Weighing the evidence: Parallel data and Greek syntactic influence

Hanne Eckhoff

hanne.eckhoff@mod-langs.ox.ac.uk

University of Oxford

3 November 2017

In the beginning was Greek

- Old Church Slavonic came into being *in order to* translate from Greek
- Moreover, the texts to be translated were objects of extreme veneration
- The translators were in all probability Greek (but bilingual)
- The translated texts then became ideals of good style for all of Slavia Orthodoxa
- Can we separate native Slavic syntactic features from Greek ones?
- A certain pessimism in the literature

Can detailed corpus data help?

- Increasingly, we now have historical/diachronic digital corpora for early Slavic
- Also parallel Greek-Slavic corpora
- Detailed corpus annotation on many levels
- Some phenomena are frequent enough for statistical modelling
- Can we use these resources and methods to reassess the relationship between Greek and native syntax?
- I will report from two case studies using the PROIEL/TOROT treebanks

The case studies

- Case 1: Word order
  - Objects and verbs
  - Subjects and verbs
- Case 2: Choice of possessive construction
- Data from the Codex Marianus with token-aligned Greek parallel

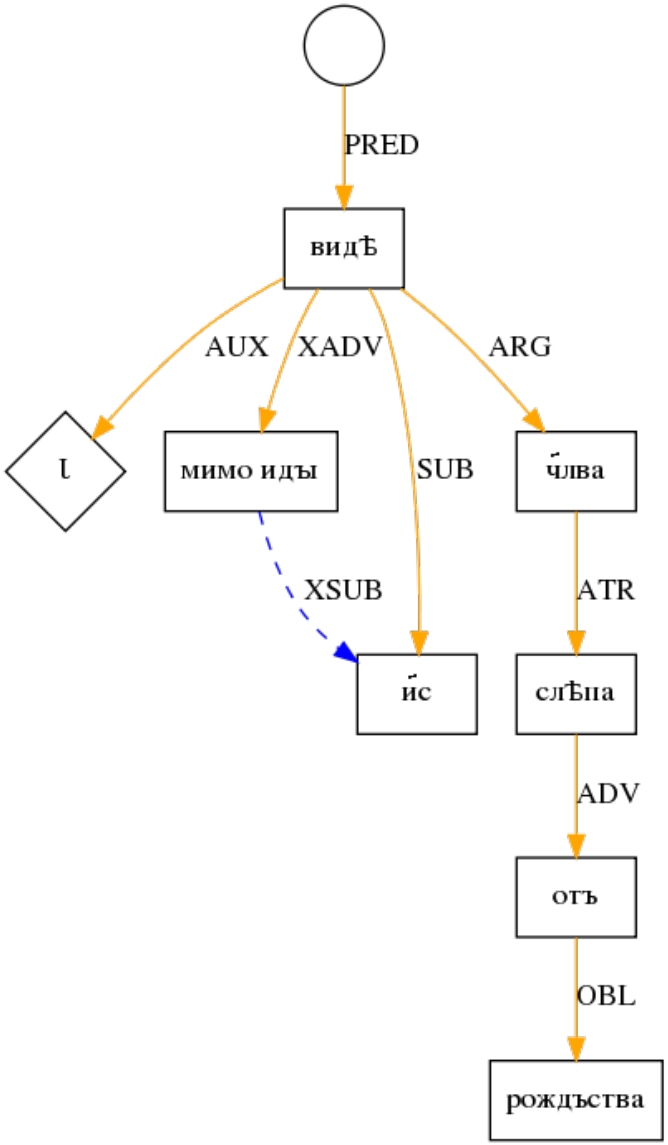
The PROIEL treebank

- PROIEL: Pragmatic Resources in Old Indo-European Languages (University of Oslo 2008–2012)
- By what linguistic means do these languages express pragmatics and information structure?
- Word order, anaphoric expressions, definiteness, participles (background events), discourse particles
- Centrepiece: An aligned parallel corpus of old Indo-European New Testament texts (Greek, Latin, Gothic, Classical Armenian and OCS (Codex Marianus))
- Focus on making the most of a limited dataset by in-depth manual annotation on many levels
- Lemmatisation, morphology, syntax, information structure, customised tagging

The TOROT treebank

- The Tromsø Old Russian and OCS Treebank (2013–)
- A daughter treebank expanding on the OCS part of the PROIEL treebank
- More OCS: Zographensis, Suprasliensis and expanding
- Later Church Slavonic (Vita Constantini, Vita Methodii and expanding)
- Large collection of Old East Slavic and Middle Russian texts (ca. 250,000 word tokens and expanding)
- Old Polish and Early Modern Bulgarian pilots

