Building and using online corpora for (historical) linguistic research

Computational linguistics and the dating of early Irish Texts, Maynooth University

Marius L. Jøhndal

15 December 2016

Overview

- Why annotated corpora?
- Building an annotated corpus for historical languages
  - Example: annotating a treebank
- Speeding up the annotation process
  - Example: rule-based and statistical morphological tagging
- Making an annotated corpus usable and useful

Why annotated corpora?

- Historical linguistics (whether synchronic or diachronic) is by definition based on corpora.

- We only have the text so we had better make the most of it!

- Traditional corpora (= collections of texts), whether printed or electronic, are good for hypothesis formation. They are less suitable for hypothesis testing.

Why annotated corpora?

Word order in declarative main clauses in the Gospel of Luke and the Acts of the Apostles according to three scholars

	Luke + Acts		Luke only	
	Rife (1933)	Davison (1989)	Rife (1933)	Kirk (2012)
VSO	15	20	9	14
SVO	50	56	19	13
SOV	9	8	8	5
VOS	3	4	2	3
OVS	6	6	1	1
OSV	1	1	0	1

Haug 2015

Why annotated corpora?

- Why are the numbers different?

- ‘The investigation was limited to main declarative clauses where both subject and object are substantives.’ (Rife 1933)

- ‘clauses…which contained at least one nominative noun, one accusative noun and one indicative verb… Verbs normally followed by a genitive or a dative were traced using a concordance’ (Davison 1989)

Why annotated corpora?

- The clause contains at least an S(ubject), V(erb) and O(bject)
- The clause is continuous
- S and O are not embedded in a participial clause
- The verb assigns accusative, genitive, or dative to an argument that is a patient or theme
- The V consists of one word (no periphrastic forms, modal embeddings or light verbs)
- S and O are determiner phrases (this includes nominalizations) or quantifier phrases, and not clausal
- S and O are continuous strings

Kirk 2012

Why annotated corpora?

- Three problems (at least):

1. Implicit assumptions: Which edition did Rife (1933) use? What is meant by ‘main declarative’, ‘subject’, ‘object’?

2. Not replicable: Davison (1989) uses an electronic text and a computer programme to locate relevant passages but neither is freely available to other academics.

3. Manual work: How likely is it that someone can get the numbers right the first time around using Kirk’s (2012) explicit but complex criteria?

Why annotated corpora?

- Hypothesis testing by hand is very error prone: Even if one includes everything that should be included, things may have been excluded that should not have been

- Replication is very time consuming: The worst-case scenario is that multiple scholars engage in unnecessary repetition of boring, error-prone clerical tasks (the good side is that we get to know our texts!)

- Part of the solution is

  - annotated and structured corpora

  - freely available resources

Why annotated corpora?

- Building an annotated corpus means making a range of decisions, which are inherently informed by theory, whether we like it or not

- General corpus linguistics is in practice not a strictly empirical endeavour

- Linguistic categorisation reflects linguistic theory: words grouped into lexemes, morphological analysis, syntactic function

- We need to be explicit about out assumptions

Building annotated corpora

- Typical problem areas
  - The overall architecture: the annotation scheme
  - The tools: the annotation process
  - The afterlife: preservation and ease of use
- The specific example I will use is treebanks, i.e. corpora with (morpho)syntactic annotation, and experience with the *PROIEL-family of treebanks*

Building annotated corpora

- Many tricky decisions to make + severe resource constraints, e.g.

    1. Decide on annotation schemes that balance theoretical concerns and level of detail

    2. Choose tools for annotation that keep the annotation speed up but the error rate down

    3. Commit to making raw data and detailed documentation easily available today and forever

    4. Make preprocessed data available for typical tasks (e.g. searching for word forms or producing an electronic dictionary)

The PROIEL-family of treebanks

- The original PROIEL Treebank (Haug and Jøhndal 2008) stems from a research project called *Pragmatic Resources in Old Indo-European Languages* (PROIEL) at the University of Oslo (2008–2012)

- Aimed at studying word order, anaphoric expressions, definiteness, background events and discourse particles cross-linguistically in ancient Indo-European languages

- Used the New Testament in its original and in translation since this is a natural parallel text

The PROIEL-family of treebanks

- The corpus was designed with these research questions in mind but was also intended to be open-ended and maintainable in the long term

- Several ‘daughter’ projects have built on this work (ISWOC, TOROT, Menotec) (Eckhoff et al. *to appear*)

- Now an integrated collection of treebanks with the same annotation system

The PROIEL-family of treebanks

Ancient Greek	246,783
Classical Armenian	23,513
Gothic	57,211
Old Church Slavonic	126,556
Latin	170,306
Old English	29,406
Old French	2,340
Portuguese	36,415
Spanish	54,661
Old Russian	180,994
	928,185

The PROIEL-family of treebanks

- Still expanding:
  - A lot more Ancient Greek and Latin in the pipeline
- Pan-Indo-European ambitions:
  - Sanskrit (*Śatapatha-Brāhmaṇa*)
  - Hittite (*New Hittite Letters*)
  - Lithuanian (Baltramiejus Vilentas, *Evangelijos bei Epistolos*)

The PROIEL-family of treebanks

- A corpus for *linguists*: linguistically relevant annotation
- Low-resourced languages: We have to do a lot manually
- Annotation must be consistent across all languages for cross-linguistic comparison to make sense
- Annotators are trained centrally
- Reviewers enforce an annotation system that encode comparable structures in the same way regardless of language

