Computational and Linguistic Issues in Designing a Syntactically Annotated Parallel Corpus of Indo-European Languages

Dag T. T. Haug — Marius L. Jøhndal — Hanne M. Eckhoff — Eirik Welo — Mari J. B. Hertzenberg — Angelika Müth

Department of Philosophy, Classics, History of Arts and Ideas
P.O. Box 1020 Blindern
N-0315 Oslo
Norway

ABSTRACT. *This paper reports on the development of the PROIEL parallel corpus of New Testament texts, which contains the Greek original of the New Testament and its earliest Indo-European translations, into Latin, Gothic, Old Church Slavic and Classical Armenian. A web application has been constructed specifically for the purpose of annotating the texts at multiple levels: morphology, syntax, alignment at sentence, dependency graph and token level, information structure and semantics. We describe this web application and our annotation schemes. Although designed for investigating pragmatic resources, the corpus with its rich annotation is an important resource in contrastive and historical Indo-European syntax and pragmatics, easily expandable to include other old Indo-European languages.*

RÉSUMÉ. *L'article décrit le développement du corpus aligné PROIEL, qui couvre le texte original grec du Nouveau Testament et les traductions latine, gotique, vieux-slave et arménienne. Pour faciliter la création du corpus, nous avons developpé une application web qui permet l'annotation des textes sur plusieurs niveaux: morphologie, syntax, alignement de phrases, syntagmes et mots, structure informationelle et sémantique. Dans l'article nous décrivons cette application web ainsi que nos schémas d'annotations. Bien que conçu pour l'étude des ph énomènes pragmaticaux, l'annotation très riche des textes a resulté à une ressource importante pour l'étude comparée and historique du syntax et pragmatique indo-européen, et le corpus pourra facilement être étendu à d'autres langues indo-européennes.*

KEYWORDS: *Indo-European, Greek, Latin, Gothic, Old Church Slavic, Classical Armenian, corpus, syntax*

MOTS-CLÉS : *indo-européen, grec, latin, gotique, vieux slave, arménien classique, corpus, syntax*

TAL. Volume 50 – n°2/2009, pages 17 à 45

1. Introduction

In this paper we describe the PROIEL corpus, which at present consists of the Greek text of the New Testament (NT), as well as the Vulgate translation into Latin, the Gothic translation traditionally ascribed to Wulfila and the Slavic translation as attested in the Codex Marianus. The Classical Armenian NT translation has also been added to the corpus, but is not yet annotated.

The corpus is developed within the project *Pragmatic Resources in Old Indo-European Languages* (PROIEL) at the University of Oslo. This project studies the syntactic and morphological means available to old Indo-European languages for the expression of information structural categories such as topicality, backgrounding, focus, etc. The particular interests of PROIEL are: a) word order, b) anaphoric expressions, c) discourse particles, d) definiteness and e) participles as expressions of background events. These are areas where the grammars of our object languages are known to diverge from each other, and a prime concern has been to faithfully represent differences and similarities in these areas. However, a serious investigation of these phenomena obviously requires a rich representation of the entire syntax.

The NT text provides a naturally occurring parallel corpus: no other text exists in old stages of more than two branches of Indo-European. In addition, the NT is the oldest attested text in Armenian, Germanic and Slavic. It therefore provides excellent data for contrastive and comparative Indo-European syntax. It is also one of the texts to have been translated into the most languages, and a version of the Greek text with rich annotation therefore provides a potentially important resource for the study of other languages as well.

On the other hand, there are several problems with using a religious text. Some of them are not directly relevant to the corpus creation (literalness of translations, differing theological conceptions), but one aspect, the sheer number of manuscript variants, which is due to the wide dissemination of the text, is a potential problem.

The phenomena studied in the PROIEL project are not purely syntactic in nature, and the texts are therefore annotated and aligned at several levels. In this article we discuss the computational and technical work on creating the corpus, as well as the annotation work in the areas of syntax, information structure and animacy, and the alignment of sentences, words and dependency graphs, and compare and evaluate our choices against previous efforts in the field.

We have developed a web application specifically for creating this corpus. In section 2 we discuss what kind of resources already existed for our languages, give some background for our choice of developing a custom application and discuss some of the factors which influenced its design. In section 3 we describe the preprocessing which was necessary before the available texts could be loaded into our application.

In the remaining sections we discuss the annotation itself, describing the annotation schemes we have adopted and how they relate to approaches adopted in other projects. Section 4 and 5 focus on the syntactic annotation and the workflow we have

adopted for it; sections 6 to 8 describe other layers of annotation, in particular alignment between the translations, information structure and animacy.

2. Corpus design and software development

2.1. *State of the art: text corpora and treebanks of old Indo-European languages*

The initial stage in digitizing old Indo-European texts consisted in making text corpora. This had the advantage of making the texts searchable and thus enabling faster and more accurate research into these texts. Non-parsed corpora are, however, of limited use to research on the grammatical structure of languages. Accordingly, several parsed corpora (treebanks) are being developed to make various types of grammatical information available. In this section, we present an overview of some of the resources which are available for the study of old Indo-European languages.

As far as the languages covered by the PROIEL project are concerned, there exist several large text corpora and some smaller parsed corpora.

2.1.1. *Greek and Latin*

For Greek, the most important is the Thesaurus Linguae Graecae (TLG).[1] The Perseus project contains a large number of texts in both Greek and Latin.[2] For Latin there is also the LASLA project[3] which offers rich annotation, but as they do not make their underlying data available, it cannot be used for computational purposes.

The TLG is basically a collection of texts, although lemmatization is being developed. The corpus has wide coverage of Classical, Hellenistic, Imperial and Byzantine Greek, but no syntactic or morphological annotation. This makes it difficult to use for linguistic purposes, since information on grammatical patterns which are essential to such highly inflected languages as Ancient Greek is not included.

