

The PROIEL treebank family: a standard for early attestations of Indo-European languages

Hanne Eckhoff (UiT The Arctic University of Norway), Kristin Bech (University of Oslo), Gerlof Bouma (University of Gothenburg), Kristine Eide (University of Oslo), Dag Haug (University of Oslo), Odd Einar Haugen (University of Bergen) and Marius Jøhndal (University of Oslo)

Abstract

This article describes a family of dependency treebanks of early attestations of Indo-European languages originating in the parallel treebank built by the members of the project *Pragmatic Resources in Old Indo-European Languages* (PROIEL). The treebanks all share a set of open-source software tools, including a web annotation interface, and a set of annotation schemes and guidelines developed especially for the project languages. The treebanks use an enriched dependency grammar scheme complemented by detailed morphological tags, which have proved sufficient to give detailed descriptions of these richly inflected languages, and which have been easy to adapt to new languages. We describe the tools and annotation schemes and discuss some challenges posed by the various languages that have been annotated. We also discuss problems with tokenisation, sentence division and lemmatisation, commonly encountered in ancient and mediaeval texts, and challenges associated with low levels of standardisation and ongoing morphological and syntactic change.

1. Introduction

This article¹ describes a family of dependency treebanks of early attestations and historical varieties of Indo-European languages originating in the parallel treebank built by the members of the project *Pragmatic Resources in Old Indo-European Languages* (PROIEL). The treebanks all share a set of open-source software tools, including a web annotation interface, and a set of annotation schemes and guidelines developed especially for the project languages. We describe the tools and annotation schemes and discuss some challenges posed by the various languages that have been annotated.

The treebanks use an enriched dependency grammar scheme complemented by detailed morphological tags. The use of dependency grammar allows us to posit structural equivalence with differing word order, which is useful both in situations of extremely free word order and diachronic change, both of which are discussed in the article. The annotation scheme follows in the tradition of the Prague Dependency Treebank scheme by using a formalism that is more expressive than the schemes that are used in most treebanks and in current statistical dependency parsing. However, unlike the PDT scheme, we integrate empty nodes and secondary dependencies in a single annotation layer.

The scheme has proved sufficient to give detailed descriptions of the richly inflected project languages, and has also been easy to adapt to new languages.

¹ Work by Gerlof Bouma was supported by the Marcus and Amalia Wallenberg foundation (MAW 2012.0146: MAPiR).

We describe the scheme and exemplify its usefulness with case studies from several of the project languages. We also discuss problems with tokenisation, sentence division and lemmatisation, commonly encountered in ancient and mediaeval texts, and challenges associated with low levels of standardisation and ongoing morphological and syntactic change. The main differences between the PROIEL annotation scheme and more “classical” examples of dependency grammar are discussed in section 9.

Given the expressivity and flexibility of the annotation schemes and tools, as well as the language coverage, we argue that the PROIEL schemes and tools serve as a standard for dependency treebanks of ancient and mediaeval attestations of Indo-European languages. The bulk of data on this format is steadily growing, and is in use both for linguistic studies and as training data for other linguistic resources. For easy accessibility and compatibility with other resources, the treebanks are convertible to the Universal Dependencies scheme, but with some loss of information.

The article is structured as follows: Section 2 describes the projects in the PROIEL treebank family. Section 3 describes the PROIEL software tools. Section 4 discusses text processing, sentence division and segmentation with case studies from the projects. Section 5 describes lemmatisation and part-of-speech assignment. Section 6 deals with morphological tagging, including morphological preprocessing and ways of dealing with case syncretism. Section 7 introduces the syntactic annotation scheme and presents several case studies demonstrating the flexibility of the scheme across the project languages. Section 8 briefly presents other levels of annotation. Section 9 discusses compatibility with other annotation schemes. Section 10 presents the conclusions.

2. The projects

The work on the PROIEL treebank family originated in the project *Pragmatic Resources in Old Indo-European Languages* (University of Oslo, PI Dag Haug).² A central aim of this project was to establish a parallel treebank³ of the oldest Indo-European New Testament translations, since this is the oldest and most extensive natural parallel corpus available for these languages. The selected project languages were Greek (the NT source text) and the earliest translations into Latin (Vulgate), Gothic (Wulfila), Classical Armenian and Old Church Slavonic (OCS; Codex Marianus). For the three latter languages, the NT translations are also the first written attestations of the language, and thus carry extra weight.

As the project title suggests, the PROIEL project sought to study the resources that the grammar makes available for structuring information in a text, such as word order, definiteness marking, pronominal reference, discourse particles and the use of participles for backgrounding purposes. In order to be able to study these phenomena in the necessary depth, the parallel treebank was created and

² <http://www.hf.uio.no/ifikk/english/research/projects/proiel/>,
<http://proiel.github.io/>

³ [foni.uio.no:3000](https://doi.org/10.1111/1365-3113.00000)

annotated not only with morphological and syntactic annotation, but also tagged for a number of different other features known to indicate discourse prominence, most importantly givenness status and anaphoric relations.

The PROIEL project at the University of Oslo ended in 2013, but group members have continued to work on the corpus and expand it. This work has focussed on the Greek and Latin part of the corpus and has aimed primarily at extending the corpus to cover the classical periods of Greek and Latin, which is what most scholars are interested in. Herodotus was chosen as a representative of the Classical period of Greek, even though this text is not in the Classical Attic dialect, because Herodotus has been the focus of many linguistic studies. For Classical Latin, samples of the two most important authors, Caesar and Cicero, have been annotated. Annotation of the comedies of Plautus and Terence is in progress. To a limited extent we have also tried to provide diachronic depth by covering texts from different time periods so as to enable historical investigations. In Greek, we have annotated Sphrantzes' *Chronicles* (an eyewitness account of the fall of Constantinople in 1453). In Latin we have annotated the late texts *Peregrinatio Aetheriae* and are now working on the *Opus agriculturae* by Palladius. These additions are obviously but short steps towards the ultimate goal of having a corpus that provides a balanced selection of texts from the histories of the Greek and Latin languages.

The PROIEL project has also inspired a series of new treebank projects, all of which have benefited from the PROIEL schemes and tools, and all of which have developed adaptations of the annotation schemes for new languages in cooperation with the original PROIEL project group.

The aim of the ISWOC project⁴ was to study the verb-second structures of older European varieties of the Romance languages French, Spanish and Portuguese and of the Germanic languages German, Norwegian and English. The ISWOC treebank⁵ includes texts from various periods of French, Spanish, Portuguese and English,⁶ which have different degrees of inflection at various stages: from rich case morphology and highly inflectional verbal paradigms in early stages, to less inflectional modern languages without case and with an increasing use of auxiliary verbs and participles. In order to study the development from highly inflectional to not-so-inflectional languages, there was a need for a flexible annotation format capable of capturing all stages of the evolution.

The Tromsø Old Russian and OCS Treebank (TOROT)⁷ is a direct expansion of the OCS part of the PROIEL parallel treebank, built by the members of the project *Birds and Beasts: Shaping Events in Old Russian* and its pedagogical companion

⁴ *Information Structure and Word Order Change in Germanic and Romance Languages*, <http://www.hf.uio.no/ilos/english/research/projects/iswoc/>.

⁵ <http://iswoc.github.io/>, also hosted with the PROIEL treebank at foni.uio.no:3000

⁶ The project has also made use of Menotec's Old Norwegian treebank and PROIEL's Gothic treebank.

⁷ <http://torottreebank.github.io/>, <https://nestor.uit.no>,

project *The Varangian Rus Digital Environment* (both at UiT The Arctic University of Norway). The TOROT contains additional texts in OCS (Codex Zographensis and Codex Suprasliensis), as well as Old East Slavic and Middle Russian texts. The treebank also contains a modern Russian part, the SynTagRus treebank converted into the PROIEL dependency format. The linguistic aim of the projects is to study the historical development of aspect and aktionsart in Russian.

Menotec was a three-year infrastructure project (2010–2012) aimed at transcribing and annotating a corpus of Old Norwegian texts from the 13th and early 14th century.⁸ While similar projects to a great extent have been using existing editions of texts, the Menotec project has based all their annotations on transcriptions from the primary sources themselves. The corpus includes a variety of genres and offers considerable stylistic variation. Some of the texts are translations, others are originally written in Old Norwegian. In addition to the work on transcribing texts, Menotec has released the first-ever treebank⁹ of Old Norwegian. The annotation of the Menotec treebank will be continued within the context of the Medieval Nordic Text Archive network¹⁰ and in cooperation with the PROIEL group.

Greinir skáldskapar¹¹ is a linguistically annotated corpus containing Old Icelandic poetry. It was created as a part of the project *Interfaces of Metrics, Phonology and Syntax* (2009–2011) and contains all poems of the Codex Regius manuscript of the *Poetic Edda* and a selection of skaldic poetry and rímur. The texts are annotated for various grammatical and metrical factors, such as lifts, syllable structure (heavy vs. light syllables), alliteration, morphology and syntactic structure. The syntactic annotation was carried out in the PROIEL Annotator using the PROIEL dependency scheme.

The project *Methods for Automatic analysis of Text in digital Historical Resources* (MAPiR, 2014–2016)¹² aims at developing NLP tools for Old Swedish. As part of this, a treebank is currently being annotated, first and foremost for evaluation purposes. The material for the MAPiR treebank is mainly taken from Fornsvenska textbanken (Delsing 2002). Lemma annotation is based on Söderwall's (1884–1918) dictionary, which contains about 28,000 entries. The lemma annotation also covers multiword units included in the dictionary, such as compound nouns and particle verbs. Since Old Norwegian and Old Swedish are very closely related, MAPiR annotation at the morphosyntactic levels can follow the guidelines for Old Norwegian of the Menotec project (Haugen & Øverland 2014) with only minor alterations. At the time of writing, fragments of several different texts have been annotated, the oldest manuscripts stemming from the middle of the 13th c, the newest from the 15th c. A particular challenge

⁸ <http://www.menota.org/menotec.xml>

⁹ Hosted with the PROIEL treebank at <http://foni.uio.no:3000> and also accessible through the INESS portal at <http://clarino.uib.no/iness> (select the treebanks for Old Norse).

¹⁰ <http://www.menota.org>

¹¹ <http://bragi.info/greinir/>

¹² <https://spraakbanken.gu.se/mathir>

in the material is the low level of standardisation and the state of flux in which late medieval Swedish resides, with increasing syncretism (e.g., reduction of case and gender distinctions on adjectives and nouns, loss of person distinctions on verbs) and changes in word order and word order flexibility.

Branch	Language stage	Project	Annotated tokens
Armenian	Classical	PROIEL	77,393
Germanic	Gothic	PROIEL	57,211
	Old English	ISWOC	29,406
	Old Icelandic	Greinir skáldskapar	32,599
	Old Norwegian ¹³	Menotec	227,985
	Old Swedish	MABiR	33,441
Greek	New Testament	PROIEL	146,172
	Ancient	PROIEL	110,263
	Byzantine	PROIEL	24,612
Romance	New Testament Latin	PROIEL	122,856
	Classical Latin	PROIEL	94,300
	Late Latin	PROIEL	55,135
	Old French	ISWOC	55,353
	Old Portuguese	ISWOC	63,336
	Old Spanish	ISWOC	95,190
Slavic	OCS ¹⁴	PROIEL/TOROT	164,835
	Old East Slavic	TOROT	123,161
	Middle Russian	TOROT	92,447

Table 1. Tokens with syntactic annotation in the PROIEL treebank family, by branch and language stage.

