Contrasts in Case Usage under Negation in Old Church Slavonic

Joseph Rhyne

Research Workshop

jtr92@cornell.edu

November 8, 2019

Introduction: Genitive of Negation

- This project started as an exploration into the diachrony of Genitive of Negation in Slavic
  - Why would Slavic not continue the Indo-European status quo of accusative objects regardless of negation?
- Started with Old Church Slavonic. A Preliminary corpus study of *Codex Marianus* using PROIEL (Haug and Jøhndal, 2008) found an interesting result:

Objects under Negation	Count
Objects (GEN)	419
Objects (ACC)	93

- Accusative objects that still occurred under negation!
  - Is there a contrast driving this use of accusative or was there an alternative explanation, like scribal errors?

Outline

Overview

Objectives

1. Explore more OCS texts to see if the pattern from *Codex Marianus* holds throughout the language
   - Eventually extend the study to all of Old Slavic
2. If the ACC/GEN contrasts persist, develop a formal analysis

- Analyses have focused on modern languages, and simply accepted GofN as an obligatory rule in older forms of the languages
- Can the results of this study give any additional insights into the origins of GofN?

Genitive of Negation in OCS

Genitive of Negation

- At it's most basic, GofN involves a genitive-marked noun licensed under sentential negation
- Can occur in many different environments and constructions, not just limited to direct objects
  - These different constructions might have different origins diachronically, but synchronically they appear to pattern together
- A unifying feature of Slavic languages
  - Every branch had it at the oldest stages
  - Modern languages continue it on in various forms

Object Genitive of Negation

- **Affirmative sentence and ACC:**

(1) ljubl̀ěaše že isusŭ **martǫ** i **sestrǫ**
loves.IPF.3RD PTCL Jesus.NOM Martha.ACC and sister.ACC
eję i **lazarě**
her and Lazarus.ACC

'For Jesus loved Martha and her sister and Lazarus.' (John 11:5; *Codex Marianus*)

- **Negative sentence and GEN**

(2) blǫdite ne vědǫšte **kŭnigŭ** ni
be.mistaken.PRES.2ND NEG knowing books.GEN nor
sily **bžiję**
power.GEN divine.GEN

'You are mistaken, not knowing the scriptures, nor the power of God.' (Matthew 22:29; *Codex Marianus*)

Object Genitive of Negation

- GofN is not always obligatory
- Lexically-selected cases, such as dative, supersede GofN:

(3) i ne da **im'** knjaz'
and NEG give.PAST.3SG them.DAT prince.NOM.SG
M'stislav'.
Mstislav.NOM.SG
'and Prince Mistislav did not allow them.' (OES)

Subject Genitive of Negation

- Genitive-marked subjects of negated existential, locatival, and unaccusative verbs instead of the expected nominative:

(4) zane ne bě ima **města** vŭ
for NEG be.IPFV.3SG them.DAT place.GEN.SG in
obitěli.
inn.LOC.SG

'for there was no room for them in the inn' (Luke 2:7; *Codex Marianus*)

(5) **okorablě inogo** ne bě tu.
boat.GEN other.GEN NEG be.IPFV.3SG there

'There was no other boat there' (John 6:22; *Codex Marianus*)

Subordinate Clauses with GofN

- GofN can also appear in non-finite complement and adjunct clauses, with matrix clause negation

(6) a. Ne uboi sę prijęti **ženy tvoję** **Mariję**.
NEG fear.IMP.2SG REFL take.INF wife.GEN your.GEN Maria.GEN
'Do not be afraid to take your wife Mary.' (Matthew 1:20; *Codex Marianus*)

b. Nikto že **světilĭnika** vŭžegŭ pokryvaetŭ i sǫdomŭ.
no.one PRT lamp.GEN light.PP.NOM.MSG hide.PRES.3SG it vessel.INST
'No one, having lit a lamp, hides it under a vessel.' (Luke 8:16; *Codex Marianus*)

- Subject-control and object-control constructions can show genitive, as well.