Morphosyntax



Derivational morphology

- Verbs tagged for prefixation, suffixation and stem (nearly complete in OCS and historical Russian)
  - сътворити: съ, твор, и
- Denominal adjectives tagged for derivational suffix
  - -ov-, -in-, -j-, -ьsk-, ьn- (Marianus)

Semantics

- Animacy tagging (OCS)
- Prepositional semantics (Greek, transferrable to the translation langugages)

Information status annotation in the PROIEL corpus

Context	Specific tag	Non-specific tag	
Discourse	OLD	NONSPEC-OLD	previously mentioned
Scenario	ACC-INF	NONSPEC-INF	accessible by inference
Encyclopaedic	ACC-GEN		acc. from world knowledge
Situation	ACC-SIT		accessible by deixis
—	NEW	NONSPEC	not previously mentioned
	KIND		kind-referring
		QUANT	quantified

Table : Contexts and tags in the PROIEL corpus, adapted from Haug et al. 2014

Greek gospels complete, automatically transferred to the other languages.
Incomplete annotation in the Old East Slavic material

Topicworthiness

- How do you find the aboutness topic in an ancient text?
- Pragmatic solution: Calculate a topicworthiness score based on features that can be reliably annotated
  - Place on the givenness hierarchy (old ranks highest, new is excluded)
  - Place on the hierarchy of syntactic relations (subjects rank highest)
  - Place on the animacy hierarchy (humans rank highest)
  - Word order (first is best)
  - Realisation (prodrops, personal pronouns, personal names rank higher)
  - Relative saliency: is the topic candidate a member of a longer and tighter anaphoric chain than the competition?
  - Properties of the immediate antecedent: does it outrank the intervening referents on the relation, animacy and givenness hierarchies?

Squeezing the empirical lemon: Classification trees

- Can complex, multilayered data to some extent compensate for the lack of native-speaker intuitions?
- The data structure calls for statistical methods that can weigh multiple factors against each other
- Classification trees
  - Well suited for data sets with a large number of predictors
  - Better suited than most models to deal with factors that are internally correlated (multicollinearity)
  - Provide intuitive visualisations of complex factor interactions
  - R: 'party' package

How does the model work?

- Recursive partitioning: an algorithm recursively subdivides the data into ever smaller sets and subsets
- At every split, the model selects the best predictor for the split
- Splits are visualised as circles representing those factors that can subdivide the data in a statistically significant way
- Terminal nodes have representations of the actual distribution in that particular subset
- In most cases there will be residual variation unaccounted for by the tree: means to assess the success of the classification

Extreme influence: word order

- What will the classification tree look like when OCS follows Greek almost completely?
- Two case studies: object-verb and subject-verb order
- Under normal circumstances we would expect the ordering to be influenced by
  - information structure
  - topicality
  - part of speech
  - referent prominence

Verb-object order: 95.5% like Greek

	OCS VO	OCS OV
Greek VO	67.5%	2%
Greek OV	2.5%	28%

Table : Direct object position in OCS and Greek, per cent, n=3747

Predictors

- Greek VO order (is the Greek object preverbal?)
- Greek definiteness (does the Greek object have a definite article?)
- Animacy (human, concrete, non-concrete, place)
- OCS part of speech
- Number
- Givenness status (simplified to old, accessible, new, non-specific)
- Topicworthiness score

The classification tree

- Correct classification rate of 95.5%
- Baseline = 69.9% (the result you get if you guess that all OCS objects are postverbal)
- Exactly the same result as if you predict the VO order by Greek VO order alone
- Obviously Greek VO order is the single best predictor, drowning out all others

Deviations when Greek has VO

- It *is* possible to read a few observations off the tree
- OCS is more likely to follow the Greek if the Greek object is postverbal (left-hand side of the split)
- Among the postverbal objects, the only deviations have to do with part of speech
- Pronouns are more likely to be prenominal than other objects
- Node 8: There are 41 OCS preverbal personal pronoun objects translated from postverbal Greek ditto
- A real difference?
- A few preferred environments (e.g. after *da* in purpose clauses, directly after a wh-word or a relative pronoun, between a finite verb and an infinitive or between auxiliary *byti* and an l-participle)

Preverbal pronominal objects

(1) *kъ tomu že ne sъměaxo ego vъprašati ničьsože*
to that ptc not dared him ask nothing
ouketi gar etolmōn eperōtan auton ouden
'they did not dare to ask him anything more' (Lk. 20:40)

