GRAMMATICAL RESEARCH ON SEMI-TAGGED CORPORA (ON THE DATA OF OSSETIC NOMINALIZATIONS)

Pavel Graschenkov,
Institute of Oriental Culture (Moscow), pavel.gra@gmail.com
Svetlana Malyutina,
MSU, s.malyutina@gmail.com
Maxim Ionov,
MSU, max.ionov@gmail.com

Overview

- Methodology
- Ossetic basics
- Creating corpora
- Extracting data
- Evaluating data
- Interpreting data
- Conclusion

What Is This All About?

- Syntactic researches are often made by questioning native speakers
- Sometimes speakers don't express clear preference for a specific surface structure
- Corpora-oriented studies could help
- Some languages have problems with corpora studies

Ossetic and corpora studies

- Problems:
  - No tagged corpora
  - No e-dictionaries or tag sets
- But:
  - Rich morphology

    ⇒ we can rely on affixation
  - Well-developed literature tradition

    ⇒ large text array

Ossetic and corpora studies

- Research strategy:
  - Searching untagged corpora
  - Subsequent supervised filtration
  - Manual tagging the results

Basics information about Ossetic

- Iranian language
- Mostly synthetic
- 9 grammatical cases
  - Marked by suffixes
  - No accusative case
- Morphosyntactic alignment: nominative-accusative

Basics information about Ossetic

- Unmarked case: nominative
- Direct Object case: nominative or genitive

- Nominalizations are formed by –yn– suffix

Ossetic Nominalization: Problems

- Theoretical problems:
  1. How much VP structure is involved in it
  2. How DP structure influences nominalization
- In Ossetic:
  - Both problems are topical, because –*yn*– forms are homonymous between infinitives and nominalizations

Ossetic Nominalization: Problems

1. Nominal:

…iron ævzag ahwyr kænyn-y

Ossetic language study-*ING*-*GEN*

raydayæn etap…

beginning stage

the first stage of studying Ossetic

2. Infinitival:

…raidydta ahwyr kænyn

he-started study-*ING*

matematikon naukæ-tæ…

mathematical science-PL

he began studying mathematical sciences

Ossetic Nominalization: Arguments

- According to native speakers' judgements:
  - Both external and internal arguments participate in nominalizations
  - Flexible word order in simple predication
  - Strict left branching in noun phrases
- So direct questioning doesn't clarify:
  - Arguments that are in argument list
  - Directionality of branching

Ossetic Nominalization: Ordering

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

All these orderings were attested by native speakers

Ossetic Nominalization: Basis

- Artemis Alexiadou, 2004:
  - Nominalizations are always merged under the same structure
- Syntactic material is the same, differences are in phi-features
- Differentiation of the phi-features is induced by external context
- Every feature set forces specific internal configuration

Ossetic Nominalization: V vs. N

- Two most prominent patterns are nominal and verbal one
- Nominal:
  - Merged under postpositions and in noun phrases
  - Acquire all properties of noun phrases
    - Able to assign Gen to their subject
    - Shouldn't exhibit word order permutation
- Verbal:
  - Merged under modals and phrase verbs
    - Do not have own subjects
    - Exhibit word order dependency on the information structure

Ossetic Nominalization: Hypothesis

We expect to observe the following distributional properties:

- No difference in number or marking of arguments
- Differentiation in surface string ordering:
  - Nominal contexts: strict left branching
  - Verbal contexts: flexible ordering

These two statements were chosen for testing by corpora method

Extracting data: Corpus & tools

- Corpus:
  - Consisted of 1.3 million words
  - Modern fiction and press
- Extraction:
  - Indexing text array
  - Querying:
    - Word1 + distance span + word2
    - Word1 and word2 are regular expressions

Extracting data: The process

- Initially extracted ~20 000 sentences
- Examples with 8 most frequent verbs were chosen

Verb	**Translation**
‘arazyn’	make
‘zuryn ’	say
‘sæwyn ’	go
‘hwydy kænyn ’	think
‘maryn ’	kill
‘særyn ’	live
‘ahwyr kænyn ’	study
‘pajda kænyn’	use

Extracting data: The process

- Distinguishing V from N:
  - Genitive forms as the examples of nominal contexts
  - Constructions with ‘start / begin’, ‘want’ and ‘need’ as the examples of verbal contexts
- ~700 contexts were left after filtering
- They were manually translated and tagged:

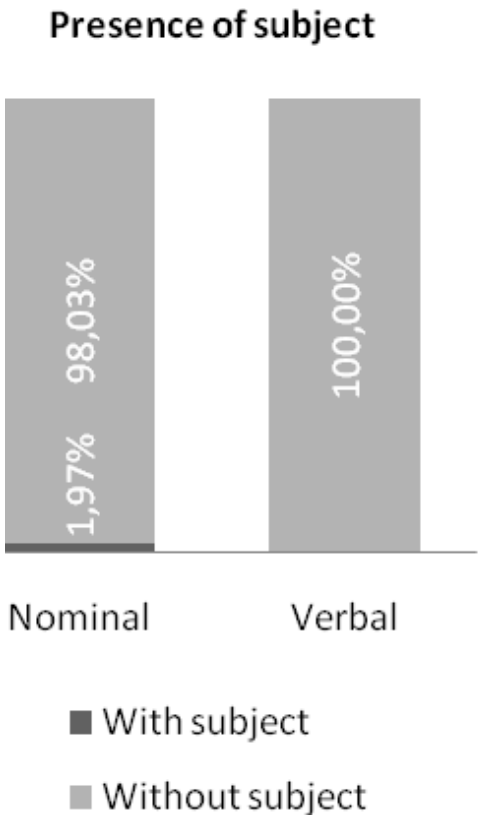
Context	Presence of subject	Presence of direct object	Directionality of branching
…	…	…	…