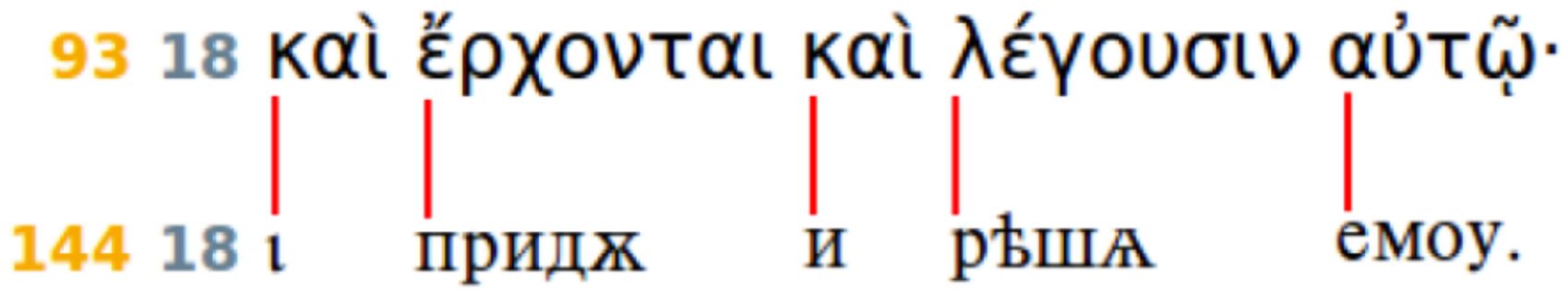
The PROIEL-family of treebanks

- Several levels of annotation
    - The text itself: normalised; split into sentences and words; translated texts aligned
    - Morphology
    - Syntax
    - Information structure
    - Some semantic annotation (e.g. aspect/Aktionsart)

The PROIEL-family of treebanks: The text

- The electronic text is generally kept the way it is, reflecting the underlying printed edition and its orthographic conventions

- Parallel texts are aligned word for word



The PROIEL-family of treebanks: The text

- The text is tokenised: paragraphs split into sentences, sentences into tokens (words/clitics/some affixes):

|97 [5]OI. Quid tibi negotist mecum?

Quid [0] *tibi* [1] *negoti* [2] *st* [3] *me* [4] *cum* [5]

- This can be complicated if the orthographic conventions of the language do not include word boundaries:

vratamupaiṣyan | antareṇāhavanīyaṃ ca gārhapatyaṃ ca prāṅ tiṣṭhannapa upaspṛśati

1.1.1.[1] [0]vratam upaiṣyan

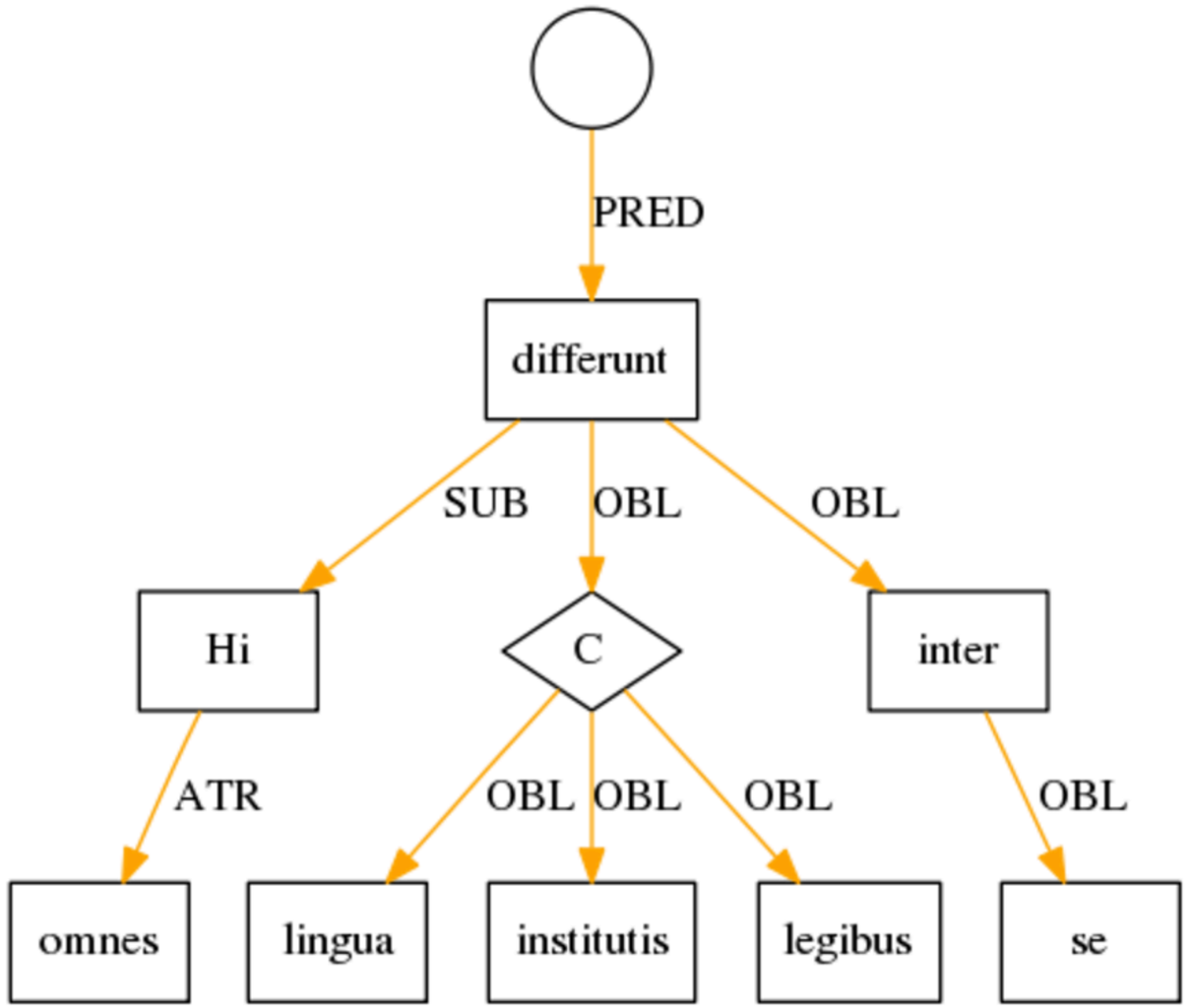
antareṇa āhavanīyam ca gārhapatyam ca prāṅ tiṣṭhan apaḥ upaspṛśati

The PROIEL-family of treebanks: Syntax

- A thorny issue because syntacticians disagree on theoretical fundamentals

- Choice of formalism and primitives have significant consequences down the line for the type of research the corpus can be used for

- PROIEL uses dependency grammar (DG)