3. The PROIEL software tools

In the course of the treebanking projects in the PROIEL family, we have built a collection of software tools that assist in creating, manipulating and analysing PROIEL treebanks. This collection, the *PROIEL framework*,¹⁵ is open-source, free for anyone to use and unencumbered by restrictive software licenses. While some knowledge of programming and familiarity with a UNIX-type platform is required to install the software and adapt it to the needs of a particular treebanking project, the software has been designed with reusability in mind.

¹³ The Menotec treebank also contains approximately 30,000 tokens of text with lemmatisation and morphological annotation, but as yet no syntactic annotation (in the law manuscript Upps DG 8 I).

¹⁴ The TOROT also contains the full text of the Codex Zographensis with automatic lemmatisation and morphological annotation, partially hand-corrected. Parts of the text also have syntactic annotation.

¹⁵ <http://proiel.github.io/framework>

The most important component in the collection, and the most mature, is the PROIEL Annotator, which is used to create a new treebank. Its core functionality is to guide an annotator through the annotation of morphology and syntax, which are the two mandatory levels of annotation in the PROIEL annotation scheme.

The guide first prompts the annotator to decide on the correct lemmatisation, part of speech and morphological tags for each token. It then asks the annotator to add a syntactic analysis to the sentence and finally to confirm the consistency of the annotation. If necessary, it is possible to interrupt annotation to adjust tokenisation and sentence segmentation, but for optimal performance tokenisation should be done before annotation starts.

During the annotation task the annotator is supported by guesses produced by the application. Morphological annotation can be supported by finite-state morphology, if available, as well as feedback from existing annotation that has been verified by a reviewer. The latter method is particularly useful for bootstrapping annotation of a language for which existing language resources are lacking. If training data is available, it is also possible to run an automatic tagger. Experiments have shown that this has significant positive impact on annotation speed and accuracy during the morphological stage of annotation (Skjærholt 2014, see further section 6.1).

Since annotators first decide on the morphology, the next stage of the annotation process can exploit morphological information to guess the most probable syntactic relationship between tokens. The application supports a simple form of this based on manually crafted implicational rules that predict the dependency relation of a token from the token's morphology and its attachment in the graph.

After annotation of morphology and syntax by an annotator, a more experienced annotator can make a second pass through the text to review the annotation.¹⁶ At this point the application also supports adding ancillary annotation, such as information-structure annotation.

¹⁶ The reviewers have generally been senior project members or very experienced annotators. The number of corrections made by reviewers vary considerably depending on the accuracy and experience of the annotator as well as on the complexity of the text. For instance, the PROIEL New Testament texts generally have few corrections due to the fact that this text is extremely well supported by translations and exegesis, as well as by the fact that analyses could be compared across languages during annotation (0.1–3.5 % of the tokens were corrected for morphology or lemmatisation errors, 1.5–11.8 % of the sentences were corrected for syntactic attachment or label errors). More complicated and less supported texts had considerably more corrections. For instance, Herodotus' *Histories* (Ancient Greek, PROIEL) had 9 % of its tokens corrected for morphology or lemmatisation, while 65.5 % of the sentences were syntactically corrected. Very similarly, the *Russkaja pravda* (Old East Slavic, TOROT) had 10.8 % of its tokens corrected for morphology or lemmatisation, and 65.1 % of its sentences were syntactically corrected.

PROIEL Annotator is a web application that runs in the annotator's web browser. No software has to be installed on the annotator's own computer and the administrator of the system has full control over any adaptations and upgrades. This facilitates distributed annotation and enables annotators to start work as soon as they have been trained.

While the application itself uses an SQL database to store annotation, the end product is a collection of XML files. The administrator can export all annotation as PROIEL XML, which serves as an interchange format and as the authoritative representation of a PROIEL treebank. During export some information is lost, such as a detailed audit trail and timestamps, but all linguistically relevant information is preserved.

From a software-development perspective the focus of the PROIEL framework is now on long-term maintainability and on integrating additional tools for manipulation of PROIEL treebanks and on analysis of existing treebank data. The framework includes several tools whose purpose is to fill this gap. The tools consume PROIEL XML and can be installed independently of the web application.

The framework includes a library for building custom analysis scripts. The library is written in the programming language Ruby and requires programming knowledge to use. For more routine tasks, such as converting a treebank to another format, the framework also includes a command-line utility. This utility also supports certain information extraction tasks. It can be used, for example, to extract morphological annotation for training a statistical tagger like TnT or Hunpos.

PROIEL Reader is an online front-end that allows end-users to browse stable, versioned releases of a treebank, and, if configured for this purpose, it can provide permanent links to individual annotated treebank objects suitable for crossreferencing and citation.

The framework also includes a query tool. The tool uses the TigerQuery formalism (König and Lezius 2003) and builds on work done by the Stockholm TreeAligner project.¹⁷

At the time of writing these tools remain less mature than the original PROIEL Annotator. As we gain more experience using them, we will document standard workflows for the maintenance and analysis phases of treebanking so that these can easily be applied by others.

4. Text processing, sentence division and segmentation

¹⁷ Currently, the PROIEL, Menotec, ISWOC and TOROT treebanks are also available for syntactic query in the INESS treebank facility, <http://clarino.uib.no/iness/page>.

The PROIEL system requires us to divide the text into sentences and word tokens. In this section we discuss some challenges to sentence division and tokenisation.

4.1 Sentence division: Introductory conjunctions vs. real coordination

In ancient and mediaeval manuscripts, sentence division is rarely signalled uniquely by punctuation, in fact, punctuation is often reserved for signalling smaller syntactic units. The annotators are therefore left to make their own decisions or to follow decisions made by previous editors, based on linguistic cues.

A problematic case is the abundant use of conjunctions in most of the texts in the PROIEL treebank family, here exemplified by OCS and Old East Slavic *i* ‘and’. In some cases, it is fairly clear that the conjunction serves to introduce a new sentence, as in (1):

- (1) *tъgda* *gla* *čvku.* *prostъri* *rъko*
 then say.AOR3SG man.DAT stretch-out.IMP2SG hand.ACC
 i *prostъrě*
 and stretch-out.AOR3SG
 “Then he said to the man: ‘Stretch out your hand.’ And he stretched out his hand.” (Mar. Matt. 12:13, 60036)¹⁸

In such cases, the sentence is split before the conjunction and the conjunction is taken to serve as an introductory particle with the relation label AUX.

However, in many other cases, it is unclear whether the conjunction serves to coordinate the two sentences or whether it just introduces the second one, for instance in a typical Old East Slavic chronicle text context such as (2).

- (2) *i* *pride* *къ* *smolenъsku* *съ* *kriviči .*
 and come.AOR3SG to Smolensk.DAT with Krivichi.INST
 i *prija* *gradъ.* *i* *posadi* *mužъ*
 and take.AOR3SG city.ACC and place.AOR3SG man.ACC
 svoi
 his
 “And he came to Smolensk with the Krivichi and took the city and placed his man there” (PVL 23.1–2, 123859)

Given the pervasiveness of null subjects in Old East Slavic, there is no principled way to decide whether this is one, two or three sentences. The projects have dealt with this problem in different ways, some leaning on the decisions of previous editors, others settling for a policy of making the sentences as short as possible, allowing coordinations only when there are clear linguistic arguments in favour of such an analysis, for instance when the two potential conjuncts

¹⁸ All examples are given with a text reference and a sentence ID in the relevant treebank, if they are publicly available.

clearly share arguments apart from null subjects. For an NLP-based approach to sentence division in the Old Swedish material, see Bouma and Adesam 2013.

4.2 Tokenisation: Mesoclis in Portuguese

Clitics are an obvious challenge to word tokenisation, since they often produce units that are rendered as a single word orthographically, but which contain more than one syntactic unit. In the PROIEL annotation, we solve the problem using the interplay between tokenisation and token presentation. Clitics are taken to be separate word tokens, and treated as such in the syntax, but their cliticness can be indicated by the fact that the preceding token is not followed by a space in its presentation form, as well as by lemmatising e.g. clitic pronouns separately from regular pronouns or by token-level customised tags. In the following, we describe an example solution from Old Portuguese.

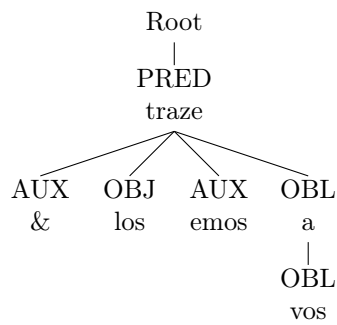
Portuguese clitics can appear in proclitic, enclitic and, more so in historical texts than in modern texts, in mesoclitic positions with future and conditional forms. Proclisis and enclisis occur in different syntactic environments, and may include one or two clitics, cliticised onto each other. Enclisis and mesoclis often involve assimilation of the final consonant of the verb and resegmentation, as in (4a), where the final *-r* has been assimilated to *-l* and *-l* is considered the onset part of the clitic. (<*dallo* < *dar(e) illu-*) The form *lho* in (3b) and (4b) is composed of the dative clitic *lhe* and the object clitic *o* and is segmented into two separate forms in the syntactic/morphological analysis.

- | | | | | |
|-----|----|-------------------|----|--------------------------|
| (3) | a. | o dar | b. | lho dar |
| | | it.ACC give | | him-it give |
| (4) | a. | dá-lo | b. | dar-lho |
| | | give-it | | give-him-it |
| (5) | a. | dá-lo-ei | b. | dar-lho-ei |
| | | give-it-FUT1SG | | give-him-it-FUT1SG |
| | | 'I shall give it' | | 'I shall give it to him' |

In historical texts, such forms are written in one word. We analyse these forms following the hyphenation of the modern forms in (4a) and (5a). Forms such as *trazelosemos* in (6a) are composed of the root (*trazer*), the clitic pronoun (*los*) and the future verbal inflection *emos* (1.pl.), historically a short form of the verb *haver* (<lat. *habere*). Without the clitic, the verbal form is *tra(ze)remos*.

- | | | | | |
|------|-----|--|----|-----|
| (6a) | & | trazelosemos | a | vos |
| | and | bring-them-FUT1PL. | to | you |
| | | 'and we will bring them to you' (F. Lopes: <i>Crónica de D. João I</i> CAP 15) | | |

We want to keep the linearity, even though the clitic is not part of the verbal inflection. The solution is to use the tokenisation function in the PROIEL application to split the word in three parts: (i) the root, which is then annotated for tense, person and number, carrying all the inflectional information, (ii) the clitic, and (iii) the verbal ending. There are separate lemmas for the clitic pronouns. The verbal ending is lemmatised as a variant of the verb *haver*, and syntactically annotated as an auxiliary word (AUX), as demonstrated in (6b).



(6b)	&	traze	los	emos	a	vos
	conj.	verb	pers. pron.	verb	prep.	pers.pron.
	NO INFL.	1PL.FUT	3PL.M	NO INFL.	NO INFL.	2PL
	e	trazer	o	haver2	a	vos

We obtain the following: (i) the original orthography is maintained at the token presentation level, as in (6a); (ii) the linearity is maintained: all searches for postverbal clitics of finite verbs will include these examples, and exclude them from searches on proclisis, since the ending is not annotated with inflection or finiteness, only as a lemma variant of *haver*; (iii) mesoclitic forms are easily searchable through the lemmatisation of the inflection clitic as an alternative form of the verb *haver*; (iv) simple searches for futures and conditionals, as well as for clitics, will find these examples without further ado.

