Creating a Parallel Treebank of the Old Indo-European Bible Translations[1]

Dag T. T. Haug, Marius L. Jøhndal

University of Oslo
daghaug@ifikk.uio.no, mlj32@cantab.net

Abstract

In this paper, we report on the creation of a syntactic annotation scheme to be used in a comparative study of the oldest extant versions of the New Testament in Indo-European languages: the Greek original, as well as the translations into Latin, Gothic, Armenian and Church Slavonic. The tag set was created in a pilot study involving syntactic annotation of the Gospel of Mark in Greek and Latin. The resulting tag set is well suited for capturing syntactic variation between these languages, particularly in areas having to do with pragmatics and information structure – as the treebank is created within a larger project in this field – but also more general syntactic differences.

1. Introduction

The project *Pragmatic Resources of Old Indo-European Languages* (PROIEL) aims to study the linguistic means of information structuring which are offered by the grammar of Greek, Latin, Armenian, Gothic and Church Slavonic, i.e. the means that the lexicon and the syntax of these languages make available[2] for expressing such categories as old and new information, contrast, parallelism, topicality and others. Five particular phenomena will be examined in the PROIEL project:

- Word order
- The definite article
- Discourse particles
- Anaphoric expressions, including zero anaphora
- Participles and absolutive constructions

These topics were chosen because they are known to be important in information structure systems cross-linguistically and because they are areas where the languages in the corpus are likely to diverge. For example, Ancient Greek is the only language in the corpus to have a grammaticalized definite article. This language is also well known for its abundance of discourse particles, which cannot be rendered directly in the target languages. Word order is notoriously free in these languages, and while this led to direct adoption of the Greek word order in many cases, there are still patterns that cannot be rendered directly. Similarly, the anaphoric and participial systems vary widely.

The most important objective for our treebank is to be able to represent these phenomena correctly with as fine-grained information as possible. On the other hand, it is likely that in the course of the project, we will find other phenomena that are relevant to the general topic of information structure, so we need to be prepared to adapt our scheme to changing requirements. Finally, it is important that the treebank is created in such a way as to be useful for a wider audience, no matter what topics they are interested in. The annotation scheme must therefore be suitable for representing the general structure of sentences in these languages.

It was decided that rather than focussing on creating coherent data from the very start of the annotation process, the best way to accomplish our objectives was to annotate a pilot text while we were developing the annotation scheme. This way we could maximize the value of feedback from annotators, gain experience with the annotation process itself, and have a readily available testbed during development of the software. The remainder of this paper describes this process and its outcome.

2. Preparing the pilot text and creating the annotation tools

Building a treebank is labour-intensive, so our initial concerns were to avoid duplication of efforts and to get our annotators started as quickly as possible. This was greatly facilitated by the availability of a morphologically annotated electronic version of the Greek New Testament (Sandborg-Petersen, 2008) and by the work done by the Perseus digital library (Crane, 1987; Crane et al., 2001) on their electronic version of Jerome's Vulgate and word-lists for Latin and Greek.

We used these resources to prepare the text for the pilot study. This text consists of the Greek and Latin versions of the Gospel of Mark – which in each language amounts to roughly 13,000 words or 10% of the complete New Testament.

Due to the complexity of Biblical textual criticism, and since the purpose of the overarching project is to do a cross-linguistic comparative study, we chose to ignore manuscript variants. Our texts are instead based on the text of a specific edition, and we only correct digitization errors, should these occur.

The preparation and annotation of the pilot text proceeded in four stages:

- Pre-processing
- Automated morphological tagging
- Manual annotation by annotators
- Manual review by a reviewer

The pre-processing stage involved segmentation, detection of sentence boundaries and sentence alignment. Segmentation is occasionally problematic as certain morphemes behave as separate entities in the syntactic model we use, but

[1]The research project described here is funded by the *Norwegian Research Council*'s YFF program, grant no. 180632. We thankfully acknowledge this support. The glossing in this paper follows the Leipzig Glossing Rules (`http://www.eva.mpg.de/lingua/resources/glossing-rules.php`).

[2]Prosody presumably also played a role which is, however, irrevocably lost for us.

form part of other words. This happens, for example, with instances of *krasis*[3] in Greek. A trickier case is presented in Latin where enclitics like *-que* cannot always be tokenized unambiguously.