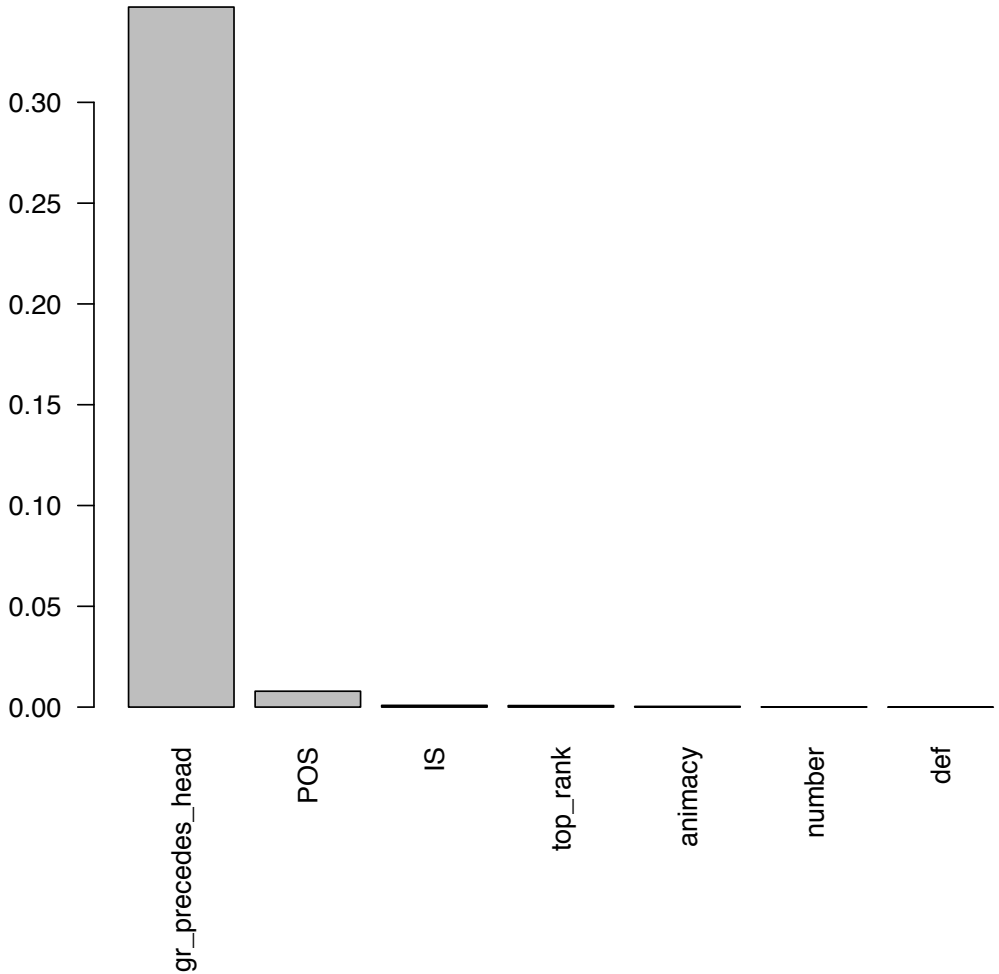
Deviations when Greek has OV

- Slightly more discrepancies when Greek has OV
- Split 2 on the right-hand side (node 10):
  - 38 Greek preverbal personal pronoun objects are rendered by postverbal ones in OCS
  - Personal pronouns are particularly prone to deviate from the Greek word order, but in both directions
- Split 3 on the right-hand side:
  - Node 12: objects that are neither personal pronouns nor numerals, and that render Greek definite objects
  - 27 deviant objects, mostly nouns

Postverbal noun objects

(2) *kto ti dastъ oblastь sijǫ da si tvoriši*
who.nom you.dat gave power.acc this.acc that these.acc do
tis soi tēn exousian tautēn edōken hina tauta poiēis
'who gave you this authority to do these things?' (Mk. 11.28)