The Perseus project focuses on syntactic annotation to a larger extent than the TLG. The project's Greek and Latin treebanks contain poetry and prose texts in XML markup.[4] Both treebanks are currently at about 50,000 words (the Greek contains only selections from the works of Homer while the selection of Latin authors is wider with respect to both genre and chronology). The texts contain information on morphology and syntax and are lemmatized. The syntax is analyzed in the format of Dependency Grammar (see further discussion in section 4).

The Index Thomisticus Treebank[5] is another Latin treebank project focusing on the texts of the medieval theologian and philosopher Thomas Aquinas. This project

1. See `http://www.tlg.uci.edu/`.
2. See `http://www.perseus.tufts.edu/hopper/`.
3. See `http://http://www.cipl.ulg.ac.be/Lasla/`.
4. See `http://nlp.perseus.tufts.edu/syntax/treebank/`.
5. See `http://itreebank.marginalia.it/`.

also features morphological annotation and lemmatization as well as syntactic analysis using a Dependency Grammar format, adopted from the Prague Dependency Treebank (PDT).

Our Greek text of the NT is based on the MorphGNT version of the Tischendorf (1869–1872) edition of the Greek New Testament, prepared by Ulrik Sandborg-Petersen.[6] This is a morphologically annotated and lemmatized version of Tischendorff's edition.

For our text of the Latin Vulgate we have used the version prepared by the Perseus project.[7]

2.1.2. *Gothic*

The Wulfila project[8] has published a digitized version of Streitberg (1919)'s edition of the Gothic Bible (prepared by Tom De Herdt and the Wulfila project). The text is aligned by verse with English, Greek, French and Latin (Clementine Vulgate) versions. The morphology has been automatically annotated and ambiguous forms have to some extent been manually disambiguated. There is no syntactic information.

The Gothic text in the PROIEL project is based on this digitized version.

2.1.3. *Armenian*

The *Leiden Armenian Lexical Textbase* (LALT)[9] has published a digitized version of the edition of the Armenian Gospels by Beda Künzle (Künzle, 1984).

The Armenian text from LALT contains morphological annotation (not disambiguated in context) and lemmatization, but this annotation was unfortunately added to the older Zohrab edition (Zôhrapean, 1805), which meant that it had to be ported between the two editions, as described in section 3. No syntactic information is provided.[10]

The Armenian text of the PROIEL project is based on the LALT version of Künzle's edition, which was put at our disposal with approval from the author and the publisher.

---

6. See `http://morphgnt.org/`.
7. See `http://www.perseus.tufts.edu/hopper/`.
8. See `http://www.wulfila.be/gothic/`.
9. See `http://www.sd-editions.com/LALT/home.html`.
10. A selection of texts in Classical Armenian have also been made available by the Titus project (`http://titus.uni-frankfurt.de/indexe.htm`). Word forms are searchable, but no morphological or syntactic analysis is provided.

2.1.4. *Old Church Slavic*

The Corpus Cyrillo-Methodianum Helsingiense[11] (CCMH) in Helsinki provides machine-readable texts of the central OCS canon. (More or less the same texts are available through the TITUS project.) No morphological or syntactic analysis is provided.

The USC Parsed Corpus of Old South Slavic[12] contains morphological analysis of several OCS texts. The corpus is not lemmatized and no syntactic analysis is provided.

Our text of the Codex Marianus is based on Jouko Lindstedt's electronic version of the Jagić (1883) edition of this manuscript[13] (which is a part of the CCMH).

2.2. *Requirements and background for design choices*

This section gives an overview of the requirements and the background for some of the design choices that were made. This is primarily intended to explain how our initial requirements influenced later decisions and how we attempted to delimit the task at hand. As they are not the focus of this paper, the details surrounding these questions and the options available will not be subject to in-depth discussion here.

An important aspect of our initial plan was to be able to recruit annotators worldwide and let them use their own computing equipment without relying on support from us to install or use the software. The most straightforward way of accomplishing this is to use a web application so that only Internet access and a modern web browser would be required. This is of particular importance for university students who do not always have the privileges necessary to install software on their workstations.

We furthermore wanted to be able to annotate the text on a range of different levels. Part of speech and morphological information with lemmatization, syntactic annotation, information structure and alignment links for sentences and tokens are all required for the study of parallel syntactic structures. It is also highly desirable to have free-form attribute-value matrices associated with tokens, sentences or lemmata to cater for the needs of individual researchers in the future. These can be used more specifically for the tagging of semantic properties such as animacy, polarity or *Aktionsart*, which are also of great relevance to the core research questions of the PROIEL project and to contrastive syntax in general.

To our knowledge there exists no tool that offers a unified interface for these actions and does not require the end-user to install software on their own workstation. Customizability was also of some importance, as we wanted to be able to easily adapt and extend the system. This requirement would further have restricted our choices among existing tools.

11. See `http://www.slav.helsinki.fi/ccmh/`.
12. See `http://www-rcf.usc.edu/~pancheva/ParsedCorpus.html`.
13. See `http://www.slav.helsinki.fi/ccmh/marianus.html`.

Since the most pressing issue was to have an interface for annotators to use to annotate text with morphological and syntactic information, we decided to develop our own interface. To limit the scope of this task, the application is only intended for annotation and not for querying or analysis. These tasks have to be done using other tools, e.g. TIGERSearch.[14]

2.2.1. *Morphology and lemmatization*

We also limited the size of the development task by not integrating a dedicated morphological analyser or guesser. Instead the application interfaces with external tools, such as the Stuttgart Finite State Transducer Tools[15] and the Functional Morphology toolkit,[16] or it relies on pre-processed word-lists or reuses annotation already existing in the treebank. Since ready-made word-lists are in fact available for most of the languages in the PROIEL corpus and since these enable us to do a basic form of morphological tagging, further exploration of this area has not been a priority.