For detecting sentence boundaries, we decided to use punctuation as a guide, as the canonical division into verses frequently fails to match what we consider to be meaningful syntactic units. A particular problem was presented in our Vulgate text as it lacks punctuation altogether. To solve this, the punctuation from a different electronically available edition, the Clementine Vulgate, was merged into our text by first aligning the orthography of the two editions, then using an implementation of the standard *diff* algorithm (Hunt and McIlroy, 1976) and finally patching the target text using only chunks that involved insertion of punctuation. This simple approach gave good results in spite of numerous textual differences, and only occasionally resulted in off-by-one errors in sentence division.

To answer the questions posed in the research project, corpus users will eventually need to be able to query the same sentence in multiple languages. This requires sentences to be aligned, and our experiments indicate that Gale and Church (1993)'s algorithm performs this task well with chapters as hard delimiters and sentence boundaries as soft delimiters.

As ready-made word-form lists were available, morphological tagging was done simply as an assignment of one or more pairs of lemma and morphological tags to each token in the text. The morphological tag set that we developed is more fine-grained than those of our source data, so for many tokens the level of detail of the assigned morphological tags was insufficient. We were able to address this by manually adding assignment rules, and harvesting additional rules from data already disambiguated by annotators. The morphological tags are positional tags and derived from the system used by the Latin Dependency Treebank (Bamman and Crane, 2006). For the purposes of the PROIEL project, more fine-grained distinctions had to be made for certain parts of speech, in particular pronouns, whose functions are primarily pragmatic. We have also deviated from the traditional grammarian's view of morphology and adopted a more 'modern' view by treating adverbs that double as prepositions as intransitive prepositions, and by merging the two categories particle and adverb (see table 1).

Major part of speech	**Minor part of speech**
Verb	
Noun	declinable common noun indeclinable common noun declinable proper noun indeclinable proper noun
Pronoun	relative pronoun interrogative pronoun indefinite pronoun demonstrative pronoun personal pronoun possessive pronoun personal reflexive pronoun possessive reflexive pronoun reciprocal pronoun
Numeral	declinable cardinal number indeclinable cardinal number ordinal number
Adjective	
Article	
Adverb	comparable adverb relative adverb interrogative adverb other non-comparable adverb
Conjunction	
Subjunction	
Preposition	
Interjection	
Foreign word	

Table 1: The parts of speech defined in the PROIEL morphological tag set.

It has furthermore been necessary to introduce a means for indicating ambiguity of form or unresolvable syncretism, e.g. for nouns that alternate between genders. This has been solved by allowing multiple inheritance within each field of a positional tag, so that, for example, the tag for masculine gender has a super-tag that indicates masculine or neuter gender, one that indicates masculine or feminine gender etc.

The two manual stages of the process – annotation and review – were performed using variants of the same graphical interface. We wanted an interface that students could use on typical campus workstations which frequently have a limited choice of installed software and restrict students' rights to run local, stand-alone applications. We therefore chose to develop a light-weight web-based interface that would function with only a modern browser and client-side scripting.

The interface is designed as an incremental 'wizard' that splits the annotation into three steps. First, annotators verify and, if necessary, adjust sentence boundaries. We have found that this ought to be restricted as annotators felt tempted to override the judgements of the text editors and therefore did excessive adjustments of sentence boundaries. This may be due to the style of our texts in which sentences tend to 'run together' and a large number of sentences are introduced by *and*. The choice between coordinating a main clause with the preceding clause or not is thus often an arbitrary one. We therefore let annotators adjust the sentence boundaries only one token at a time so that they could only correct off-by-one errors.

The second step of the 'wizard' involves morphological disambiguation. Annotators are presented with the output from the automated morphological tagging and are asked to choose the correct lemma and tag pair in cases of ambiguity. Finally, in the third step, annotators build depend-

[3]The term *krasis* refers to a contraction in which the final vowel of one word coalesces with the initial vowel of the next, and the two words are written together.

ency structures for each sentence. This is done visually and guided by a simple rule-based 'guesser' that suggests the most likely dependency relation based on the morphology of head and dependent.
In addition to the interface for annotation, and an interface for text browsing, we added functionality for tracking change history and for inserting cross-references to other information sources such as dictionaries. In particular, we have made use of *Strong's Concordance* (Strong, 1890) and the *Analytical Lexicon of the Greek New Testament* (Friberg et al., 2000), as these were the basis for lemmatization in our Greek text.
The system is based on Ruby on Rails with a database backend. As a by-product of this choice, the system offers not only a traditional web-interface to the corpus, but also exposes a RESTful XML interface that can be used by clients to query the database. This should facilitate interchange of data and direct reuse of our work in other contexts.

