Breaking down and putting back together: analysis and synthesis of New Testament Greek

Dag T. T. Haug, Hanne M. Eckhoff, Marek Majer, Eirik Welo

University of Oslo, Norway

Abstract

In this paper we first briefly describe the design of a corpus containing the Koine Greek original text of the New Testament and its translations into Gothic, Latin, Old Church Slavic and Armenian. We then discuss extensively the annotation that we have applied in each layer of annotation: morphology and syntax, information structure, animacy, and token alignment. For each type of annotation we provide some preliminary results and applications that draw on it, often in combination with other layers of annotation.

Keywords

corpus linguistics, information structure, New Testament Greek, pragmatics, syntax

1. Background

Pragmatic Resources in Old Indo-European Languages (PROIEL) is a project based at the Department of Philosophy, Classics, History of Art and Ideas, University of Oslo. The main objective of the project is to investigate what morphological and syntactic resources different older Indo-European (IE) languages utilize for expressing categories related to information structure. In particular we focus on how word order, definiteness, anaphoric expressions, discourse particles and participles are employed in information packaging.

The goal of PROIEL is to study comparatively how information packaging works in old IE languages. In order to do this, we have developed a richly annotated corpus consisting of the Koine Greek original of the New Testament.
(NT) and its translations into Latin (Jerome's Vulgate), Gothic (Wulfila's translation), and Slavic (as found in the Codex Marianus). The NT provides the most suitable material for conducting comparative research since it is a naturally occurring parallel corpus and covers the oldest texts of substantial length written in Germanic, Slavic and Armenian.

Apart from that, however, texts from outside the NT are also added; we wish to expand the corpus so as to achieve diachronic depth for each of the language families, but also to provide non-religious texts so that we can correct for the special language of the NT with its possible influence from Semitic and its stylistic limitations. For the translation languages we also want to include works written originally in a given language, so that the text may function as a contrastive background to the translation from Greek. So far, one Latin text has been added (Julius Caesar’s *Commentaries on the Gallic War*); the addition of further texts is planned.

Even though the structure of the corpus is designed so as to fit the needs of the research interests of the PROIEL project, its potential uses are far greater. One of the purposes of the project is to offer the scientific community a valuable tool suitable for various kinds of research. The rich annotation our corpus involves makes it of interest both to historical linguists – Indo-Europeanists as well as specialists in the respective languages – and to scholars interested in purely synchronic issues concerning any of the languages. What is more, the data may be of interest to researchers interested in Bible scholarship, translation studies, computational linguistics and numerous other fields. For this reason, the structure of the corpus is at all times open - there always exists a possibility of adding further layers of annotation should it be required for a given purpose.

Basing linguistic research on publicly available corpora leads to replicability. Openness and access to raw data allow other researchers to question the underlying assumptions of our analyses. In the traditional philological literature, it is not uncommon to find wildly divergent estimates on, say, the order of verb and object in one and the same author or text, because scholars have included different kinds of clauses, used different texts and excerpts, worked with different definitions of subject and object and so on. These assumptions are almost never made entirely explicit, and even if they were, it would not be possible to change them without doing the entire data sampling anew.

Corpora do not eliminate such problems, but they do make them more tractable. Since the raw data are available, everyone can inspect exactly what was counted as a subject. And to the extent that the annotation permits, counts can be redone with changed assumptions. For example, since the clause type can be recovered in the PROIEL annotation, it is easy to distinguish complement and adverbial clauses and provide separate counts of their word order.
The PROIEL corpus is a useful resource for scholars working specifically on Hellenistic Greek or doing diachronic work on Greek, particularly on the syntax/semantics/pragmatics interface. While this article is a general presentation of the architecture of the corpus, the main focus of attention is the Greek part, with demonstrations of how the various layers of information in the corpus can be used in quantitative analyses of specifically Greek problems within the PROIEL project’s area of research interests.

2. Corpus design

This section provides some background on the design of the PROIEL project. For more details, especially on the technical issues, we refer the reader to Haug et al. (2009).

2.1. The texts

The first electronic research infrastructures for ancient languages were digitized texts. In fact, traditional philologists were pioneers in this area: Roberto Busa started work on digitizing Thomas of Aquinas in 1949, the Gothic bible was digitized in the 1960s, in 1972 Thesaurus Linguae Graecae started their work on making editions of Greek texts from before 1453 electronically available, and from the mid-1980s textual and other material has been digitized by the Perseus project.

Therefore the availability of electronic editions of ancient material is relatively good, not least when it comes to the New Testament, which for obvious reasons, has attracted much attention. We decided that we would do no digitization within the project, but rather would use extant texts. Several NT texts are available, but we settled on Tischendorf’s 8th edition (1869-1872), which had been digitized by Ulrik Sandborg-Petersen. This text seemed to offer a reasonable balance between a modern edition and a text which is both closer to those used for the various translations in our corpus and available in the public domain.

For the texts other than the Greek NT, the choice of editions was relatively straightforward: we use digitizations of Streitberg (1919) by the Wulfila project

---

3 http://www.corpus Thomisticum.org/
4 http://www.wulfila.be
5 http://www.tlg.uci.edu/
6 http://www.perseus.tufts.edu/hopper/
for Gothic, of Künzle (1984) by the Leiden Armenian Lexical Textbase for the Armenian text, whereas we use the electronic edition of the Perseus project for the Latin Vulgate. For the Slavic version, we decided to use the Codex Marianus, since this is the largest extant manuscript; our text is a digitization of Jagić (1883) by Jouko Lindstedt.

Our approach implies a vast oversimplification of the editorial situation, not least in that the Armenian, Latin, Gothic and Slavic translations are linked to the same Greek original, but it offers considerable practical advantages, as a corpus aiming at a representation of the textual tradition of the Greek NT and the translations would demand its own - much larger - project. For scholars who study rare forms or individual passages of the NT, consultation of modern editions remains necessary, but for the study of more frequent phenomena and for statistical analysis of the texts, the PROIEL corpus is fully adequate.

However, for such statistical analysis to be interesting, much more than a simple electronic text is required: We need structured markup of the text with linguistic information. In the PROIEL corpus we can distinguish several layers of such annotation:

1. lemmatization
2. morphological analysis
3. syntactic analysis
4. information-structure analysis

In addition we have semantic tags, which are atomic attribute-value features (i.e., full or partial categorizations) of lemmata and/or tokens.

Linguistic markup of ancient texts is not nearly as well developed as pure text corpora, probably because these texts have always attracted more attention from philologists than from linguists. Lemmatization, which is essential for the study of word usage, has seen the most work and currently TLG offers lemmatized queries over their whole corpus, although this lemmatization is automatic and therefore without ambiguity resolution. Perseus offers automatic morphological analysis, but again without disambiguation.

Fully automatic and sometimes ambiguous lemmatization and morphological analysis is probably the only feasible solution for such large-scale corpora as TLG and Perseus. The advantage of a smaller corpus, such as the New Testament texts, is that is possible to do much more fine-grained annotation. For that reason, and because of the special interest of the NT text, morphological analyses of the Greek NT already existed, again due to Ulrik Sandborg-Petersen.

---

The annotation was slightly modified for PROIEL purposes, but as far as the Greek text is concerned, the syntactic and pragmatic annotation has demanded the most work, as described in sections 3 and 5.

For the other texts, fewer data were available, but we have profited from morphological analyses provided by the Perseus project (Latin), LALT (Armenian), the Wulfila project (Gothic) and the USC South Slavic Corpus.  