5. Lemmatisation and parts of speech

During the annotation process, every overt token is associated with a lemma. Each lemma is stored with a form, a language tag (ISO code) and a part-of-speech tag (see table 2) in the SQL database.

A-	adjective	Mo	ordinal numeral
Df	adverb	Pp	personal pronoun
S-	article	Pk	personal reflexive pronoun
Ma	cardinal numeral	Ps	possessive pronoun
Nb	common noun	Pt	possessive reflexive pronoun
C-	conjunction	R-	adposition
Pd	demonstrative pronoun	Ne	proper noun
F-	foreign word	Py	quantifier
Px	indefinite pronoun	Pc	reciprocal pronoun
N-	infinitive marker	Dq	relative adverb
I-	interjection	Pr	relative pronoun
Du	interrogative adverb	G-	subjunction

Pi	interrogative pronoun	V-	verb
-----------	--------------------------	-----------	------

Table 2. Part-of-speech tag inventory.

If two otherwise similar lemmas differ on one of these counts, they will be stored as two separate lemmas. In particular, there are many examples of lemma forms that are stored with multiple part-of-speech tags. For example, Latin *ut* is stored as a subjunction, a relative adverb, and an interrogative adverb. Since it is also possible to store lemmas with variant numbers, *ut* is also stored as an adverb in three varieties: *ut#1* ‘thus’, *ut#2* ‘as, such as’ and *ut#3* ‘as for example’, since these have been deemed sufficiently different to merit three separate lemmas. An advantage of positing separate lemmas is that it may ease the retrievability of certain constructions: *ut#2*, for instance, signals comparison constructions.

The projects in the PROIEL family have dealt with the possibility of positing homonymous lemmas in different ways. While the PROIEL and TOROT treebanks have posited multiple homonymous lemmas quite liberally, the Menotec treebank of Old Norwegian has chosen a stricter line.

While traditional grammars and dictionaries of Old Norse (covering Old Icelandic and Old Norwegian) operate with a heterogeneous class of pronouns, modern grammars tend to draw a distinction between pronouns proper and determiners. This is well motivated by Old Norwegian morphology, since pronouns have inflections of their own, rich in suppletivism, while determiners basically have the same inflection as adjectives. In this situation, it was necessary to make a choice between the practicality of following the grammatical tradition and the temptation of metaphorically turning the leaf and establishing a new standard. The latter option was chosen, but as a consequence, it was necessary to specify the part of speech for all words belonging to the traditional pronoun category, and introduce some other modifications (cf. Haugen and Øverland 2014, ch. 2).

A fair number of words are assigned to more than one part of speech in grammars and dictionaries of Old Norse. While it is unquestionable that some words have to be analysed in more than one way, the Menotec group decided to reduce the number of homonymous lexemes as far as practically possible. For example, the word *engi* ‘no one’ would in several grammars be classified as a pronoun in *þar var engi* ‘no one was there’, and a determiner in *þar var engi maðr* ‘no man was there’. Based on the latter usage, the Menotec group chose to classify *engi* as a determiner in both cases, considering *þar var engi* as a case of deleted head. In the guidelines to the annotation, 125 truly homonymous words are identified, which is a rather short list considering that the language has between 40,000 and 50,000 entries in the dictionaries (cf. Haugen and Øverland 2014, ch. 3).

A particularly difficult word is *einn* ‘one’, which in most cases can be analysed as a determiner, but at some point in the development of Norwegian also became an indefinite article in certain contexts. In *Maria gat ein son ok er mærr* ‘Mary gave birth to a/one son, and [she] is a virgin’ (*Homily book*, ca. 1200–1225) one might

argue that *einn* is an article rather than a determiner, since the meaning probably is ‘a son’ rather than ‘one son’. However, this may be a distinction forced on the material from the viewpoint of Modern Norwegian. In Old Norwegian, one could argue that the distinction between article and determiner simply was not established, and that the language was neutral between the interpretations ‘a son’ and ‘one son’. For this reason, *einn* is classified only as a determiner in these contexts. In some contexts, however, *einn* has to be classified as an adjective, e.g. in *Æin man Asta huila ser i nott, sagðe hann* ‘Alone, Asta will sleep tonight, he said’ (*Óláfs saga*, ca. 1225–1250). This usage is marginal, so with few exceptions the word *einn* is analysed as a determiner in the Old Norwegian treebank.

6. Morphology

PROIEL Annotator allows detailed morphological analysis. The morphology is stored in the database in a ten-place positional tag with the features listed in table 3. The tagset was synthesised from the tagsets in the available tagged sources when the original PROIEL corpus was built, i.e. the morphGNT¹⁹ and the Wulfila project,²⁰ but has since been expanded to support the needs of new languages.

1. Person	1, 2, 3, x (uncertain)
2. Number	s (singular), d (dual), p (plural), x (uncertain number)
3. Tense	p (present), i (imperfect), r (perfect), s (resultative, i.e. l-form), a (aorist), u (past), l (pluperfect), f (future), t (future perfect), x (uncertain tense)
4. Mood (combined mood and finiteness)	i (indicative), s (subjunctive), m (imperative), o (optative), n (infinitive), p (participle), d (gerund), g (gerundive), u (supine), e (indicative or subjunctive), f (indicative or imperative), h (subjunctive or imperative), x (uncertain mood)
5. Voice	a (active), m (middle), p (passive), e (middle or passive)
6. Gender	m (masculine), f (feminine), n (neuter), p (masculine or feminine), o (masculine or neuter), r (feminine or neuter), q (masculine, feminine or neuter), x (uncertain gender)
7. Case	n (nominative), a (accusative), o (oblique), g (genitive), c (genitive or dative), d (dative), b (ablative), i (instrumental), l (locative), v (vocative), e (accusative or dative), x (uncertain case), z (no case)
8. Degree	p (positive), c (comparative), s (superlative), x (uncertain degree)

¹⁹ <https://github.com/morphgnt>

²⁰ <http://www.wulfila.be/gothic/>

9. Strength ²¹	w (weak, i.e. long form), s (strong, i.e. short form), t (weak or strong)
10. Inflection	n (non-inflecting), i (inflecting)

Table 3. Morphological tags

6.1 Morphological preprocessing

The level of detail in the morphological annotation results in a high number of distinct morphological tags. For instance, the OCS annotation in the TOROT treebank uses 1003 different tags. Combined with the part-of-speech tags, the number becomes even higher. Nonetheless it is possible to train fairly successful statistical morphological taggers even with such diversity in the training data. As mentioned in Section 3, an important application of a morphological tagger is in the manual annotation process itself, as a pre-tagger. Fort & Sagot (2010, on POS-tagging English) and Skjærholt (2011, on morphological tagging of Latin) show that the speed – and to a lesser extent also the accuracy – of annotation is improved already by a pre-tagging accuracy above 80%. Depending on the corpus and language, it may take as little as 10–20,000 annotated tokens of training data to produce automatic morphological pre-annotation that eases the work of the annotator considerably.

Preannotation is regularly used in the annotation of Greek, Latin, OCS, Old East Slavic and Middle Russian, using the TnT tagger (Trigrams’n’Tags, Brants 2000). For the Slavic texts, which do not use normalised orthography, this is especially useful, since both the training data and the new text can be normalised before training and tagging, without affecting the representation in the treebank (for a detailed description of the procedure, see Eckhoff and Berdičevskis 2015, Berdičevskis et al. 2016). The success rate of the tagging varies considerably with the size of the training set and the similarity of the new text to the texts in the training set. Berdičevskis et al. (2016) report a success rate of 89.5 % for part-of-speech labels and 81.5 % for ten-place morphology tags²² in tagging a 15th century Russian Church Slavonic text with highly idiosyncratic spelling based on a training set of 166,000 Old East Slavic and Middle Russian tagged word tokens. Birnbaum and Eckhoff (to appear) report a success rate of 91.3 % for part-of-speech tags and 94% for ten-place morphology tags in tagging a Byzantine Greek vita based on a training set of 295,000 Ancient, Koine and Byzantine Greek tagged word tokens from the PROIEL corpus. The latter tagger had the advantage both of a larger training set and of being trained and used on orthographically regular material.

In a setting where there isn’t any annotated material available, a sufficiently accurate tagger may be quickly developed during the annotation process itself. Adesam & Bouma (2016) illustrate this for the annotation of the Östgötalagen

²¹ In the Slavic languages and Gothic, this feature is used for the distinction between long and short adjective forms. In the Old Scandinavian languages, it is also used for definiteness marking on nouns, such as in *hestr-inn* ‘the horse’, so that definite nouns are encoded as weak, indefinite ones as strong.

²² A further 7.5% of the tokens off-by-one errors, i.e. only one of the ten morphological fields had the wrong value.

Old Swedish provincial law, where as little as 1000 tokens were sufficient to create a POS-tagger with an accuracy of over 80%, and 7000 tokens to train a similarly accurate morphological tagger, assuming only sentence and token segmentation as preprocessing, without the use of external data sources and in spite of the lack of an orthographic standard. The crux here is that the annotation process allows for an extremely in-domain application of the tagger: although the tagger is applied to new material, it stems from the same document as the tagger was trained on.

6.2 Case syncretism in Old Norwegian

Case syncretism is a challenge to the morphological annotation of Old English, Old Swedish and Old Norwegian. For example, Old Norwegian had more or less the same morphology as Modern Icelandic, but unlike Icelandic, a number of distinctions began to merge during the Middle Norwegian period 1350–1500 (as was the case in the other Scandinavian languages), and even as early as in the early 13th century, it had a slightly higher degree of syncretism than Icelandic. One example is the inflection of the present participle, which began to converge towards a single form as early as in 13th century Norwegian. Based on the size of other nominal paradigms, the present participle has no less than 24 potentially distinct forms – three genders, four cases and two numbers. In practice, it had three, two or even just one distinct form, i.e. the endings *-i*, *-a* and *-um*. From the point of view of morphological annotation, it is in most cases possible to ascribe the full set of inflectional forms to the participle based on concordance with other words in the sentence. In cases of ambiguity, however, unspecified features had to be introduced. This applies especially to syncretism in case (the oblique cases accusative, dative and genitive), in gender (particularly in genitive and dative plural) and in mood (indicative and subjunctive), in the entire nominal domain.

While the grammar of Old Norwegian cannot in any practical way be used for the description of Modern Norwegian, there is no clear cut-off point between the two stages during the Middle Norwegian period, other than some time between ca. 1400 and ca. 1500. For a historical study of the language, it makes sense to use an older system as a reference as long as possible, so that mergers of grammatical categories can be described and timelines can be established. Case merging began earlier in adjectives than in nouns, and mergers often seemed inconsistent; in some noun classes the oblique form became the standard form, while in other classes the nominative form became the standard. Since the texts have been annotated according to the older and fuller system, we believe that the complexity of these mergers can more easily be traced and understood.

7. Syntax

Using dependency grammar is especially useful in annotating languages with rich morphology and relatively free word order (often conditioned by information structure rather than syntax). For many of the languages in the PROIEL treebank family, there is simply not sufficient evidence to use a phrase-structure scheme. Instead, word order information is stored in a separate layer, and can be combined with dependency information in complex syntactic queries.