The PROIEL-family of treebanks: Syntax

- PROIEL's version of DG is influenced by Lexical Functional Grammar

- Concepts like *subject* and *object* are primitives

Label	Function
PRED	Predicate
SUB	Subject
OBJ	Object
OBL	Oblique
AG	Agent
ADV	Adverbial
ATR	Attribute
APOS	Apposition
NARG	Nominal argument
XADV	Free predicative
XOBJ	Open complement
AUX	Auxiliary
XOBJ	Open complement clause
COMP	Complement clause
PART	Partitive
PARPRED	Parenthetical
VOC	Vocative

The PROIEL-family of treebanks: Syntax

- DG is dominant in computational linguistics due to its simplicity and efficiency

- DG is a good choice for early Indo-European, many of which have less rigid word order, because it does not embed phrase-structural information in the annotation

  - Does Latin have a VP? Not possible to test this if the annotation already assumes it does

  - Fewer difficult decisions for annotators to make

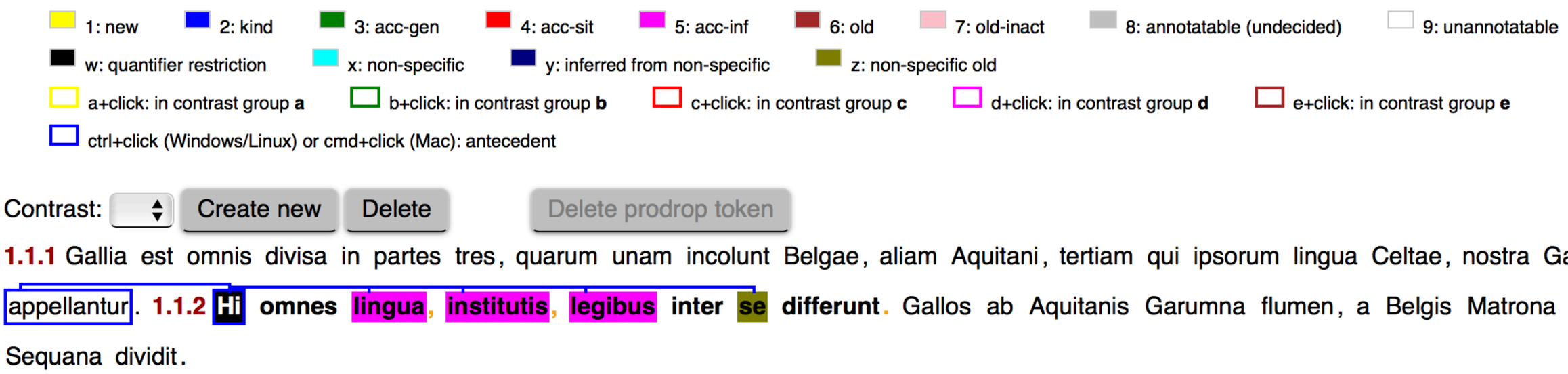
The PROIEL-family of treebanks: Morphology

Hi	***omnes***	***lingua***	***institutis***	***legibus***	***inter***	***se***	***differunt***
dem. pron.	indef. pron.	common noun	common noun	common noun	prep.	pers. refl. pron.	verb
nom., pl., m.	nom., pl., m./f.	abl., sg., f.	abl., pl., n.	abl., pl., f.	non-infl.	acc., 3rd p., pl., m./f./n.	ind., pres., act., 3rd p., pl.
hic	*omnis*	*lingua*	*institutum*	*lex*	*inter*	*se*	*differo*

- Morphological analysis is less controversial
- PROIEL uses *part of speech* (POS), a positional *morphological tag* and a lemma
- Here too important decisions: What constitutes a unique lemma?

inflection	mood
tense	voice
degree	case
person	number
gender	strength

The PROIEL-family of treebanks: Information structure



- Givenness tags, which are based on which context the hearer uses to establish reference

- Links to antecedents for NPs whose givenness tag implies an anaphoric relationship

The annotation process

- The rationale for a custom, web-based annotation tool:

1. PROIEL needed international expertise (often students)

   - Annotators should not have to install software and should be able to work whenever and wherever they want

2. We had to build the tools while annotating

   - The software should be continuously updated without interrupting annotators

The annotation process

- How do we speed up manual annotation?
  - Experiments tend to show that annotators work faster if given help by predictive tools
    - Annotators also make fewer mistakes this way
- Two general approaches:
  - Rule-based methods
  - Statistical methods

Rule-based morphological analysis

```
swan:latmorph-foma(master) $ flookup latin.bin
amatus
credo
armiger
mentem
amatus   amō+VERB+ppp+masc+sg+nom

credo    credō+VERB+pres+ind+act+sg+p1

armiger armiger+ADJ+masc+sg+nom
armiger armiger+ADJ+masc+sg+voc

mentem   mens+NOUN+fem+sg+acc
```

Rule-based morphological analysis

Enumerated inflectional affixes

```
64
65  LEXICON MorphPNPerfect          ! perfect person-number endings
66    +act+sg+p1:+ī                 #;
67    +act+sg+p2:+istī              #;
68    +act+sg+p3:+it                #;
69    +act+pl+p1:+imus              #;
70    +act+pl+p2:+istis             #;
71    +act+pl+p3:+ērunt             #;
72    +act+pl+p3+VAR:+ēre           #;
73
```

Morphological and phonological replace rules

```
181
182  define VowelDeletion  [a|e|o] -> 0 || _ %+ V ;
183
```

Rule-based morphological analysis

Word-derivation rules

```
LEXICON COMPior1
  +comp:0 COMPior1B;
LEXICON COMPior1B
  0:ior InflACCS;
  0:ius InflACNS;
  0:ior InflACL;
LEXICON SUPLimus1
  +supl:0 SUPLimus1B;
LEXICON SUPLimus1B
  0:im InflA12;
  0:um InflA12;
```

…and a *large* lexicon

Rule-based morphological analysis

- This rule-based approach uses *finite-state transducers*, which are well-understood, scalable and fast

- Building them is very work intensive

- The system can guess unknown words based on what a likely stem is

- But this particular method gives all possible analyses; it does not *disambiguate* analyses in context

Machine learning in NLP

- *Machine learning* can be used for many tasks in Natural Language Processing (NLP):

    - Tokenisation: splitting a paragraph into sentences, a sentence into words, a word into morphemes

    - Part-of-speech (POS) and morphological tagging

    - Named-entity recognition: identifying people, places etc.