We have followed conventional practice and divided the morphological annotation into part of speech, inflectional tag and lemma, but lemmatization is subject to a uniqueness constraint on 4-tuples of language, dictionary base form, part of speech and a variant number. This is a deviation from the normal practice in dictionaries, as multiple parts of speech are commonly treated under the same headword. We consider it necessary to distinguish lemmata based on part of speech if lemmatization is to be done consistently. The judgments of annotators, reviewers and dictionary editors are not likely to converge on the same lemmatization if the subjective assignment of multiple parts of speech to a headword is permitted. The need for a device to distinguish homographs, in our case the variant number, is also reduced as only homographs with identical part of speech must be distinguished. To preserve links between lemmata and digitized dictionaries despite our unconventional lemmatization principle, we also maintain separate links with dictionary resources for each lemma.

2.2.2. *Annotation process*

The annotation process we devised consists of three steps. After the initial preprocessing of the text, all text is flagged as unannotated. In the first annotation step annotators verify that the sentence has been correctly divided or flag a bad sentence division, in which case no further annotation is done before the sentence division has been corrected.

In the next step the application retrieves morphological analyses and presents annotators with the results for disambiguation or asks for confirmation should disambiguation be unnecessary. Annotators are then asked to enter the syntactic annotation

14. See `http://www.ims.uni-stuttgart.de/projekte/TIGER/TIGERSearch/`.
15. See `http://www.ims.uni-stuttgart.de/projekte/gramotron/SOFTWARE/SFST.html`.
16. See `http://www.cs.chalmers.se/~markus/FM/`.

in an editable tree-view of the dependency structure.[17] The manual actions in the process are supported by predictive methods that, for example, guess dependency relations based on the morphology of the head and the dependent. Finally the annotator is presented with a tree structure and can confirm the analysis.

After the annotation, there is a second stage where an independent reviewer inspects the annotator's work and corrects mistakes. The team of reviewers consists of members of the core project and specialists in the relevant languages. Discussions between reviewers aim at keeping consistency between the analyses of the different languages.

To ease the review work we gradually build ad hoc validation rules for syntactic structures (see section 5 for details) and maintain a log of all annotation activity with time-stamps and deltas so that individual changes can be reverted.

2.3. *Data model*

The printed editions of the texts that are part of the PROIEL corpus contain a significant amount of editorial information indicating interpolations, lacunae, corrupted text, etc. Adequate representation of such information is typically realized by using structural markup as defined by a suitable XML schema such as TEI.[18] For the purposes of annotation, on the other hand, only the sequence of tokens grouped into sentences with basic structural information for reference purposes is required, and there are XML schemas for the representation of such annotated texts as well, e.g. TigerXML. Such a simple representation also makes it trivial to process, so that it can be stored in a relational database, and the process can also trivially be reversed.

If a representation with little structural detail is deemed sufficient, then the choice between manipulating it as some form of XML or using a relational database is in reality a matter of practical considerations and individual preferences. Depending to some extent on the size of the data set and the complexity of the annotation scheme, indexing and validity or referential integrity can be deciding factors, but at least in theory it is possible to achieve very similar results with either approach.

The deciding choice in the case of the PROIEL corpus was the availability of state of the art frameworks for web applications, such as Ruby on Rails, the one we ended up using, which are tightly integrated with relational databases and facilitate a rapid, and thus low-cost, development cycle.

It is, however, of some use to researchers to be able to see a faithful reproduction of the original text with editorial information and other structural information intact. Furthermore, it is desirable to preserve the original orthography of a digitized text

17. Due to the fact that reliable statistical parsers are not available for the languages of the project, the syntactical annotation of the project relies more on manual analysis than e.g. (Brants *et al.*, 2003).

18. See `http://www.tei-c.org/index.xml`.

even when this has to be normalized to some extent for annotation purposes, as this allows for easier cross-referencing with the printed original and thus enables users to locate and correct errors in digitization.

This is, however, still a secondary concern, and since this more structurally complex version of the text is not subject to editing, it can be maintained in parallel to the tokenized text prepared for annotation. This amounts, in other words, to a two-level representation: one level with full structural markup and one without.

Examples of differences between the two levels include abbreviations and contractions, which are expanded in the annotation representation, editorial markup, which is removed, and organizational elements such as chapter headings or explicit paragraph dividers, which are also removed.

Practically this is achieved by keeping an XML representation of each sentence in a very simplified version of TEI. This representation is processed using XSL and fed to the tokenizer and to a normalizer that removes undesired orthographic variations. The normalized tokens are then saved in the database before being subjected to annotation.

3. Pre-processing

The text to be annotated has been imported from external sources after appropriate conversion. Rather than compiling manuscript variants or cataloguing variations between text editions, we tried to find the linguistically most suitable, non-copyrighted electronic text version available and used this single text edition alone. While this approach does raise methodological questions, the practical benefits are substantial.

In two cases our text combines information from two different electronic sources which were synthesized. This is the Latin text, where punctuation was imported from the Clementine edition into our text to identify potential sentence breaks, and the Armenian text, where morphology and tokenization were projected from one edition to another.

The projection of punctuation from one text onto another was done using a variation of the *diff* algorithm (Hunt and McIlroy, 1976). The algorithm was run on token sequences, and for each comparison of word token pairs, minimal edit distance was used to test if the word pair could be taken to be a pair of textual variants that should be treated as identical. The product of the process – a sequence of 'chunks' each containing a number of tokens that were found to be different in each text – was filtered by rejecting all non-punctuation tokens and the chunks then applied as patches to the target text. Finally, a subset of the introduced punctuation tokens were used to identify candidates for sentence breaks.