The treebanks in the PROIEL family use an enriched dependency grammar scheme inspired by and convertible to Lexical-Functional Grammar's F-structures (see Haug 2011).²³ It is more expressive than the dependency schemes used in most treebanks and statistical dependency parsing in several respects. In many ways it follows in the tradition of the full multilayer annotation scheme of the Prague Dependency Treebank (PDT),²⁴ which is an important standard in the field, and which is the basis of the annotation of e.g. the Ancient Greek and Latin Dependency Treebank (AGDT, LDT)²⁵ and the Index Thomisticus Treebank (IT-T).²⁶ All the discrepancies between the two schemes are motivated by the wish for more expressivity in a single layer of annotation in the PROIEL scheme. There are three important differences: use of empty nodes, use of secondary dependencies, and a more fine-grained set of syntactic relations. (For further discussion of the compatibility between the PROIEL and AGDT schemes, see section 9.)

Empty verb and conjunction nodes are systematically employed to model ellipsis, null copulae, gapping and asyndetic coordination. The tectogrammatical layer of the PDT scheme also includes use of empty nodes to account for ellipsis,²⁷ but the analytical layer of the scheme does not. The tectogrammatical layer is often not implemented in PDT-style treebanks, this is for instance the case for AGDT and LDT (but not the IT-T). Including this information in the main annotation layer thus makes for more economical annotation. Note that this comes at the price of making the annotated data less useful as training data for a syntactic parser,²⁸ but in a scheme designed specifically for sometimes very scarcely attested language stages, the priority should be preserving structure. Large-scale automatically parsed treebanks are out of the question for most of the project languages.

The PROIEL scheme also employs secondary dependency edges to indicate structure sharing, for instance to indicate the external subjects of conjunct participles (see section 7.1), but also to indicate shared dependents in

²³ For a fuller description of the scheme, see Haug et al. 2009. For an exhaustive documentation of the scheme, see the PROIEL guidelines for syntactic annotation (http://folk.uio.no/daghaug/syntactic_guidelines.pdf). For documentation of the application of the scheme to Slavic, see the TOROT guidelines (<http://folk.uio.no/hanneme/torot.pdf>). For documentation of the application of the scheme to Old Norwegian, see Haugen and Øverland 2014. For documentation of the application of the scheme to Old English, see http://folk.uio.no/krisbec/OE_guidelines.pdf.

²⁴ <https://ufal.mff.cuni.cz/pdt2.0/>

²⁵ https://perseusdl.github.io/treebank_data/

²⁶ <http://itreebank.marginalia.it/>

²⁷ <http://ufal.mff.cuni.cz/project/pdt2.0/doc/manuals/en/t-layer/html/index.html>

²⁸ For a discussion of , with empty nodes, see Seeker et al. 2012. For an experiment using MaltParser on OCS data from TOROT, see Berdičevskis 2015, for a pre-parsing experiment on TOROT data, see Eckhoff & Berdičevskis 2016.

coordinations. Again, this is a choice motivated by the wish to capture more structure: Indicating external subjects with secondary dependencies yields rich data that may e.g. be of great use in studies of agreement. Indicating shared dependents yields considerably richer argument structure data than an annotation style that omits this information.

Finally, the PROIEL scheme has a richer set of syntactic relation labels (table 4) than that found in the analytical layer of the PDT scheme. For instance, direct objects (OBJ) are distinguished from oblique arguments (OBL) and passive agents (AG), again yielding richer argument structure data than the PDT scheme.²⁹

adnom	adnominal (supertag ³⁰)	obl	oblique argument
adv	adverbial	parpred	parenthetical predication
ag	passive agent	part	partitive
apos	apposition	per	peripheral (supertag)
arg	argument (supertag)	pred	predicate
atr	attribute	rel	relative clause (supertag)
aux	auxiliary	sub	subject
comp	complement	voc	vocative
expl	expletive	xadv	adverbial with external subject
narg	adnominal argument	xobj	argument with external subject
nonsub	non-subject (supertag)	pid	predicate identity (secondary)
obj	direct object	xsub	external subject (secondary)

Table 4. Syntactic relation label inventory.

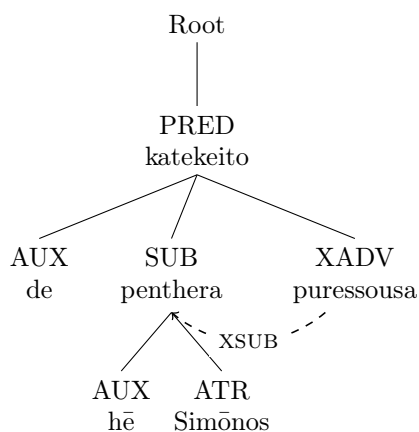
In the following sections, we give some examples of how the PROIEL dependency scheme has been employed to annotate various structures in some of the project languages. In several of these case studies, we demonstrate how several layers of annotation (word order, part of speech, morphology and syntax) can be combined to give detailed and retrievable analyses of complex structures.

7.1 Control structures in Greek

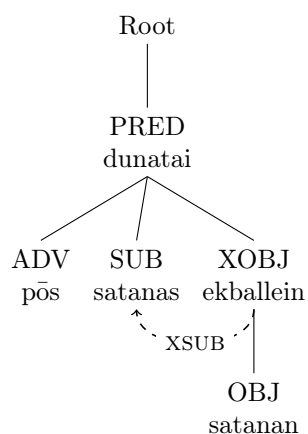
²⁹ For further discussion and motivation of the differences between the two schemes, see Haug and Jøhndal 2008, Haug et al. 2009.

³⁰ Supertags are tags to be used by annotators in cases of doubt: if it is not clear whether an adnominal dependent is an atr, apos, narg or part, the supertag adnom can be used.

One well-known limitation of standard dependency analysis is the unique head principle, which says that every word is a dependent of a single head. Formally, this is a very attractive constraint because it means that the dependent-head relation is functional and that the resulting data structure is a rooted tree, which makes it easier to handle and store, but, more importantly, is easier to reason about when we want to validate or query an annotated resource. On the empirical side, though, it is well known that the unique head principle is not adequate for natural language. One case in point comes from control structures, which involve non-finite verbs with an implicit subject which is obligatorily identified with some argument in the matrix clause. Such structures can occur both in complementation and in adjunction. Examples are shown in (7–9)

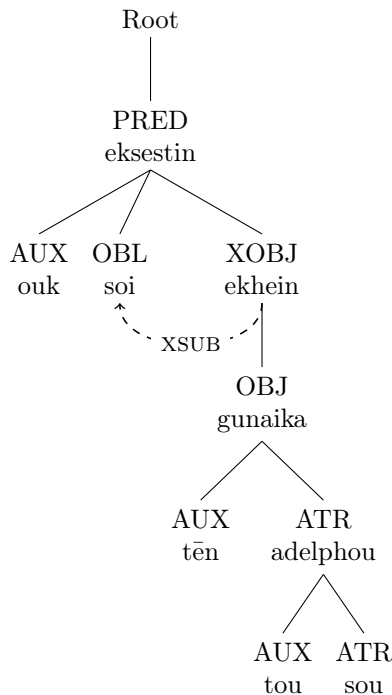


- (7) hē de penthera Simōnos katekeito
 the PTC mother-in-law.NOM Simon.GEN lie-down.IMPERF3SG
 puresousa
 suffer-with-fever.PRES.PTCP.F.NOM.SG
 ‘Now Simon’s mother-in-law was in bed with a fever.’
 (Greek New Testament, Mark 1.30, 6542)



- (8) pōs dunatai satanas satanan ekballein
 how be able. PRES3SG Satan.NOM Satan.ACC drive-out.INF

‘How can Satan throw out Satan?’
(Greek New Testament, Mark 3.23, 6637)



- (9) ouk eksestin soi ekhein tēn
 not be-proper. PRES3SG you.DAT have.INF the
 gunaika tou adelphou sou
 wife.ACC the brother.GEN you.GEN
 ‘You are not allowed to have your brother's wife.’
 (Greek New Testament, Mark 6.18, 57368)

In the first example, the participle *puressousa* ‘suffering with a fever’ modifies the main event *katekeito* ‘lie down’ and expresses an ‘attendant circumstance’ of that event. As such, the participle should be seen as an adverbial modifier of that event, which translates into a syntactic dependency between *katekeito* and *puressousa*. Equally obviously, *hē penthera Simōnos* ‘Simon’s mother-in-law’ is a subject dependent of *katekeito*. However, there is also a syntactic dependency between *hē penthera Simōnos* and *puressousa*, which is encoded through agreement in case, number and gender. Case agreement in particular signals a syntactic relationship, as it is not found in semantic coreference relationships in Ancient Greek. Moreover, notice that this syntactic relationship is one of selection, i.e. the participle requires this argument to be present. If for example the matrix verb is impersonal (does not take a nominative argument), then it would be ungrammatical to add a nominative adjunct participle.

This relation is not expressible under the unique head constraint, but can be easily captured with a secondary edge. In the PROIEL scheme, this is done by making *hē penthera Simōnos* depend additionally on *puressousa* via the relation XSUB, i.e. external subject. This is the standard analysis in Lexical-Functional

Grammar, see Andrews 1982, see also Andrews 1971 for an early discussion of the relevance of the Ancient Greek data.

The same approach is taken in the latter two examples, which involve coindexation. *satanas* ‘Satan’ and *soi* ‘you’ are taken to have two grammatical roles: one as the subject (*satanas*) or the oblique (*soi*) in the matrix clause, and one as the external subject of the infinitive.³¹ Again, we have a double syntactic relationship which is widely recognized in the syntactic literature as a control structure. Traditional dependency grammars with their unique head constraint have no analysis of such structures (neither in Ancient Greek or in other languages). But notice that the Universal Dependencies standard (UD, see section 9) now mark these structures with a special relation (XCOMP, comparable to our XOBJ), although they do not force the use of a secondary edge to express the relationship between the infinitive and its subject in the matrix clause, because secondary edges are as of yet not standardized at all in UD. However, we see from these examples that it is not predictable which matrix argument serves as the external subject of the non-finite verb, so an explicit annotation is required for preserve linguistic information. Our approach allows us to do this in an easy way, while also allowing conversion to other linguistic analyses which involve empty elements such as PRO.

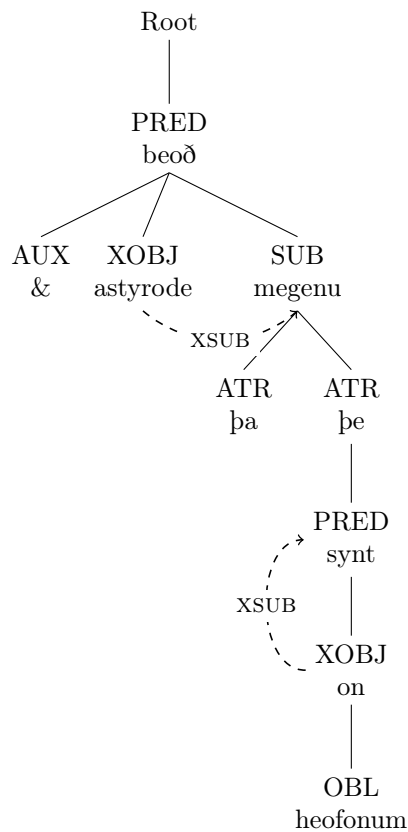
7.2 Old English relative clauses

Old English relative clauses are constructed in five different ways: with the indeclinable relativiser *þe*, with the demonstrative *se* as the relative pronoun, with a combination of the demonstrative and the relativiser: *se þe*, of which there are two subtypes, or as headless relatives (see Haugland 2007:305–312 for an overview; Mitchell 1985:vol. II for an extensive discussion; Traugott 1992:223–233). These possibilities are reflected in the syntactic analysis, some of which are illustrated here.

