditional activity (1,615 words), (c) a path description (1,427 words), and (d) an account of modern life (4,111 words).

3. RESEARCH QUESTION

Our concern is how syntactic projections are linearized in Urum. We examine two types of embedded constituent: (a) embedded noun phrases and (b) objects of transitive verbs. The question is whether the embedded constituent in these constructions appears to the left or the right side of the head. The challenge is to draw conclusions exclusively from observational data by interpreting the frequencies of these linearization alternatives and their distribution in the speaker sample.

The background is that the languages at issue have different properties in the relevant constructions. Turkish is consistently head-final, (i.e., embedded NP precedes head NP, object precedes verb, embedded V precedes V-head), while Russian is predominantly head-initial (i.e., embedded NP follows head NP, object follows verb, embedded V follows V-head). The observations in the Urum lexicon indicate that the basic substrate of the language is Anatolian Turkish, while Russian has had a substantial impact on the subset of the lexical inventory that is more likely to be borrowed (following the borrowability indices established by Haspelmath & Tadmor 2009). Hence, the prediction for the syntactic facts is that head-finality will be the core pattern, while head-initial structures are more likely to occur in constructions that are more likely to be influenced in a contact situation. Moreover, since the materials come from sixteen speakers who are exposed to Russian to different degrees, we have to allow for the possibility of idiolectal variation and its correlation with the degree of exposition of the individuals to the contact language.

4. EMPIRICAL FINDINGS

4.1. Embedded noun phrases

Here we examine noun phrases that are embedded within higher noun phrases. In Turkish, there are two types of nominal heads for these constructions:

(a) possessed nouns, in constructions of the type ‘the door of the house’ and
(b) secondary postpositions, in constructions of the type ‘in the inner side of the house’.

In both constructions, the embedded noun phrase bears the genitive case (optional with lexical nouns), is cross-referenced by a personal agreement suffix on the head, and is not obligatory. The crucial property for our purposes is that the embedded noun phrase in Turkish canonically occurs at the left side of the head, e.g., *kitab-ın iç-in-e* (book-GEN inside-POSS.3.SG-DAT) ‘to the interior of the book’ (see Lewis 2000:239, Kornfilt 1997:101). It is possible to realize the embedded noun phrase in an adjunct position, which results in ‘stranding’ the nominal head in its basic position (see Kornfilt 1997:101), however this is contextually restricted. In Russian, the embedded noun phrase is marked with a genitive case and canonically occurs at the right side of the head, e.g., *dom brata* (house brother:GEN) ‘the house of the brother’ (see Wade 2011:106).

The Urum data in our corpus are very close to the Turkish pattern. The head noun bears person agreement with the embedded noun phrase, as shown in examples (1a)

and 1(b). The predominant linearization pattern (177 tokens, 94.1%) is head-final, as illustrated in (1a). However, we observe some instances of head-initial noun phrases (11 tokens, 5.9%), as illustrated in (1b). [1]

(1) (a) *bän-ım* *baba-m-ın* *deduşka-si*
1.SG-GEN father-POSS.1.SG-GEN grandfather-POSS.3
'… the grandfather of my father'
(*Ancestor Story*, speaker 25)

(b) *ğat-ier-lär* *maya-sın-i* *peinır-ın*
add-PROG-3.PL whey-POSS.3-ACC cheese-GEN
'they add cheese's whey'
(*Cheese Manufacturing*, speaker 24)

The head-initial configuration in (1b) is not ungrammatical in spoken Turkish, though it is contextually restricted. A first hypothesis is that the frequency of this configuration in our data relates to the exposure of the speakers to Russian, i.e., to a language with head-initial noun phrases. This hypothesis would be empirically supported if a subset of the speakers in our sample had a preference for head-initial noun phrases. Such a distribution in our sample is presented in Table 1. This Table shows that head-final noun phrases are the dominant pattern for all speakers, hence idiolectal variation does not have a significant impact.