Adjunct Genitive of Negation

- Genitive-marked temporal and locatival adjuncts under negation instead of the normal accusative

(7) Tako li ne vŭzmože **edinogo časa**
thus QU NEG can.PAST.2SG one.GEN hour.GEN
pobĭděti sŭ mŭnojǫ.
keep.watch.INF with me
'Thus could you not keep watch with me for one hour?'
(Matthew 26:40; *Codex Marianus*)

Partitive Origins

Proposed Origins

- How did Slavic get this construction?
- Different origins for GofN have been proposed
  - Indo-European ablative
  - **Original partitive construction** (Meillet, 1897; Pirnat, 2015; Pesetsky, 1982)
- The latter has become the dominant perspective

Partitive Origins

- Certain predicates took partitive genitives that referred to an indefinite quantity, e.g. 'a piece of cake'
- Expanded and extended to contexts beyond the limited set of predicates that took partitives
- Eventual association with negation in a Jespersen Cycle-like process
  - Partitive meaning became a marker of emphatic or pronounced negation
  - "I did not eat cake" $\rightarrow$ "I did not eat any cake whatsoever"
- Loss of emphatic negation and emergence of GEN-ACC contrast with negation

Partitivity in OCS

- OCS still had predicates that took “partitive" genitives synchronically, in both affirmative and negative contexts, e.g. *vŭkusiti* ‘taste’, *jĭmǫtŭ* ‘take, receive’, etc.
  - jako že *vŭkusi* arxitriklinŭ **vina byvūšaego** otŭvody “When the ruler of the feast tasted the wine made from water"
  - priętŭ **xlēba** “he took (some?) bread"
- Moreover, OCS and other Slavic languages(e.g Modern Russian) still had partitive genitives
- How would this contrast under negation spread when you still have access to partitives not under negation?

Partitive Origins and Definiteness

- There is a link between definite/indefinite interpretations and partitivity
- Some modern languages have optional application of GofN
  - Russian: ACC mapped to definite and GEN mapped to indefinite readings (Bailyn, 1997). Relatively recent innovation (Krasovitsky et al. 2011)
  - Polish: GEN in general can have partitive and indefinite interpretations
  - Croatian: GofN is generally more archaic, but can be associated with emphasis, indefiniteness and partitivity (Menac, 1979)
- Example from Russian (Pirnat, 2015):
  - Dima ne našel **sledy** (ACC) 'Dima did not find the traces'
  - Dima ne našel **sledov** (GEN) 'Dima did not find any traces'
- Given the exploratory results from *Codex Marianus*, did the older stages have any such contrasts?
  - Could help explain the rise of GofN while still maintaining partitives in other contexts

Definiteness and GofN

- Given the connection between indefinite/definite readings and the partitive origins, we need to explore potential definiteness contrasts in OCS
- OCS lacks a definite article but have other ways to overtly mark definiteness
  - Long-form vs. short-form adjectives:
    - slěpaja žena 'the blind woman' vs. slěpa žena 'a blind woman' (Lunt, 1974)
  - Possessive pronouns
  - Demonstratives
- Ways to overtly mark indefiniteness:
  - Indefinite pronouns and short-form adjectives (Willis 2013)
- Need to look at the distribution of these with GEN and ACC under negation

Corpus of Old Slavic

Corpus-study

- Descriptions of GofN in old Slavic assert that it is an obligatory morphosyntactic rule, e.g. (Lunt, 1974)
  - Want to test whether this is really the case or if there is variation
  - If there is variation, is there a meaningful contrast?
- To do this, we need a large corpus of the available data
- But there is no tagged corpus that covers all of the oldest stages of Slavic
- This study motivated the creation of such a corpus

Building an Old Slavic Corpus

- Some good resources are already available:
  - *Pragmatic Resources of Old Indo-European Languages* Treebank (PROIEL; Haug and Jøhndal (2008)): One OCS text
  - *TromsøOld Russian and OCS Treebank* (TOROT; Eckhoff and Berdiceviskis (2015)): Only OCS and OES texts
- But these lack the depth and breadth that we need for a comprehensive view of Old Slavic
- Consequently, a new corpus needs to be compiled
- These previous resources form the foundation, as they are morphologically-, syntactically-, and semantically-tagged

Composition of the Corpus

- All texts gathered from the internet in an electronic format
- Converted to plain text files, with a standardized orthography for each language
- Not all texts are relevant to the current investigation, but it is still important to create a general corpus that is conducive to multiple different analyses and investigations
- Following the example of PROIEL and TOROT, this corpus uses dependency grammar: with dependency relational, morphological, POS, and information structure tags
- Not exhaustively complete: Still many texts that can be added and tagged.