    - Chunking and parsing

Machine learning in NLP

- The canonical method for statistical tagging and parsing uses *supervised machine learning*:

    1. The system is given a *training set* which consists of an input with *features* and their correct *labels*
    2. The system, using a machine-learning algorithm, produces a classifier that can assign labels to new inputs with features

- In other words: The system is given the correct answers for part of the data, uses this to induce a model that can generalise to new, unseen data

Statistical tagging

State-of-the-art POS tagging for English
(per-token accuracy) using neural networks

Model	News	Web	Questions
Ling et al. (2015)	97.44	94.03	96.18
Andor et al. (2016)*	97.77	94.80	96.86
Parsey McParseface	97.52	94.24	96.45

https://github.com/tensorflow/models/tree/master/syntaxnet

Statistical tagging

Full morphological tagging of Latin using *TnT tagger*

Experiment	TA	OOV	IV
Poudat and Longrée (2009)[a]	84.3	?	?
Poudat and Longrée (2009)[b]	63.7	?	?
Poudat and Longrée (2009)[c]	77.2	?	?
Skjærholt (2011)[d]	84.3	60.7	88.9
Skjærholt (2011)[e]	62.8	33.3	85.0
Vulgata & *BG* on *Att*	76.9	50.0	85.7

[a] LASLA, *BG* books 1–2,4–7 on book 3
[b] LASLA, *BG* and *Bellum Civile* on *1st Catilinarian*
[c] LASLA, historical texts on *1st Catilinarian*
[d] PROIEL, *BG* 10-fold cross-validation
[e] PROIEL, trained on *BG*, tested on *Vulgata*

Table 4: Tagging accuracy (in percent) on Latin. Token accuracy (TA), out-of-vocabulary (OOV) and in-vocabulary (IV) accuracy.

Skjærholt (2011: 160)

Statistical tagging

- The differences are due to
  - the model used (i.e. the tool and training method),
  - the annotation system used (e.g. granularity),
  - the size of the training set

Statistical tagging

State-of-the-art POS tagging (and parsing)
using neural networks and *Universal Dependencies*

Language	No. tokens	POS	fPOS	Morph	UAS	LAS
Ancient_Greek-PROIEL	18502	97.14%	96.97%	89.77%	78.74%	73.15%
Latin-PROIEL	14906	96.50%	96.08%	88.39%	77.60%	70.98%

https://github.com/tensorflow/models/blob/master/syntaxnet/universal.md

Statistical tagging

- Historical corpora tend to be small
  - All Latin until c. AD 600: c. 10,000,000 words
- Diachronic depth can also be an issue
  - All Greek until AD 1453: c. 100,000,000 words
- An annotated corpora will be a much smaller subset

Statistical tagging

- Ways to squeeze more out of small training sets:
    - Normalise spelling: map spelling variation to some form of normalised spelling
        - Train using modern form of the language, then apply to the historical form
- These solutions do not appeal to everyone, and one can choose more or less extreme approaches

Statistical tagging

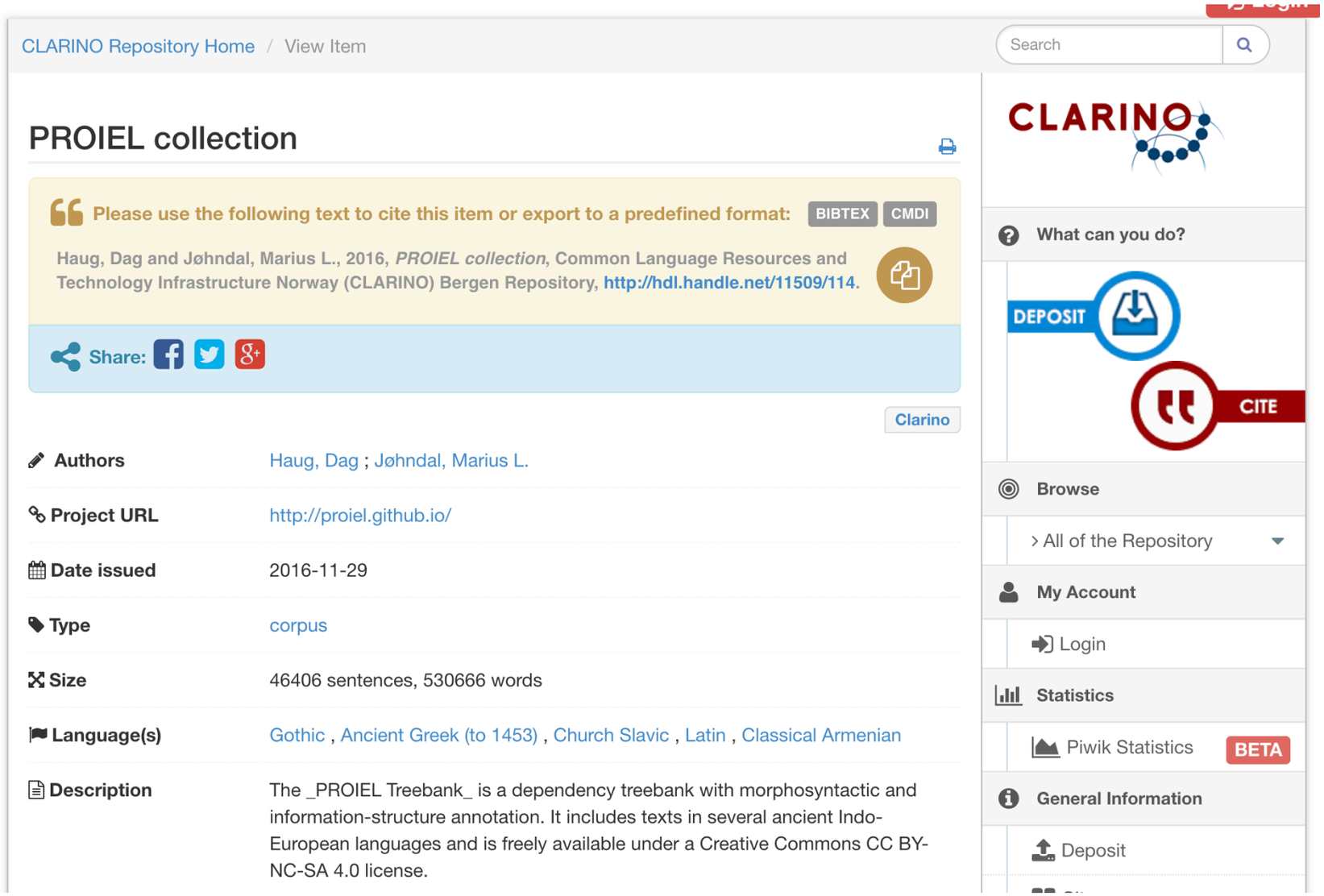
- For texts, whose orthography show a lot of variation, normalisation before training and tagging improve results:
  - Slavic: 89.5% for POS; 81.5% for ten-field morphology (Berdičevskis et al. 2016)
- Enlarging the training set also helps despite internal variation:
  - Byzantine Greek trained on Ancient Greek, Koine and Byzantine Greek: 91.3% for POS tagging; 94.0% for ten-field morphology (Birnbaum and Eckhoff *to appear*)

