ГРАММАТИЧЕСКОЕ ИССЛЕДОВАНИЕ НА ПОЛУРАЗМЕЧЕННОМ КОРПУСЕ ТЕКСТОВ (НА МАТЕРИАЛЕ НОМИНАЛИЗАЦИЙ В ОСЕТИНСКОМ ЯЗЫКЕ)

П. Гращенков (pavel.gra@gmail.com)
Институт Востоковедения, Москва, Россия

М. Ионов (max_ionov@mail.ru)

С. Малютина (i-am-stupid@list.ru)
МГУ, Москва, Россия

В настоящей работе мы предлагаем метод лингвистического исследования на полуразмеченном корпусе, предназначенный для изучения грамматической структуры тех языков, где создание полноценного корпуса текстов невозможно. В качестве примера применения нашего метода мы приводим исследование структуры номинализаций в осетинском языке. Грамматическое исследование было осуществлено в три основных этапа. Во-первых, было определено множество интересующих нас грамматических конструкций. Далее, была сформулирована гипотеза о вероятной грамматической структуре данных конструкций. Наконец, выдвинутая гипотеза была апробирована на корпусе текстов. Создание корпуса осуществлялось в два этапа. Во-первых, на основании выборки осетинских текстов была собрана значительная текстовая коллекция. Во-вторых, данный массив текстов был снабжен специальным средством поисковых запросов. В результате, наша исходная гипотеза подтвердилась, что позволило нам уточнить результаты проведенного ранее полевого исследования и сформулировать новые предположения о грамматическом устройстве номинализаций в осетинском языке.

Ключевые слова: номинализации, осетинский язык, полуразмеченный корпус, неполный корпус.

SEMI-TAGGED CORPORA METHOD EXEMPLIFIED WITH A STUDY OF OSSETIC NOMINALIZATION

P. Grashchenkov (pavel.gra@gmail.com)
Institute of Oriental Studies, Moscow, Russian Federation

M. Ionov (max_ionov@mail.ru)

S. Maliutina (i-am-stupid@list.ru)
Lomonosov Moscow State University, Moscow, Russia

We propose the method of Semi-Tagged Corpora (STC) for grammar research in languages that are not expected to have corpora in the nearest future. We exemplify this method with an STC study of internal structure of nominalization

in Ossetic. The research was implemented in three major steps: 1) a set of valid surface structures was established; 2) theoretical predictions were made; 3) the initial hypothesis was tested on the text corpora. The corpora were created in two steps. First we selected a signifi cant amount of texts available for Ossetic and merged them in a single text collection. Then we supplied the collection with specifi c search tools. The initial hypothesis was confi rmed that made our fi eld results more accurate and allows a further elaboration of the syntactic structure that we proposed for Ossetic nominalizations.

Key words: semi-tagged corpora, nominalization, ossetic naguage, Ossetic nominalization.

1. Introduction

Syntactic research is generally conducted via native speakers questioning. However when a speaker doesn't express clear preference for one surface structure in the set of possible structures, the questioning is not satisfactory. For instance, it was claimed in (Chelliah 2001, Brody1982, a. o.) that elicitation approach has quite limited scope and can not be applied to e. g. word order study.

Corpus-oriented researches (see Sinclair 1991 a. o.) were recently implemented on "major" languages like English (Biber et al. 1995) or Chinese (Huang 1994) and gave important output for the grammatical theory. But the enterprise of using corpora and quantitative study of minor or endangered languages seem strange at first. Indeed, languages like Ossetic seem not good candidates for corpus study. First, there are no corpora but only large text collections. Second, there are no electronic dictionaries or ready tag sets for them.

However, rich morphology of Ossetic allows to skip tagging and rely on affixation in corpus research. At the same time, Ossetic possesses a good collection of fiction and paper/magazine articles, sufficient for creation of a large text array.

We supplied our previous field study[1] of Ossetic syntax with a corpora study that favors up one of some initial hypothesis. We used a method of morphology-based search on the untagged corpora. Search results were subsequently filtered and tagged manually. We called this research strategy Semi-Tagged Corpora (STC) study. STC helped us to fill some theoretical gaps in syntactic structure analysis of Ossetic.