[1] The abbreviations are: 1 = first-person, 3 = third-person, ABL = ablative, ACC = accusative, DAT = dative, GEN = genitive, NEG = negation, PL = plural, POSS = possessor, PROG = progressive, PST = past, SG = singular.

Table 1
Linearization of Noun Phrases (per speaker)

speaker	*h-final*	*h-initial*	*total*	*% h-final*
21	10	1	11	90 .9
22	10	1	11	90 .9
23	10	2	12	83 .3
24	6	1	7	85 .7
25	13	1	14	92 .9
26	4	0	4	100 .0
27	19	0	19	100 .0
28	16	0	16	100 .0
29	4	0	4	100 .0
30	9	2	11	81 .8
31	18	0	18	100 .0
32	14	0	14	100 .0
33	3	0	3	100 .0
34	10	0	10	100 .0
35	19	3	22	86 .4
36	12	0	12	100 .0
Total	177	11	188	94.1

A further possibility is that linearization depends on the origin of the head constituent. Our data contains a substantial number of Russian words (14.9%), either as integrated loanwords or as instances of code-switching. It is known that the linearization of syntactic projections is determined by the language of the head constituent (Mahootian & Santorini 1996), hence our hypothesis is that the proportion of head-initial noun phrases increases if the head is of Russian origin. Example (2a) illustrates exactly this for a complex noun phrase. The first genitive *biz-ım* is embedded in a noun phrase with a head noun of Turkish origin and is realized at the left of the head. The second genitive is embedded under a Russian head (*polojenie*) and is projected to the right of the head noun.

(2) ... *polojenie* *biz-ım* *urum-lar-ın*
situation 1.SG-GEN Urum-PL-GEN
'... the situation of our Urum people'
(*Modern Life*, speaker 30)

However, this pattern is not obligatory, as illustrated in example (3) where the Russian head noun is morphologically integrated, i.e., it bears a person agreement suffix cross-referencing the embedded noun phrase. The order of the subconstituents of the noun phrase follows the head-final pattern.

(3) ... *uşağ-ın* *denrajdenya-si*
child-GEN birthday-POSS.3
'... the birthday of the child'
(*Modern Life*, speaker 25)

Table 2 summarizes our observations according to the lexical origin of the head and the embedded noun. The low frequency of constructions with a Russian noun does not allow for reliable conclusions. However, the tokens recorded suggest that the head-initial noun phrases are most likely when both the head noun as well as the embedded noun are of Russian origin. This observation is weak evidence for the assumption that the lexical origin of the head determines the linearization of the syntactic projection. The crucial issue is that if both nouns are of Russian origin, we cannot distinguish between a construction affected by the substrate language and an instance of code switching.

Table 2
Linearization of Noun Phrases and lexical origin of the NP-subconstituents

head NP	*embedded NP*	*h-final*	*h-initial*	*total*	*% h-final*
Turkish	Turkish	170	7	177	96 .1
	Russian	4	0	4	100 .0
Russian	Turkish	3	1	4	75 .0
	Russian	0	3	3	0 .0
Total		177	11	188	94 .1

4.2. Objects of transitive verbs

The order within verbal projections is clearly different in Turkish and Russian. Turkish is unequivocally an OV language (see Lewis 2000:239, Kornfilt 1997:9, Göksel & Kerslake 2005:337). Under restricted contextual conditions, the object may follow the verb constituent, in particular if it contains non-accented background information in spoken or informal written Turkish (see Kornfilt 1997:206f., İşsever 2003:1041, Kiliçaslan 2004:718, Göksel & Kerslake 2005:345). The word order in Russian is very flexible, however the most frequent order in discourse is SVO (46% in a corpus study reported in Timberlake 2004:451). The object constituent in SVO sentences is typically part of the new information, however this is not obligatory, i.e., SVO order is not restricted to particular contexts. SOV order is the second most frequent order in Russian (30% according to Timberlake 2004:451) and is particularly frequent when the object is pronominal or in cases of verb-focus (see Timberlake 2004:451, Wade 2011:525f.).