Breakdown of Old Slavic Corpus

Composition of the Texts

Language	Pre-tagged	Untagged	Total
Old Church Slavonic	10	36	46
Old Slovene	0	5	5
Old Croatian	0	1	1
Old Polish	0	20	20
Old Czech	0	4	4
Old Sorbian	0	2	2
Old East Slavic	32	3	35
Totals	42	71	113

Using the Corpus

- With a tagged corpus, we can conduct a thorough investigation of GofN in Old Slavic
- Not just limited to GofN, since the new corpus includes morphological and syntactic tags
- More texts still being processed and added to the corpus

Corpus-Study Results

Corpus Results

- Looking only at object GofN and only in OCS
- See how definite expressions interact with GofN
  - Possessive Pronouns
  - Adjectives
  - Leave demonstratives aside as a possible way to express definiteness
- 11,071 total occurrences across the OCS texts:

Type of entry	Count	Type of Entry	Count
GofN objects	11,071	Short adjs. (GEN), GofN	1,493
Poss Pronouns, GofN	1,107	Long adjs. (GEN), GofN	31

Results: Object Genitive of Negation

(8) a. kŭ tomu otŭ tebe vŭ věkŭ
to that.DAT from 2nd.SG.GEN into age.ACC
niktože **ploda** ne sŭněstŭ
no.one.NOM fruit.GEN NEG eat.PRES
'May no one ever eat fruit from you' (Mark 11:14; *Codex Marianus*)

b. μηκέτι εἰς τὸν αἰῶνα ἐκ σοῦ
NEG into the.ACC age.ACC from 2nd.SG.GEN
μηδεὶς **καρπὸν** φάγοι
no.one.NOM fruit.ACC eat.AOR.OPT
'May no one ever eat fruit from you' (Mark 11:14; *Greek New Testament*)

Results: Adjectives and GofN

- OCS uses short-form adjectives to give an indefinite reading:

(9) a. ni vĭlivajǫtŭ vina **nova** vŭ
NEG pour.PRES.3rd wine.GEN new.GEN into
měxy **vetŭxy**
wineskin.ACC old.ACC
'Neither do they put new wine into old wineskins' (Matthew 9:17; *Codex Marianus*)

b. οὐδὲ βάλλουσι **οἶνον** **νέον** εἰς
NEG throw.PRES.3rd wine.ACC new.ACC into
ἀσκοὺς παλαιούς
wineskin.ACC old.ACC
'Neither do they put new wine into old wineskins' (Matthew 9:17; *Greek New Testament*)

Results: Possessive Pronouns and GofN

- OCS uses possessive pronouns with a similar distribution to Greek definite articles

(10) a. ne iskusiši gospodi boga **tvojego**
NEG tease.PRES lord.GEN god.GEN 2nd.SG.GEN
'Do not test the Lord your God' (Luke 4:12; *Codex Marianus*)

b. οὐκ ἐκπειράσεις κύριον **τὸν** θεόν
NEG test.FUT lord.ACC the.ACC god.ACC
'Do not test the Lord God' (Luke 4:12; *Greek New Testament*)

Use of the Accusative in OCS

- Alongside the genitive, we also still accusative objects under negation

Type under NEG	**Count**
Objects (GEN)	11,071
Short adjs. (GEN)	1,493
Long adjs. (GEN)	31
Possessive Pronouns (GEN)	1,107
Objects (ACC)	**2,465**
Long adjs.(ACC)	**231**

Use of the Accusative in OCS

(11) bezumĭni ne iže li estŭ
foolish.VOC NEG who.NOM INTERROG be.PRES.3RD
sutvorilŭ **vŭnestinee** i **vŭnotrinee** sŭtvori
make.PART.NOM outside.ACC also inside.ACC make.AOR.3RD
'O foolish people, did the one who made the outside not also make the inside?' (Luke 11:40; *Codex Suprasliensis*)

(12) ne može utaiti **ję**
NEG be.able.AOR hide.INF 3RD.PL.ACC
'And he was not able to hide them.' (*Codex Zagrophensis*)

Preliminary Results

Takeaways

1. GEN marking of objects under negation is not obligatory
   - ACC objects can occur under negation and tend to be definite
   - GEN objects tend to be indefinite, but some are definite
2. OCS use long-form adjectives to mark definiteness, especially for ACC objects.
   - Majority of short-form adjectives occur with GEN under negation
3. Both GEN and ACC can occur with possessive pronouns to mark definiteness

Tentative Conclusions

- GofN does not seem to be just a hard-and-fast rule even in the oldest attested stage
- There is at least a partial definiteness contrast in OCS under negation
  - GEN-marked objects can have a definite or indefinite interpretation
  - ACC-marked objects seem to only have definite (and never indefinite) interpretations
- Could explain this contrast in a few ways:
  - A byproduct of the development from partitive constructions
  - Remnant of old use of Accusatives
  - In the process of syntactic leveling back to ACC-marked objects
- Need to verify these results:
  - Might be errors in the tagging and parsing that lead to inaccurate results
  - Contrast might not hold in all instances
  - Need to Compare the chronology of the texts and the distribution of GofN/ACC under negation in each of the texts