Preservation

- A finished corpus needs to be available somewhere. Universities love to reorganise their web pages, but dead links help nobody.
- Researchers need access to the right version of the corpus to replicate a study or make corrections.
- The raw data must be available and readable.
- These are mostly solved problems!

Preservation

- Finished corpora should be deposited with trusted third parties along with metadata



Preservation

- Open-source software processes work very well for versioning annotated corpora…

12 cic-att.conll View

@@ -14731,7 +14731,7 @@

14731	14731	1	Terentiae	Terentia	N	Ne	NUMBs|GENDf|CASEd	2	obl	_	_
14732	14732	2	pergrata	pergratus	A	A-	NUMBs|GENDf|CASEn|DEGRp	3	xobj	_	_
14733	14733	3	est sum	V	V-	PERS3|NUMBs|TENSp|MOODi|VOICa	0	pred	_	_	
14734		-4	adsiduitas	adsiduitas	N	Nb	NUMBs|GENDf|CASEn	6	sub	_	_
	14734	+4	adsiduitas	assiduitas	N	Nb	NUMBs|GENDf|CASEn	6	sub	_	_
14735	14735	5	tua tuus	P	Ps	PERS2|NUMBs|GENDf|CASEn	6	atr	_	_	
14736	14736	6	et et	C	C-	INFLn	3	sub	_	_	
14737	14737	7	diligentia	diligentia	N	Nb	NUMBs|GENDf|CASEn	6	sub	_	_

@@ -15163,7 +15163,7 @@

15163	15163	14	non non	D	Df	INFLn	15	aux	_	_	
15164	15164	15	oportuit	oportet	V	V-	PERS3|NUMBs|TENSr|MOODi|VOICa	11	apos	_	_
15165	15165	16	aliquando	aliquando	D	Df	INFLn	17	adv	_	_
15166		-17	revertamur	reverto	V	V-	PERS1|NUMBp|TENSp|MOODs|VOICp	10	pred	_	_
	15166	+17	revertamur	revertor	V	V-	PERS1|NUMBp|TENSp|MOODs|VOICp	10	pred	_	_
15167	15167										

Preservation

- …and for making regular, scheduled releases

Latest release

20160607
07b7445

20160607 release

Edit

mlj released this on 7 Jun

This release updates the whole collection to PROIEL XML 2.1 and adds alignments to the New Testament texts in the collection.

It also adds some sentences missing from previous releases of Sphrantzes' *Chronicles* and Cicero's *Letters to Atticus*, corrects minor inconsistencies in the Latin and Greek lemmatisation and a few errors in *Codex Marianus*.

Downloads

Source code (zip)

Source code (tar.gz)

20160118
1568452

20160118 release

Edit

mlj released this on 18 Jan · **2 commits** to master since this release

This release adds the remaining parts of Sphrantzes' *Chronicles* along with a few annotation corrections to other texts.

Downloads

Source code (zip)

Source code (tar.gz)

Preservation

- There are many things to keep in mind to ensure that data remains readable in the future.

- Rules of thumb:

1. *Never* use closed, proprietary file formats; *always* use open, standardised file formats

2. Prefer raw data over derived data

3. Follow *de facto* conventions; nobody cares about your personal preferences

4. Keep things simple!

Desiderata for freely reusable corpora

1. Raw data can be downloaded
2. Comprehensive documentation freely available online
3. Available without user registration, signing of contracts etc.
4. Developed using free/open source software to allow for transparent replication
5. Developed openly using an online version control system
6. Regular, scheduled releases with numbered versions
7. Can be modified and improved on by anyone without special permission
8. Free for academic use
9. Free for commercial use
10. Released under a free and standard license such as GPL, LGPL or CC

Rögnvaldsson et al. (2012: 1982)

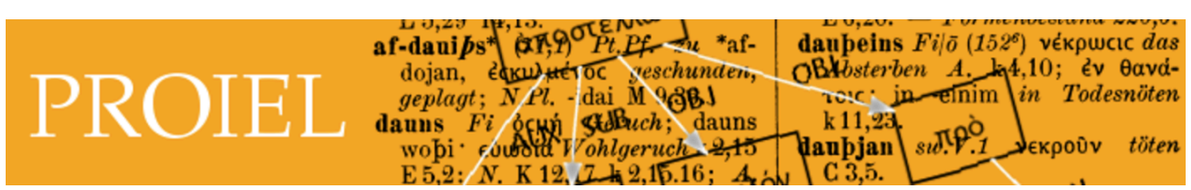
Treebanks for historical linguists

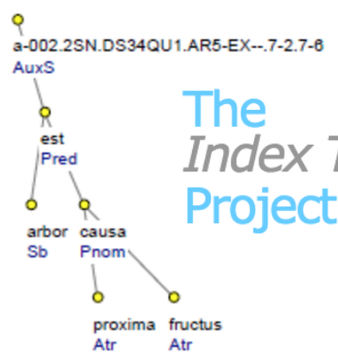
Penn Parsed Corpora of Historical English

The Ancient Greek and Latin Dependency Treebank

Icelandic Parsed Historical Corpus (IcePaHC)

The Parsed Old and Middle Irish Corpus (POMIC)