3. The development of the annotation scheme

As noted above, all the languages in our corpus have a 'free' word order, i.e. the word order does not indicate syntactic dependencies or grammatical functions, but serves pragmatic purposes. Therefore, while word order data are important for PROIEL, they cannot be conflated with information about grammatical function as is done in a phrase structure grammar. For this reason, it was decided to base the annotation scheme on dependency grammar (DG). This also had the advantage that other projects developing treebanks of Latin, e.g. the Latin Dependency Treebank (LDT), are based on DG, using a faithful adaption of the well-documented Prague Dependency Treebank (PDT) (Hajič, 1998).
We began our work using the Greek and Latin versions of the New Testament, since these exist publicly available in electronic form with morphological annotation. We expected the syntax of most old Indo-European languages to be sufficiently similar to be captured within a single annotation scheme and our experienced with the Greek and Latin texts have confirmed this. There are diverging constructions, of course, but they can all be captured using our primitive syntactic relations, and we do not expect Gothic, Armenian or Church Slavonic to be different in this respect.

3.1. General presentation

While we wanted to keep the option to automatically convert our treebank to a more general format, we soon realised that the level of granularity of the PDT annotation scheme or the LDT annotation scheme (Bamman et al., 2007) would not be sufficient for PROIEL. Table 2 shows the general outline of our annotation scheme in comparison with that used by LDT. It is more fine-grained than the LDT scheme, both in the domain of verbal arguments and that of adnominal functions. To study the interaction between syntax/argument structure and pragmatics in determining word order, we need to be able to separate objects (OBJ) from other arguments of the verb (OBL). Furthermore, agent expressions (AG) are particularly interesting for the syntax-pragmatics interface, because they are both optional and receivers of a thematic role from the verb. [4]
In the adnominal domain, it is well known that there are interesting correlations between types of genitives and information structure. For example, possessive genitives tend to be old information in a text and are typically used to access new referents, whereas object genitives are more often new information. Partitive genitives are special as they, and not their syntactic heads, introduce the discourse referent of a noun phrase: 'two of the disciples' refer to a group of disciples, and not to some kind of 'twoness', unlike 'the teaching of the disciples'. It is therefore essential for PROIEL to distinguish these uses of the genitive.
There is one notable exception to the general pattern that our tags are more fine-grained than those of the LDT; the LDT scheme provides 9 subtypes of auxiliary relations: AuxP for prepositions, AuxC for conjunctions, AuxR for the reflexive passive etc. In our opinion all items with the relation AuxX in the LDT can be conflated to a single relation as instances can still be differentiated based on lexical information when the need arises.

3.2. Granularity

By asking annotators to do fine-grained classification of the data, we run the risk of more inconsistencies in the application of the scheme. For this reason, we have introduced some 'super-tags', i.e. tags that we ask the annotators to use whenever they are in doubt. For example, it can be hard to tell whether a given relative clause is restrictive (ATR) or not (APOS). We provide a tag REL for such cases, so that the annotators do not simply guess.
However, in the case of adnominal tags, we purposefully did not provide any such super-tag, in order to test the viability of making distinctions within this domain. The results were mixed. In the beginning, we asked annotators simply to distinguish attributes and appositions. After a couple of weeks, we introduced more granularity by means of the tags PART, to be used for partitive expressions, and OBL, to be used whenever an expression is an *argument* of the noun – typically an object genitive as in *amor fati* 'love of faith'. OBL was chosen because this is the relation we use for non-object arguments in the verbal domain.
When the pilot was finished, we studied how annotators had used these tags. Although the results are not statistically significant, they were valuable in guiding our development of the annotation scheme. In general, the annotators coped well with the PART relation: of 42 uses of this relation, only 3 were wrong – not too bad a result at such an early stage in the annotators' training. Moreover, the errors could easily be detected automatically, since they did not involve any uses of PART with an adnominal genitive that should have had another relation, but rather the generalisation of PART to other contexts with partitive semantics, i.e. a genitive object and an object of the Greek preposition *apo*.
The concept of arguments of nouns was harder to apply. This relation was used 22 times for items dependent on

[4] The decision to include the AG relation, which combines syntactic function and semantic role, was a pragmatic choice motivated by the fact that we do not expect to have the resources to do a full tectogrammatical annotation as in the PDT.