This method was technically fairly successful for the Latin text in the sense that 'off-by-one' errors, i.e. sentence divisions where the sentence boundary is offset by only one or two words, were avoided in most cases. Still, the overall number of badly divided sentences is significant and constitutes about 20–25% of the cases.

Most of the errors are due to the principles of punctuation employed in the Clementine edition. Subordinate clauses, multi-word appositions and preposed relative clauses are separated from the rest of the sentence by a colon in the Clementine. In cases of so-called 'mixed speech,' i.e. where direct speech is introduced by a subjunction (which would normally indicate indirect speech), the subjunction is always grouped together with the direct speech, not the speech verb (see below).

The clearest tendency is for subordinate clauses to be separated from the main clause to which they belong, whatever their syntactic function:

(1) si in digito Dei eicio daemonia / profecto praevenit in vos regnum Dei
if by finger God throw out demons no doubt come to you kingdom God
'If I with the finger of God cast out devils, no doubt the kingdom of God is come upon you.' (*Lk.* 11:20)

Apparently, this happens most frequently with subordinate clauses introduced by *quia*, which may be a result of the fact that *quia* at this time could introduce different kinds of subordinate clauses, and therefore occurs frequently. The use of colons before *quia* also causes errors in cases of mixed speech, where the subjunction is grouped with the direct speech, as here:

(2) scriptum est / quia non in pane solo vivet ...
written is that not by bread alone will live
'It is written that (man) shall not live by bread alone.' (*Lk.* 4:4)

However, since the speech verb selects a complement sentence, we want to have the subjunction in the same sentence, for the purpose of extracting valency information from the corpus.

The few cases of 'off-by-one' errors mentioned above are triggered by textual differences between the two texts:

(3) nolite condemnare et non condemnabimini dimittite / et dimittemini
do not condemn and not be condemned forgive and be forgive
'Condemn not, and ye shall not be condemned: forgive, and ye shall be forgiven.' (*Lk.* 6:37)

The correct sentence division is before *dimittite*, but instead of this word form, the Clementine edition has *Dimitte*. With the parameters used for distance measurement, these forms were deemed too different, and the projection resulted in an 'off-by-one' error:

$$\left.\begin{array}{l}\text{condemnabimini dimittite et}\\ \text{condemnabimini . Dimitte , et}\end{array}\right\} \left[\begin{array}{l}\text{condemnabimini}\\ +\,.\\ -\,\text{dimittite}\\ +\,\text{Dimitte}\\ \text{et}\end{array}\right] \rightarrow \left[\begin{array}{l}\text{condemnabimini}\\ +\,\texttt{DIVISION}\\ \text{et}\end{array}\right]$$

For the Slavic text, erratic use of punctuation constitutes a significant problem. Under the assumption that the Greek original text is a reliable predictor of useful sentence breaks, we introduced sentence divisions using all available punctuation, and used sentence alignment to find the optimal divisions. For this purpose, we used the Gale-Church algorithm (Gale and Church, 1993) with word counts as the similarity measure, and text structure, i.e. chapters and verses, as hard delimiters.

A similar method was used to port markup into our Armenian text. The LALT edition of Künzle's Armenian text unfortunately does not contain any markup except manuscript line numbers and verse and chapter references. LALT had instead done morphological analysis and lemmatization on the basis of the much older Zohrab edition (Zôhrapean, 1805), which they had also digitized. We decided that a new corpus should be built on the more recent edition, so we ported the markup from the Zohrab text to the Künzle text. As in the projection of punctuation in Latin, we used an implementation of the *diff* algorithm, testing each compared token pair for minimum edit distance and transposing annotation between tokens that were judged similar enough.

Porting the markup was particularly important because Classical Armenian has a rich array of one-letter clitics: a) the proclitic prepositions *z-*, *y-*, *č-*; b) the enclitic deixis or definiteness markers *-s*, *-d*, *-n*; c) the so-called *nota accusativi z-*, which often marks accusative forms and is graphically and phonetically identical to the preposition *z-*. Modern editions do not normally indicate word division in such cases. Therefore, since Classical Armenian has many words with initial *z-*, *y-*, *č-* or final *-s*, *-d*, *-n*, it is often not obvious how to distinguish word initial or final consonants from clitic elements. Additionally, the case suffix for the acc.loc.pl. is *-s* as well. That means that a form like *zawr* is ambiguous between nom.acc.sg. of the noun *zawr* 'power' or the acc.sg. (with *nota accusativi z-*) of the noun *awr* 'day', and the form *zawrs* can either be acc.pl. 'powers, hosts' (without *nota accusativi*) or nom.acc.sg. *zawr-s* 'this power (here)' followed by the clitic element *-s*.

In the LALT edition of the Zohrab, the correct tokenization is indicated, and this information was imported to the Künzle text along with the morphology.

To sum up: in building the PROIEL corpus we were able to make use of several available electronic resources. These resources were primarily useful in two areas: first, we were able to secure machine-readable texts for the different versions of the New Testament included in the project; secondly, for some of these texts, high-quality morphological annotation and lemmatization had already been done. The amount of syntactic annotation available was negligible as far as the project languages are concerned. Consequently, it is in this area that the project has been able to contribute most to the expansion of knowledge about these old Indo-European languages.

4. Syntactic annotation

4.1. *Choice of syntactic model*

The languages in our corpus have a rather free word order: word order serves pragmatic and information-structural purposes rather than marking grammatical function. For this reason, word order has to be represented independently of grammatical function.