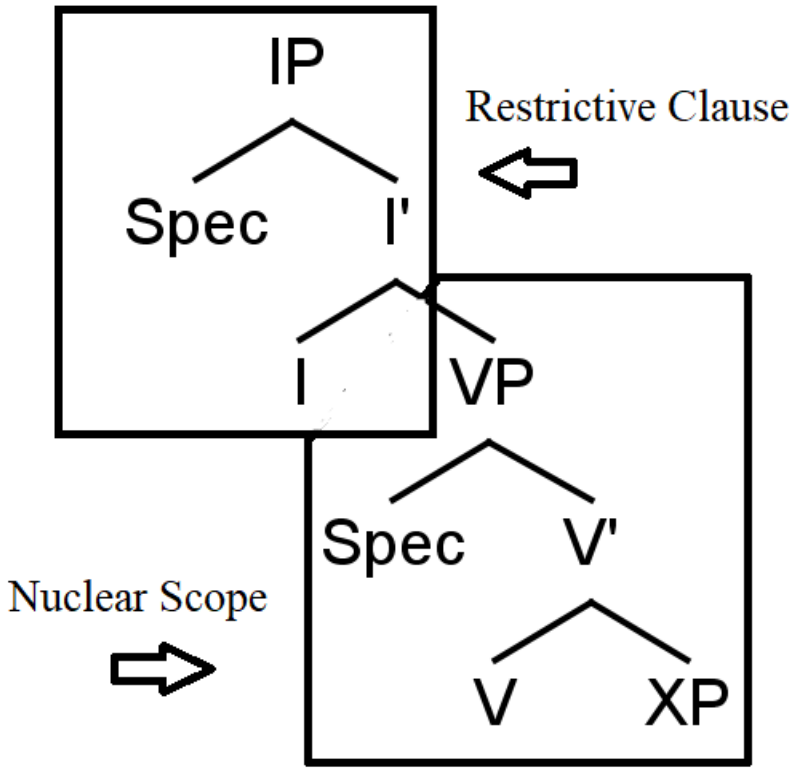
Possible Analysis

Potential Analyses

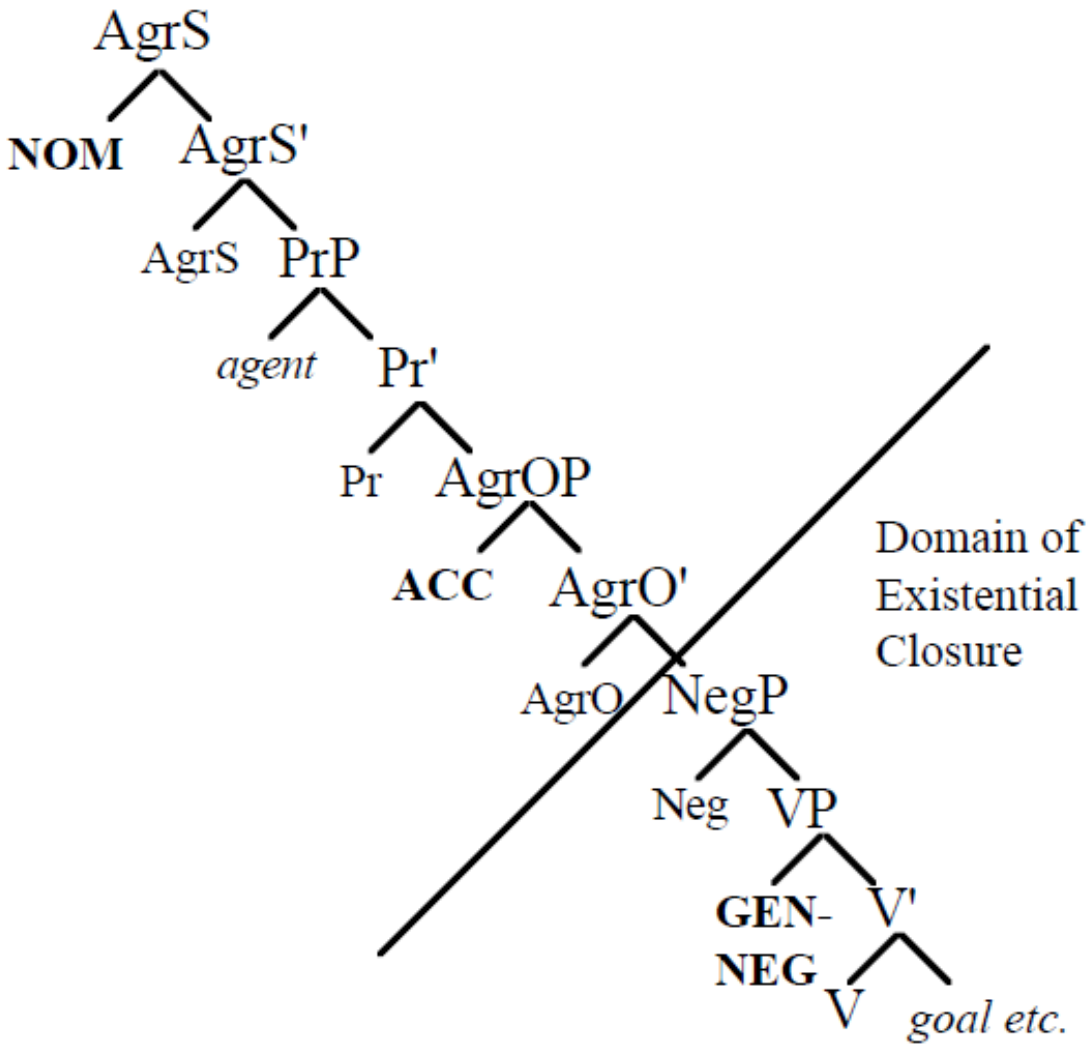
- How do we explain this contrast in definiteness?
- Potential analyses:
  - Diesing (1992)'s Tree-Splitting hypothesis
  - Schwarz (2009, 2013)'s two types of definites
- These are tentative proposals that might not account for all of the data