In the Urum text collection we observe that the order within the verb phrase displays substantial variation. Both OV and VO occur under similar discourse conditions, as exemplified in (4). The examples in (4a, b) were produced at the very beginning of a narrative describing a traditional activity, namely the way Urum people manufacture cheese. The referent 'cow' (the object constituent) is in both

cases new information. However, speaker 26 realizes this configuration in a head-final VP, while speaker 33 uses a head initial VP under exactly the same discourse conditions.

(4) (a) ... *inäg-i* *sağ-iyer-lär*
cow-ACC milk-PROG-3.PL
'...they are milking the cow'
(*Cheese manufacturing*, speaker 26)

(b) ... *sağ-iyer-ıx* *inäg-i*
milk-PROG-1.PL cow-ACC
'...we are milking the cow'
(*Cheese manufacturing*, speaker 33)

The examples in (4) motivate the hypothesis that the OV~VO alternation in Urum depends on differences between speakers, probably due to their different degrees of exposure to the head-initial language at issue, Russian. Table 3 shows the frequencies of OV and VO orders per speaker.

Table 3
Linearization of Verb Phrases (per speaker)

speaker	*head-final*	*head-initial*	*total*	*% h-final*
21	17	5	22	77 .3
22	14	6	20	70 .0
23	17	6	23	73 .9
24	6	4	10	60 .0
25	24	4	28	85 .7
26	14	4	18	77 .8
27	28	12	40	70 .0
28	16	5	21	76 .2
29	11	3	14	78 .6
30	17	4	21	81 .0
31	15	9	24	62 .5
32	18	5	23	78 .3
33	4	9	13	30 .8
34	14	4	18	77 .8
35	18	8	26	69 .2
36	9	4	13	69 .2
Total	242	92	334	72 .5

The data in Table 3 show that the frequencies of head-final verbal projections per speaker form a unimodal distribution, as in Figure 1 (excluding speaker 33). In other

words, we are dealing with a homogeneous sample: the fact that the majority of speaker-frequencies are distributed around a central value (average 72.5%) indicates that the variation per speaker may be due to chance.

Figure 1
Number of speakers and proportions of head-final VPs

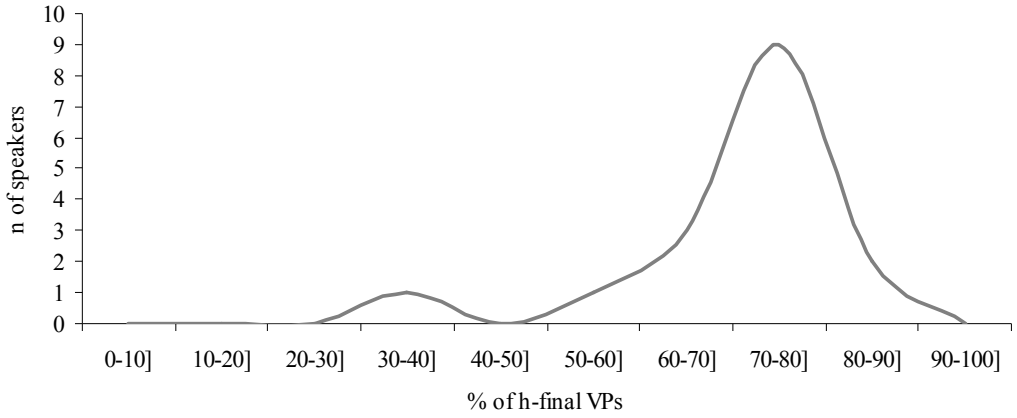

Figure 1 indicates that chance is a *possible* source of the variation in our data. However, this does not exclude the possibility that speaker-specific proportions depend on a genuine preference for a particular directionality of syntactic projections. If this hypothesis is true, then the proportion of head-final VPs (per speaker) should correlate with the corresponding proportion of head-final NPs. This can be examined by a linear regression using the proportions of head-final NPs as predictor and the proportions of head-final VPs as dependent variables. The value of R^2 is particularly low (0.02), which means that the amount of variance explained by our model is only 2%. An analysis of variance on the residuals reveals a low F-value ($F_{1,15} = 0.288$) that is associated with a non-significant probability level. i.e., a regression model on the proportions of head-final VPs does not account for the variation in our data.