All texts had to be preprocessed and fitted to our own needs. They were converted to Unicode 5.1, which provides adequate support for Old Church Slavic. For syntactic annotation, we also needed to break the texts up in syntactic units, i.e. sentences. Since the style of the NT is rather paratactic, and sentence-initial καί ‘and’ is extremely frequent, the division into sentences will to some extent be arbitrary. Our approach was to use the punctuation in Tischendorf’s edition to determine sentence boundaries in the Greek text, and then to transfer this to the other versions using an implementation of the Gale-Church algorithm (Gale and Church 1993).

2.2. The corpus application

A corpus is essentially a collection of data that can be stored and distributed in various ways, as a flat text file, a structured XML file, or as a relational database. Our approach has been to create a web application which uses a relational database to store data, but also generates export files in various standard formats which can be browsed with tools such as TIGERSearch. The corpus web site and application is therefore essentially a tool for annotating texts and not primarily for retrieving data.

The web application has allowed us to use student assistants from various countries to annotate our texts and save their analyses, as described in the next section. Once sentences have been reviewed by project members, the morphological and syntactic annotation can also be browsed, but for the moment, morphological and syntactic queries must be done using external tools. However, the interface does allow queries (with wild cards) for word forms and lemmata, and the lemma pages show concordances and statistics on the morphological forms of that lemma as found in the corpus.

As a complement to the current application which focuses on annotation, we are currently developing a web based query interface which will also be

---

8 http://www.rcf.usc.edu/~pancheva/ParsedCorpus.html
9 http://www.ims.uni-stuttgart.de/projekte/TIGER/TIGERSearch/
linked to the database. This will allow searches in the syntactic structures (combined with other layers of the annotation) using the Tiger graph description language.

2.3. Workflow

The parsing of the corpus is the joint effort of a team of project members and collaborators—annotators, reviewers and administrators. The process of annotating a sentence comprises several stages, which are described below.

2.3.1. Annotation and reviewing

When starting the annotation of a ‘raw’ sentence, the annotator is first asked to check and possibly correct the sentence boundary. Since the text has been segmented automatically, it is sometimes the case that the boundary has been misplaced (for example, a subordinate clause got assigned to a different governing clause than the one it actually belongs to), in which case it must be corrected before the sentence is annotated.

The annotator then proceeds to enter the morphological data. This is done by manually marking each item using the relevant grammatical categories as well as providing the lemma. However, the annotation is usually based on what the system application guesses relying on the already existing reviewed annotations and/or specially written rules. The accuracy of the guesses varies from language to language and from sentence to sentence, but often the annotator will only have to disambiguate a limited number of ambiguous cases and to confirm the correctly inferred forms. In the annotation of Greek, in particular, morphological annotation mostly involves making distinctions that were not made in Sandborg-Petersen’s original text. For example our part-of-speech information is more fine-grained and involves the distinction of proper and common nouns, as well as several types of pronouns.

In the next step, the annotator proceeds to enter the syntactic annotation. This is carried out by creating a tree depicting the syntactic dependencies between the items. The structure of the tree is created entirely manually, by marking a head and clicking on its dependent. The application then provides a guess on the label of that relation, based on the morphology of head and dependent (i.e., accusatives dependent on verbs are guessed as objects). The decision to rely on manual procedures for the generation of the tree structure is dictated by the fact that no reliable parsers are currently available for any of the languages included in our corpus.

When the annotator has entered the morphological tags and created the dependency tree, the sentence is marked as ‘annotated’. At a later stage, it is
analyzed by an independent reviewer. If the annotation is found to be correct, the sentence is flagged as ‘reviewed’; if not, the reviewer may introduce the necessary corrections first. The reviewers are members of the PROIEL project and specialists in the respective languages. In particularly difficult or important cases, the reviewers discuss the annotation together in order to keep the annotation policy for the different languages as consistent as possible. The rules agreed upon after such consultations are then added to the annotation guidelines and applied consistently throughout the corpus.

2.3.2. Customized tagging

Even the detailed annotation of morphology, syntax and information structure that the corpus provides does not suffice for all kinds of research. Project members and others will want to classify nouns for animacy or relationality, verbs for aktionsart and so on. To make these data also available to other researchers in the interests of replicability and further research, the database allows for customized tagging where scholars can define their own attribute-value features and mark up linguistic material. Such tags can be set at the lemma level and overridden at the token level, a system which allows for efficient, but accurate tagging. An example from the domain of animacy is described in section 7. Whenever we want to apply a certain tag to all languages in the corpus, the starting point is a manual annotation of the Greek lemmata; subsequently, the tags can be transferred to the corresponding items in the remaining languages. The tagging automatically inferred from the aligned Greek lemmata is then checked and corrected manually by project members.
After this brief presentation of the corpus framework, we now go into more details concerning the linguistic annotation schemes themselves in separate sections on syntax (3), information structure (5), animacy (7) and token alignment (9), with an eye on specifically Greek issues. For each layer of annotation we also present some preliminary results which draw on that information (often in combination with other layers): in particular we discuss data on predicative participles (section 4 and 10) and the realization of discourse referents as prodropp, pronouns or full NPs (section 6 and 8). It should be stressed that these results are preliminary and draw on data that have not yet been finalized. Nevertheless, we believe they show the usefulness of corpora in general, and in particular of the multi-layered annotation we are doing. We also believe that the preliminary results point to the rich possibilities the corpus gives us for a quantitative study and a deeper understanding of the grammar of Hellenistic Greek.

3. Syntactic annotation

3.1. Introduction

One of the most important desiderata for linguistic research into earlier stages of Greek is syntactic annotation. Lemmatization and especially morphological annotation will make many queries possible, but does not make it possible to separate e.g. objects from complements of prepositions demanding the accusative, or subjects from subject complements, or adnominal participles from predicative ones. Large-scale studies on the order of major constituents, or on valency frames of verbs, also require syntactic annotation.

However, syntactic categories are not as clear-cut as morphological ones, and there are several theories and annotation schemes on offer. The most basic theoretical distinction is that between word-based and constituent-based syntax. The first approach focuses on dependencies between words, whereas the second one focuses on how words make up constituents that build up the structure of the sentence. Also, at least traditionally, the first approach gives prominence to grammatical relations, such as subject, object, predicate etc., which hold between words: this means that graph edges are typically labelled. The phrase structure approach typically focuses on grammatical categories, such as noun phrase, prepositional phrase and others, instead of relations between words. This means that nodes are labeled with category information and the syntactic relations are often held to be derivable from the phrase structure.
A theory that spans both approaches is Lexical Functional Grammar (LFG, Bresnan 2001, Dalrymple 2001), where syntax is represented in two mutually constraining structures: a tree structure representing constituency and an attribute-value matrix representing dependencies and grammatical relations, but also other grammatical features such as tense, aspect, gender etc.

Little is known for certain about the word order, and particular the constituency facts, of the Ancient Greek sentence, despite studies such as Dover 1960; Dik 1995, 2007; Devine and Stephens 2000. We believe that one of the main contributions our corpus can offer is a more secure empirical footing for studies of Greek word order. For this reason it did not seem advisable to impose an interpretation on the data, but rather to leave the annotation of word order facts at the purely descriptive level of linearization. In other words, in a PP εἰς ἀσκοὺς παλαιοῦς ‘into old wineskins’ the order of words is represented in our corpus, but not the bracketed structure PP [εἰς NP [ἀσκούς AP [παλαιοῦς]]], since it is entirely unclear how a discontinuous construction like ἀσκοὺς εἰς παλαιοῦς should be represented. Although such structures have become rather rare by the time of the New Testament, they still occur. Therefore we rejected the idea of a phrase-based syntax, but opted for a dependency structure enriched with some features from LFG. It should be noted, though, that phrase structure facts can be reconstructed from the combination of word order facts and dependency relations.