2. Linguistic object: Ossetic nominalization

Ossetic is an Iranian language with 0,5 mln of speakers. It has GenN, SOV word order and accusative case system. Cases are marked overtly except nominative and human direct object (unmarked). Non-human direct objects receive marking

[1] The original study of nominalization was provided during the MSU field research trips to Northen Ossetia in 2007–2010. We are very grateful to all our collegues and especially to the chiefs of the expedition, Sergei Tatevosov and Ekaterina Liutikova, for their assistance both in and outside linguistics.

phonologically identical to genitive. There are two nominalization strategies in Ossetic, *-yn* nominalization described here is more regular and productive one.

Two most prominent linguistic problems concerning nominalization in some particular language are the following. First — how many lexical and functional VP material receives nominal distribution. In particular — which arguments are involved in nominalization. Second — how DP structure influences nominalization, i. e. what are the way(s) of marking verbal argument(s), do they receive cases from a verb (accusative) or from nouns (genitive), etc. In Ossetic both these problems are relevant since Ossetic *-yn* forms are homonymous between nominalizations and infinitives. According to native speakers' judgments, both external and internal arguments are valid in the context of nominalization. Moreover, whereas the noun phrase displays strict left branching, the order of arguments in both simple predication and nominalization is quite flexible, see the Table 1. So, direct questioning of native speakers doesn't clarify which arguments (external, internal, both) are present on the argument list and what is the directionality of branching in nominalization.

The hypothesis that may help us to reveal the structure of Ossetic nominalization was reported in (Alexiadou 2004). According to Alexiadou's proposal, nominalizations, even if they allow different distribution and display distinct internal properties, are always merged[2] under the same structure. We can technically elaborate this proposal as follows: the syntactic material merged into enumeration is always the same, and what differs depending on context are phi-features[3] (see Chomsky 1999 and its developments). Differentiation of phi-features is induced by the external context where nominalization is merged. Every particular feature set forces specific internal syntactic configuration, see the Table 2.

Table 1. Constructions with nominalization attested during native speakers' questioning. External and internal arguments accepted under different orderings. Meaning: ***father's sharpening a scythe***[4]

fyd-y	sævæg	daw-yn-
father-GEN	scythe	sharp-*ING*
fyd-y	daw-yn-	sævæg
father-GEN	sharp-*ING*	scythe
daw-yn-	fyd-y	sævæg
sharp-*ING*	father-GEN	scythe
daw-yn-	sævæg	fyd-y
sharp-*ING*	scythe	father-GEN

[2] Merge is an operation that combines two items of the lexicon into a single unit with a label borrowed from one these items.

[3] Informally, phi-features are grammatical categories associated with particular nodes in syntactic structure, functional heads.

[4] In case if the DO is animate the patient will be marked with the genitive (=accusative) in both finite clause and nominalization, *fyd-y fyrt-y wyn-yn-* is a nominalization *father's seeing of his son*.

Table 2. Influence of the distribution of nominalization on its internal structure

1. Enumeration:
{Adjuncts, IntArg, Verb, D, (ExtArg, v,…)}

2. Nominalization merged as DP:
$[_{DP}\ [_{DP}\ Subj_{GEN}]\ \ldots\ XP\ \ldots\ [_{VP}\ Verb]\ \ldots\ D]$
$*[_{DP}\ [_{DP}\ Subj_{GEN}]\ \ldots\ [_{VP}\ Verb]\ \ldots\ \mathbf{XP}\ \ldots\ D]$

3. Nominalization merged as infinitive:
$[_{DP}\ \ldots\ XP\ \ldots\ [_{VP}\ Verb]\ \ldots\ YP\ \ldots\ D]$
$*[_{DP}\ [_{DP}\ \mathbf{Subj}]\ \ldots\ [_{VP}\ Verb]\ \ldots\ D]$

Then, two most prominent nominalization patterns are nominal and verbal ones. Merged under the postpositions and in noun phrases, nominalizations acquire all properties of noun phrases: they get able to assign genitive case to their subjects and should not exhibit word order permutation. Merged under modals and phase verbs they do not have their own subjects and exhibit word order dependency on the information structure as schematized in the Table 2.