Latin Dependency Treebank	PROIEL Corpus	Explanation
PRED	PRED	Main clause predicate
*	PRED	Subordinate clause predicate
SBJ	SUB	Subject
OBJ	OBJ	Object
	OBL	Oblique
	AG	Agent
	XOBJ	Open complement clause
ADV	ADV	Adverbial
ATR	ATR	Attribute
	NARG	Nominal argument
	PART	Partitive
ATV	XADV	Free predicative
PNOM	XOBJ	Subject complement
OCOMP	XOBJ	Object complement
COORD	*	Coordinator
APOS	APOS	Apposition
AuxX (X defines the subtype of Aux)	Aux	Auxiliary
ExD	*	External dependency
	VOC	Vocative

Table 2: Sentential functions in LDT and PROIEL. An asterisk in one of the columns indicates that the two annotation schemes diverge in some other way than by one simply being more specific than the other.

nouns, 6 times erroneously. Apparently the possibility of using OBL adnominally tempted annotators into analysing verbs with an object and a PP complement as if the prepositional phrase were dependent on the object, e.g. so that *super* is a dependent of *manus* in the participial construction

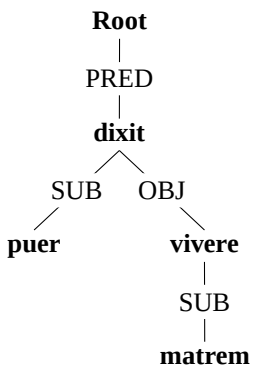
(1) *imponens* *manus* *super*
put.PRS.PTCP.NOM.SG hand.ACC.PL upon
illos
they.ACC.PL.
'laying his hands upon them'

Such errors cannot be detected automatically. Moreover, since the OBL tag is used in more contexts, we run the additional risk of contaminating the entire set of OBL-relations. Not only was OBL used in cases where another relation should have been used, there were also cases where PART and OBL were not used when they should have been. As part of our analysis of the data from the pilot annotation, we examined the 123 cases of genitive nouns dependent on another noun that had been annotated after the introduction of PART and OBL as adnominal tags. 17 of these were given an incorrect analysis, and in 16 cases this was because ATR was used when PART or OBL would have been correct.
The period of pilot annotation has taught us that it is difficult for annotators to distinguish different functions in the adnominal domain. Still we will continue to make these distinctions, but we no longer use the relation OBL, but rather a separate relation NARG (nominal argument) which is devoted to arguments of nouns. In this way, we have an 'exit strategy' in case the the quality of the annotation remains low, since we can merely convert all NARGs to ATRs. Also, since we have now had the opportunity to test the annotators' ability to make fine-grained distinctions in the adnominal domain, we have introduced a super-tag ADNOM so that we no longer force the annotators to choose when they are in doubt.

3.3. Dealing with covert elements

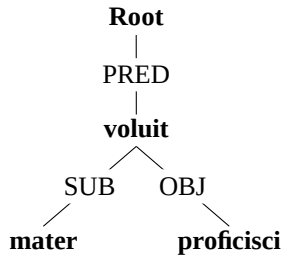
A dependency grammar is well equipped to deal with the free word order of ancient languages. However, it is less well adept at representing another feature typical of old Indo-European languages, namely ellipsis. The DG formalism has difficulties with all constructions without a clear syntactic head, e.g. ellipsis, coordination (and in particular asyndetic coordination) and sentences lacking a verb (most often the copula).
Different solutions have been devised to these problems; in the following we describe our solution, which tries to capture the facts in a theory-neutral manner.
First consider a less well-known problem for dependency grammar, namely ellipsed *dependents*. Ellipsis of dependents is much more frequent than ellipsis of heads and, while it is generally easier to deal with, it can sometimes lead to problems. Consider the treatment of the accusative with infinitive (AcI) (example 2) and the complement infinitive in the LDT (example 3):

(2)



puer *dixit* *matrem*
boy.NOM.SG say.3SG.PFV mother.ACC.SG.
vivere
live.PRS.INF
'The boy said his mother was alive.'

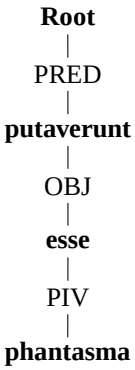
(3)



mater *voluit* *proficisci*
mother.NOM.SG. can.3.SG.PFV travel.PRS.INF
'Mother wanted to travel.'

The fact that we here have two different constructions is signalled only by the presence of a subject daughter in example 2. However, Latin being a pro-drop language,[5] this subject is optional:[6]

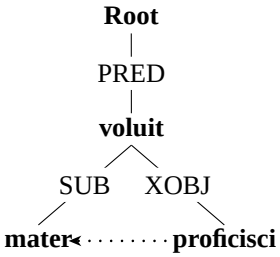
(4)



putaverunt *esse* *phantasma*
think.3.PL.PFV be.PRS.INF ghost.ACC.SG
'They thought [he] was a ghost.'