3.2. Dependency syntax

The core idea of dependency grammar (DG) is that there are hierarchical relations, dependencies, between the words of sentences. Although the relation of dependency is hard to define properly, the intuitive idea is that a word X is dependent on a word Y just in case, if you remove Y from the sentence, you also have to remove X. In normal sentences, then, the finite verb is the head on which all other words depend: if you remove it, you also have to remove any subjects or objects (for example).

Typically, dependencies are labeled with grammatical relations corresponding to traditional grammatical functions such as subject, object, attribute etc. In other words, a simple transitive sentence will look like that in Fig. 2. Notice that dependency structures do not represent word order facts, so the linear ordering of two bottom nodes is not significant.

---

10 On the other hand the various translations of the New Testament tend to follow the word order of the Greek text very closely, so that in general little can be known about the native word order in those languages.
Much computational representation of syntax is done within pure dependency grammar. This format has several advantages: the result is always a well-formed tree, which has a simple structure where each node has a unique head, thus creating a structure which is easy to query; and from a parsing perspective it is useful that the nodes of the tree to build are already given by the words of the sentence.

However, natural language facts are more messy than dependency structures and therefore there are phenomena that can only artificially be captured in a pure dependency format. Consider first the analysis of a coordinated object NP in (1):

(1) ἐσθίων ἀκρίδας καὶ μέλι ἄγριον
‘eating locusts and wild honey’ (Mk 1:6)
To capture the fact that the whole coordinated NP is the object of ἐσθίων ‘eating’ we need to take the conjunction καὶ ‘and’ as the head of the coordination, and then make the coordinated nouns dependent on it. This accurately depicts the most important fact, since the subgraph dominated by καὶ is represented as the object of the verb, but the analysis undoubtedly has an unnatural flair. A treatment in terms of unification (as in LFG), where two lexical words could correspond to one node, would be more intuitive, but was rejected because it would be more difficult to implement.

3.3. Structure sharing, ellipsis and secondary edges

While coordination can at least be simulated in a dependency graph, there are other phenomena that cannot really be captured at all, such as control and raising. Consider the sentence in (2):

(2) η δὲ πενθερὰ Σίμωνος κατέκειτο πυρέσσουσα

‘Now Simon’s mother-in-law was in bed with a fever’ (Mk 1:30)

![Figure 4. Secondary edges.](image)

Here, η πενθερὰ Σίμωνος ‘Simon’s mother-in-law’ is the subject not only of the finite verb κατέκειτο ‘lay’, but also of the participle πυρέσσουσα ‘having fever’. This fact cannot be captured in standard dependency grammar because of the unique head constraint: every word is supposed to have only one head. Our approach in the PROIEL corpus has been to introduce secondary edges in our graphs: thus, η πενθερὰ Σίμωνος is the primary subject dependent of κατέκειτο, but also the secondary (external) subject of πυρέσσουσα. This
represents an extension of dependency grammar in the direction of LFG, which also allows this kind of structure sharing in functional structure.

The same analysis is used for control and raising infinitives (see Fig. 5). In this way we can easily retrieve all subjects of the verb ἐκβάλλω ‘I throw out’, for example for a study of selectional preferences of this verb, irrespective of whether the verb appears as a finite verb with a “normal” subject, or as a particle or an infinitive with a controlled or raised subject.

(3) πῶς δύναται σατανᾶς σατανᾶν ἐκβάλλειν  
‘How can Satan cast out Satan?’ (Mk 3:23)

Figure 5. Raising/control.

Another problem for dependency grammar is ellipsis. DG requires that every word has a head, but sometimes this head is not present in the structure. Consider the following sentence:

(4) ἐγερθήσεται γὰρ ἔθνος ἐπ’ ἔθνος καὶ βασιλεία ἐπὶ βασιλείαν  
‘For nation will rise against nation, and kingdom against kingdom’ (Mk 13:8)

If we just made ἔθνος ‘nation’ and βασιλεία ‘kingdom’ subjects and ἐπ’ ἔθνος ‘against nation’ and ἐπὶ βασιλείαν ‘against kingdom’ oblique dependents of ἐγερθήσεται ‘will rise’, there would be no way of seeing that the ἔθνος and ἐπ’ ἔθνος belong together. So we need two independent predications, but there is only one verb. To remedy this situation we introduce a limited use of empty nodes to represent “missing heads”, as it were. These come in two types: empty verbal nodes (for gapping, nominal sentences and similar
constructions), and empty conjunctions (for asyndetic coordination where there is no conjunction available as head). As Figure 6 shows, we can also exploit our secondary edges to model the fact that the empty verbal node represents a predicate identical with ἐγερθήσεται. Such secondary edges are labelled PID, for predicate identity.

We only use empty nodes to represent missing heads, not missing dependents such as prodropped arguments, since we want our syntax to be as theory-neutral as possible. In other words, whenever no overt subject is present, we simply do not represent it:

(5) καὶ λέγει αὐτοῖς
    ‘And he tells them’ (Mk 1:38)

However, since prodropped arguments can clearly have referential properties in Greek, they are represented in the information structure annotation. Also, structure sharing via secondary edges allows us to capture the facts whenever
the argument is in fact present somewhere else in the sentence. Thus in figure 8, the secondary edge from ἥκολοουθηκαμέν ‘we have followed’ to ἡμεῖς ‘we’ shows that ἡμεῖς is also the subject of that verb, something which would not be possible in pure dependency grammar.

(6) ἡμεῖς ἄφηκαμεν πάντα καὶ ἥκολοουθηκαμέν σοι
‘We have left everything and followed you.’ (Mk 10:28)

In principle, this kind of annotation would allow us to distinguish between sentence coordination and VP-coordination, and we could have reserved the
analysis in Fig. 8 for cases where we can be reasonably certain that we have VP coordination, so that a single word is actually the subject of both verbs.

But in prodrop languages such as Ancient Greek, it is notoriously hard to make the distinction between VP and sentence coordination, where the second sentence has a prodrop subject which is coreferential with the subject of the first sentence. To us it seemed impossible to make this distinction in a proper way, so we opted instead for using secondary edges not only for argument sharing in the strictest sense, but also for anaphoric relationship.

This illustrates another factor in the design of our syntactic annotation scheme, namely the constant tradeoffs between information that we would like to preserve, and what is possible to for annotators to disambiguate during annotation. We present more such examples in the section on the labeling of the syntactic dependencies.

3.4. The relations

In the PROIEL annotation we use three non-clausal argument functions: subject, object and obliques. The latter are arguments in oblique cases or in prepositional phrases; they are distinguished from objects by their inability to appear as subjects in passives, although it should be stressed that it is not always clear whether accusative arguments can be subjects of passives or not: in the normal case, annotators distinguish between non-subject arguments based on case so that accusatives are taken as objects and non-accusatives as obliques.

We have adopted a conservative approach to grammatical relations and do not recognize e.g. oblique case subjects for finite verbs. The guiding principle has been to impose as little interpretation as possible on the data. We believe this will make them more useful for future research also by scholars who do not share our theoretical assumptions.