There already existed two treebanks of Latin based on dependency grammar (DG), Perseus' Latin Dependency Treebank (LDT)[19] and the Index Thomisticus (IT).[20] These treebanks have developed common annotation guidelines based on those of the Prague Dependency Treebank.[21] We therefore settled for a dependency-based formalism where information about word order is kept out of the syntactic model and instead preserved by the logical organization of the annotated text as a sequence of tokens.

We saw several problems with the PDT annotation style. First, to avoid empty nodes, it relies on annotating meta-linguistic material such as punctuation and on complex labeling of the syntactic relations. Second, the granularity of the relations was not fine enough for our purposes. Third, and most importantly, the 'unique head' constraint which PDT has adopted from traditional dependency grammar limits the expressiveness of the formalism. In the following sections we discuss these three issues in more detail. We are aware of the drawbacks of deviating from the choice made by other treebanks, but believe that it was justified in this case.[22]

4.1.1. *Empty nodes*

One attraction of DG, particularly for computational purposes, is its reliance on overt elements: the nodes of the structure to be built for a sentence are given by its words, whereas a phrase structure grammar needs additional phrasal nodes.

But sometimes the reliance on overt elements becomes a problem. In DG every node must have a head, but sometimes the 'natural' head is not available. Two cases are particularly frequent in the ancient languages: asyndetic coordination and 'eclipsed' verbs.

Asyndetic coordination (i.e. without a conjunction) is solved by LDT/IT by letting e.g. the first comma in *lingua, institutis, legibus* be the head coordinating the conjuncts. We believe such an annotation to be linguistically unrealistic. Moreover, it depends on punctuation that a modern editor has introduced into the text; variation between the languages is not always indicative of differing interpretations but can simply reflect editorial practices. An extreme case is our Latin text, which has no punctua-

19. See `http://nlp.perseus.tufts.edu/syntax/treebank/`.
20. See `http://itreebank.marginalia.it/`.
21. See `http://ufal.mff.cuni.cz/pdt2.0/`.
22. Some of the problems with the Dependency Grammar format described in this section are noted also by Brants *et al.* (2003), in particular elliptical and coordinated structures.

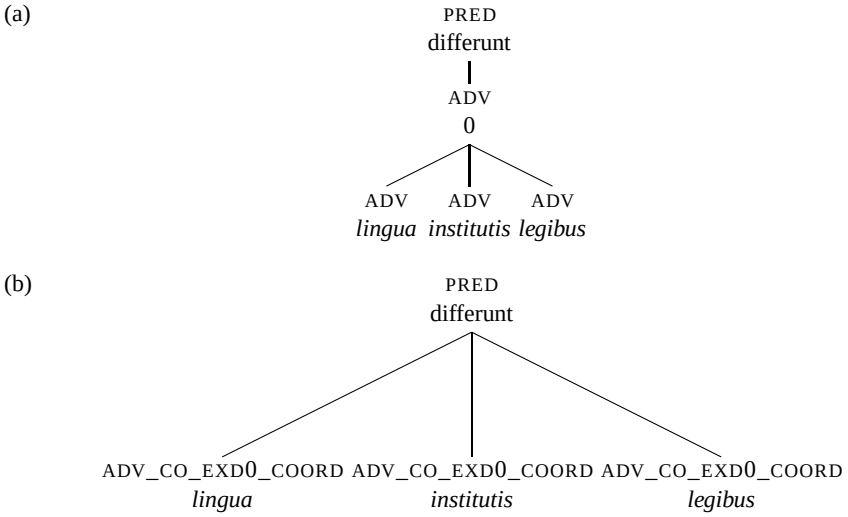

Figure 1. *Coordination with empty node*

tion whatsoever. To solve the problem of the missing head we therefore explicitly add an empty node to the dependency structure (figure 1a): In such cases, where there is punctuation which can serve as the head, LDT and PDT attach orphaned coordinated nodes to their grandmothers or the sentence root and extend the relation name with an index referencing the ellipsis followed by the relation the eclipsed element would have if it were present (figure 1b).

The same system is used whenever there is an eclipsed verb. We believe there are several disadvantages to this: it leads to a large (in principle infinite) number of syntactic relations, amounting to several hundreds in a corpus of about 50.000 words, and they are not directly interpretable but must be 'parsed.' We therefore think it is better to explicitly introduce an empty element.

4.1.2. *Granularity of relations*

We have increased the granularity of the syntactic relations compared to those of the PDT. This is true in both the adnominal and the adverbal domain. Table 1 shows the differences between the annotation schemes. This granularization of the OBJ and ATR relations makes available information which is highly relevant to a study of pragmatic categories.

Data from the corpus (see table 2) show that objects and obliques pattern differently with respect to the use of the definite article. Furthermore, elements that belong to the valency of the verb tend to take the article much more often than 'free' elements: nouns in preposition phrases that are oblique arguments take the definite article more often than nouns in adjunct PPs. Interestingly, agent expressions in passive construc-

LDT	The PROIEL corpus
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	AUX (auxiliary)
EXD	* (external dependency), VOC (vocative)

Table 1. *Specificity of functions in LDT and PROIEL. An asterisk indicates that the annotation schemes diverge in some other way than by one being more specific than the other.* x *in* AUXx *indicates that a number of subtypes are defined.*

		Definite	Indefinite
Adverbal relations	SUB	67.1% (2498)	32.9% (1227)
	OBJ	58.0% (1913)	42.0% (1386)
	OBL	73.9% (627)	26.1% (221)
Nouns in PPs, per P relation	OBL	64.7% (1569)	35.3% (855)
	ADV	52.6% (1061)	47.4% (958)
	AG	66.3% (65)	33.7% (33)
Adnominal relations	PART	71.3% (97)	28.7% (39)
	NARG	51.0% (74)	49.0% (71)
	ATR	57.5% (1453)	42.5% (1074)

Table 2. *Definiteness data from the Greek part of the PROIEL corpus*

tions, which are on the borderline between arguments and adjuncts, pattern with the arguments, though the data set is much smaller. If we had collapsed OBJ, OBL and AG to one tag, as in the PDT, we would have missed these differences.