Evaluating results

- Total: 668 instances
  - 355 nominal contexts
  - 313 verbal contexts

Evaluating results: Subjects

- Only 7 examples
- All in nominal contexts
- ⇒ They are pragmatically introduced participants, not true arguments

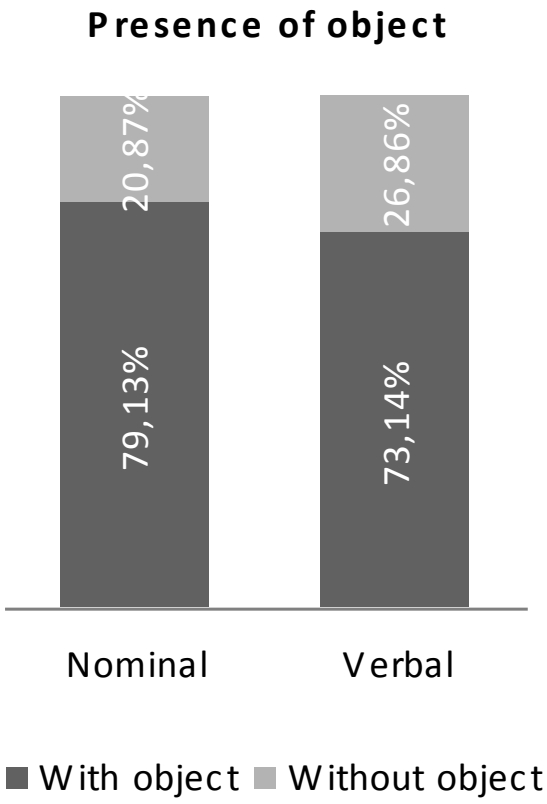


Evaluating results: Direct objects

- Total: 291 context
- Nominal: 163 = 79%
- Verbal: 128 = 73%
- Paired t-test (amount of subjects of each verb in nominal vs. verbal contexts):

  $t(5) = 0.34, p > 0.1$

  ⇒ no significant difference



Evaluating results: subject with DO

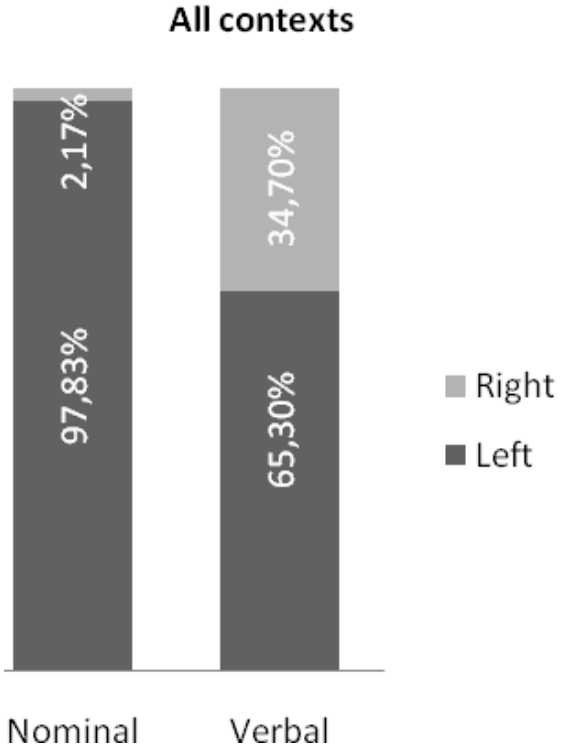
No nominalizations with both subject and direct object have been attested

Evaluating results: branching

- Left branching was met:
  - In 98% of nominal contexts
  - In 65% of verbal contexts
- Yates-corrected chi-square test (nominal and verbal context in the amount of examples with left vs. right branching):

  $p<.001$

  ⇒ significant difference



Interpreting data: argument structure

- Two observations can be done:
  1. Both types of nominalizations lack subject on argument list
  2. Direct objects are equally frequent in both types of nominalization

  ⇒ Argument structures are the same

Interpreting data: word order

- Nominal contexts are strictly left branching
- > 1 / 3 of infinitival contexts are right branching

- Explanation:
  - Nominalizations in nominal contexts do not allow pragmatically driven scrambling (like in regular DPs)
  - Infinitival nominalizations are not restricted in this option

  - Branching directionality depends on phi-features supplied by external context, internal structure is the same

Conclusion

- Ossetic nominalizations do not project external arguments
- Their argument structure can include only direct object
- The internal structure of nominalization is a function of the context where it was merged

Acknowledgements

We are very grateful to all our colleagues and especially to Anastasia Garejshina and Lidia Kirpo for their help on collecting text corpora and to the chiefs of the expedition, Sergei Tatevosov and Ekaterina Lyutikova, for their assistance both in and outside linguistics.

Thank you!