The indeclinable relative particle *þe* is analysed as a subjunction, on a par with the Old Norwegian relativiser *er* in the Menotec corpus. Thus, the annotation scheme for Old English does not operate with a tag for relative pronouns: since *þe* is indeclinable, we do not posit a role for it within the relative clause. The relativising subjunction is thus the head of the relative clause, as shown in (10), where *þe* is the head, and the relative clause verb *synt* ‘are’ is dependent on *þe*.³²

³¹ In PDT-style treebanks, the dependent infinitive would be analysed as a subject. In the PROIEL scheme, argument infinitives are never analysed as SUBs unless they are nominalised by way of definite articles, since such structures are often ambiguous. Instead, they are analysed as COMP or XOBJ depending on whether they have an external subject.

³² Note that the secondary dependency from *on* ‘in’ to *synt* ‘are’ indicates that the dependent shares its subject with its head verb, but that this subject is not overtly expressed. Moreover, the dependent’s subject may be any (non-overt) argument of the verb.

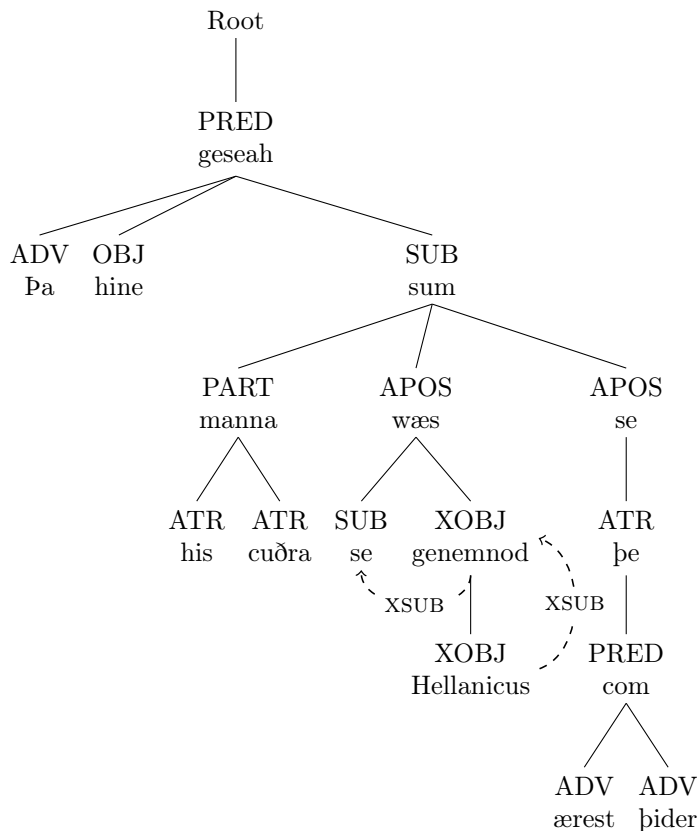


- (10) and beoð astyrode þa megenu þe on
 and are shaken the powers.NOM REL in
 heofonum synt
 heavens are
 ‘and the powers that are in heaven shall be shaken’ (*West-Saxon Gospels*,
 Mark 13.25, 103341)

A relative clause can also be introduced by a form of the demonstrative determiner *se*, or by a combination of *se* and *þe*. (11) is an example of both.³³ In the first relative clause *se* functions as subject of the relative clause and is dependent on the verb *wæs* ‘was’, which is the head. In the second clause, *se* is the head, and the particle *þe* attaches to it as a subjunction, taking the relative clause verb as a dependent. In such cases, *se* gets case from its antecedent, and the meaning is ‘the one who’. There is also another, rarer, type of *se þe* relative, not exemplified here, in which *se* gets case from its function in the relative clause, and *þe* is merely a reinforcing particle, rather than a subjunction. In such cases, *se* is a dependent of the verb and analysed according to its function in the relative clause, and *þe* is attached to *se* as an AUX element. In other words, *se þe*

³³ Note that the APOS relation label is used not only for the usual type of nominal appositions, but also to mark non-restrictive relative clauses, as seen in the tree for example (11). Restrictive relative clauses are ATR dependents of their antecedents. In this the PROIEL scheme differs from the annotation in PDT-style treebanks, where restrictive and non-restrictive relative clauses are not distinguished.

relatives of the first type are analysed similarly to *þe* relatives, and *se þe* relatives of the second type are analysed like *se* relatives.



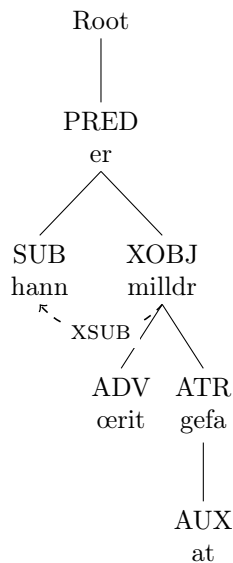
- (11) þa geseah hine sum his cuðra manna **se**
 then saw him one his known.GEN men.GEN who
wæs Hellanicus genemnod, se þe ærest þider com
 was Hellanicus called, he þe first thithercame
 'Then one of his men who was called Hellanicus, who had first come
 thither, saw him' (*Apollonius of Tyre*, 118763)

The flexibility of the annotation system makes it easy to accommodate the different clause types in the annotation, as well as instances of discontinuous relative constructions, where other sentence elements intervene between the antecedent and the relative clause.

7.3 Permutations and interpolations in Old Norwegian

As in other early Germanic languages, the word order in Old Norwegian was comparatively free. For example, the subject and the verb were allowed to switch places, e.g. *hann stígr af hestinum* and *stígr hann af hestinum*, both meaning 'he alights the horse'. The latter would have been a question in Modern Norwegian. Likewise, the adjective and the noun could change their order, *hestir hvítr* and *hvítr hestr*, both meaning 'a white horse', as could the determiner and the noun, e.g. *køttr minn* and *minn køttr* 'my cat'. Also, words and phrases were frequently fronted, as in *hana ælskaðu marger* 'her.ACC loved many.NOM'. The case marking made the meaning clear: *hana* is accusative (and thus the object) and *marger* is

nominative (and thus the subject). There are also examples of discontinuous fronting, as seen in (12), in which the adverb *œrit* ‘greatly’ is a dependent of the adjective *milldr* ‘gracious’.³⁴



- (12) *œrit* *er* *hann* *milldr* *at* *gefa*
 greatly is.PRES3SG he.NOM gracious.NOM to give.INF
 ‘he is very generous’ (*Guruns lioð* in *Strengleikar* (ms. dated ca. 1270),
 223810)

While these examples testify to the comparatively free word order in Old Norwegian (and Old Icelandic) prose, word order in poetry could be unusually free. An extreme case is offered by the skaldic poems, such as the *dróttkvætt* poetry. Within a single stanza, one sentence could be interpolated in another, and each could be discontinuous. A case in point is the first half of st. 6 of Sigvat Thordsson’s *Víkingarvísur* (dated to 1014–1015, preserved in a manuscript from ca. 1225–1250). Here, the interpolated sentence, referred to as a *stál* ‘beak’, is identified by round brackets (13a). The last word of the interpolated sentence, *at*, and the isolated word *ygs* in the primary sentence should be taken together as the phrase *at Yggs* ‘Odin’s conflict’, a poetical expression (*kenning*) for ‘fight’.

- (13a) *Rett* *er* *at* *socn* *en* *setta*
 true.NOM is.PRES3SG that attack.NOM the.NOM sixth.NOM
 (*snarr* *pængill* *vann* *ænglum*)

³⁴ As seen in the tree for example 12, the relation XOBJ is also used for nominal predicates in copular constructions. The reasoning is thus that nominal predicates are arguments of the copula and have external subjects which are identical with the copula’s subject. In this the PROIEL scheme deviates from the PDT scheme, which has a separate label PNOM for nominal predicates. PNOMs are also deemed to be dependents of the copula, but there is no direct indication of the external subject.

(swift.NOM	king.NOM	fought.PRET3SG	English.DAT
at)	þar er	Olafr	sætte
conflict.ACC)	there when	Olaf.NOM	sought. PRET3SG
(ygs)	lunduna	brygjum	
(Odin's)	London.GEN	bridges.DAT	

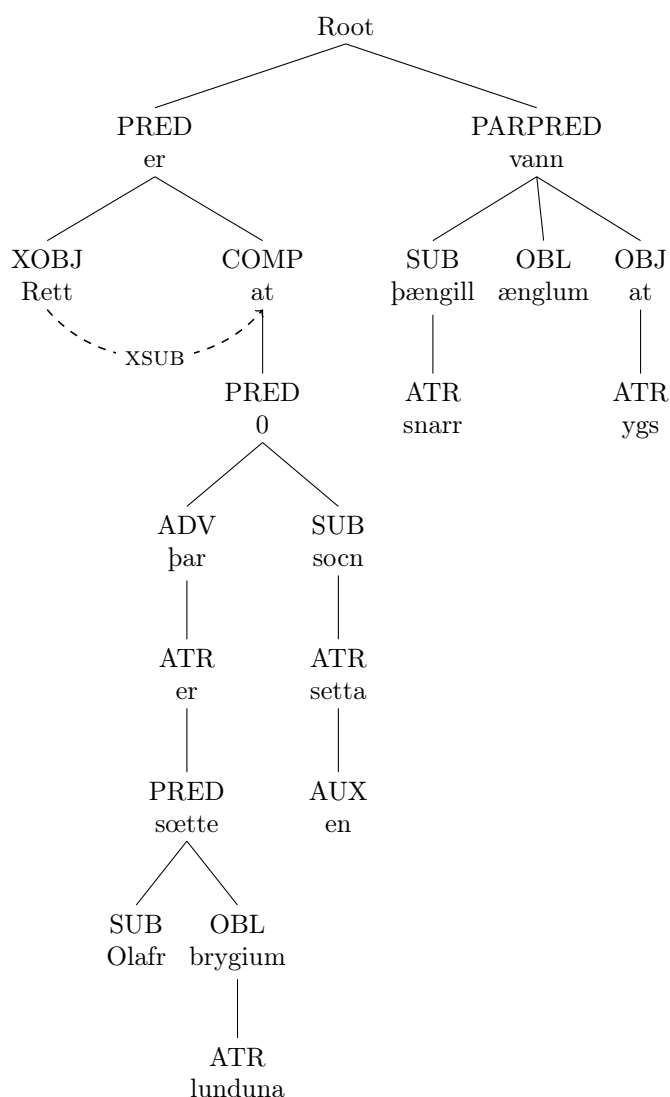
(St. 6 of Sigvat Thordsson's *Víkingarvísur* (ms. dated ca. 1225–1250),
218891)

In prose and in regularised orthography, this stanza might be paraphrased as follows:

(13b) Rétt er at en setta sókn [vas] þar er Óláfr sótti Lundúna bryggjum.
 Snarr þengill vann Englum Yggs at.
 'It is true that the sixth attack [was] when Olaf made against London's
 bridges. A swift king fought the English.'