We recognize two different forms of clausal arguments, complement clauses (COMP) and open complement clauses (XOBJ). The difference is whether the clause contains its own subject or not. Finite clauses are thus always COMP, but nonfinite clauses can have both functions according to whether the construction admits an internal subject or not. While the distinction is clear in theory, it is not always easy to maintain in practice, because subjects...
are easily prodropped. Although the distinction we are trying to capture corresponds to a traditional distinction, between accusative with infinitives (with potentially prodropped accusative subject) and complement infinitive (where an accusative subject would not be grammatical), there are no reliable lists of which verbs govern what kind of complement. The annotator must rely on this knowledge of Greek grammar in general, since the cues to the correct annotation are not present in the overt sentence: when she encounters \( \text{θέλω} \, \text{ποιεῖν} \) ‘I want to do’, she needs to decide whether \( \text{θέλω} \, \text{ἐμὲ} \, \text{ποιεῖν} \) would be possible or not. In this case the correct analysis is given away by the existence of sentences such as in Fig. 9, where the nominative \( \text{πρῶτος} \) ‘first’ shows that we do not deal with an elided accusative pronoun.

\[
\varepsilon \iota \, \text{τις} \, \text{θέλει} \, \text{πρῶτος} \, \varepsilon \iota \nuαι \, \varepsilon \iota\sigmaται \, \pi\nu\Delta\nuων \, \varepsilon \iota\sigmaχωτος
\]

‘Whoever wants to be first must be last of all’ (Mk 9:35)

As Fig. 9 shows, XOBJ is used also for predicative complements, such as \( \text{πρῶτος} \) in this example. Such complements also get a secondary edge pointing to their subject: however, in the present sentence, the subject is not realized in the local context (as an argument of \( \varepsilon \iota\nuαι \)), so the secondary edge points to the head verb instead, indicating the subject is an element in the argument structure of \( \varepsilon \iota\nuαι \) but is not overtly realized.
The difficulty of resolving cases like this is the main reason we adopted the non-standard workflow of having a student annotator and a reviewer who corrects the student’s annotation, rather than two independent student annotators: as we are building our treebank and developing our annotation guidelines, we are actually also creating a formal grammar of ancient Greek, notably making decisions on valency frames and argument structures. Full uniformity will probably not be achieved until the corpus is finished: at that point we may analyze the frames that occur with different verbs and make our final decisions on their argument structures.

Corresponding to the distinction between open and closed complement clauses, we distinguish between open and closed adverbials (XADV and ADV). The latter relation is used for adverbials that are non-clausal, e.g. instrumental, temporal and locational adverbs, as well as for adverbial clauses. The corresponding open relation is used for adverbial uses of nonfinite forms such as the infinitive of purpose (figure 10) or conjunct participles. Like XOBJs, XADVs have a secondary edge pointing to their (external) subject.

(8) ἰδοὺ ἐξῆλθεν ὁ σπείρων σπειραί
   ‘Listen! A sower went out to sow.’ (Mk 4:3)

![Diagram of structure-shared participle subject.](Figure 10)
In the adnominal domain we also distinguish several relations: the most basic distinction is that between appositive and non-appositive modifiers: in addition to normal appositions, the APOS relation is also used for non-restrictive relative clauses. Among non-appositive relations we attempt to distinguish partitive attributes (PART), and arguments of nouns (NARG) from normal attributes. While we see much agreement between annotators on prototypical cases of PART (e.g. ἔσχατον πάντων ‘the last of all’) and NARG (e.g. ἁμαρτίαν ἁμαρτιῶν ‘forgiving of sins’), there is much variation in the application of these tags to the more atypical cases. Again, it is likely that we will not achieve uniformity until the annotation has been completed and the material analyzed in toto.

Table 1 shows all the normal labels used in the PROIEL corpus. In addition, we have a number of ‘supertags’ to be used whenever annotators or reviewers are unsure about the correct analysis. For a discussion of these tags, as well as the relationship to the annotation scheme used by the Perseus Greek and Latin Dependency Treebanks which are currently under development, we refer to Haug and Jøhndal (2008).

Table 1. The PROIEL syntactic annotation scheme

<table>
<thead>
<tr>
<th>Label</th>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>PRED</td>
<td>Predicate</td>
</tr>
<tr>
<td>SUB</td>
<td>Subject</td>
</tr>
<tr>
<td>OBJ</td>
<td>Object</td>
</tr>
<tr>
<td>OBL</td>
<td>Oblique</td>
</tr>
<tr>
<td>AG</td>
<td>Agent</td>
</tr>
<tr>
<td>ADV</td>
<td>Adverbial</td>
</tr>
<tr>
<td>COMP</td>
<td>Complement clause</td>
</tr>
<tr>
<td>XOBJ</td>
<td>Open complement clause, predicative complement</td>
</tr>
<tr>
<td>XADV</td>
<td>Open adverbial, secondary predication</td>
</tr>
<tr>
<td>VOC</td>
<td>Vocative, exclamation</td>
</tr>
<tr>
<td>AUX</td>
<td>Auxiliary (used for ‘grammatical words’)</td>
</tr>
<tr>
<td>ATR</td>
<td>Attributes</td>
</tr>
<tr>
<td>APOS</td>
<td>Apposition</td>
</tr>
<tr>
<td>NARG</td>
<td>Argument of noun</td>
</tr>
<tr>
<td>PART</td>
<td>Partitive</td>
</tr>
</tbody>
</table>
4. Sample study: predicative participles

In this section we show how a combination of the morphological data with the word order and dependency data can shed new light on the use of predicative participles in Hellenistic Greek.

As is well known, the Greek of the New Testament, like the classical language, abounds in participles. This is such a distinct trait of Greek that the Roman grammarians dubbed the Greeks *philometochoi* ‘lovers of participles’ (Schwyzer 1950: 386n.2). Participles have several functions, and traditional grammars in particular recognize three: the attributive use, the nominalized use, and the predicative use.

Traditional grammar does not normally subdivide the predicative use into different categories but rather lumps them all together as expressions of “circumstantial events”. However, as we see below, the group is far from unitary; we must recognize at least three different types of predicative participles: frames, which act as a background and contextualize the main predicate; elaborations, which offer more information about the event denoted by the main predicate; and independent rhemes, which are on a par with finite verbs regarding their information structure. And just like finite verbs, participial independent rhemes can either share their frame and theme with the other verb, which gives us something akin to VP-coordination; or they can have their own frame and theme, in which case the result is reminiscent of sentence coordination. As a matter of fact, participles and finite verbs are sometimes coordinated in such cases.

(9) καὶ ἐβαπτίζοντο ὑπ’ αὐτοῦ ἐν τῷ Ἰορδάνῃ ποταμῷ ἐξομολογούμενοι τὰς ἁμαρτίας αὐτῶν.
   ‘They were baptized by him in the river Jordan, confessing their sins.’
   (Mk 1:5)

(10) καὶ ἰδὼν ὁ Ἰησοῦς τὴν πίστιν αὐτῶν λέγει τῷ παραλυτικῷ:
   ‘And when he saw their faith, Jesus said to the paralytic.’
   (Mk 2:5)

(11) Αὐτὸς γὰρ ὁ Ἰρώνης ἀποστείλας ἐκράτησεν τὸν Ἰωάννην.
   ‘For Herod himself had sent out (men) and arrested John.’
   (Mk 6:17)

The three different constructions differ notably in their temporal semantics, as shown by Haug (2008) who discusses the differences between framing participles and event modifiers.

For this paper, though, we would like to show how the different participle uses may be traced in the corpus: in particular we show that different positions of the participles are associated with different behavior in several respects, thus lending support to the hypothesis that there is indeed a distinction to be made. In particular we look at how aspectual morphology and lexical
variation differs between sentence-initial participles, directly preverbal participles, and postverbal participles. Since participles that are both sentence-initial and directly preverbal are ambiguous within this classification, they have been treated as a separate group. As becomes clear, the data lend support to the hypothesis that framing participles occur sentence-initially, independent rhemes preverbally and event modifiers postverbally. In all the statistics we have ignored participle forms of λέγω ‘I say’, which are extremely frequent. This high frequency is likely due to Semitic influence, and the participle often seems to mean little more than a colon, serving simply to introduce direct speech.