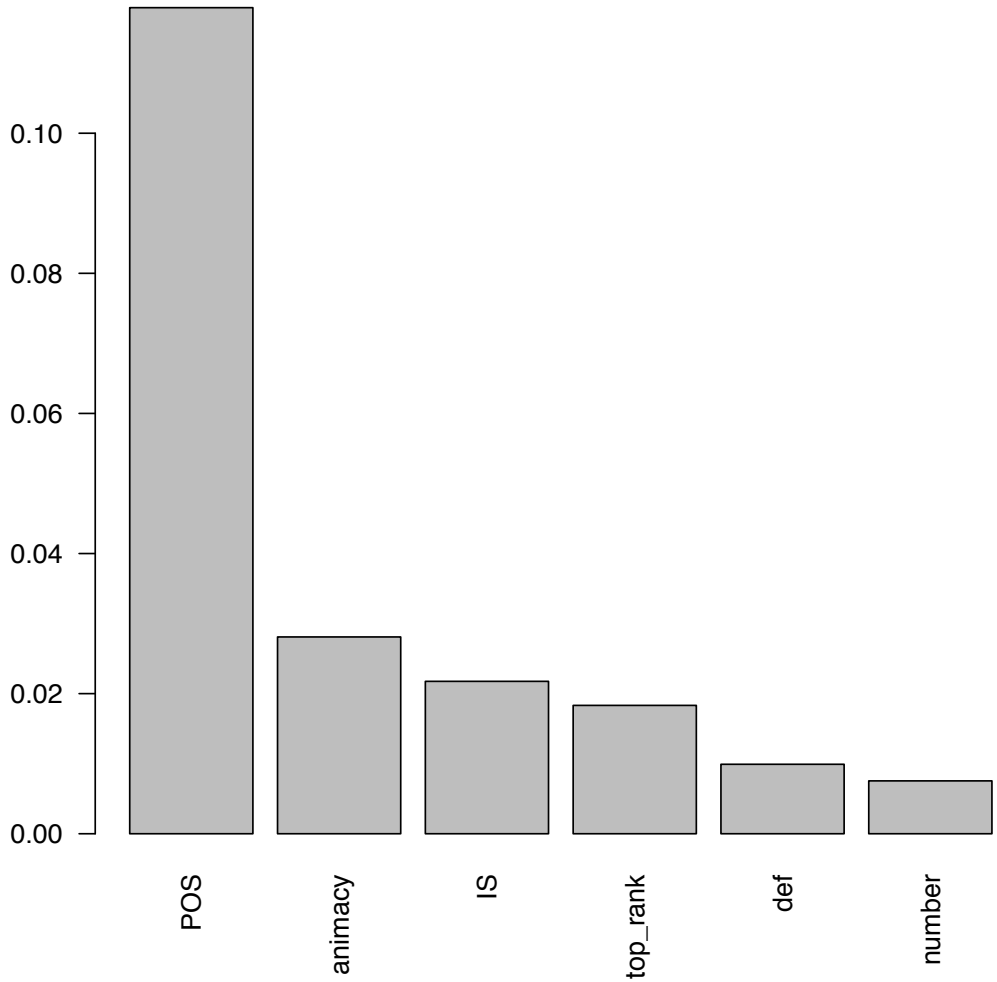
Relative importance of predictors (Random Forest)



What if we take Greek word order out of the equation?

- Verb-object order conditioned by a rich interplay of predictors
- Part of speech, givenness status, animacy
- Correct classification rate: 79.3% (baseline 69.9%)
- A description of NT Greek?
- We can hope that OCS and Greek had similar conditioning patterns, but we have little proof