Tree-Splitting hypothesis (Diesing, 1992)

- Maps semantics onto the syntax
- Restrictive Clause gets definite interpretation
- Nuclear Scope gets indefinite interpretation
- Could be used to explain the variation between GEN and ACC



Bailyn (1997, 2004)'s Analysis of Russian



Different Definites (Schwarz, 2009, 2013)

- Different uses of definites:
  - Anaphoric Use: 'John bought a book and a magazine. The book was expensive.'
  - Immediate situation: 'the desk' (uttered in a room with exactly one desk)
  - Larger situation: 'the prime minister' (uttered in the UK)
  - Bridging: 'John bought a book. The author is French.'
- Weak article definites based on uniqueness
- Strong article definites involve an anaphoric link

Different Definites (Schwarz, 2009, 2013)

- This type of analysis has been extended to Lithuanian (Šerikaitė, 2018)
- Very similar situation to Old Slavic
  - No definite articles
  - Contrast between long- and short-form adjectives
  - Demonstratives
  - Possessive pronouns
- Could be used to explain variation in interpretation

Different Definites in Lithuanian (Šerikaitė, 2018)

(13) a. Praėjus dviem savaitėm po rinkimų, prezidentas turi
Passed two weeks after elections president has
teisę atleisti **naują ministrą pirmininką** tik
right fire new minister prime only
išskirtiniais atvejais.
exceptional cases
"Two weeks after the election, the president has a right to fire the new prime minister only in exceptional cases."

b. **Knyga** "Lietus" sulaukė neįtikėtino populiarumo,
Book 'Rain' received incredible popularity
nepaisantto, kad **talentigas-is** rašytojas nusprendė
despite that talented-DEF$_{strong}$ writer decided
likti anonimas.
remain anonymous
"The book 'Rain' became incredibly popular despite the fact that the talented writer decided to remain anonymous.'

Different Definites (Schwarz, 2009, 2013)

- In Lithuanian, there is a contrast between short- and long-form adjectives:
  - Short-form adjective typically receive indefinite readings, but can receive weak definite interpretations
  - Long-form adjectives receive strong definite interpretation
- We can extend this to OCS
- Under negation, GEN-marked objects can receive indefinite and weak definite readings, while ACC can receive strong definite readings
  - Some ACC-marked objects might also be able to receive weak definite interpretation
  - Need to go through all of the thousands of data points

Weak Definites in OCS

- Where we see *definite* GEN-marked objects under negation, they are weak definites built on uniqueness.
- This is in addition to the indefinite readings