A search for all AcI structures would have to sacrifice precision (by matching all infinitive objects) or recall rate (by matching all infinitives that have a subject daughter). This problem is particularly important to PROIEL, since the subject of the infinitive in such examples as example 4 can only be left out because it is given information in the context.
At first we tried to solve the problem by *not* letting the verb stand in for the whole sentence, but rather let sentences (including AcIs) be represented by an empty node that dominated the verb and its arguments, so that the defining feature of these empty nodes was the possibility (but not necessity) of dominating a subject. However, this quickly leads to problems: the empty elements are hard to deal with computationally and result in an unmotivated distinction between verbs in full sentences, which would be sisters of their arguments, and verbs in participial constructions, which would dominate their arguments. Although this system provided an intuitive way of dealing with so-called 'gapping' (the absence of the verb in the second conjunct, see example 12), we quickly abandoned it.
Inspired by Lexical-functional grammar, we instead chose to represent the structural difference between infinitives in AcIs and complement infinitives as two contrasting relations, OBJ and XOBJ. The latter function is by definition one which cannot have an overt subject, but shares its subject with another element in the clause. We designate this structure-sharing by what we call 'slash notation'.[7] The full representation of example 3 is therefore:

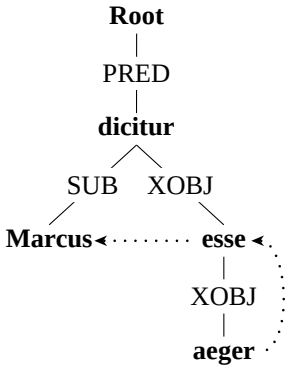
(5)



mater *voluit* *proficisci*
mother.NOM.SG. can.3.SG.PFV travel.PRS.INF
'Mother wanted to travel.'

The arrow in this example should be interpreted as a secondary dependency relation. In this case it shows that *mater* is the subject of both *proficisci* and *voluit*. This accounts for case agreement with predicate nominals in the dependent infinitive construction, as in the following example, which also shows how we deal with instances where the subject of the XOBJ is not overtly realized:

(6)



Marcus *dicitur* *aeger*
Marcus.NOM.SG say.3.SG.PRS.PASS ill.NOM.SG.
esse
esse.PRS.INF.
'Marcus is said to be ill.'

In traditional grammar the subject of *aeger* is supplied by the verb *esse*. We incorporate this by letting the slash arrow point to the head verb whenever it 'ought' to pointed to a 'pro-dropped' argument. This has the further advantage of

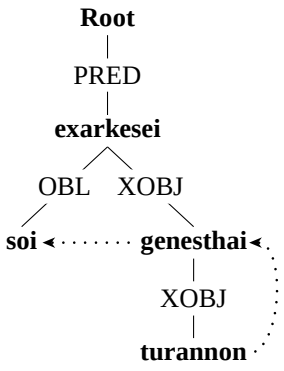
[5]The term *pro-drop language* refers to languages in which some pronouns may be omitted when they can be inferred pragmatically.

[6]In this tree, we use PIV for the traditional notion of 'predicative complement', which actually does not exist in our scheme; see below.

[7]This designation is in turn inspired by the vaguely similar SLASH-lists of Head-driven phrase structure grammar.

making it easier to validate annotations; we can enforce the principle that every XOBJ or XADV relation should have one slash arrow and that this arrow should point towards the head verb or an element dominated by the verb.
esse in turn gets its subject from *Marcus*. Notice that we treat the traditional category of predicative complement as XOBJ, seeing that the facts are the same: the element is subcategorized for by the verb and does not have a direct relation to its subject.
Our representation is neutral between *control* and *raising* analyses. Compare the example above to the following example:

(7)



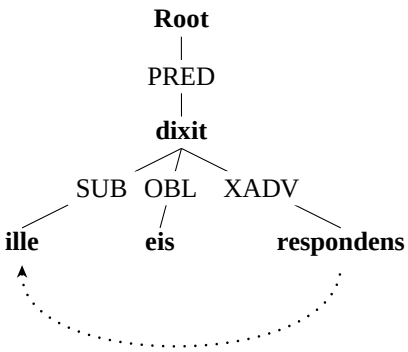
exarkesei *soi* *genesthai*
suffice.3.SG.FUT you.DAT become.PFV.INF
turannon
tyrant.ACC.SG
'It will suffice for you to become a tyrant.'