Pretty, but unrealistic graph



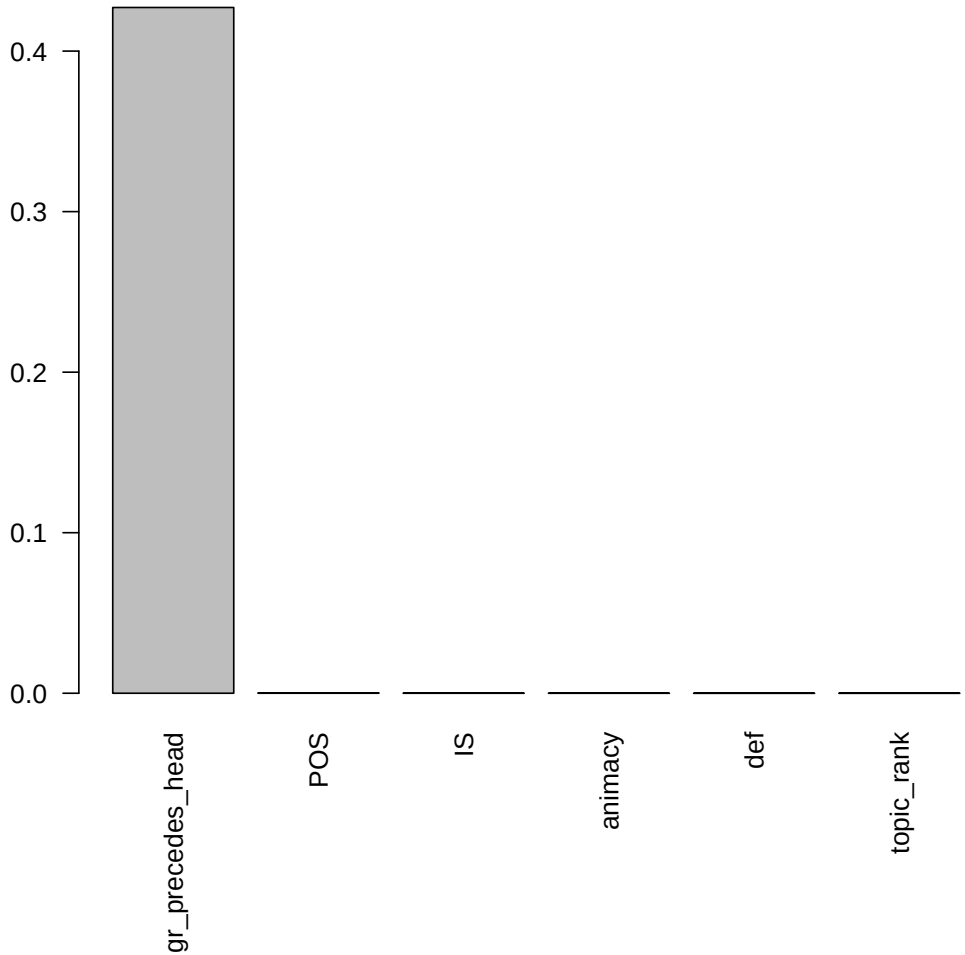
Subjects and verbs

	OCS VS	OCS SV
Greek VS	35.5%	1.4%
Greek SV	1.4%	61.7%

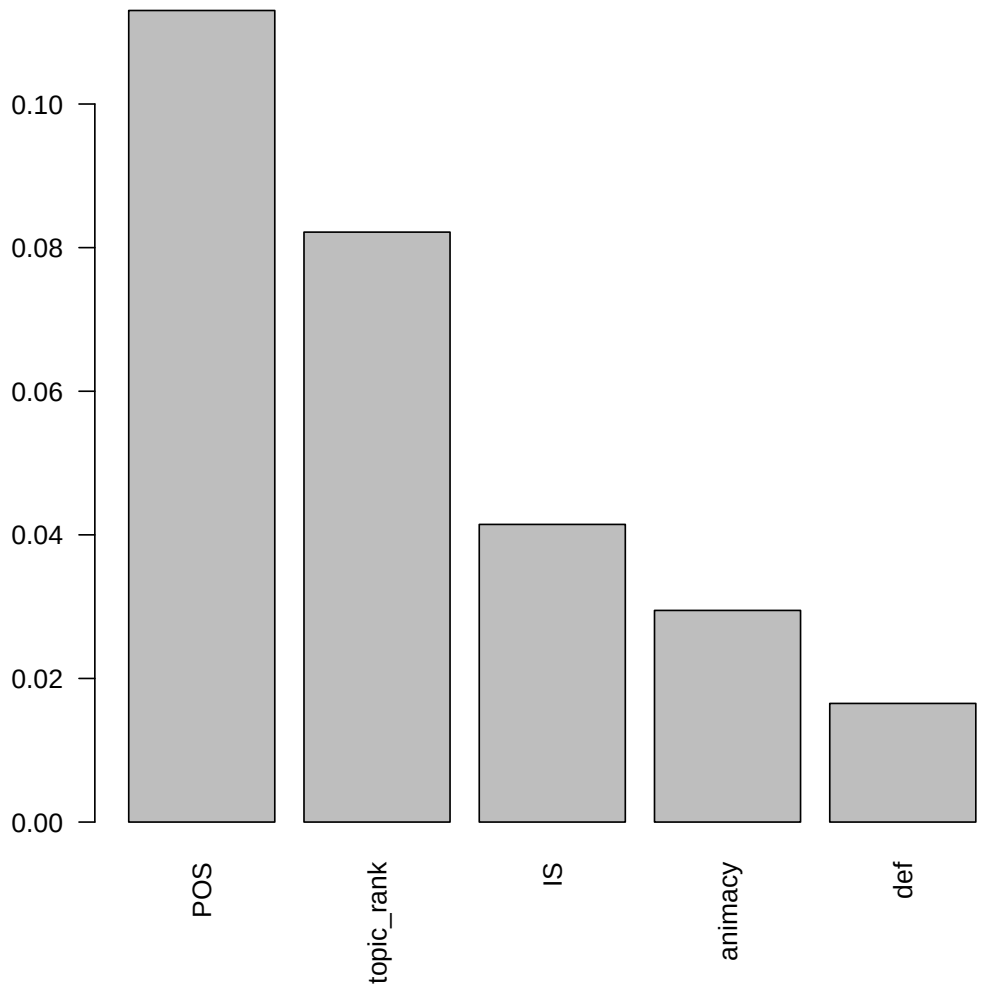
Table : Subject position in OCS and Greek, per cent, n=3915

- Things are no better if we try subject-verb order
- Follows Greek in 97.2% of the cases
- The classification tree shows up a small group of prenominal pronouns translating Greek postnominal ones

Greek word order the sole predictor



But dropping Greek word order . . .