Table 2 shows the distribution of the aorist and present aspect stems in the different positions. As we see, the distribution is very different in preverbal and postverbal position. To our knowledge, this fact has never before been noticed. However, the distribution actually makes sense if the participles have different functions in the narrative; by far the most common framing function of participles in the narrative of the Greek NT is that of linking an event to a previous event: “After he had …”. Thus the aorist participle clearly dominates in sentence-initial position. Similarly, participles in the preverbal position are typically independent rhemes. They share the information structure of the finite verb and also tend strongly to be perfective: in fact, the distribution of aspects in the preverbal positions is similar to what we find among finite verbs, where the distribution of the two preterite forms is 19.6% imperfectives and 80.4% perfectives (ignoring presents, perfects and futures). Event modifying participles, on the other hand, normally express events that are simultaneous with the main event: for this reason, they are much more likely to occur in the imperfective aspect.

The main use of participles as frames is to link the main event to some known, previous event; their information is not asserted, but presupposed.

Table 2. Correlation of aspect and position

<table>
<thead>
<tr>
<th></th>
<th>Present participle</th>
<th>Aorist participle</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initial, not directly preverbal</td>
<td>6.5% (25)</td>
<td>92.7% (354)</td>
<td>0.8% (3)</td>
</tr>
<tr>
<td>Initial, directly pre-verbal</td>
<td>13.1% (74)</td>
<td>86.0% (485)</td>
<td>0.9% (5)</td>
</tr>
<tr>
<td>Internal, pre-verbal</td>
<td>14.8% (91)</td>
<td>82.6% (509)</td>
<td>2.6% (16)</td>
</tr>
<tr>
<td>Postverbal</td>
<td>74.4% (450)</td>
<td>16.2% (98)</td>
<td>9.4% (57)</td>
</tr>
</tbody>
</table>
In other words, the event denoted by the participle is not presented as new information, but rather ‘taken for granted’ - either because it has actually been mentioned in the previous discourse, or because it is easily inferable, as in (12).

(12) *(They were wondering at Zacharia’s delay in the temple.)*

 ἐξελθὼν δὲ οὐκ ἐδύνατο λαλῆσαι αὐτοῖς

‘When he came out he could not speak to them’ (Lk 1:22)

Since framing participles refer to events that are known or inferable from the context, they should be more ‘predictable’, i.e., we expect them to be drawn from a smaller set of lemmata denoting ‘typical’ events in the narrative, to be less complex and have fewer dependents.

As for lexical variation, we can look at how many of the examples of participles in each position belong to the 10 most frequent lemmata in Table 3.

As Table 3 shows, we see much less lexical variation among sentence-initial participles. The verbs we find express typical ‘linking events’, such as movement from one location to another, or perception (of a previous event by another participant in the narrative).

Another way of looking at predictability is to study the length of the participle constructions. If initial participles are typically frames linking the main event to an event known or inferable from the preceding context, they should typically be shorter. And this is indeed what we find.

As Fig. 11 shows, initial participles are the ones that most frequently have no dependents as indicated by the size of the lowermost ‘bean’. Postverbal participles are on average much longer: this probably indicates that we find not only event-elaborating participles in this position, but also independent rhemes, e.g. afterthoughts expressed in participles.

<table>
<thead>
<tr>
<th>Type</th>
<th>Examples</th>
<th>10 most frequent lemmata</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initial, not directly preverbal</td>
<td>382</td>
<td>213</td>
<td>55.8%</td>
</tr>
<tr>
<td>Initial, directly preverbal</td>
<td>564</td>
<td>182</td>
<td>32.3%</td>
</tr>
<tr>
<td>Internal, preverbal</td>
<td>616</td>
<td>203</td>
<td>33.0%</td>
</tr>
<tr>
<td>Postverbal</td>
<td>605</td>
<td>121</td>
<td>20.0%</td>
</tr>
</tbody>
</table>
5. Information structure: givenness, anaphoric distance and topics

Information structure annotation can be added to text portions that have already been completed and reviewed. It is essential that the morphological and syntactic annotation of the given text is fully complete before the information tags are added, since any further changes in the morphology and syntax may alter the input for information structure tagging, therefore rendering the existing annotation inadequate. The application will provide no guesses, but it will automatically choose the items which are annotatable for information structure. The choice is usually correct, but it may happen that the annotator has to introduce changes manually; also, pro-drop arguments (subjects, objects and obliques) must be identified and entered by the annotator. The annotation itself is carried out by marking the items for specific categories and linking them to their antecedents.

Figure 11. Length of participle constructions per position.
To ensure the highest possible level of consistency in this layer of annotation, which cannot be codified using strict rules as easily as in the case of morphology and syntax, the information structure of each portion of text is first tagged independently by two annotators. Next, the annotators compare their versions and discuss the cases in which they differ. The controversial and difficult points are then analyzed at meetings involving all project members, so that rules and policies for dealing with more complex issues may gradually be worked out.

The field of information structure is untidy, controversial and full of partially overlapping perspectives and definitions. Many concepts of information structure are also particularly difficult to work with in historical linguistics, as they rely crucially on intonation and/or are subtle and require native speakers’ intuitions to be applied properly. In dealing with ancient texts, of course, neither are available. Our annotation schemes must therefore be devised to make the most of bad data and concentrate on the information structure categories most objectively available to us in such material. The corpus will therefore be annotated for a) givenness, b) anaphoric distance and c) contrast. The notion of topic is difficult in general, and in particular in dealing with historical texts. We have therefore chosen to deconstruct the notion and look for aboutness topics computationally by combining features from multiple annotation levels. The annotation of contrast is at an experimental stage and is not discussed in this article; the following three subsections deal with givenness annotation, the use of anaphoric links and the topic problem.

5.1. Givenness

Our approach to givenness is to find the answer to the question “How does the hearer establish the referent of an NP?” In our case, the hearer would be a member of the first intended audience of the New Testament: a first-century AD Hellenized Jew holding the current religious beliefs and with a fair knowledge of the geography of the areas where the NT takes place. 14 In direct speech, on the other hand, the hearer would be the actual hearer, who must then be aware of the preceding conversation and things in the concrete discourse situation, in addition to having the general world knowledge expected from a person of that period.

Most approaches to givenness use a fundamental tripartition into new – mediated / accessible – old (as in Prince’s classic 1981 article, for corpus

---

14 This is not the only possible approach, other options include looking for the referent’s discourse status (has it been mentioned in the text?) or assessing its state of cognitive activation (cf. Gundel et al. 1993).
annotation schemes see Nissim et al. 2004 and Dipper et al. 2004), where the accessible group is generally divided into various subgroups. We decided to base our subdivisions on the insights in Riester and Lorentz’s (2009) annotation scheme, which draws on Discourse Representation Theory (DRT, Kamp and Reyle 1993) and its notion of different contexts where discourse referents may be identified by the hearer.

The givenness annotation scheme is the result of a long collaborative process among the project members and in cooperation with other projects doing information structure annotation. An important goal was to have a scheme that was simple enough to give good intersubjective agreement. Since the scheme has so far been under constant development, and new annotators have been trained in the process, the interannotator agreement has lately not reached a higher kappa value than about .6. We are confident that this level will rise – at the end of a first pilot annotation round, we saw kappa values around .8.