Many theories would treat example 6 as raising and example 7 as control, to explain the differences in case agreement. Our representations are meant only as input to such discussions, so we make no commitment regarding the status of our slash dependencies, nor do we ask our annotators to make decisions based on linguistic theory. The annotators should simply identify the subject of infinitives and predicative complements: if the subject of the XOBJ is not present in the subtree dominated by the governing verb, the slash arrow should point to the verb, which 'supplies the subject' according to traditional grammar.

3.4. Other uses of the slash notation

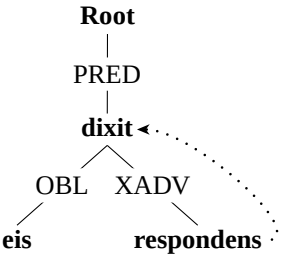
The slash notation was first introduced to separate AcI's from complement infinitives even in cases where the subject of the AcI has been 'pro-dropped'. It is a modest but powerful addition to our data-model that allows us to obtain a richer annotation for many structures. The slash notation most manifestly translates to infinite predications that are not arguments of the main verb, i.e. predicative (conjunct) participles:

(8)



ille *respondens*
he.NOM.SG answer.PRS.PTC.NOM.SG
dixit *eis*
say.3.SG.PFV they.DAT.PL
'Answering them, he said'

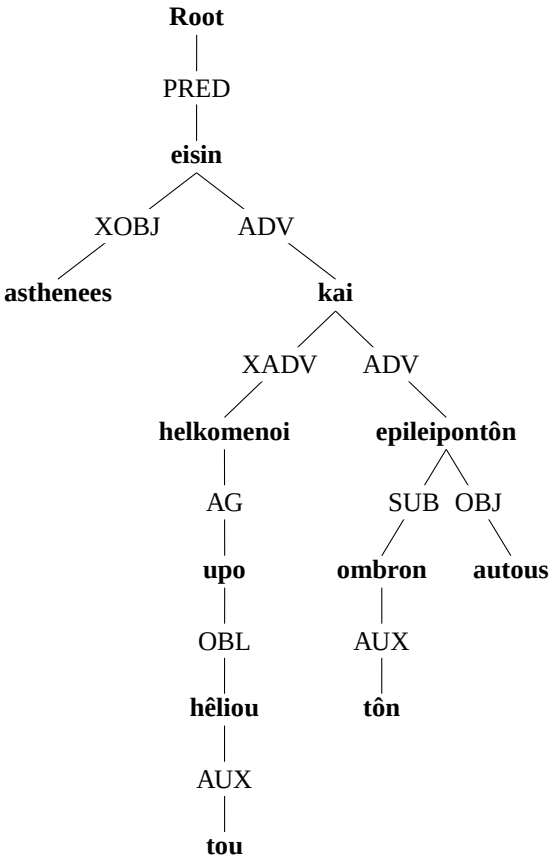
(9)



respondens *dixit*
answer.PRS.PTC.NOM.SG say.3.SG.PFV
eis
they.DAT.PL

Note that this preserves the structure even when the subject of the predicative participle is 'pro-dropped'. This is otherwise hard to achieve: If the participle were to depend on its subject, the structure would be different when the subject was omitted. We could have used complex tags to denote the subject of the participle, but then we would have had to deal with predicative participles that depend on main clause adjuncts, of which there may be several. The slash notation solves this by pointing to the subject of the participle and not to the relation which the subject of the participle has.
Dependency grammars generally speaking only allow coordination of elements related to the same head via the same relation. It is, however, not hard to find examples with a conjunct participle coordinated with an adverbial element, e.g. an absolute genitive:

(10)

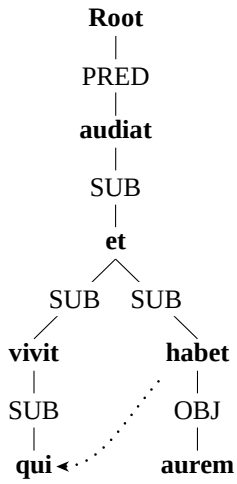


tôn *ombrôn*
the.GEN.PL rain.GEN.PL.
epileipontôn *autous* *kai*
leave behind.PRS.PTCP.GEN.PL they.ACC.PL and
upo *tou* *hêliou*
by the.GEN.SG sun.GEN.SG
helkomenoi *asthenees*
draw up.PRS.PTCP.NOM.PL weak.NOM.PL
eisi
be.3.PL.PRS
'With the rain leaving them being and drawn up by the sun, they [sc. the rivers] are weak.'