A further plausible hypothesis is that the proportion of head-finality in our data correlates with the degree of exposure of the speakers to Russian. A predictor drawn from the primary data is the set of proportions of Russian words per speaker. The average of Russian words in all texts is 14.9%, while the proportion per speaker ranges from 2.8% (speaker 32) to 58.1% (speaker 29). However, a linear regression with the proportions of Russian words as predictor and the proportion of head-initial VPs as dependent variable again reveals that a regression model does not account for our data ($R^2 = 0.089$; $F_{1,15} = 1.37$, $p <$ *). In conclusion, the statistical data indicates that we do not have evidence that the variation in our sample reflects a dynamic process, i.e., the transition between different states of the grammar seen in the idiolects of several individuals.

Thus, we do not have evidence that the variation per speaker reflects a process of change in the directionality of syntactic projections. If we now look at the lexical origin of verb and object nouns in our corpus, we observe that there is an asymmetry in the data, such that head-final VPs are less frequent if the head of the VP is of

Russian origin, as shown in Table 4. However, the frequency of VPs with a Russian verb is generally low in our data, hence the high frequency of VO in our corpus is not explained by the occurrence of VPs with a Russian head.

Table 4
Linearization of Verb Phrases and lexical origin of the VP-subconstituents

verb	*object*	*head-final*	*head-initial*	*total*	*% h-final*
Turkish	Turkish	204	71	275	74 .2
	Russian	30	14	44	68 .2
Russian	Turkish	1	2	3	33 .3
	Russian	7	5	12	58 .3
Total		242	92	334	72 .5

Assuming a homogeneous speaker sample, we will now examine what determines the choice of OV and VO in language production. Both orders are possible in Turkish and Russian, however under different discourse conditions. In Turkish, postverbal material is only used for background information, while in Russian postverbal objects are not contextually restricted, i.e., they can be either new or given. Furthermore, SOV is particularly frequent in Russian with pronoun objects.

It is clear that the Urum data do not show a one-to-one correspondence with particular information structure, since we observe different orders under identical conditions. Example (5) illustrates the possibility of free variation: both clauses in this example have a Turkish verb and a loan noun as object. Both noun phrases denote referents that are not previously mentioned in the discourse. However, the first clause displays VO order and the second OV order. Hence, the order of words in the utterance cannot be fully predicted by information structure. Other factors, e.g., rhythm, may play a role.

(5) *al-dı-lar* *danio*, *kvartira-lar* *al-dı-lar*
take-PST-3.PL credit flat-PL take-PST-3.PL
'They took credits, bought flats.'
(*Modern Life*, speaker 28)

In order to examine whether the factors known from Russian and Turkish to have an influence on word order also apply to Urum, we coded object constituents for two factors:

(a) realization: lexical or pronominal
(b) discourse status: given or new, where a noun phrase were coded as 'given', if the referent was already introduced in the discourse.