Word order and classification trees

- The OCS ordering of subjects, objects and verbs follows the Greek order extremely closely
- Unsurprisingly, Greek word order is by far the best predictor if we want to classify the OCS occurrences
- Most deviations involve personal pronouns, and thus possibly clitic ordering rules
- If we remove Greek word order as a predictor, the classification trees indicate a rich interplay of POS, givenness status, topicworthiness and animacy
- But we have no way of telling if this is just a description of NT Greek
- . . . unless we tag early Slavic original texts in the same ways and run the models on them as well

Extreme independence: Possessive constructions

- OCS expressions of adnominal possession are notably independent of Greek
- Greek: Adnominal genitive dominates completely
- OCS: At least five construction types in complex interaction
- But even here, Greek influence has been suggested (no possessive genitive in Proto-Slavic?)
- The distribution is likely to be influenced by many of the same predictor types as argument word order
- How will the classification models perform?
- Fun fact: Possessor-possessee order follows Greek as much as SV and OV order does, regardless of construction!

What counts as possessive?

- Classification from Eckhoff 2011: wide definition of possession
- Anchoring/identification (John's head)
- Classification/labelling (camel's hair, Adam's apple, the Skull Place)
- Elaboration (a piece of bread, the forgiveness of sins, the virtue of abstinence)

OCS possessive types

- Type 1 adjectives, derived from nouns denoting humans with the suffixes *ov, in, j, n'*, also the adjective *božij* 'God's': *oučenici ioanovi* 'John's disciples' (abbreviated '1' in figures)
- Type 2 adjectives, derived from nouns with the suffixes *ьsk, ьn, ij*: *otъ oučenija fariseiska* 'from teaching of the Pharisees', also the rare adjectives derived from non-human nouns with the Type 1 suffixes (abbreviated '2' in figures)
- Adnominal genitive: *domy vъdovicъ* 'the houses of widows' (abbreviated 'g' in figures)
- Adnominal dative: *srъdьca otьcemъ* 'the hearts of the fathers' (abbreviated 'd' in figures)

OCS vs. Greek possessive constructions

OCS	agreeing adjective/participle	dat	gen	other
adj1	0	0	323	2
adj2	36	0	357	1
dat	0	4	102	1
modified gen	0	0	167	4
unmodified gen	2	0	151	7

Table : OCS possessive type by Greek type, n=1157

Modified genitives and nominalisations excluded

(3) *glavǫ* *ioana* *krstitelja*
head.acc Ioan.gen.sg baptist.gen.sg
tēn kefalēn lōannou tou baptizontos
'the head of John the Baptist' (Mk. 6:24)

Can we predict the OCS distribution by Greek syntactic factors?

- What should count as a Greek syntactic factor?
- Greek part of speech and case
- Not entirely a failure: correct classification rate of 50.6%
- Baseline 41.62% (guess type 2 adjectives every time).

But the classification reflect OCS-internal facts

- The first split sets Greek proper nouns apart from other possessors
- Type 1 adjectives are very frequently derived from (human) proper nouns
- A distinction that OCS makes, but not Greek

Type 1 adjectives from personal names: Correctly classified

(4)

sěmę	*avramle*	*esmъ*
seed	Abraham-j.neut.sg	are
sperma	Abraam	esmen

'we are Abraham's seed' (Jh. 8:33)

Type 2 adjectives from toponyms: Misclassified (30%)

(5) *pride isъ otъ nazareta galileiskaago*
came Jesus from Nazareth.gen Galilee-ьsk.gen.sg.
ēlthen Iēsous apo Nazaret tēs Galilaias
'Jesus came from Nazareth in Galilee' (Mk. 1:9)

The rest of the tree

- The second split is the only one that could indicate direct Greek influence
- If the Greek possessor is an agreeing adjective, the OCS will always render it with a Type 2 adjective
- But OCS uses denominal adjectives much more frequently than Greek
- At the bottom of the tree we are left with a large lump of common nouns, where all construction types are well represented

Adjective to adjective

(6)
otecъ *vašъ* *nbsky*
father.nom.sg. your heaven-ьsk.nom.sg.
ho patēr humōn ho ouranios
'your heavenly Father' (Mt. 6:32)

Full set of predictors

- Restricting the analysis to Greek predictors is clearly artificial
- The limited success of the model is due to OCS-internal facts
- The literature (Huntley 1984, Eckhoff 2011) suggests that we should rather use a wide range of properties of the possessor and possessed noun as predictors

Greek predictors

- Greek part of speech
- Greek case
- Greek possessor-possessee order
- Greek possessor definiteness
- Greek head definiteness
- Greek number