(14) a. ne iskusiši gospodi boga **tvojego**
NEG tease.PRES lord.GEN god.GEN 2nd.SG.GEN
'Do not test the Lord your God' (Luke 4:12; *Codex Marianus*)

b. Ne uboi sę prijęti **ženy** **tvoeję**
NEG fear.IMP.2SG REFL take.INF wife.GEN your.GEN
Marije.
Maria.GEN
'Do not be afraid to take your wife Mary.' (Matthew 1:20; *Codex Marianus*)

Strong Definites in OCS

- ACC-marked objects all receive a definite reading
- Most are based on an anaphoric link either within the sentence or within the preceding discourse

(15) ne može utaiti **ję**
NEG be.able.AOR hide.INF 3RD.PL.ACC
'And he was not able to hide them.' (*Codex Zagrophensis*)

(16) bezumĭni ne iže li estŭ
foolish.VOC NEG who.NOM INTERROG be.PRES.3RD
sutvorilŭ **vŭnestinee** i **vŭnotrinee** sŭtvori
make.PART.NOM outside.ACC also inside.ACC make.AOR.3RD
'O foolish people, did the one who made the outside not also make the inside?' (Luke 11:40; *Codex Suprasliensis*)

Conclusions and Further Directions

Conclusions

- Genitive of Negation is not obligatory
  - Partial definiteness contrast between ACC and GEN under negation
- This contrast follows from the proposed partitive origins
  - See a connection between partitivity and indefiniteness across Indo-European
  - Gothic and OHG perfective verbs: ACC objects receive definite interpretation while GEN objects receive indefinite (Abraham 1997)
  - Compare Modern French:
    - je bois du vin “I drink (some) wine"
    - je bois le vin “I drink the wine"
  - Might point to cross-linguistic phenomena
- These contrasts can be accounted for using the Split-Tree Hypothesis (Diesing, 1992) and Strong/Weak Definites (Schwarz, 2009, 2013)
  - The former is necessary to account for the distribution of cases and plays a key role in accounting for Subject GofN
  - The latter is necessary to account for the contrasts in interpretation

Further Directions

- Despite covering lots of ground, still plenty that can be done!

Further Directions

1. Expand the scope of this study
   - Include other GofN contexts
   - Look at other Slavic languages
2. Go through all 13,000 data points!
3. Refine and expand the corpus
4. Look at Finnish, Baltic, and other Indo-European languages

Bibliography I

Bailyn, J. F. (1997). Genitive of negation is obligatory. In W. Browne, E. Dornisch, N. K. and Zec, D., editors, *Formal Approaches to Slavic Linguistics, The Cornell Meeting*, pages 84–114, Ann Arbor. Michigan Slavic Publication.

Bailyn, J. F. (2004). The case for q. In et. al., O. A., editor, *In Formal Approaches to Slavic Linguistics 12*, pages 1–36, Ann Arbor. Michigan Slavic Publication.

Diesing, M. (1992). *Indefinites*. MIT Press, Cambridge.

Eckhoff, H. M. and Berdiceviskis, A. (2015). Linguistics vs. digital editions: The tromsø old russian and ocs treebank. In *Scripta and e-Scripta 14-15*, pages 9–25.

Bibliography II

Fang, M. and Cohn, T. (2017). Model transfer for tagging low-resource languages using a bilingual dictionary. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 587–593, Vancouver, Canada. Association for Computational Linguistics.

Georgi, R., Xia, F., and Lewis, W. (2012). Improving dependency parsing with interlinear glossed text and syntactic projection. In *Proceedings of COLING 2012: Posters*, pages 371–380, Mumbai, India. The COLING 2012 Organizing Committee.

Georgi, R., Xia, F., and Lewis, W. (2015). Enriching interlinear text using automatically constructed annotators. In *Proceedings of the 9th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH)*, pages 58–67, Beijing, China. Association for Computational Linguistics.

Bibliography III

Haug, D. and Jøhndal, M. (2008). Creating a Parallel Treebank of the Old Indo-European Bible Translations. LaTeCH 2008.

Lunt, H. G. (1974). *Old Church Slavonic Grammar*. Mouton, Paris.

Meillet, A. (1897). *Recherches sur l'emploi du gentif-accusatif en vieux-slave*. Ministere de l'instruction publique, Paris.

Menac, A. (1979). Slavenski genitive u suvremenom hrvatskom književnom jeziku. *Jezik: časopis za kulturu hrvatskoga književnog jezika*, 26.3:65–67.