Thus in case of Ossetic *-yn* nominalization we expect to observe the following distributional properties: (i) no difference in the number or marking of arguments in nominal and verbal nominalizations; (ii) differentiation in surface string ordering: strict left branching under the nominal external context and flexible ordering in verbal context. These two statements were chosen for testing by corpora method.

3. Creating corpora

Modern Ossetic has a status of a minor language (<0.5 million of speakers, the absolute majority of which reside on the North Caucasian Mountain) with a well-developed literary tradition. We collected and included into the corpora texts of modern Ossetic newspapers and writers of 20-th century with total volume of 1.3 million words. After that we supplied the text array with the search tools that allow to extract sentences including two words defined in the search query (with the regular expressions option) at some distance also specified in the query.

4. Extracting data and tagging results

Writing search queries, we relied on the rich morphology of Ossetic which made possible to select *-yn* nominalizations and distinguish between the nominal and verbal type of nominalizations.

Different case contexts of nominalizations of all verbs in the selected corpora provide about 20 thousand sentences. We chose eight of the most frequent verbs: *arazyn* ‘make’, *zuryn* ‘say’, *sæwyn* ‘go’, *hwydy kænyn* ‘think’, *maryn* ‘kill’, *særyn* ‘live’,

ahwyr kænyn 'study', *pajda kænyn* 'use'. To distinguish nominal uses from verbal ones we chose genitive forms of nominalizations as instances of the first type and contexts with the verbs 'start/begin', 'want' and 'need' as examples of the second type of representation, see examples in the Table 3. All such instances of the selected eight verbs provide about seven hundreds of contexts.

Then all these contexts were translated and tagged with respect to following properties: presence of subject, presence of direct object, directionality of branching of internal material (obliques and adjuncts considered as well).

Table 3. Corpora examples of nominal and infinitival contexts

1. Nominal:				
…iron	ævzag	ahwyr kænyn-y	raydayæn	etap…
Ossetic	language	study-*ING*	beginning	stage
the first stage of studying Ossetic				
2. Infinitival:				
…raidydta	ahwyr kænyn	matematikon	naukæ-tæ…	
he-started	study-*ING*	mathematical	science-PL	
he began studying mathematical sciences				

5. Evaluating results

The number of nominal contexts consists of 355 and verbal contexts — of 313 examples, 668 instances in total.

Concerning subjects, there were only 7 examples, all of them used in nominal contexts, see Figure 1. Paired t-test performed on the amount of subjects of each verb in nominal vs. verbal context revealed no significant difference between nominal and verbal contexts ($t(7) = 1{,}80$, $p > 0.1$).

Direct objects are met 291 times. Nominal contexts have 163 examples that represent 79% of 206 items of nominalizations of transitive verbs. Direct objects in verbal contexts are met 128 times which is 73% of 128 items, see Figure 2. Again, paired t-test performed on the amount of objects of each transitive verb in nominal vs. verbal context revealed no significant difference between nominal and verbal contexts ($t(5) = 0{,}34$, $p > 0.1$).

No nominalization with both subject and direct object has been attested.

Branching directionality is distributed as follows. Left branching is met in 98% of nominal and 65% of verbal contexts, Figure 3. Yates-corrected chi-square test revealed a significant difference between nominal and verbal context in the amount of examples with left vs. right branching ($p < 0.001$).