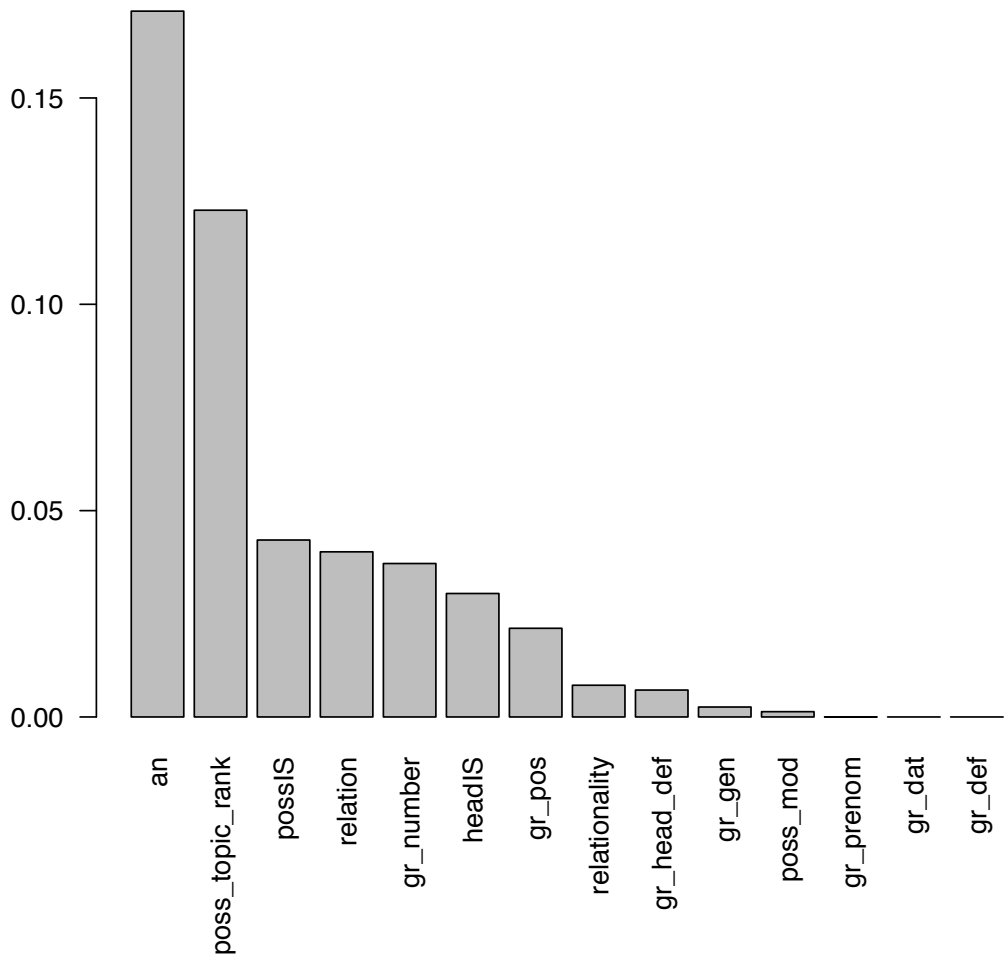
Semantic, pragmatic, syntactic predictors

- Animacy (human, concrete, place or non-concrete?)
- Possessor modification: is the possessor complex or does it consist of a single word?
- Relationality: is the head noun relational?
- Possessor givenness status (old, accessible, new, non-specific)
- Head givenness status
- Syntactic relation: ATR (attribute), APOS (apposition), PART (adnominal partitive) or NARG (object-like argument of relational noun)?
- Topicworthiness score

The full model

- A much more successful classification tree: correct classification rate 80.1% (baseline 41.62%)
- Most successful in predicting adjective constructions
- Has at least *some* success with predicting datives and genitives (33 of 107 datives and 54 of 118 genitives)
- The most powerful predictor is animacy, followed by topicworthiness

Variable importance of possessive predictors



Greek factors still around

- We see that quite a few Greek factors make at least some contribution to the model
- Greek number: plural possessors typically realised as Type 2 adjectives
- Greek part of speech: proper nouns (personal names) almost exclusively realised as Type 1 adjectives
- Greek features as 'semantic tags'

Plural possessors

(7) *bljuděte sę otъ kvasa fariseĭska*
watch refl from leaven.gen.sg. Pharisee-ьsk.masc.gen.sg.
blepete apo tēs zumes tōn pharisaiōn
'beware of the leaven of the Pharisees' (Mk. 8:15)

Split 1: Animacy

- Human possessors are much more likely to be realised as Type 1 adjectives
- Type 1 adjectives are not used with plural possessors (nodes 18, 34)
- But the classification tree is also able to identify human possessor subclasses where Type 1 adjectives are much less likely to occur
- Especially if the possessor has non-specific or generic reference, Type 2 adjectives or even datives are a better choice
- Genitives and datives also turn up when the 'possessor' is partitive or a nominal argument (node 19)

Non-specific human possessor

(8) *se sь člvkъ ědъca i vinopiica. mytaremъ*
behold this man glutton and wine-drinker tax-collector.dat.pl
drugъ i grěšьnikomъ
friend and sinner.dat.pl
idou anthrōpos phagos kai oinopotēs, elōnōn philos kai hamartōlōn
'behold, a gluttonous man and a drunkard, a friend of tax collectors and sinners!'