Our notation solves this problem. The daughter nodes of XADV relations *always* have a slash arrow, so the 'X' merely serves to indicate the presence of the slash.[8] This means that we can coordinate XADV and ADV without distorting the analysis. The advantage of our notation would become even clearer if the participle had an overt subject, as the two conjuncts in this case would have to have different heads in a 'classical' analysis.

Once introduced, the slash notation can be exploited for richer annotation of other structures that involve ellipsis or structure-sharing. Since the verb substitutes for the whole sentence, we treat coordination of two verbs as sentence coordination and use the slashes to indicate double dependencies (i.e. subject sharing):

(11)

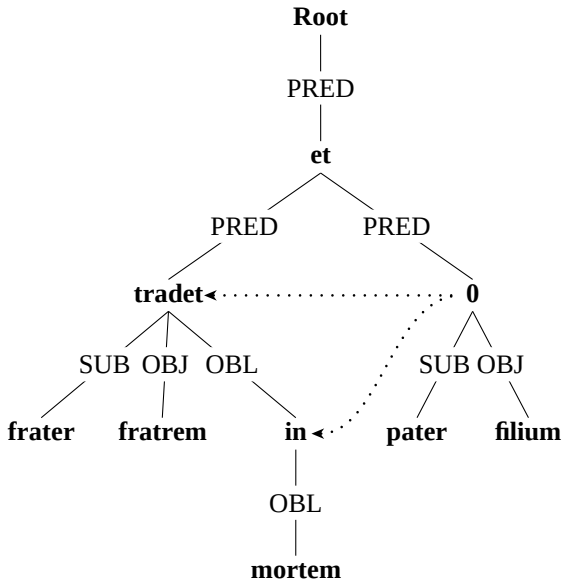


qui *vivit* *et* *aurem*
who.NOM.SG live.3.SG.PRS and ear.ACC.SG
habet *audiat*
have.3.SG.PRS hear.3.SG.PRS.SBJV
'Whoever lives and has ears shall hear.'

The advantage of this notation is evident in gapping constructions where the predicate is omitted in the second conjunct:

[8]The relation between XOBJ and OBJ is of another nature since verbs subcategorize differently for OBJ and XOBJ.

(12)



tradet *frater* *fratrem*
deliver.3.SG.PRS brother.NOM.SG brother.ACC.SG
in *mortem* *et* *pater*
to death.ACC.SG and father.NOM.SG
filium
son.ACC.SG
'The brother shall betray the brother to death, and the father the son.'

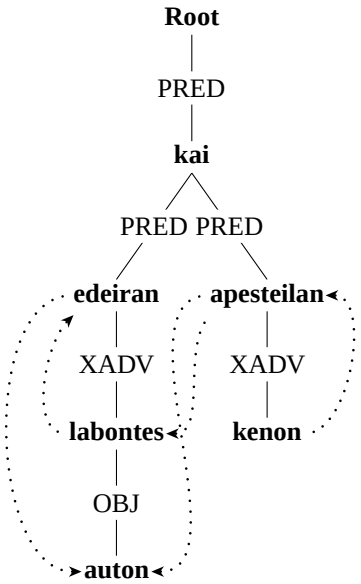
The combination of a restricted use of empty nodes and the slash notation makes it possible to preserve the structure of the tree. We also capture the fact that the argument *in mortem* is shared between the two conjuncts. The two slash arrows have rather different interpretation: the one from the empty node to the verb indicates sharing of lexical material, whereas the one from the empty node to the preposition *in* indicates a double dependency.

Since the slash relation is not labelled, it is important that the relation can be interpreted based on other information in the sentence. And, in fact, this remains possible. We can distinguish three uses of the slash notation.

- Slash arrows from an empty node to a sister node signal predicate identity
- Slash arrows from an XOBJ or XADV node to a mother or sister node indicate the subject of the infinite verb
- Slash arrows from other verbal nodes signal a shared argument

The first case is not a dependency relation at all, so there is no need to infer a label. In the second case, the slash arrow always indicates a SUB relation: there is widespread typological support for 'controlled' functions always being subjects, and this holds for the old Indo-European languages as well. Only the third kind of slash arrow may have different labels. We therefore constrain such arrows to cases where the shared arguments have the same function in both conjuncts. This is by far the most frequent case. The following example illustrates how unambiguous interpretation is possible even in complex cases:

(13)



kai	*labontes*	*auton*	*edeiran*
and	take.PRS.PTCP	he.ACC.SG	beat.3.PL.PFV
kai	*apesteilan*	*kenon*	
and	send away.3.PL.PFV	emtpy.ACC.SG.	