5.2. Anaphoric links

Figure 12 shows a sample sentence with givenness tags (black for OLD, dark gray for ACC-gen, light gray for NEW) and anaphoric links. Unlike the givenness annotation, where a high degree of intersubjective agreement is quite difficult to achieve, anaphoric distance can be tagged and measured with relative objectivity. In every case where a referent (overtly realized or prodrop) has been mentioned within the 13 last sentences, we set an anaphoric link from anaphor to antecedent. The limit of 13 sentences is to some extent arbitrary, and may be adjusted as the work proceeds. For referents with a last mention outside this limit, but which seem not to be ‘reintroduced’ but simply picked up again, we use the tag OLD-inactive.

The anaphoric links can be used to calculate not only exact distances between anaphors and antecedents, but also the length and density of entire anaphoric

<table>
<thead>
<tr>
<th>Context</th>
<th>Tag</th>
</tr>
</thead>
<tbody>
<tr>
<td>Discourse (anaphora)</td>
<td>OLD</td>
</tr>
<tr>
<td>Situation (deixis)</td>
<td>ACC-sit</td>
</tr>
<tr>
<td>Scenarios (inferences)</td>
<td>ACC-inf</td>
</tr>
<tr>
<td>Encyclopaedic knowledge</td>
<td>ACC-gen</td>
</tr>
<tr>
<td>No context (no extra-NP info)</td>
<td>NEW</td>
</tr>
</tbody>
</table>

Table 4. Givenness tags in the PROIEL corpus
chains, and thus also the relative saliency of individual discourse referents. Several pilot studies have shown that anaphoric distance may be an equally or even more important predictor than givenness in many cases, and it is a tool that is well suited to the special style of the New Testament, where so many of the discourse referents are old due to the small circle of main participants and the repetitive and somewhat formulaic narrative.

To capture distinctions between various OLD elements, it is useful to calculate saliency. A brief look at the properties of the most salient discourse referent of all in the New Testament, Jesus himself, makes this clear: In the Gospels, Jesus is so immensely salient that he can be picked up with a prodrop subject even though he has not been mentioned for a long stretch of text. This is not the case with any other referent.

As we show in section 7 on animacy annotation, the anaphoric links can also be useful devices to transfer information from antecedent tokens to anaphors.

5.3. Topics

The notion of topic is an important one in the information structure literature, but also one of the most difficult to define, and not least to apply in practical data analysis. There are many definitions available, some more similar

Table 5. Givenness tags in the Gospel of Mark

<table>
<thead>
<tr>
<th>Tag</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>NEW</td>
<td>16.6% (694)</td>
</tr>
<tr>
<td>ACC-gen</td>
<td>8.7% (363)</td>
</tr>
<tr>
<td>ACC-sit</td>
<td>2.1% (90)</td>
</tr>
<tr>
<td>ACC-inf</td>
<td>6.3% (266)</td>
</tr>
<tr>
<td>OLD-inact</td>
<td>1.0% (41)</td>
</tr>
<tr>
<td>OLD</td>
<td>65.3% (2739)</td>
</tr>
</tbody>
</table>
than others, and common to most of them is that they are non-rigorous and not necessarily intended as guidelines for actual classification of potential topics. As historical linguists, we are at an additional disadvantage, in that the notion of topic is also closely linked to prosody, and that topichood diagnostics generally require native speaker intuitions. Obviously, neither prosody nor native speakers are available to researchers working on ancient texts.

Nonetheless, we believe that aboutness topics are “real” in the sense that they shape the way we structure our discourse. Continuing topics are clearly treated differently from shifting topics, and topic shifts seem to influence both argument realization and use of definite articles, phenomena that are central to the PROIEL project’s main interests. We have therefore experimented with annotation of frame-setting and aboutness topics. However, the results were rather poor: the inter-annotator agreement was low, and annotators agreed that they tended to stick to a small set of criteria: preverbal, animate, definite subjects would generally be taken as topics. If we were then to use the annotated topics to look at the animacy, definiteness and word order properties of topics, we would clearly run into circularity.

This problem has been noted before: in their discussion of principles for topic annotation of Old High German texts, Petrova and Solf (2009) solve the problem by decomposing the topic notion into

• givenness
• referentiality
• definiteness
• early position in the sentence

An optimal topic candidate would be given/old, referential, definite and positioned early in the sentence. The decomposition is, however, only meant as a help to the annotator, to facilitate clear decisions on topichood. We agree that decomposition is probably necessary in working with old texts. However, if topichood must be decomposed into multiple weighted features, it is better to annotate for each of these independently, and do the topic selection computationally. In this way we can also include more features and be sure that the weighting is not biased by different human judgments.

To establish the likely topic of each sentence we first find all main verbs in each sentence (verbs in dependent clauses are excluded) and extract their dependents. All dependents with an information status other than new are considered topic candidates. These are then ranked according to various parameters, and the one with the highest score is selected as the likely topic.

The program includes all of Petrova & Solf’s parameters, in addition to several others, and combines information from the givenness and anaphoric
link annotation with information from all the other annotation layers: syntax, morphology, word order and semantics. Old referents are considered better topics than other, less given referents; subjects are better than objects or obliques; humans outrank concretes, whereas times, places and non-concretes are hugely disfavoured; definite referents, i.e. common nouns with articles, proper nouns and personal pronouns, are rewarded, and so are prodrop tokens. Furthermore we take saliency into account: topic candidates that are members of long and tight anaphoric chains are more likely topics than others. Finally, we consider the properties of the immediate antecedent in the case of old referents: does it outrank the intervening referents on the relation, animacy and givenness hierarchies?

Taking all these properties into account we are able to predict topichood in a satisfactory manner. The main difficulty is to predict topicless sentences; for this we will probably need to take verb semantics into account. The inherent uncertainty of the notion of topic does not disappear, of course, but we can at least apply a consistent definition of it and even change this definition without having to annotate the material anew.

6. Sample study: realization of discourse referents

In this short study we look at how the realization of discourse referents as nouns, pronouns or prodrop tokens is conditioned by their syntactic function and their information status. In discourse, reference is typically made time and again to the same referents, creating a sense of coherence and narrative/argumentative flow: a small set of known referents interact in new ways, steadily advancing the narrative. New referents may be introduced and described in more detail.

Observing how this process of reference interacts with the grammatical system of the language used to express it is important for understanding the structure of the discourse. The morpho-syntactic realization of discourse referents gives the hearer implicit clues to the interpretation of this structure.

The referents in discourse may be realized in various ways, depending on the morpho-syntactic possibilities of the language in question. In Ancient Greek, the possible realizations range from zero anaphora (prodrop) through pronominal expressions and demonstratives to full NPs, also including modifying adjectives or relative clauses.

By using information from all levels of annotation in the PROIEL corpus, we want to gain insight into the discourse factors which influence the realization of discourse referents. Earlier studies suggest that several different factors
are involved in determining this realization. Accordingly, it is of great interest to see whether we can estimate the relative importance of various factors to each other.

In a preliminary experiment, this was modeled by means of a classification tree. We looked at the following set of predictors (others may be relevant as well):

- syntactic relation of the discourse referent;
- information status of the discourse referent;
- syntactic relation of the antecedent if any;
- information status of the antecedent if any;
- whether the antecedent had a definite article or not;
- whether the antecedent and the anaphor had the same syntactic relation;
- referential distance between antecedent and anaphor in sentences;
- referential distance between antecedent and anaphor in words;
- morphosyntactic realization of the antecedent.

As the above list shows, some of the predictors concern features of the referent being mentioned (syntactic relation and information status), while others concern properties of the antecedent, while yet others concern features calculated on the basis of both (distance measures).