The subclasses of PDT's adnominal ATR-relation also behave differently with respect to definiteness. Partitives (PART) are predominantly definite, whereas definiteness is more evenly distributed in nominal arguments (NARGs) with (74 vs. 71 cases). Still, both of these relations are much less common than the general ATR-category, where definites predominate slightly.

4.1.3. *Secondary edges*

Our most important deviation from the PDT scheme alters the graph structure itself. Dependency grammar traditionally enforces a 'unique head' principle according to which each word can only have one head. While this provides a restricted and computationally convenient model, there are a number of well-known problems, mostly associated with nonfinite structures where the subject of a nonfinite verb is coreferent with an element of the matrix clause, as in (4) and (5):

(4) ille dixit eis respondens
he.NOM said them.DAT answering.PTCP.NOM
'He told them answering.'

(5) hoc potest fieri
this.NOM can happen.INF
'This can happen.'

In (4), *ille* is the subject of both *dixit* and *respondens*. *respondens* is in turn often analysed as a modifier of the main verb, hence the name 'adverbial participle,' which is sometimes used in traditional grammar.

In generative grammar such structures have been analyzed as 'control' of an empty PRO subject of the participle by the main clause element (*ille* in (4)) or as 'raising' from the subject position of the subordinate clause (the infinitive *fieri* in (5)) to the subject position of the matrix clause. This analysis allows *hoc* to have several functions in the course of the syntactic derivation, but is computationally unattractive.

We analyze both these constructions using a structure sharing mechanism, which is similar to that of Lexical Functional Grammar (LFG). The idea is that single words can have multiple syntactic functions, e.g. we take *ille* in example (4) at face value as both the subject of *dixit* (as indicated by number agreement on the verb) and as the subject of *respondens* (as indicated by case agreement) as shown in (figure 4.1.3).

To deal with such phenomena we enrich our graphs with secondary edges, which are also employed for other types of shared arguments and to indicate predicate identity in the case of ellipsis. The principles behind this annotation are presented in Haug and Jøhndal (2008). Compared to the PDT scheme, we see the following advantages: a) arguments can be encoded as dependents of several heads, b) the subject–predicate relationship can be uniformly represented as a dependent–head relationship, and c) we can often indicate which verbal lemma hides behind an ellipsis These are considerable advantages, for example for the extraction of valency information.

The introduction of secondary edges and the added granularity in the syntactic relations have the bonus effect of bringing the PROIEL annotation closer to LFG. We are working on implementing an algorithm that converts PROIEL dependency graphs to the Prolog representation of LFG f-structures used by XLE.[23] This will be of

23. See `http://www2.parc.com/isl/groups/nltt/xle/`.

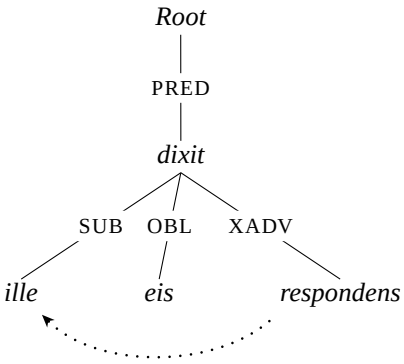

Figure 2. *Shared subjects*

significant help in developing rule-based LFG parsers for our languages, by providing benchmark analyses to test grammars against.

It should also be noted that the introduction of secondary relations means that the dependency structures in the corpus are in fact labeled, directed graphs with the potential for cycles. The cycles can, however, be removed, should that be necessary. The complexity facing annotators is also kept down because the labels of secondary edges can be synthesized from context and therefore do not have to be provided by annotators.

4.2. *Porting syntactic annotation schemes*

The syntactic model was developed on the basis of Greek and Latin and then transferred to Gothic and Slavic. The syntax of the four languages is very similar, broadly speaking, but the transfer raised some issues.

A particular challenge came from the fact that the Old Church Slavic (OCS) and Gothic NTs are the first (and, in the case of Gothic, only extant) fixations of the languages. The texts are not the products of a long and stable writing tradition, as the Greek and Latin versions are, and display much more variation. The orthography is unstable, which is a challenge to lemmatization. An OCS word may have multiple orthographical variants: consider the lemma *дьпьсь* ‘today’, which also occurs as *дъпесь*, *дьпесъ* and *дьпесь*. In these cases we follow the lemmatization of Cejtlin *et al.* (1994).

Two emergent phenomena in OCS syntax required principled decisions: a) The incipient genitive-marking of animate objects with transitive verbs, and b) the clitic third person reflexive pronouns that also serve as markers of reflexive and passive verb forms.

4.2.1. *Genitive objects*

The annotation scheme distinguishes between objects (OBJ) and obliques (OBL). Only arguments that can be subjects in a corresponding passive construction can be OBJs, whereas OBL is a wide category, including e.g. prepositional phrases indicating goal arguments of motion verbs. A challenge to this distinction is found in the extensive use of the genitive to mark arguments in OCS. OCS attests the earliest stages of a change that has affected all Slavic languages to a smaller or greater extent – animate objects of regular transitive verbs are genitive-marked instead of accusative-marked; see e.g. Klenin (1983). In OCS this only affects nominals denoting male human beings, preferably adult and high-status, as seen in (6), but even such nouns can have the regular accusative, as in (7).