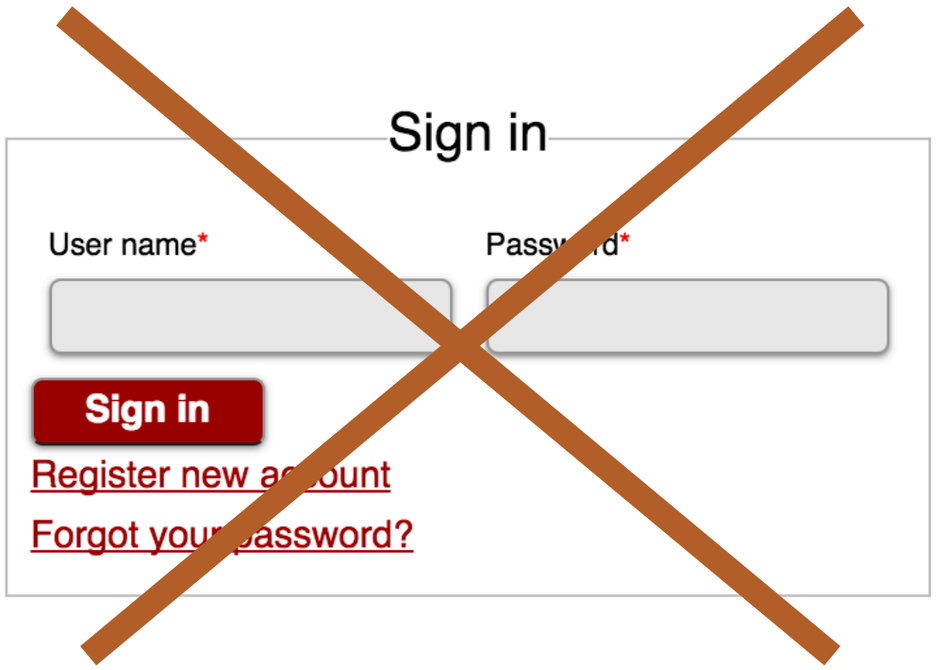


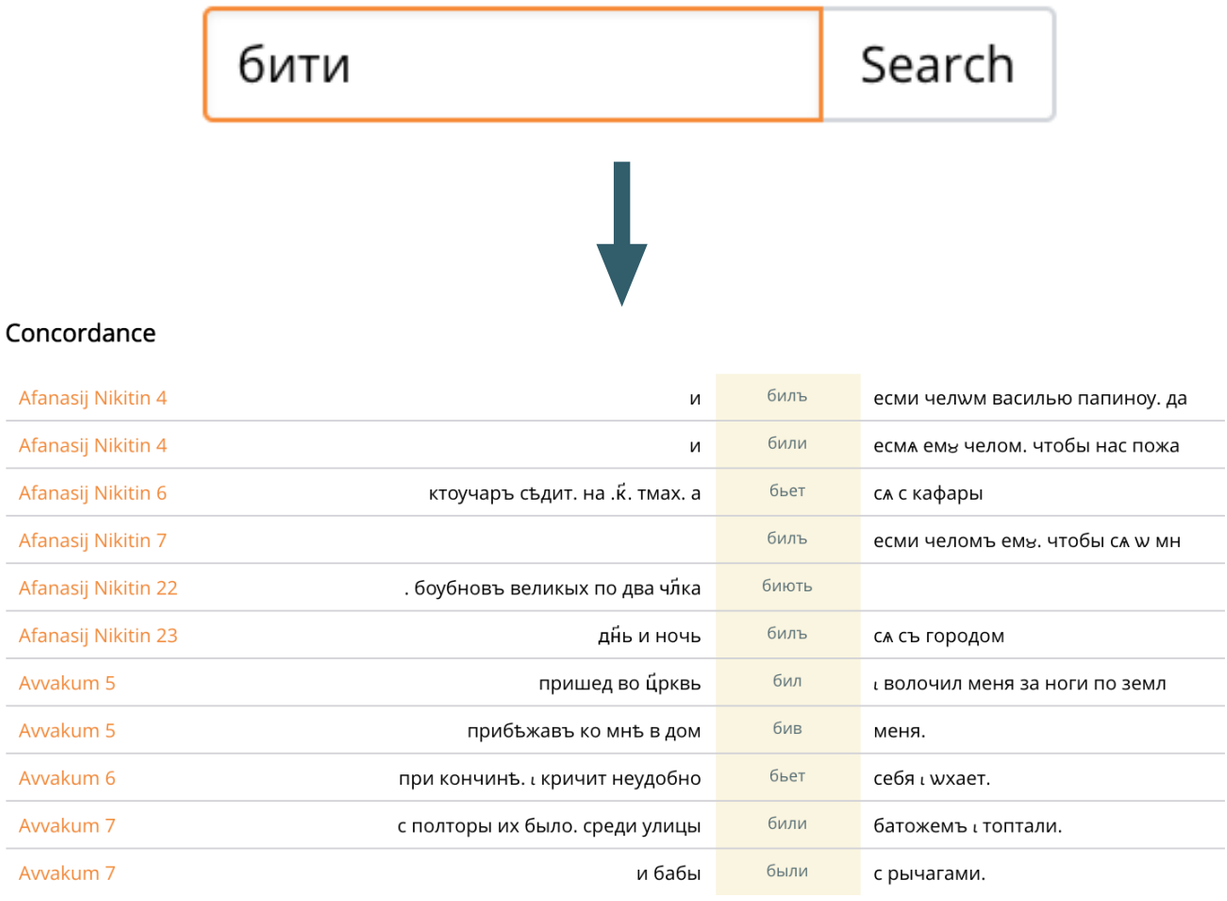
The *Index Thomisticus* Treebank Project

'arbor est causa proxima fructus'