In the first part of the experiment, we looked only at referents being realized as the subject of their clause. The dataset consisted of 1760 subjects taken from the Gospel of Mark. Considering the predictors listed above, we produced the classification tree in Fig. 13. To avoid overfitting the data, the resulting tree was pruned using a cost-complexity algorithm which eliminates less important predictors.

Fig. 13 provides an example of a simple classification tree. Non-terminal nodes are labeled with a condition. If the condition is true, one should follow the left branch, otherwise the right branch. Terminal nodes are labeled with the most common realization in that group of referents, and with the frequency of each type of realization: subject referents with info status new or (some kind of) accessible are realized as nominals in 291 cases, prodrops in 35 cases and pronouns in 39 cases. Other (i.e. old) subject referents are most often realized as prodrops (1017 cases), followed by nominal realizations (263) and pronominal ones (115).

The classification algorithm recursively chooses the predictors which give the most extreme partition of the data set. The present tree classified 74% of the subjects correctly. The baseline established by always choosing the most frequent realization is 60%, so the predictors considered are sufficient to improve the classification considerably.
What the classification tree tells us is that the main distinction is between old and non-old subjects: non-old subjects are predominantly realized as nominals; old subjects, on the other hand, are mainly realized as zero anaphora (prodrop). What about subject pronouns (anaphoric and demonstrative)? Even though these mostly fell into the OLD category, there is reason to believe that another predictor is relevant for this realization possibility, viz. contrast. Since the textual frequency of contrast is rather low including contrast as a predictor would presumably not change the big picture.

In the second part of our experiment, we looked at referents which are realized as (direct) objects in their clause. The data set consisted of 799 direct objects taken from the Gospel of Mark. Again, the same procedure was applied to the data and the pruned classification tree in Fig. 14 was produced.

This tree classified 81% of the objects correctly against a baseline of 58%. Again, the chosen predictors improved considerably on the classification. As with subjects, the main distinction is between non-old and old referents. However, referential distance and syntactic relation of the antecedent also play a part:

We have not yet done extensive annotation of contrast, but this will be part of the information structure annotation at a later stage of the project.
• if the referential distance is 12 words or more, subjects are easier to pick up with a pronoun than objects and obliques;
• if the referential distance is 11 words or less, anaphors are predominantly pronominal.

With the objects, then, a distance effect is visible which did not show up with the subject data.

7. Semantic annotation: animacy

As seen in section 2.3.2, the corpus application allows for customized tagging by individual researchers working on specific problems. So far, these tags have mainly been used for semantic classifications, but they are not restricted to semantics. We exemplify this annotation layer by describing our approach to animacy annotation.

Animacy is a semantic category that has been shown to have impact on and to interact with a series of grammatical phenomena cross-linguistically, such as argument realization, word order and definiteness, all areas of interest to the PROIEL project. In Slavic, animacy has developed into a formal grammatical category. In OCS, we only see this category at an emergent stage, which made animacy annotation essential to the PROIEL corpus.
We wanted an animacy annotation scheme that was sufficiently detailed to give interesting distinctions, but simple enough to make it easy to apply. We therefore chose a slightly simplified version of the annotation scheme in Zaenen et al. 2004 (see Table 6). The scheme is considerably more detailed than a simple human:non-human dichotomy or e.g. the classic extended animacy hierarchy (Dixon 1979), but much simpler than e.g. the 20-way distinction employed in the Russian National Corpus.\footnote{See their guidelines for semantic annotation at http://ruscorpora.ru/en/corpora-sem.html.} We consider the distinction between various kinds of inanimates to be important: in particular we expect there to be interesting differences between concrete and non-concrete inanimates. It is also convenient to be able to retrieve nominals denoting time and space by way of the animacy annotation.

For a maximally efficient annotation, we started by annotating all Greek nouns for animacy at the lemma level. The annotation was done by one project member and then reviewed by another. In most of the cases the two agreed on the annotation. The tags are meant as token-level tags, so the annotator and reviewer kept an eye on what would be the most common meaning among the tokens actually represented in the corpus, not on the lemma’s “basic meaning”. For example, none of the occurrences of καρδία ‘heart’ referred to a

<table>
<thead>
<tr>
<th>Animacy tag</th>
<th>Brief description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td>things that look and act like humans, including deities and spirits</td>
</tr>
<tr>
<td>Org</td>
<td>a collectivity of humans with some degree of group identity</td>
</tr>
<tr>
<td>Animal</td>
<td>non-human animates</td>
</tr>
<tr>
<td>Veh</td>
<td>Vehicles</td>
</tr>
<tr>
<td>Concrete</td>
<td>“prototypical” concrete objects or substances, excluding intangibles</td>
</tr>
<tr>
<td>Nonconc</td>
<td>anything that is not prototypically concrete but clearly inanimate, e.g. events</td>
</tr>
<tr>
<td>Time</td>
<td>expressions referring to periods of time</td>
</tr>
<tr>
<td>Place</td>
<td>nominals that normally serve as locations for human actions</td>
</tr>
</tbody>
</table>

Table 6. Animacy annotation scheme
physical heart, but always to the minds or opinions of people. The lemma was therefore annotated as NONCONC, not CONCRETE. The annotation is still in need of adjustment at token level, but even the non-adjusted annotation is of good quality.

The animacy annotation also provides a good example of how the various components of the corpus can be put to use. Naturally, it is not sufficient to have animacy annotation only for nouns: we want it for other referential expressions too, for pronouns and even for prodrops. However, it is not necessary to do manual annotation of all referring expressions. Instead the anaphoric links in the information structure annotation can be put to work since every member of an anaphoric chain necessarily has the same animacy as the starting-point of that chain.

Similarly, there is no need to do separate animacy annotation for each of the project languages: since the OCS, Latin, Gothic and Armenian texts are all close translations of the Greek original, we can use the token alignments to do an automatic transfer of the annotation to the other languages. A test transfer to OCS showed over 95% success for this method.

8. Sample study: Animacy and discourse referent pickup

As seen in section 6, the realization of discourse referents is clearly sensitive to syntactic relation, information status and anaphoric distance. However, when we look at how already mentioned discourse referents are picked up again, we see that animacy must also play a part. We took out all the nouns in the Gospel of Mark that were picked up again, and checked the animacy of the noun and the realization of the anaphoric expression.

Table 7. Animacy of antecedent noun, realization of anaphor

<table>
<thead>
<tr>
<th></th>
<th>Noun</th>
<th>Pronoun</th>
<th>Prodrop</th>
</tr>
</thead>
<tbody>
<tr>
<td>human</td>
<td>112</td>
<td>29.30%</td>
<td>145</td>
</tr>
<tr>
<td>org</td>
<td>4</td>
<td>14.81%</td>
<td>14</td>
</tr>
<tr>
<td>animal</td>
<td>7</td>
<td>50.00%</td>
<td>5</td>
</tr>
<tr>
<td>veh</td>
<td>8</td>
<td>88.89%</td>
<td>1</td>
</tr>
<tr>
<td>concrete</td>
<td>49</td>
<td>66.22%</td>
<td>12</td>
</tr>
<tr>
<td>nonconc</td>
<td>40</td>
<td>61.54%</td>
<td>10</td>
</tr>
<tr>
<td>place</td>
<td>34</td>
<td>80.95%</td>
<td>3</td>
</tr>
<tr>
<td>time</td>
<td>6</td>
<td>100.00%</td>
<td>0</td>
</tr>
</tbody>
</table>
As Table 7 shows, there is a rather dramatic difference between humans and non-humans. All inanimates are preferably picked up by nouns, whereas humans (human, org) are much more likely to be picked up by pronouns or prodrops. This fits in with the idea that the animacy hierarchy is one of several measures of natural saliency: humans care about humans and put them at the centre of their communication, and thus also expect them to be retained in the hearer’s mind for longer. Places and times, on the other hand, as natural frame-setters, are only rarely picked up by anything but nouns. Consider e.g. the following passage (Mk. 13:19–20):

\[\text{ἔσονται γὰρ αἱ ἡμέραι ἑκεῖναι θλίψις ὅπως οὐ γέγονεν τοιαύτη ἀπ' ἀρχῆς κτίσεως, ἣν ἔκτισεν ὁ θεὸς ἐως τοῦ νῦν καὶ οὐ μὴ γένηται. καὶ εἰ μὴ ἐκολόβωσεν κύριος τὰς ἡμέρας, οὐκ ἂν ἔσωθη πᾶσα σάρξ· ἀλλὰ διὰ τοὺς ἐκλεκτοὺς οὓς ἐξελέξατο ἐκολόβωσεν τὰς ἡμέρας.}\]