'Having captured him, they beat him and sent him away empty-handed.'

In this graph, we capture the information that *auton* is an object of all verbs in the sentence; that the subjects of the free predicatives *labontes* and *kenon* are elided arguments of the verbs *edeiran* and *apesteilan*; and that *labontes* is an adverbial adjunct (here, in fact, equivalent to a subordinate temporal clause) relevant to both main verbs.

Thus the simple addition of an extra binary relation in our data model enables us to capture a wide variety of facts about structure sharing without introducing a plethora of empty nodes. Notice also that our two levels of annotation are not interdependent: while the slash arrows cannot be interpreted without the dependency tree, the opposite does not hold. If in some processes (such as automated parsing) we are forced to exclude slash relations, the dependency tree can still be drawn independently and the slash relation added by other means.

The annotators made two kinds of errors in dealing with the slash arrows: sometimes they forgot to use them where they should have been used, and sometimes they attached arrows indicating double dependencies to an empty node. In example 12, they introduced an empty OBL-node under the empty verb in second conjunct, and a slash arrow from the OBL-node to the preposition *in*. In this way, they enforced a more consistent interpretation of the slash notation as an indication of identity of lexical material. We considered this option, but rejected it due to the proliferation of empty nodes it leads to.

Fortunately, both these errors are easily detectable, and as future extension the annotation interface will enforce validation constraints that prohibit dependency graphs that have an XOBJ or XADV nodes lacking a slash arrow, or in which a slash arrow exits an empty node which has been assigned a non-PRED relation.

4. Conclusion

While work on the complete PROIEL corpus is still in its infancy, we feel that the pilot stage of the project has enabled us to establish a relatively firm base for the annotation scheme to be used and the accompanying tools that annotators will rely on. In the course of our work, we have been confronted with many of the 'classical' difficulties that syntactic theory still struggles with, e.g. the difficulty of strict morphological categorisation and the analysis of ellipsis in Dependency Grammar, but also encountered novel problems that arise in ancient Indo-European languages. Our solutions to these problems should enable us to address the needs of the PROIEL project, but still be sufficiently theory neutral to ensure that the corpus will be useful for others, and the technologies used should enable an open exchange of data and eliminate many obstacles for potential reuse of our data.

5. References

David Bamman and Gregory Crane. 2006. The design and use of a Latin dependency treebank. In Jan Hajič and Joakim Nivre, editors, *Proceedings of the Fifth International Treebanks and Linguistic Theories*, pages 67–78, Prague. Data available from `http://nlp.perseus.tufts.edu/syntax/treebank`.

David Bamman, Marco Passarotti, Gregory Crane, and Savina Raynaud. 2007. Guidelines for the syntactic annotation of Latin treebanks. Technical report, Tufts Digital Library, Medford. Version 1.3.

Gregory Crane, Robert F. Chavez, Anne Mahoney, Thomas L. Milbank, Jeffrey A. Rydberg-Cox, David A. Smith, and Clifford E. Wulfman. 2001. Drudgery and deep thought: Designing digital libraries for the humanities. *Communications of the ACM*, 44(5):34–40.

Gregory Crane. 1987. From the old to the new: Integrating hypertext into traditional scholarship. In *Hypertext '87: Proceedings of the 1st ACM conference on Hypertext*, pages 51–56. ACM Press.

Timothy Friberg, Barbara Friberg, and Neva F. Miller. 2000. *Analytical Lexicon of the Greek New Testament*. Baker, Grand Rapids.

William A. Gale and Kenneth W. Church. 1993. A program for aligning sentences in bilingual corpora. *Computational Linguistics*, 19(1):75–102.

Jan Hajič. 1998. Building a syntactically annotated corpus: The Prague dependency treebank. In Eva Hajičová, editor, *Issues of Valency and Meaning. Studies in Honor of Jarmila Panevová*, pages 12–19. Charles University Press, Prague.

James W. Hunt and M. Douglas McIlroy. 1976. An algorithm for differential file comparison. Computing Science Technical Report 41, Bell Laboratories.

Ulrik Sandborg-Petersen. 2008. Tischendorf's 8th edition Greek New Testament with morphological tags. Version 2.0. `http://morphgnt.org/projects/tischendorf`.

James Strong. 1890. *The exhaustive Concordance of the Bible : showing every word of the text of the common English version of the Canonical Books*. Methodist Book Concern, New York.