Making the PROIEL-family of treebanks easier to use

- We are removing all registration walls…

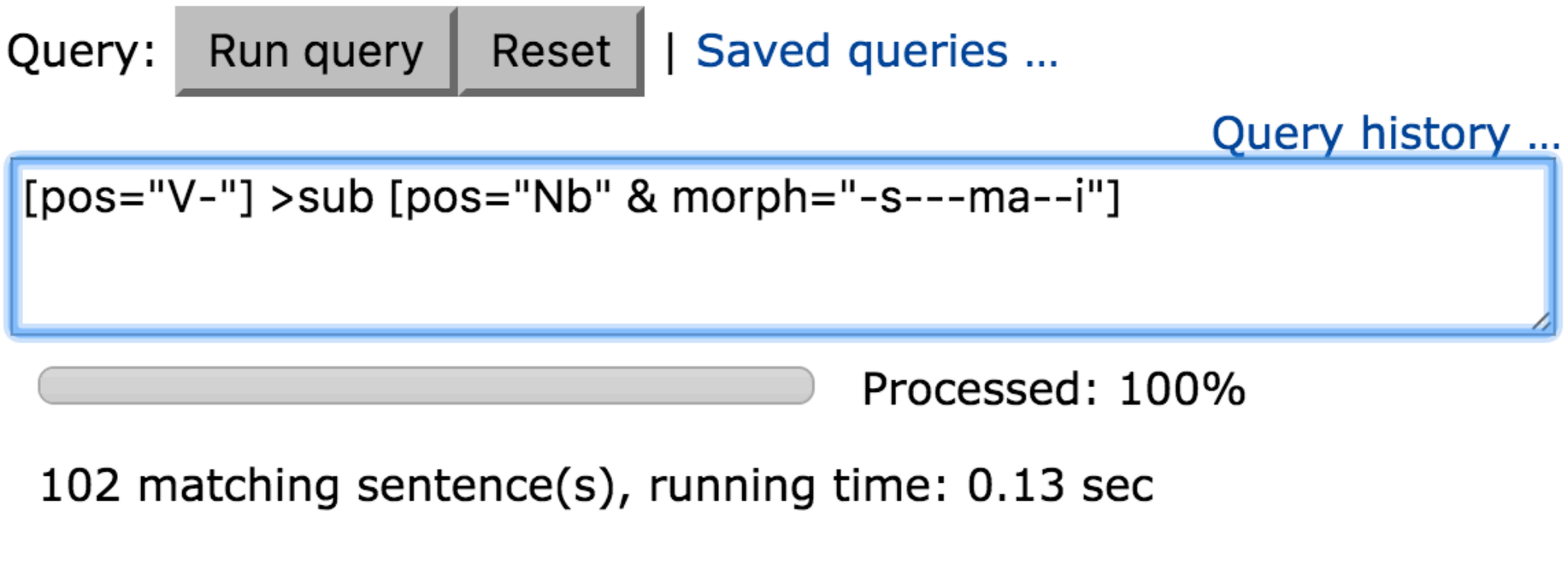




- …and replacing them with a simple search box

Making the PROIEL-family of treebanks easier to use

- Also integrating advanced queries in the same web-frontend
- The current method involves using *INESS Search*



Making the PROIEL-family of treebanks easier to use

- We also serve pre-processed, derived data
  - Automatically generated dictionaries
  - Paradigms with actually attested forms
  - Chronological charts
  - Valency lexica

Making the PROIEL-family of treebanks easier to use

везти

Old Russian, verb

- Definition
- Concordance
- Paradigm
- Chronology
- Valency

Valency

Arguments	Non-reflexive	Reflexive
(none)	6	
OBJ (genitive)	2	
OBJ (accusative)	2	
OBJ (accusative) OBL (dative)	1	
OBJ (accusative) OBL (preposition *къ* + dative)	1	
OBL (preposition *до* + genitive)	1	
OBL (adverb *тудꙑ*)	1	
OBL (preposition *отъ* + genitive) OBL (preposition *до* + genitive)	1	

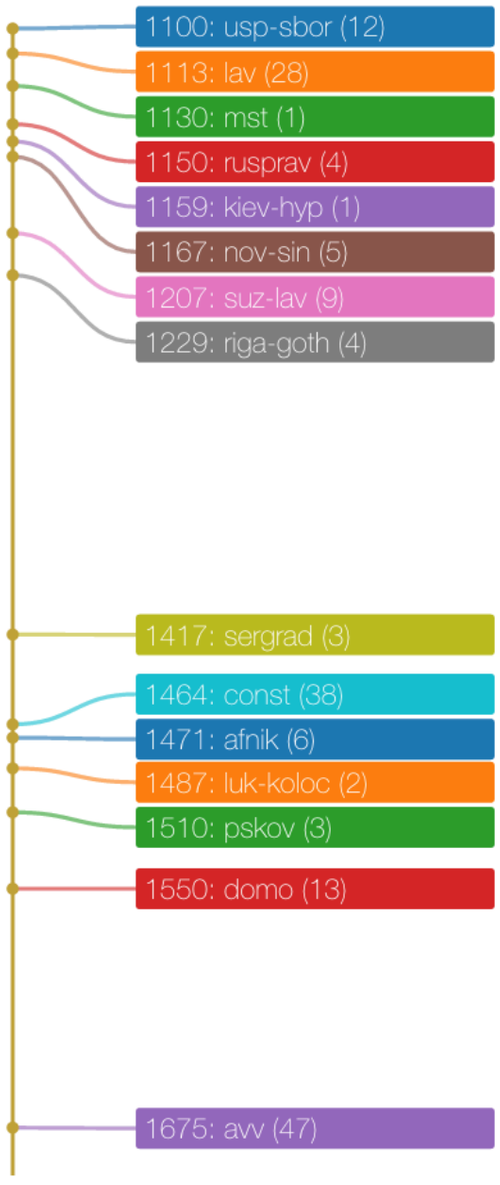
Making the PROIEL-family of treebanks easier to use