Pesetsky, D. (1982). *Paths and Categories*. PhD thesis, Massachusetts Institute of Technology.

Pirnat, Z. (2015). Genesis of the genitive of negation in balto-slavic and its evidence in contemporary slovenian. *Slovene Linguistics Studies*, 10:3–52.

Bibliography IV

Reimers, N. and Gurevych, I. (2017). Reporting Score Distributions Makes a Difference: Performance Study of LSTM-networks for Sequence Tagging. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 338–348, Copenhagen, Denmark.

Schwarz, F. (2009). *Two Types of Definites in Natural Language*. PhD thesis, University of Massachusetts, Amherst.

Schwarz, F. (2013). Different types of definites crosslinguistically. *Language and Linguistics Compass*, 7.10:534–559.

Šerikaitė, M. (2018). Strong vs. weak definites: Evidence from lithuanian adjectives. pages 83–111.

Thank you!

Old Slavic Corpus

Corpus of Old Slavic

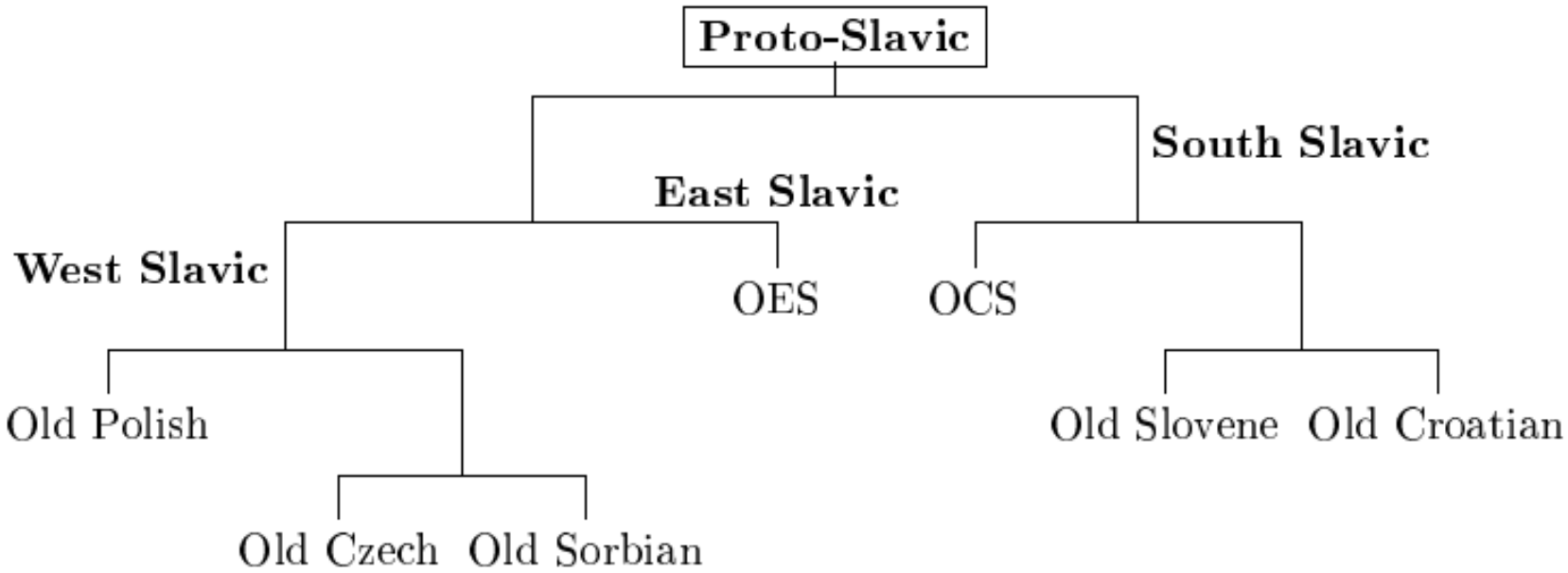
- This is part of a larger project looking at the diachrony of GofN in Slavic, but this relies on data from all of Slavic.
- While there are some good resources out there, e.g. PROIEL and TOROT, but none covered the depth and breadth of Slavic languages needed

Solution

Create a corpus of Old Slavic to verify these claims (and for future use in other linguistics investigation)

Corpus of Old Slavic

- Need representation across all branches of Slavic
- These are the current languages represented