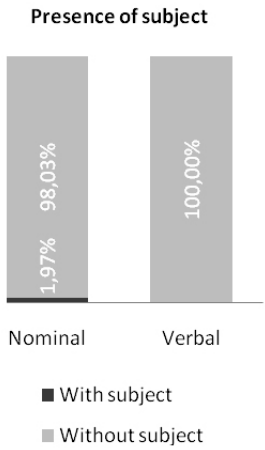

Figure 1. Subject DPs attested in nominal and infinitival contexts

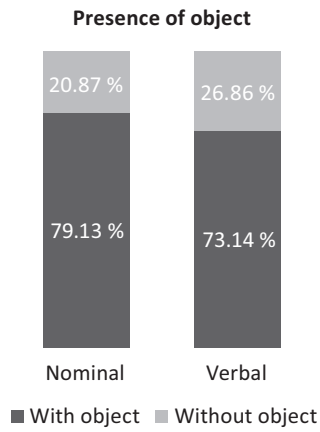

Figure 2. Direct object DPs in nominal and infinitival contexts

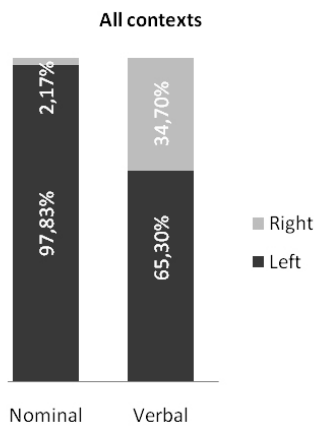

Figure 3. Word order directionality in nominal and infinitival contexts

6. Interpreting results

From the point of view of argument structure, two important observations can be done.

First, we can argue that both nominal and infinitival nominalizations lack subjects on their argument list. Attested 7 cases of subjects as well as artificial subject examples in the Table 1 should be addressed to as pragmatically introduced participants, not true arguments, cf. the traditional treatment of oblique agents. Genetive case can be assigned to such non-argumental DPs as a dummy case marker (see analysis

in Chomsky 1986 for English *of*). Verbs (both transitive and intransitive) other than those that we take for our study also exhibit less than 2% frequency of nominalized subjects. Based on such a low frequency, we argue that Ossetic nominalizations do not really have subject on their argument list.

Second, direct objects are equally frequent in nominal and infinitival nominalizations.

These two observations clearly show that the argument structure in both types of nominalizations is the same.

Concerning word order directionality, nominal contexts do not display any permutations — they are strictly left branching. At the same time more than one third of infinitival nominalizations display right branching. The explanation here is that nominalizations in nominal contexts (as well as in regular DPs) do not allow pragmatically driven scrambling. Infinitival nominalizations and simple clauses are not restricted in this option.

Thus branching directionality depends on phi-features "supplied" by external context, whereas other items that constitute nominalized structures are the same in different instances of nominalizations, see the Table 2. We can further speculate that only nominal phi-features create a phase, opaque for external syntactic processes but this statement comes beyond the scope of current research.

7. Conclusion

As we showed basing on our STG study of Ossetic, nominalizations in this language do not project external arguments. Their argument structure can include only direct objects (that may be marked genitive or nominative). Then, the internal structure of nominalization is a function of the context where it was used.

These results, that seem us quite interesting from the theoretical point of view, could hardly be achieved without quantitative corpora-based investigation of syntactic structure. And corpora creation for languages like Ossetic looks much more realistic under STC-methodology.

References

1. *Abney P. S.* 1987. The English Noun Phrase in its Sentential Aspect. Ph. D. Dissertation.
2. *Alexiadou Artemis.* 2004. Argument Structure in Nominals.
3. *Biber D., Johansson S., Leech G., Conrad S., Finegan E.* 1995. Longman Grammar of Spoken and Written English.
4. *Brody Jill.* 1982. Some Problems With the Concept of Basic Word Order. Linguistics, 22: 711–36.
5. *Chomskii, N.* 1999. Derivation by Phase. MIT Occational Papers in Linguistics. 18.

6. *Chomskii, N.* 1986. Knowledge of Language: Its Nature, Origin, and Use.
7. *Sinclair J.* 1991. Corpus Concordance and Collocation (Describing English Language).
8. *Chelliah Shobhana L.* 2001. The Role of Text Collection and Elicitation in Linguistic Fieldwork. Linguistic Fieldwork : 152–165.
9. *Huang Chu-Ren.* 1994. Corpus-based Study of Chinese: Preliminary Results. In Honor of William SY. Wang: Interdisciplinary Studies on Language and Language Change.