‘For in those days there will be suffering, such as has not been from the beginning of the creation that God created until now, no, and never will be. And if the lord had not cut short those days, no one would be saved; but for the sake of the elect, whom he chose, he has cut short those days.’

The whole chapter is about the end of times, the days of the apocalypse. These days are referred to throughout the chapter and are thus highly salient in the context, and in our example they occur three times in two verses. Nonetheless they are picked up by a full noun each time.

9. Token alignment

To profit fully from the fact that we have the same text in several languages, it is necessary not only to analyze and tag each version separately, but also to align the versions in such a way that it becomes possible to compare two versions computationally. As we see below, this not only makes it possible to single out constructions in the translation languages that differ from the Greek and thereby to shed light on their native syntax, but also lets us use the various translations to cluster subgroups of Greek constructions which appear unitary on the surface.

The NT also offers the advantage that all versions are marked up with Robert Estienne’s division of the text into verses. While this system is not perfect - the units are not always syntactic units, and there are some divergences in its application among the different translations - it does in most cases provide a correct and reasonably fine-grained alignment. Furthermore, this ‘chunk alignment’ can be exploited to automatically create alignments at the token level.
To do this, we first generated a dictionary of candidate translation pairs by looking at the frequency of their co-occurrence in the same Bible verse. In principle, each word in the Slavic text of, say, Mark 3:18 is possibly the translation of each word in the Greek Mark 3:18:

(13) καὶ Ἀνδρέαν καὶ Φίλιππον καὶ Βαρθολομαίον καὶ Μαθθαίον καὶ Θωμᾶν καὶ Ἰάκωβον τὸν τοῦ Ἀλφαίου καὶ Θαδδαίον καὶ Σίμωνα τὸν Καναναῖον

‘And Andrew, and Philip, and Bartholomew, and Matthew, and Thomas, and James son of Alphaeus, and Thaddeus, and Simon the Cananaean.’

(14) тъ Андрѣ и Филип. тъ вѣртологмѣ. тъ матѣе. тъ томѣ. тъ иѣкова алфѣова. тъ тадеа и симона. кананѣ.

However, words that are actual translation equivalents are likely to co-occur in the same verse with more than chance frequency: there are more bible verses where Θωμᾶς and тома occur in the same verse of the Greek and Slavic versions than those where Ἰάκωβος and тома cooccur when we also take the total frequency of Θωμᾶς, тома and Ἰάκωβος into account. By calculating the relative frequencies of cooccurrences we are therefore able to rank pairs of Greek and Slavic lemmata according to their likelihood of being translation equivalents.

Notice that the dictionary generation only uses the lemmatization: other linguistic information is brought to bear in the next step, where likely translation equivalents inside each verse are ranked using the morphological and syntactic annotation as well as the linearization, which is particularly important, since the early Bible translators aimed at preserving the order of words.

The token alignments are thus created entirely automatically. To evaluate the procedure we manually aligned a single chapter to use as a benchmark. The alignments created by the automatic procedure are about 97% correct when compared to this benchmark.

10. Sample study: Latin translations of Greek participles

The most obvious use of the token alignment is for the study of the translations. Alignment permits us to separate what is pure imitation of the Greek original from what is potentially native syntax. But the alignments can also be put to use for the study of Greek grammar by itself.

Particularly useful are the cases where the target grammar does not offer an exact equivalent of the Greek construction. In such cases the translators had to make independent choices: sometimes they opt for a uniform strategy, using...
Table 8. Jerome's translation of aorist active participles per position

<table>
<thead>
<tr>
<th></th>
<th>Predicative participle</th>
<th>Absolute participle</th>
<th>Subordinate clause</th>
<th>Main clause</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initial, not directly preverbal</td>
<td>55.6% (139)</td>
<td>11.6% (29)</td>
<td>22% (55)</td>
<td>7.6% (19)</td>
<td>3.2% (8)</td>
</tr>
<tr>
<td>Initial, directly preverbal</td>
<td>52.0% (195)</td>
<td>14.7% (55)</td>
<td>17.6% (66)</td>
<td>12.5% (47)</td>
<td>3.2% (12)</td>
</tr>
<tr>
<td>Internal, preverbal</td>
<td>47.9% (175)</td>
<td>11.0% (40)</td>
<td>14.2% (52)</td>
<td>20.3% (74)</td>
<td>6.6% (24)</td>
</tr>
<tr>
<td>Postverbal</td>
<td>54.5% (36)</td>
<td>13.6% (9)</td>
<td>6.1% (4)</td>
<td>6.1% (4)</td>
<td>19.7% (13)</td>
</tr>
</tbody>
</table>

one and the same (sometimes quite artificial) construction to render each exemplar of the Greek construction. But often the translators, and especially Jerome, who was translating into an already established literary language, adopt more flexible translation strategies and choose translations that fit the present context.

One case in point is the translation of Greek aorist active participles. The perfect participle of Latin is only passive, except in the case of deponents. So Jerome cannot render the Greek participle exactly; either he uses a conjunct present participle, or an absolutive construction (with a present participle or with a passive), or a subordinate clause, or a main clause, as shown in Table 8.

The ‘other’ column includes some cases where Jerome used an adverb, but also some misalignments, cases where the automatic token alignment yielded the wrong result.

From the perspective of what is seen in section 4, the most interesting fact is perhaps that Jerome translates 20.3% of all internal, preverbal participles as main clauses. These are the ones that are highly likely to be independent rhemes that cannot be put into subordinate clauses for pragmatic reasons. Notice also that there is much more variety in the translation of postverbal participles, and we find a number of ‘other’ translations.

11. Conclusions

In this article we have provided a brief overview of the design of the PROIEL corpus itself, but we have focused primarily on the linguistic annotation schemes that we apply. There is a large amount of sophisticated information...
available in such a way as to be open for all to inspect. We have provided some glimpses into possible applications of the corpus, although the results must at this date remain preliminary.

Still we believe that such studies show the usefulness of basing linguistic research on corpus data. When the corpus is finished it will provide a unique resource for research into Hellenistic Greek syntax and pragmatics. The availability of data that can be studied quantitatively will, we believe, take the study of Greek further by promoting replicability and intersubjectivity in research and make it possible for researchers to test their hypotheses against the data on a larger scale than what has been possible before. Ideally, scholars should disagree about which theories cover the data in the most adequate way, not about the data themselves.

Currently, the corpus documents a single, though important, text in the history of the Greek language. Once we are able to include texts from before and after the New Testament, it will also be possible to follow the history of the Greek language much more closely than hitherto. Since historical linguistics is, by necessity, always corpus-based, we believe that there is much to gain from enriching our corpora with as much structured data as possible.

References