It is very difficult to analyse a stanza of this type in a phrase structure model, but in a dependency structure model the individual words can surprisingly easily be brought together, as shown in (13c).

(13c)



The interpolated, shorter sentence should be regarded as the secondary sentence in this stanza and thus analysed as a parenthetical predicate (PARPRED), on equal footing with the primary sentence. The fact that the object *at* ‘conflict’ and its attribute *ygs* (i.e. *Yggs*) ‘Odin’s’ have been split does not create any problem in the dependency structure. Here, *ygs* is simply moved from its isolated and discontinuous position in the primary sentence to its position as an object in the secondary sentence. The primary sentence has to be analysed with an empty verb *vas* ‘was’ in the *at* ‘that’ clause,³⁵ but apart from that, the whole construction is a rather plain statement, *rett er at ...* ‘true is that...’. In editions of

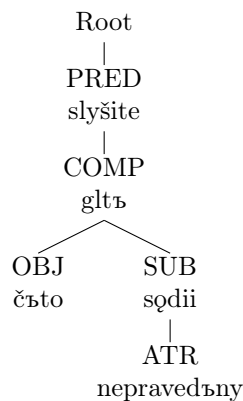
³⁵ Note that, unlike in the PDT-based schemes, subjunctions are given the dependency label of the whole subordinate clause (in this case COMP). This is a general principle in the PROIEL scheme: the head of a subtree should always carry the relation label of the whole subtree, regardless of its form. In the PDT-based schemes, the subjunction is also the head of the subordinate clause, but it is labeled AuxC, and its dependent verb carries the relation label of the whole subordinate clause.

skaldic poetry, the poems are regularly paraphrased in prose word order, as shown above. The somewhat abstract rendering of the poem in the dependency analysis is in fact also a helping hand for the interpreter, as it removes the text from the confines of the word order. Since the word order is stored in a separate layer, it is always retrievable.

7.4 Syntactic annotation and diachronic change: the history of *čъto* in East Slavic

The PROIEL annotation scheme largely requires the annotator to settle for a single analysis. As seen in section 6, it is possible to leave morphological features underspecified, and as table 4 shows, there are also a few underspecified syntactic relation labels (“supertags”). However, when it comes to part of speech and dependency attachment sites, underspecification is not an option. A diachronic treebank poses particular challenges to such a policy. How does one deal with ambiguities that are due to syntactic change in progress? In some cases, the best option is to stick to a conservative analysis that reflects the situation in the earliest annotated stage, thus rendering the structure easily retrievable at all stages (cf. section 6.2). However, this is not always possible.

A case in point is the development of structures containing the word *čъto* (and a number of other wh-words) as documented in the various stages of Slavic in the TOROT treebank. In OCS, it is possible to posit only two lemmas for this word: an interrogative pronoun ‘what?’ and an indefinite pronoun ‘something’. Of these two lemmas, only the former can signal subordination, and only in indirect questions, such as in (14). The latter can occur in subordinate clauses too, but it does not signal subordination. It typically occurs in conditional clauses such as (15), where the subordination is signalled by the subjunction *ašte*.

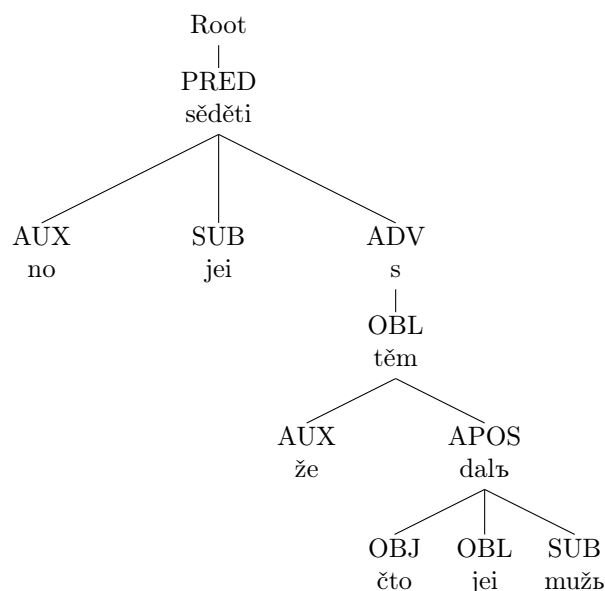


- (14) slyšite čъto sōdii nepravedъny gltъ
hear.IMP2PL what judge.NOM unjust.NOM say.PRES3SG
‘Hear what the unjust judge says’ (Codex Marianus, Luke 18.6, 51590)

- (15) i ašte esmъ kogo čimъ obidělъ.
and if be.PRES1SG someone.GEN something.INST offended
vъzvraštъ četvoricejъ
return. PRES1SG fourfold

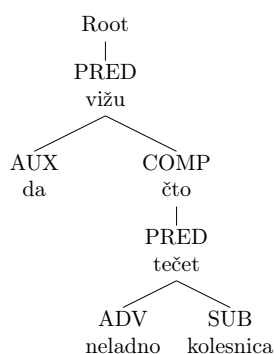
‘and if I have cheated someone of something, I will return it fourfold’
(Codex Marianus, Luke 19.8, 41151)

In Old East Slavic and especially Middle Russian, however, it is no longer possible to analyse all subordinate clauses containing *čto* as indirect questions. Even in the very earliest texts, there are examples where *čto* must be considered a relative pronoun, or even a subjunction in a complement clause. In (16), it is clear that *čto* is a relative pronoun in an adnominal relative clause.



- (16) no čto jei dal'ь muž'ь. s
 but what.ACC her.DAT gave husband.NOM with
 těm že jei s'ědět'i
 that.INST PTC she.DAT sit.INF
 'but what her husband gave her, that she can keep' (Russkaja pravda 102,
 172305)

In (17), we must consider *čto* to be a subjunction, since it clearly has no syntactic role in the complement clause.



- (17) da vižu što neladno kolesnica tečetъ
 and see.PRES1SG that not-right wagon.NOM run.PRES3SG
 ‘and I see that the wagon is not running correctly’ (Avvakum 26, 187402)

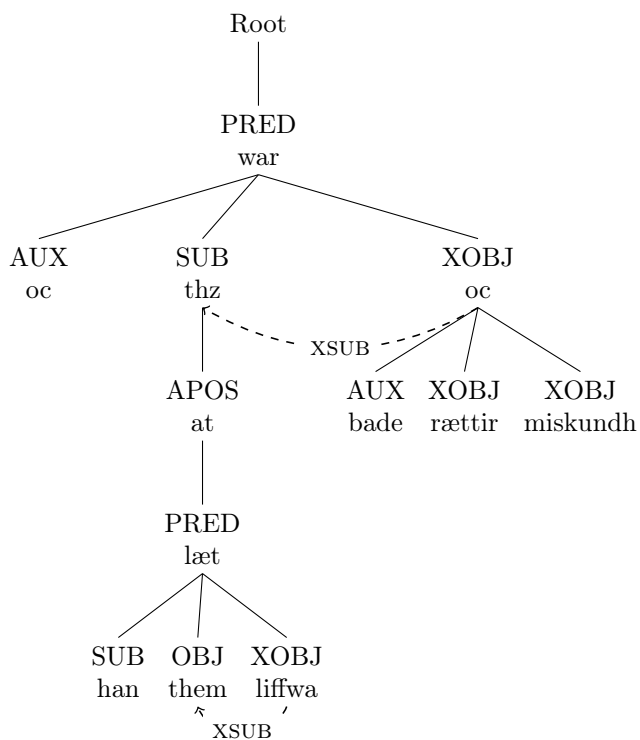
As soon as we open for these two types of analyses, however, many examples become ambiguous. It is not always clear whether something is a headless relative clause argument or an indirect question complement clause, nor is it always clear whether *što* has a syntactic role in the subordinate clause.³⁶ Since the PROIEL scheme requires the annotator to choose a single analysis in such situations, the TOROT group made the decision to analyse all unclear examples as relative clauses. In such a scenario, all dependent clauses containing *što* are still taken as complement clauses if they are dependent on speech, thought or emotion verbs (or nouns). Within this group, if *što* clearly has no syntactic role in the subordinate clause, it is analysed as a subjunction, otherwise the clause is analysed as an indirect question. Outside this group, dependent clauses containing *što* are taken to be relative clauses if there is no other signal of subordination. In these cases, *što* is lemmatised as a relative pronoun, and the dependent clause is given the relevant syntactic relation label.

Thus, we are left with two groups of clear examples and one group of both clear and ambiguous examples that must be sifted through manually by interested scholars. There is also the possibility to tag ambiguous sentences as such in the customisable annotation layers, see section 8.

7.5 Extraposition with correlate in Old Swedish

Extraposition with a clause-internal correlate is a very pervasive phenomenon in Old Norwegian and Old Swedish, but similar structures are found in several of the other PROIEL family languages. Such structures are analysed as a case of (possibly discontinuous) apposition. We will use Old Swedish as an example. For instance, in (18) we have a postposed nominal clause with an in situ subject pronoun as correlate. In the analysis, the pronoun *thz* ‘it’ will be a SUB of the finite copula *war* ‘was’, and the nominal clause an APOS of the pronoun.

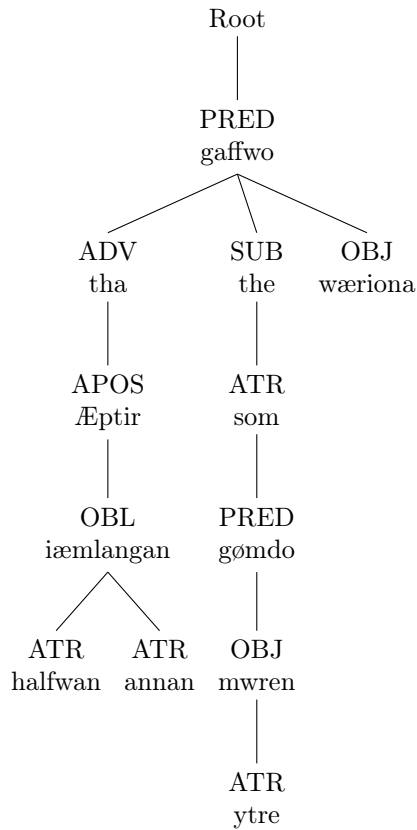
³⁶ An additional confounding factor is that conditional clauses, a common environment for the *indefinite* pronoun *što*, often lack subordinators in Middle Russian texts, which can cause even more ambiguities.



- (18) thz war oc bade rættir oc miskundh
 it.NOM was.PAST3SG also both justice.NOM and mercy.NOM
 at han læt them liffwa
 that he.NOM let.PAST3SG them.DAT live.INF
 'It was also both just and merciful that he let them live.'
 (Pentateukparafrasen, Ms B)

This way, the annotation captures the observation that the pronoun behaves like a regular subject whereas the clause appears to be adjoined to the sentence, the intuition that the pronominal and the clausal realisations share a referent, and the fact that the clause has the form of a dependent (that is, it is a subordinate rather than a main clause).