The translation of human common nouns depends on topicworthiness (nodes 24, 25, 28)

(9) *i* *vъšedъ* *vъ* *domь* *farisěovъ* *vьzleže*
and entered in house.acc.sg Pharisee-ov.masc.acc.sg reclined
kai eiselthōn eis ton oikon tou Pharisaiou kateklithē
'(One of the Pharisees asked him to eat with him.) And he went to the Pharisee's house and reclined at the table' (Lk. 7:36)

Topicworthiness score 27 (discourse-old, human)

Non-human possessors

- The first split (node 2) sets adnominal partitives apart from the rest
- As expected, they are solidly genitive
- The next split (node 3) surprisingly separates old and new from accessible and nonspecific
- Genitives and datives are preferred with singular non-partitives (often with relational head nouns)
- Datives especially frequent as 'objects' of deverbal nouns (node 4)
- Type 2 adjectives still the best choice for plural possessors (node 9)

Adnominal partitives

(10) *ěko i kaplę krъve kapljǫšta na zemljǫ*
like even drops blood.gen.sg dripping on ground
hōsei thromboi haimatos katabainontos epi tēn gēn
'(and his sweat became) like drops of blood falling on the ground'
(Lk. 22:44)

Singular old ‘possessor’ with relational head noun

(11) *i rečeta gnu domu*
and tell master house.gen.sg
kai ereite tī̄ oikodespotēi tēs oikias
‘(Follow him to the house that he enters,) and say to the owner of the house’ (Lk. 22:11)

Dative ‘object’ of deverbal noun

(12) *otъstǫpite otъ mene vьsi dělatele nepravьdě*
step-away from me all doers evil.dat.sg
apostēte ap’ emou pantes ergatai adikias
‘Depart from me, all you workers of evil!’ (Lk. 13:27)

What could the tree do for us?

- Captures many of the complex interactions of factors in the distribution of OCS possessive constructions
- Can predict all construction types to some extent, and has a decent correct classification rate
- Greek high-ranking predictors are 'semantic tags'
  - Part of speech: animacy, uniqueness and specificity
  - Number: important as a referential feature, not as a Greek syntactic influence
- Supports the findings of Huntley 1984 and Eckhoff 2011 without using direct semantic classification of possessive expressions
- Both the referential properties of possessor nouns and the valency properties of the possessee turn up in the tree

What was the tree not able to do?

- The classification tree probably can't capture all the intricacies of the system
- The generalisations in the lower splits of the tree seem more arbitrary than in the higher splits
- There is a lack of generalisations cross-cutting human and non-human possessors

Native syntax and syntactic influence

- In languages only attested in translated form, how can we distinguish between native syntax and syntactic influence?
- What can we do with richly annotated and aligned parallel Greek and OCS data?
- Morphology, syntax, givenness status and semantic features
- Statistical classification models to evaluate the relative weight of predictors
- Greek syntactic features must always be among the predictors!

Replication vs. independence

- Deliberate choice of extreme case studies
- In the word order studies, Greek word order was the supremely best predictor
- Only marginally did other predictors surface in the trees
- Only a comparative study of Slavic original texts could tell us whether Greek and OCS had the same conditioning patterns
- In the possessive study, semantic, pragmatic and referential features outrank Greek morphosyntactic predictors
- Greek morphosyntactic predictors function as 'semantic tags'
- OCS system clearly independent of Greek
- Analysis supports earlier research on OCS possessives

Intermediate cases

- What would the classification trees look like for a phenomenon where OCS syntax was *partially* independent of Greek?
- Both a Greek syntactic factor and an internal factor would both have high variable importance and be involved in the higher splits
- The Greek factor would not serve as a 'semantic tag'
- The tree would reveal robust groups of examples that did (not) conform to the Greek pattern

All I want for Christmas is more data

- Analyses of this kind require large and sophisticated data sets
- Some phenomena will never be frequent enough (accusatives with infinitives?)
- Some phenomena may be frequent enough if we annotate and align a lot more data (dative absolutes?)
- Some phenomena clearly require both parallel data studies and contrastive original-text studies (conjunct participles?)
- This type of annotation is work-intensive and time-consuming
- Full-coverage annotation is likely to be of better quality than partial annotation for a particular research purpose
- Full-coverage, multi-layer annotation should be shared and published for all scholars to use

Browse the PROIEL and TOROT treebanks
http://foni.uio.no:3000/
https://nestor.uit.no

Download reviewed data
https://proiel.github.io/
https://torottreebank.github.io/

Try the experimental Syntacticus open browsing interface
http://syntacticus.org/