Paradigm

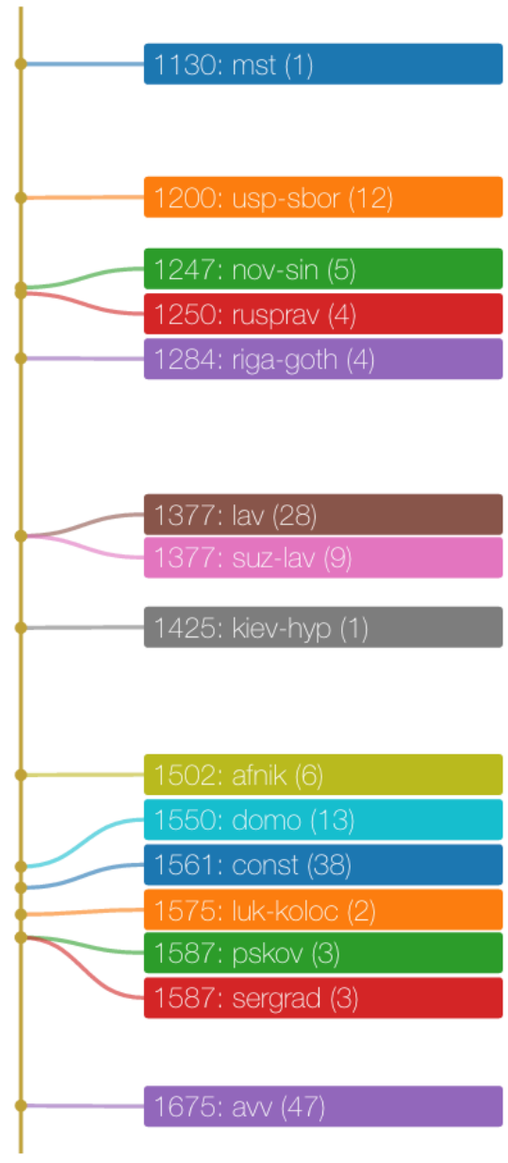
	Present	Imperfect	Aorist
1st p. sg.	бью (2)		
2nd p. sg.	бьеш (1) бїеши (1) бїи (1)		
3rd p. sg.	бьет (4) биєть (4) бъет (2) биет (2) бьетъ (2) бьєт (1) бїетъ (1) бьєть (1)	бьаше (2) бьѧше (1) бияшѣ (1)	би (2)
1st p. du.	бьевѣ (1)		
2nd p. du.			
3rd p. du.			
1st p. pl.	бьем (1) бьемъ (1) биємъ (1)		
2nd p. pl.	беите (2) биите (1)		
3rd p. pl.	бьют (2) бьютъ (2) бьють (2) биють (1)	бьѧхѫ (8) бїахѫ (2) бьахѫ (1) бьяхут (1)	биша (9)

Making the PROIEL-family of treebanks easier to use

Chronology by composition



Chronology by manuscript



Making the PROIEL-family of treebanks easier to use

- Web technology is advancing very rapidly
- These things are much easier to make today than they were just a couple of years ago
- But we are still nowhere near having off-the-shelf tools
- You *will* need a programmer on your team

REFERENCES

- Berdičevskis, Eckhoff and Gavrilova, T. 2016. ‘The beginning of a beautiful friendship: rule-based and statistical analysis of Middle Russian’. In *Computational Linguistics and Intellectual Technologies. Proceedings of Dialogue* 16. Moscow.

- Birnbaum and Eckhoff (to appear). *Machine-assisted multilingual alignment of the Codex Suprasliensis*.

- Davison. 1989. ‘New Testament Greek Word Order’. *Literary and Linguistic Computing* 4, 19–28.

- Eckhoff, Bech, Bouma, Eide, Haug, Haugen and Jøhndal (University of Oslo) (to appear). *The PROIEL treebank family: a standard for early attestations of Indo-European languages.*

- Haug. 2015. ‘Treebanks in historical linguistic research’. In C. Viti (ed.). *Perspectives on Historical Syntax.* Amsterdam: Benjamins. 185—202.

- Haug and Jøhndal. 2008. ‘Creating a Parallel Treebank of the Old Indo-European Bible Translations’. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*. European Language Resources Association.

- Kirk. 2012. *Word order and information structure in New Testament Greek*. Ph.D.dissertation, Leiden.

- Piotrowski. 2012. *Natural Language Processing for Historical Texts.* Synthesis Lectures on Human Language Technologies. San Rafael, California: Morgan and Claypool.

- Rife. 1933. ‘The mechanics of translation Greek’. *Journal of Biblical literature* 52, 244–252.

- Rögnvaldsson, Ingason, Sigurðsson and Wallenberg. 2012. ‘The Icelandic Parsed Historical Corpus (IcePaHC)’. In Calzolari, Choukri, Declerck, Doğan, Maegaard, Mariani, Moreno, Odijk and Piperidis (eds.). *Proceedings of LREC, Istanbul 2012*. Istanbul, Turkey: European Language Resources Association. 1977–1984.

- Skjærholt. 2011. ‘More, faster: Accelerated corpus annotation with statistical taggers’. *Journal for Language Technology and Computational Linguistics* 26 (2): 153—165.

ONLINE CORPORA, TREEBANKS & TOOLS MENTIONED HERE

- Corpus of Historical Low German: http://www.chlg.ac.uk/
- Penn Parsed Corpora of Historical English: https://www.ling.upenn.edu/hist-corpora/
- Icelandic Parsed Historical Corpus: http://www.linguist.is/icelandic_treebank/
- The Parsed Old and Middle Irish Corpus: https://www.dias.ie/celt/celt-publications-2/celt-the-parsed-old-and-middle-irish-corpus-pomic/
- The Ancient Greek and Latin Dependency Treebank: https://perseusdl.github.io/treebank_data/
- The Index Thomisticus Treebank: http://itreebank.marginalia.it/
- The PROIEL Treebank: http://proiel.github.io
- The ISWOC Treebank: https://iswoc.github.io/
- The TOROT Treebank: http://torottreebank.github.io/
- Foma (finite-state compiler and library backwards compatible with the proprietary Xerox Finite-State Tools): https://fomafst.github.io/
- Graphviz (graph visualiser often used in computational linguistics): http://www.graphviz.org/
- TnT tagger (statistical POS tagger often used for historical languages): http://www.coli.uni-saarland.de/~thorsten/tnt/
- SyntaxNet (state-of-the-art neural network framework for TensorFlow): https://github.com/tensorflow/models/tree/master/syntaxnet
- Universal Dependencies (a dependency-grammar standardisation effort): http://universaldependencies.org/

Appendix: Annotation speed

- *Realistic* estimates of *average* annotation speeds are given in the table below

Latin (`lat`)	**125 tks/hr**
Ancient Greek (`grc`)	**125 tks/hr**
Old Norwegian (`non`)	**110 tks/hr**

- Speeds vary substantially between experienced and inexperienced annotators and depend on the complexity of the text and the extent to which annotators are assisted by automatic tagging.s