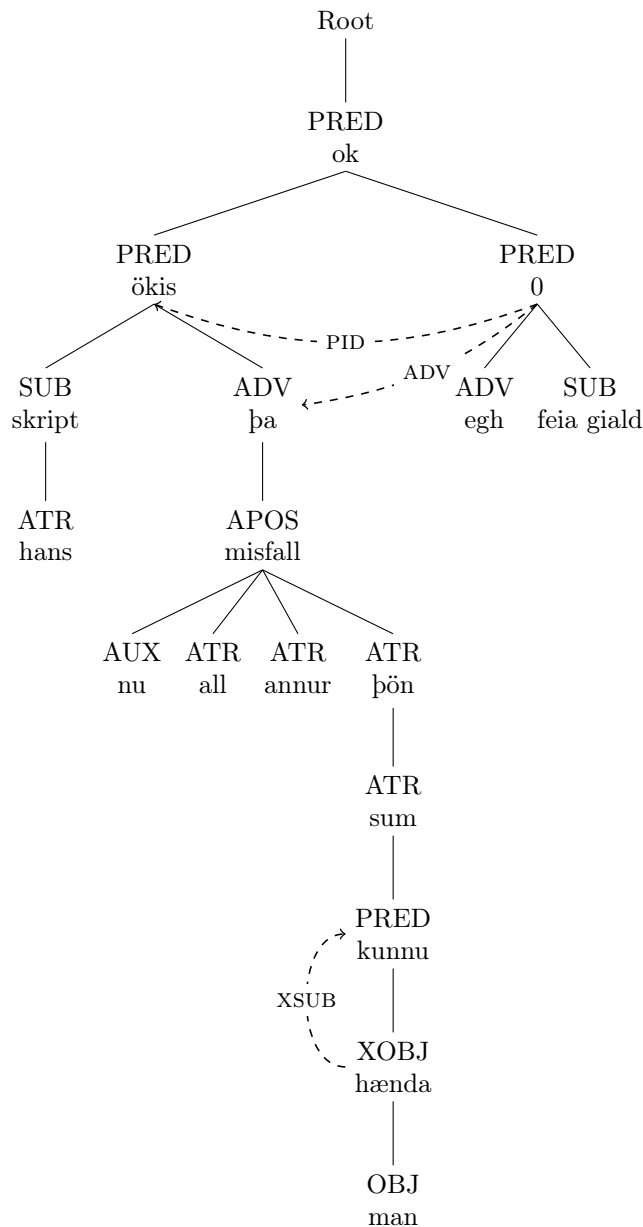
A related construction, rather common in Old Swedish, involves a preposed adverbial and an internal *þa* 'then', which appears at the front of the main clause. Here too, we use the APOS label to link the extraposed material to the clause internal pronominal realisation:



- (19) Æptir halfwan annan iæmlangan tha
 After half.ACC other.ACC equally long.ACC then
 give.PAST3PL they.NOM
 gaffwo the wæriona som ytre
 give.PAST3PL they.NOM defenses.ACC.DEF that outer
 mwren gømdo
 wall.ACC.DEF defend.PAST3PL
 'After one and a half years, the guards at the outer wall gave up defenses.'
 (ibid)

In Old Swedish law texts this construction is frequently found in combination with the use of the discourse particle *nu* 'now', which serves to introduce a new scenario or part of a scenario. This particle may attach freely, for instance to adverbial NPs (20a),³⁷ PPs (b) and clauses (c).

³⁷ Example (20a) also illustrates how predicate identity (PID) and shared dependents can be indicated by way of secondary dependencies.



(20a) Nu all annur misfall
now all.NOM/ACC other.NOM/ACC misfortunes.NOM/ACC
þönn sum man kunnu hænda
those.NOM/ACC that man.ACC can.PRES3PL happen.INF
þa ökis skript hans ok egh
then increases.SG.SUBJ.PASS penance.NOM his and not
feia giald.
fine.NOM
‘Now (for) all other sins man may happen to commit, his penance is to be increased but not his fine.’(Östgöotalagen, Ms A)

(20b) Nu firi allum andrum malum. ælla andrum
now for all.DAT other.DAT incidents.DAT or other
fallum þa hialpær lönda skript firi
cases.DAT then help.PRES3SG private penance.NOM for

fea bot.
 fine.DAT/ACC
 ‘Now in all other cases, private penance prevents fines.’ (ibid)

(20c) Nu æn siax uilia uæria ok siax fælla.
 now if six want.PRES3PL acquit.INF and six convict.INF
 þa aghu þe uitzs orþ sum uæria
 then have.PRES3PL they statement that acquit.INF
 uilia:
 want.PRES3PL
 ‘Now if six want to acquit and six convict, then those who want to acquit
 decide.’ (ibid)

As in the treatment of introductory conjunctions (see Section 4.1), we use the AUX dependency for such instances of *nu*, and attach them to the head of the preposed material.

Often, the *nu-þa* discourse structuring strategy is applied to independent syntactic units. In (21), we have two main clauses:

(21) Nu ær kirkia giorþ: þa
 now be.PRES3SG church.NOM make.PAST.PART.F.NOM then
 skal skötninga til kirkiu giua:
 shall.PRES3SG land.ACC to church.GEN give.INF
 ‘Now a church has been built. Then (the people) must give land to the
 church.’ (ibid)

Here too, we use the AUX-relation to attach *nu* to *ær*. However, as there is no sign of subordination in cases like these, we annotate two independent syntactic units and the antecedent is thus not syntactically linked to *þa*.

8. Other levels of annotation

In addition to morphological and syntactic annotation, the PROIEL software tools offer the possibility of enriching the treebank data with additional annotation at several levels.

There is a separate interface for annotating texts for givenness status and anaphoric relations. For a detailed description of the annotation scheme and its theoretical roots, see Haug et al. 2014. In the PROIEL treebank, the Greek Gospels are tagged for givenness status and anaphoric relations in their entirety. Since the parallel data in the treebank are aligned at token level, this annotation can easily be transferred from Greek to the other Gospel parallels. A number of published studies have used these data (Eckhoff 2011, Eckhoff 2015, Hertenzenberg 2011, Hertenzenberg 2014, Lindberg 2013, Müth 2015).

Individual scholars also have the opportunity to add customised tags at the token, lemma and sentence level. The tags can be used for semantics (e.g. animacy, spatial semantics, verb class), but have also been used for derivational

morphology, such as prefix, stem and suffix tags for verbs (used in Eckhoff and Haug 2015) and for information on syntactic binding for Latin reflexives (used in Jøhndal 2012). The customised tags enable individual scholars to contribute directly to the treebanks, make their studies directly replicable and avoid duplication of effort.

9. Compatibility with other schemes

Annotation of syntax is much more theory-dependent than annotation of morphology, and there is much less agreement on what information should be included in the annotation and how it should be structured. At a very basic level there is a distinction between phrase structure annotation and dependency-based annotation. Phrase structure puts the emphasis on constituency, which is a problematic notion in languages with relatively free word order, as is the case for most of the PROIEL family languages. For this reason, almost all attempts to annotate Ancient Greek and Latin are based on dependencies.³⁸ But even within a dependency-based approach there are several ways one could go, and for Ancient Greek there are in fact three different models of dependency annotation in active use: in addition to the PROIEL corpus, there are the Greek New Testament Syntax Trees³⁹ developed by the Global Bible Initiative, and the Ancient Greek Dependency Treebank (AGDT),⁴⁰ which is connected to the Perseus project. The AGDT scheme in turn builds on the joint annotation scheme developed for Latin by the Index Thomisticus Treebank (IT-TB)⁴¹ and the Latin Dependency Treebank (LDT). Their joint annotation scheme is described in Bamman, Crane, Passarotti and Raynaud (2007).

In the following we focus on the differences between the PROIEL dependency scheme (as applied to Greek) and AGDT, and efforts to convert between them and make them converge. Our reason for focusing on the AGDT is that the LDT is by now largely an abandoned effort while the IT-TB contains very specialized data, works by Thomas of Aquinas. AGDT therefore offers the most comparable data to PROIEL, but most of what we say in the following applies to LDT and IT-TB too.

Differences between treebanks may arise in several ways. When two annotation teams work separately, they will inevitably develop different conventions and analyses even when the underlying annotation scheme is the same. Such differences can arise out of a different understanding of the text, or simply out of need to deal with idiomatic or deviant constructions in some way, without a claim to linguistic adequacy. Such differences between corpora typically need to

³⁸ The Penn Parsed Corpus of Historical Greek (PPCHiG) used a constituency-based annotation, but is no longer in active development.

³⁹ <https://github.com/biblicalhumanities/greek-new-testament/tree/master/syntax-trees/sblgnt>

⁴⁰ <http://www.dh.uni-leipzig.de/wo/projects/ancient-greek-and-latin-dependency-treebank-2-0/>

⁴¹ <http://www.sfs.uni-tuebingen.de/en/ascl/resources/corpora/index-thomisticus-treebank.html>

be resolved manually and will not be further discussed here. Instead we look at differences that follow directly from the annotation scheme.

Since a dependency analysis is essentially a labelled directed graph, the differences can arise either in the edges (attachments) or in the labels. By contrast, since dependency analyses (unlike phrase structures) use the words themselves as nodes, the nodes will be the same across corpora, except to the extent that the tokenisation differs. Fortunately, the tokenisation problem is not too difficult for our languages, and there are no differences between PROIEL and AGDT in this regard. One difference worth mentioning, however, is that the PROIEL scheme uses explicit empty nodes (for asyndetic coordination and verbal ellipsis), while AGDT follows the Prague Dependency Treebank in using implicit empty nodes that are synthesised in the labels.⁴² The two solutions are equivalent, though, and both can be converted to the other without loss of information.

Attachment differences in dependency corpora typically arise in two different areas: coordination and the treatment of function words. In coordinations, one can either take the conjunction as the head, or one of the conjuncts (typically the first). Both PROIEL and AGDT use the conjunction as the head. With function words, either the function word or the lexical word it belongs with can be taken as the head. For example, in *John went to Berlin*, *to* can be taken to be a dependent of *went* and the head of *Berlin*, so that it “mediates” the relation between *went* and *Berlin*. Alternatively, *Berlin* can be taken to depend directly on *went*, with the preposition *to* serving as a modifier of *Berlin*. Similar analyses are possible for other function words such as subjunctions, auxiliary verbs and determiners. Again, PROIEL and AGDT agree in taking prepositions and subjunctions as heads, but auxiliary verbs and determiners as dependents.

There remains one substantial attachment difference between AGDT and PROIEL, concerning secondary predicates. In AGDT, these are attached to their subject, while in PROIEL they are attached to their governing verb, with the subject instead indicated via a secondary edge as discussed above. Most of the substantial differences between PROIEL and AGDT, then, reside in the labels, which are more fine-grained in PROIEL than in AGDT. For example, PROIEL attempts to differentiate between argument PPs and adjunct PPs, whereas all PPs are treated as adjuncts in AGDT. Another distinction is found in infinitives: some of these have an implicit subject determined by the governing verb (control infinitives), while others have their own explicit subjects (accusative with infinitive construction). These are given different labels in PROIEL, but not in AGDT. Moreover, the implicit subject of a control infinitive is made explicit with a secondary edge in the PROIEL corpus, as discussed above.