(6) ašte žena pouštъši mǫža posagnetъ za inъ
if woman having-let-go husband.GEN marries after other.ACC
'If a woman divorces her husband and marries another . . . ' (*Mk.* 10:12)

(7) idi prizovi mǫžь tvoi
go call husband.ACC your.ACC
'Go, call your husband.' (*John* 4:16)

However, arguments of verbs can also be genitive-marked for other reasons: a) the verb requires the genitive, b) the verb is negated, c) the object is only partially affected.

These distinctions need to be preserved in the annotation. We rejected the idea of having a separate morphological tag for 'genitive-shaped accusatives,' since, for example, there are many cases of negated verbs with animate objects and verbs with fluctuating case requirements and animate objects.

We also rejected the idea of tagging all such genitives as accusatives because of their function, and leaving it to the semantic animacy tagging (see section 8) to single out the ones that might be genitive-shaped: there are too many accusative-marked human referents for this to be a practical solution.

Instead, we chose to annotate all genitive-shaped nominals as morphological genitives. By combining the morphological information with the syntactic tags OBJ and OBL, and also the supertag ARG, and using valency information from (Cejtlin *et al.*, 1994), most of the distinctions are preserved:

– verbs that always take the genitive take genitive-marked OBLs, also when they are negated;

– verbs that are regular transitives take OBJs, also when the object is genitive-marked whether this be due to animacy, negation or partitivity;

– verbs that can take either the genitive or the accusative take a) OBJ when the argument is clearly accusative-marked; b) OBL when the argument is genitive-marked and the genitive-marking cannot be due to negation, partitivity or animacy; c) the

supertag ARG when the argument is genitive-marked and this may be due to negation, partitivity or animacy, i.e. when we cannot determine whether the argument is an object or an oblique.

In this way, information on genitive with negation and genitive-marked animate objects is well preserved, since the case tag and argument tag can be crossed with information on the presence or absence of a negation, or with the semantic animacy tags (see section 8). The partitives will be the genitive OBJs that are due neither to animacy nor to negation.

4.2.2. *Reflexives*

Both in OCS and Gothic, reflexive pronouns in the accusative or dative have become markers of reflexive or even passive verbs (8). The same enclitic pronoun, *sę* in OCS, may also serve as a regular accusative object (9) or in other functions where the accusative is possible.

(8) i eže imatъ vъzьmetъ sę otъ nego
and what.ACC has will-take REFL.ACC from him
'And what he has will be taken from him.' (*Mt.* 13:12)

(9) sъpasi sę samъ
save REFL.ACC self.NOM
'Save yourself.' (*Mk.* 15:30)

Again, we use the interplay between morphological tags and syntactic tags to make the distinction: all enclitic reflexive pronouns are annotated as such morphologically, regardless of function. Syntactically, however, those that serve as reflexive or passive markers get the tag AUX, whereas those that serve regular syntactic functions get the appropriate tag, usually OBJ.

In cases where the reflexive pronoun is ambiguous between a reflexive and an argument reading, we annotate them as arguments, for the practical reason that the group of argument reflexives is much smaller than the group of AUX reflexives. The borderline cases should be placed in the smaller group where they are easier to retrieve.

5. Consistency issues

The PROIEL corpus is developed using an international team of annotators. The annotators were chosen on the basis of academic experience with one or more of the languages relevant for the project. The nature of the annotation tool enabled annotators to do their work using the Internet, and all annotators received initial training and discussed problems with the project members by email or on a web forum.

This work-flow was combined with a bottom-up perspective on the syntactic analysis: instead of creating a model first and then applying it to the data, the model was

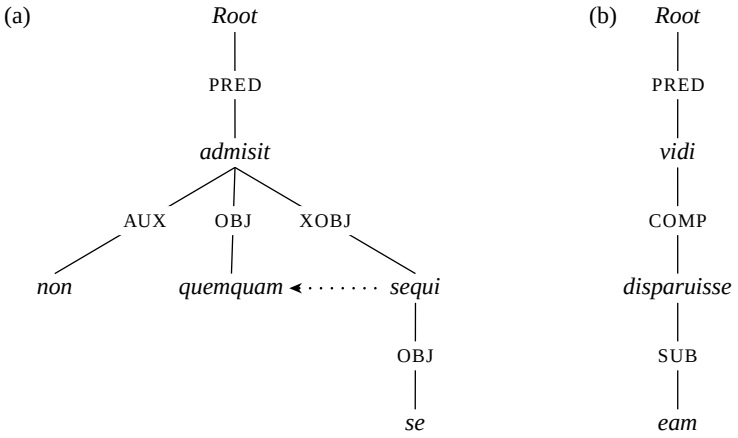

Figure 3. *Accusative with infinitive structures*

built and refined as we encountered new data. While the former approach undoubtedly leads to more consistency early on, it is in practice hard to apply to dead languages for which we often lack sufficiently precise syntactic accounts in reasonably modern frameworks. It is, for example, clear that in some structures of main verb + accusative + infinitive, the accusative gets a thematic role from the main verb as well as from the infinitive (10), while in others, it only gets a role from the infinitive (11):

(10) non admisit quemquam se sequi
not permit.3SG.PFV anyone.ACC. REFL.ACC sequi.PS.INF
'He did not permit anyone to follow him.'

(11) vidi eam disparuisse
saw.3SG.PFV her.ACC disappear.PST.INF
'He saw that she had disappeared.'

As shown in figure 3, these structures can be treated differently in our annotation scheme: In (10) we can take *quemquam* as the object of *admisit* as well as the subject of the infinitive (which is an XOBJ) via a secondary edge, whereas in (11) *eam* does not get a role from the main verb, but is simply the subject of the infinitive (which is therefore a COMP).