Breakdown of Old Slavic Corpus

- Distribution of texts(pre-tagged and untagged) across the languages

Composition of the Texts

Language	Pre-tagged	Untagged	Total
Old Church Slavonic	10	36	46
Old Slovene	0	5	5
Old Croatian	0	1	1
Old Polish	0	20	20
Old Czech	0	4	4
Old Sorbian	0	2	2
Old East Slavic	32	3	35
Totals	42	71	113

Building the Corpus

- Pre-tagged texts were taken from PROIEL (Haug and Jøhndal, 2008) and TOROT (Eckhoff and Berdiceviskis, 2015)
  - Set the standard to achieve for morphological, syntactic, and semantic tagging
  - Good training data for new tagging models
- Untagged texts were gathered from the internet
- How do we get the untagged texts in a usable format for linguistic investigation?
- Important to keep the corpus domain general and not just tailored to the current investigation

Work-flow for Tagging

1. Train BiLSTM-CRF neural net models
2. Gather digital forms of texts
3. Preprocessing of digital texts
    1. Standardize orthography
    2. Sentence-division/word-lemmatization
    3. Conversion to .conll format
4. Apply models to processed texts:
    1. Part-of-speech tagging
    2. Morphological tagging
    3. Relational tagging and parsing
5. Output tagged files
6. Query tagged files for specific investigations

Automatic Tagging

- This works for the languages that have pre-tagged data, but we still need a way to tag new texts, especially for those without any training data
- This is very difficult, as most modern methods require millions of tokens of training data to achieve fairly accurate results.
- This is a problem shared between limited historical data and low-resource languages: there is a lack of model-ready data
- How can we address this problem?
  - Created a number of different models and use the best one for each language

Models Trained

- *Extended Historical models*:used the pre-tagged historical texts with newly-made word-embeddings from all of the untagged and tagged texts. OCS extended to South Slavic; OES extended to West Slavic.
- *Historical Models*: same as *Extended models*, but used automatically tagged Old Polish data from Morfeusz as training for the rest of West Slavic

Models Trained

- *Model Transfer*: Use a bilingual dictionary, monolingual corpora in both the high- and low-resource languages, and a small annotated corpus for the low-resource language to train a model for the low-resource language (Fang and Cohn, 2017).
- *Modern embeddings*: took embeddings for the modern descendants of the older languages (Bulgarian for OCS and Russian for OES) and mixed it with the models trained for OCS and OES
- *Related Modern languages*: took both the word-embeddings and the training data from the modern descendant languages and applied them to the older stages' test sets.

Model Results

Differences between models

1. Training data
2. Word-embeddings

- Same kind of NN used each time: a BiLSTM-CRF Neural Network (Reimers and Gurevych, 2017) that takes training sets and word-embeddings
  - Word embeddings are distributed representations of text in an n-dimensional space. Words that share common contexts are located closer together
  - New word-embeddings trained using Word2Vec
- Test data for each language remained the same across the different models:
  - PROIEL and TOROT for OES and OCS
  - Test set of manually annotated tokens for languages without tagged data (∼500 words)

Model Results

- Accuracy for each model in correctly applying all tags to test set
  - Highest accuracy for each language in bold

Lang.	*Ext.H.Mod.*	*Hist.Mods.*	*Mod.Trans.*	*Modern Embeds*	*Modern langs.*	*UDs*
OCS	**75.63**	**75.63**	70.54	65.32	63.64	87.40
Old Sl.	57.21	57.21	N/A	55.42	**60.26**	88.70
Old Cr.	61.79	61.79	N/A	59.57	**63.81**	84.23
OES	69.60	69.60	N/A	68.98	**71.47**	83.91
Old Pol.	56.84	**70.56**	N/A	59.34	61.30	84.64
Old Cz.	49.53	**60.13**	N/A	58.03	56.33	92.50
Old Sorb.	30.94	**50.79**	N/A	43.29	N/A	N/A

Model Results

- Accuracy for the models is not ideal, but not terrible
- Increasing the amount of data used in training, even if it is from the modern languages, improves results
- Shows a potential link between different stages of languages and being able to use them in an all-inclusive model
- Ideally we want a fully diachronic model that can handle any stage of a given language: This might be a first step in that direction.
- Might be able to improve performance without scrapping models (Georgi et al., 2012, 2015)
- Inaccuracies might not be relevant to current investigation, e.g. many errors for information status, less for POS and morphology

Using the Corpus

- With a tagged corpus, we can investigate of GofN in Old Slavic
- Not just limited to GofN, since the new corpus includes morphological and syntactic tags
- More texts still being processed and added to the corpus
- Manual annotation will be necessary in the future
- But what do we need to look for in all of this data?