Finally, PROIEL uses a more fine-grained set of labels for dependents of nouns by distinguishing arguments of (relational) nouns as well as partitive genitives from other, typically attributive modifiers. Even if the mapping from AGDT to PROIEL labels is one-to-many, it is possible to achieve decent results in automatic

⁴² Note that AGDT also annotates punctuation.

conversion by extracting cues from the target treebank. This work is described in Lee and Haug 2010, who report 91.9% label agreement in their Greek test set on agreeing edges, which account for 81.0% of edges. The 19.0% of disagreeing edges presumably reflect different conventions that do not directly follow from the annotation scheme, as described above.

If the PROIEL and AGDT treebanks could be fully harmonised, the resulting resource would be much more useful for the scholarly community. However, both treebanks are in active development, which means that an automated procedure is necessary. Instead of attempting wholesale conversion of one treebank to the other's format, recent work has focused on converting both treebanks to a third format, that of Universal Dependencies (UD),⁴³ which offers the additional advantage of facilitating typological comparisons with the many other treebanks in that format.

UD is an initiative that was not available when treebanks for Greek and Latin were first started, but it is now emerging as a *de facto* standard for dependency treebanks. The project aims to develop a universal annotation standard – a universal tagset with the possibility of language-specific extensions – and a set of treebanks following this standard. As of version 1.4, UD contains 64 treebanks, four of which are conversions of underlying PROIEL treebanks by Dag Haug. The conversion software is made available through the command-line interface to the PROIEL corpora.

UD has an LFG heritage just like PROIEL, so there are numerous points of convergence. For example, both UD and PROIEL distinguish control infinitives (XOBJ/XCOMP, see section 7.1) from other complement clauses (COMP), whereas these are not distinguished in AGDT. Nevertheless the UD format differs from PROIEL (and AGDT) in several ways, also structurally, since it consistently takes function words as dependents and never as heads, and also in taking the first conjunct to head a coordination, rather than the conjunction. However, these differences are well defined and easily retrievable based on the parts of speech of the involved words, so the necessary conversions are comparably easy to perform. On the side of labelling, the universal dependencies scheme is closer in granularity to AGDT, which means that most PROIEL labels can be easily converted (with loss of information). However, it is important to notice that Universal Dependencies offer a “metascheme” with principles and tags intended to apply to multiple languages. Still, it does allow for language-specific extensions. Future work will therefore aim at keeping the conversions from AGDT and PROIEL to Universal Dependencies compatible in order to offer a single, coherent resource. There is a substantial challenge on the AGDT side because Universal Dependencies, like PROIEL, employ secondary edges in order to capture e.g. implicit subjects in control structures.

Finally, it is worth mentioning that we have also worked on converting the PROIEL dependency structures into phrase structures. This is a different kind of challenge which can be best described as hypothesis testing rather than pure

⁴³ <http://universaldependencies.github.io/docs/>

conversion. As was mentioned above, it is not really clear what constituent structures should be assumed for free word order languages like Greek and Latin. There is therefore no objective truth to aim for in the conversion. Instead, what we are doing is to “inject” (presumed) linguistic knowledge into the source data – things like projection levels, adjunction structures and derivations of discontinuous structures – in a way that is theoretically plausible and empirically adequate. Some preliminary results are reported in Haug (2011), but this remains an active field of research.

10. Conclusions

In this article we have presented the PROIEL family of treebanks. The treebanks are all developed using the same annotation schemes and the same set of open-source software tools. The schemes and software tools were created especially for the needs of the project languages, all early attestations of languages from several Indo-European branches. An enriched dependency scheme was especially tailored for these languages, which all have very rich morphology and relatively free word order, and do not lend themselves easily to phrase structure analysis. Capturing as much structure as possible in typically very limited data sets was considered more important than using a scheme which would make it easy to train a syntactic parser, which lead us to use both empty nodes and secondary dependencies. The dependency scheme used in combination with syntactically defined parts of speech, very detailed morphology tags and word order information stored in a separate layer yield detailed descriptions of the richly inflecting project languages, and have proved easy to adapt to new languages. The PROIEL-style data have been converted successfully to several other annotation schemes, typically with loss of information since most target schemes are less detailed.

Whenever the annotation scheme was adapted to a new language, it was usually done in close cooperation between the project groups. As a result, the treebanks are unusually compatible and allow easy comparison of these closely related languages. They also provide good coverage of early attestations of several Indo-European branches. As such the PROIEL schemes and tools already function as a standard for dependency treebanks of languages of this kind, and the bulk of data on this format is steadily growing.

An important priority has been to provide detailed, many-layered annotation in order to make the most of sparse data, which is a real concern for most of the project languages. The projects have also accumulated valuable experience with annotating unstandardised texts and texts from periods of syntactic and morphological change. The PROIEL-style data have already been used in many published linguistic studies, and have also served as training data for other electronic resources and as linguistic input data for digital edition of early manuscripts.

References

Adesam, Y. & Bouma, G. (2016). Part-of-Speech tagging Old Swedish. In *Proc of Language Technology for Cultural Heritage, Social Sciences, and Humanities*. Berlin.

- Andrews, A.D. (1971) Case agreement of predicate modifiers in Ancient Greek. *Linguistic Inquiry* 2(2), pp. 127–151.
- Andrews, A.D. (1982) Long distance agreement in Modern Icelandic. In P. Jacobson & G. K. Pullum (eds.), *The nature of syntactic representation*. Dordrecht: D. Reidel, pp. 1–33.
- Bamman, D., Crane, G., Passarotti, M. & Raynaud, S. (2007). *Guidelines for the Syntactic Annotation of Latin Treebanks*. Technical report. Boston: Tufts Digital Library.
- Berdičevskis, A. (2015). Estimating Grammememe Redundancy by Measuring Their Importance for Syntactic Parser Performance. In *Proceedings of the Sixth Workshop on Cognitive Aspects of Computational Language Learning*, 65–73. Association for Computational Linguistics.
- Berdičevskis, A., Eckhoff, H. & 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, D. & Eckhoff, H. (to appear). Machine-assisted multilingual alignment of the *Codex Suprasliensis*.
- Bouma, G. & Adesam, Y. (2013). Experiments on sentence segmentation in Old Swedish editions, *Proceedings of the workshop on computational historical linguistics at NODALIDA 2013*; May 22-24; 2013; Oslo; Norway. NEALT Proceedings Series 18
http://www.ep.liu.se/ecp_article/index.en.aspx?issue=087;article=002
 (accessed 8/24/2015)
- Brants, T. (2000). TnT – a statistical part-of-speech tagger. *Proceedings of the Sixth Applied Natural Language Processing Conference ANLP-2000*, Seattle, WA.
- Delsing, L. (2002). Fornsvenska textbanken. In *Svenska språkets historia i Östersjöområdet*. Studier i svensk språkhistoria 7. Konferensrapport från Den sjunde sammankomsten för svensk språkhistoria. Svante Lagman, Stig Örjan Ohlsson and Viivika Voodla (eds.). Tartu
- Eckhoff, H.M. (2011). *Old Russian Possessive Constructions: A Construction Grammar Approach*. Berlin: Mouton de Gruyter.
- Eckhoff, H.M. (2015): Animacy and differential object marking in Old Church Slavonic. *Russian Linguistics* 39(2).
- Eckhoff, H. & Berdičevskis, B. (2016): Automatic parsing as an efficient pre-annotation tool for historical texts. *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH, held in conjunction with COLING)*. <http://de.clarin.eu/en/current-issues/lt4dh/lt4dh-proceedings>
- Eckhoff, H. & Berdičevskis, B. (2015): Linguistics vs. digital editions: The Tromsø Old Russian and OCS Treebank. *Scripta & e-Scripta* 14–15.
- Eckhoff, H. & Haug, D. (2015): Aspect and prefixation in Old Church Slavonic. *Diachronica* 32(2), pp. 186–230.
- Fort, K. & Sagot, B. (2010): Influence of pre-annotation on POS-tagged corpus development. In *Proceedings of the Fourth Linguistic Annotation Workshop*, ACL, pp. 56–63.

- Haug, D., Eckhoff, H. & Welo, E. (2014). The theoretical foundations of givenness annotation. In K. Bech and K. Eide (eds.), *Information Structure and Syntactic Change in Germanic and Romance Languages*. Amsterdam: John Benjamins.
- Haug, D. 2011. From dependency structures to LFG representations. In M. Butt & T. Holloway King (eds.), *Proceedings of LFG12*, CSLI Publications, p. 271–291
- Haug, D.T.T., Jøhndal, M., Eckhoff, H.M., Welo, E., Hertenberg, M.J.B. & Muth, A. (2009). Computational and Linguistic Issues in Designing a Syntactically Annotated Parallel Corpus of Indo-European Languages. *Traitement Automatique des Langues* 50.
- Haug, D.T.T & Jøhndal, M.L. (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 (ELRA).
- Haugen, O.E. & Øverland, F.T. (2014). Guidelines for Morphological and Syntactic Annotation of Old Norwegian Texts. *Bergen Language and Linguistics Studies* 4(2). <https://bells.uib.no/bells/issue/view/158>
- Haugland, K.E. 2007. *Old English Impersonal Constructions and the Use and Non-Use of Nonreferential Pronouns*. Ph.D. dissertation, University of Bergen.
- Hertenberg, M.J.B. (2014): "The valley" or "that valley"? Ille and ipse in the Itinerarium Egeriae. In P. Molinelli, P. Cuzzolin & C. Fedriani (eds.): *Latin vulgaire – Latin tardif X. Actes du Xe colloque international sur le latin vulgaire et tardif. Bergamo, 5-9 septembre 2012*. Bergamo: Sestante edizioni.
- Hertenberg, M.J.B. (2011). Classical and Romance usages of ipse in the Vulgate. *Oslo Studies in Language (OSLa)* 3(3), pp. 173–188
- Jøhndal, M. (2012). *Non-finiteness in Latin*. PhD dissertation. University of Cambridge
- König, E., and Lezius, W. (2003). *The TIGER language - A Description Language for Syntax Graphs, Formal Definition*. Technical report. IMS, University of Stuttgart, Germany.
- Lee, J. and Haug, D. (2010). Porting an Ancient Greek and Latin Treebank. In *Proc. Conference on Language Resources and Evaluation (LREC)*.
- Lindberg, R. (2013). *Definiteness in Old Church Slavonic : A Study of How Long and Short Form in Adjectives Reflect Information Status*. Master's thesis. University of Oslo.
- Mitchell, B. (1985). *Old English Syntax*. 2 vols. Oxford: Clarendon.
- Muth, A. (2015). *Indefiniteness, Animacy and Object Marking: A Quantitative Study Based on the Classical Armenian Gospel Translation*. PhD thesis. University of Oslo.
- Seeker, W., Farkas, R., Bohnet, B., Schmid, H. & Kuhn, J. (2012). Data-driven dependency parsing with empty heads. In *Proceedings of COLING 2012: Posters*. 1081-1090.
<http://www.aclweb.org/anthology/C12-2105>
- Skjærholt, A. (2011). More, faster: Accelerated corpus annotation with statistical taggers. *Journal for Language Technology and Computational Linguistics* 26(2).
- Söderwall, K.F. (1884–1918). *Ordbok över svenska medeltids-språket*. (A–L, M–T, Þ (TH)–Ö. Samlingar utgivna av Svenska fornskriftsällskapet. Serie 1. Svenska skrifter. 27.)

Traugott, E.C. (1992). Syntax. In R.M. Hogg (ed.), *The Cambridge History of the English Language*, vol. 1: *The Beginnings to 1066, 168–289*. Cambridge: Cambridge University Press.