The results are given in Table 5. The data in this table shows that the frequency of preverbal realization of the object constituent increases for pronominal noun phrases ($\chi^2 = 5.9, p < 0.01$). The impact of the discourse status of the object may be observed with lexically realized noun phrases. The high frequency of postverbal new objects

does not fit with statements in the literature about the occurrence of postverbal objects in spoken Turkish. The data shows that postverbal placement of the object is not only used for background information that is postponed in an extra-clausal domain. Finally, we may test the hypothesis that given information more frequently occurs in the postverbal domain: descriptively, given objects are more frequently final (73.7%) than preverbal (68.2%). However, the observed difference is below chance level ($\chi^2 = 0.9$, $p < *$).[2]

Table 5
Linearization of Verb Phrases, lexical/pronominal realization, and discourse status

realization	*discourse-status*	*head-final*	*head-initial*	*total*	*% h-final*
lexical	new	131	61	192	68 .2
	given	70	25	95	73 .7
pronominal	new	14	1	15	93 .3
	given	27	5	32	84 .4
Total		242	92	334	72 .5

5. CONCLUSIONS

We have presented an empirical account of the linearization of syntactic projections in a documentation corpus based on a repeated-observation design, which contains balanced speaker-specific subcorpora.

The linearization of noun phrases reveals a strong tendency for head-final configurations. There are a substantial number of head-initial noun phrases in our corpus (11 tokens, 5.9%). The linearization of a subset of head-initial noun phrases (4 tokens) can be explained by the lexical origin of the head (Russian). The crucial point is that our speaker sample shows homogeneous behaviour, i.e., it displays a consistent preference for head-final noun phrases.

The linearization of verb phrases involves much more variation. Though there is a general preference for head-final verb projections (72.5%), object constituents are very frequently realized in the postverbal position. An inspection of the discourse status of the referents of object noun phrases shows that postverbal objects are not restricted to background information. This finding is relevant because it shows that postverbal position is not an extra-clausal configuration that hosts postponed material, as we would expect for a V-final language of the Turkish type. Examination

[2] A further factor that is known to determine the order of the verb phrase in Turkish is accusative marking of the object constituent: bare objects are expected to occur preverbally, while accusative-marked objects are free to occur in any position in the clause. Accusative-marking does not have a significant impact in the Urum data. Restricting the dataset to the lexically realized objects of Turkish origin, we observe that the OV order occurs in the 75.6% of tokens with bare objects and in the 67.1% of tokens with accusative-marked objects.

of the distribution of speaker-specific proportions and their lack of correlation with exposure to Russian shows again that the behaviour of the speakers is generally homogeneous and the differences may be accounted by chance variation.

In conclusion, we observe a homogeneous speaker sample that does not show the properties of diachronic transition between different states of the grammar. The linearization of noun phrases and verb phrases in Urum shows a general preference for head-final configurations. The order of the subconstituents of the verb phrase shows much more flexibility than the order within the noun phrase, which opens the possibility of selecting the optimal order under the influence of several factors (e.g., information structure), however without displaying a strict correspondence to any particular discourse configuration.

REFERENCES

Clark, Herbert. H. 1973. The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior* 12:335–359.

Göksel, Aslı & Celia Kerslake. 2005. *Turkish: A comprehensive grammar.* London: Routledge.

Haspelmath, Martin & Uri Tadmor (eds.). 2009. *Loanwords in the world's languages: A comparative handbook.* Berlin: Mouton De Gruyter.

İşsever, Selçuk. 2003. Information structure in Turkish: the word order-prosody interface. Lingua 113:1025–1053.

Kiliçaslan, Yılmaz. 2004. Syntax of information structure in Turkish. *Linguistics* 42(4):717–765.

Kornfilt, Jaklin. 1997. *Turkish*. London: Routledge.

Lewis, G. L. 2000. *Turkish grammar*, 2nd edn. Oxford: Oxford University Press.

Mahootian, S. & B. Santorini. 1996. Code-switching and the complement/adjunct distinction. *Linguistic Inquiry* 27(3):464–479.

Podolsky, B. 1986. Notes on the Urum language. *Mediterranean Language Review* 2:99–112.

Wade, Terence. 2011. *A comprehensive Russian grammar* (revised and updated by David Gillespie, 2nd edition). Oxford: Wiley-Blackwell.