However, extant grammars often treat both these structures as accusatives with infinitive (AcI) and do not supply enough information for us to create lists of verbs and their constructions. In the annotation stage we rely in part on annotators' intuitions and in part on providing template solutions, like 'prefer the AcI analysis.' In other cases, where traditional grammar *does* provide guidelines, such as on Slavic verbs governing the genitive instead of the accusative, we ask annotators to follow these. The principle we adhere to is to impose as few of our own interpretations as possible.

Many inconsistencies are simply errors. To avoid these we have implemented certain validation strategies. Analyses that violate certain criteria cannot be saved and the annotator is presented with an error message. We have taken a pragmatic approach to this, and instead of writing rules that would be the beginnings of formal grammars of our languages we simply check for common mistakes. For example, DG requires that the verb is taken as the head of relative clauses. Annotators tended to take the relative pronoun as the head (as in phrase structure analyses), so we implemented an ad hoc rule banning such analyses. In the same vein, it is easy to forget a secondary edge to the subject, so we require that all XOBJs and XADVs should have a secondary edge.

Apart from simple errors, inconsistency can actually be a good thing in the sense that it points to debatable issues. Simple misunderstandings may indicate that the instructions given to the annotator are unclear or ambiguous. In other cases, inconsistencies suggest that there is a better analysis than the one adopted at the start, or that the full range of properties of a given construction has not been taken into account.

When a category straddles the border between two functions, e.g. the border between OBL (obliques) and ADV (adverbials), similar constructions may receive completely different analyses. This may indicate that the border between the two functions is not well-defined (as is clearly the case with the argument-adjunct distinction in theoretical syntax).

In addition to blocking certain analyses, we aim to reduce the amount of inconsistency by using *supertags*. These tags allow the annotators to indicate that there is uncertainty about the correct analysis. For example, the REL supertag covers the functions ATR (attribute) and APOS (apposition) with respect to relative clauses. Whenever the annotators are in doubt whether a relative clause functions as an attribute or as an apposition, they are supposed to use the more general supertag.

Using supertags in the annotation process is profitable because it minimizes the amount of bad raw data. As long as annotators use the specific tags only in the clear cases, we avoid contamination of the data. Instead, the team of reviewers can isolate and discuss difficulties and decide on a consistent analysis.

In practice, the use of supertags has not, however, been entirely successful as supertags are used rarely. There may be several reasons for this. First, the annotator may feel certain of the correct analysis. Second, a more general psychological effect may be at work. Annotators want to do as good work as possible, but the use of a supertag would indicate that the annotator was uncertain.

To conclude, we aim to make our data as consistent as possible by different means: first, by good training on the basis of explicit guidelines, second, by not allowing certain analyses to be saved in the database, and finally, by using general supertags for unclear cases. The amount of inconsistency left by these measures, which are all applied in the *annotation* phase, is further reduced in the *revision* phase, in which classes of problematic examples are discussed and the general treatment is decided upon.

6. Alignment

Alignment is possible both for tokens, sentences and dependency sub-graphs. The latter is useful in cases where translations are faithful word-by-word translations but still structurally different; this more complicated, manual task is described in Jøhndal *et al.* (forthcoming).

Token and sentence alignments are a mixture of automatically generated alignments, manual alignments and alignment hints. In general, we align one translated text with the Greek original, and the alignment is gradually refined from sentence alignment to token alignment to dependency sub-graph alignment.

Sentence alignments are generated using the Gale-Church sentence alignment algorithm (Gale and Church, 1993). The text is split into blocks delimited by the chapter and verse numbering of the text. Block pairs are aligned and can be inspected by a reviewer who can supply hints in the form of forced alignments or 'black-listed' alignments. The automatic alignment is then recalculated to take into account any manual intervention. When the reviewer is satisfied that the alignments are correct, this iterative process is completed by committing all alignments to the database.

Once sentence alignments have been generated and reviewed, token alignment can be performed. To automate this process we have created dictionaries where, for each lemma in the target language, candidate lemmata in the source language (i.e. Greek) are ranked based on maximum likelihood of their co-occurring in the same Bible verse. This is done using a process described in Cysouw *et al.* (2007).

Candidate translation pairs within aligned sentences are then scored using the dictionary as well as the linearization numbers within the sentence, and the morphological and syntactic information available. The process is repeated, with each iteration accepting 'worse' equivalents, but penalizing alignments that imply a transposition of word order. Because the translators have aimed at keeping the original word order, this approach gives good results. Experiments on the Slavic translations show well over 90% success.

7. Information structure

The PROIEL corpus will eventually contain annotation for information structure (IS) along three dimensions: a) *information status* of discourse referents (accessibility), b) *referential distance* (anaphoric links) and c) *contrast*. These dimensions are important for explaining, among other things, patterns of word order and the lexical realization of discourse referents. The annotation of contrast is still experimental and will not be discussed in detail. Following Rooth (1992), we understand the notion of contrast in terms of *alternatives* and will distinguish between explicit and implicit contrast as well as paired contrasts (contrastive topics).

The adequacy of the annotation scheme has been tested out on text selections from the Greek NT. At a later stage, the IS annotation of the Greek text will be transferred to

the other languages using the token alignments described in section 6. Although we do not know of other attempts at porting IS annotation across languages, we expect such a transfer to be possible between our translations, due to their literalness and especially their faithfulness in rendering the Greek word order. The successful test transfer of animacy tagging from Greek to Slavic, involving the same types of referents, supports this expectation (cf. section 8).

The annotation of IS focuses on *nominal* elements, i.e. noun phrases. In particular, referential noun phrases are selected for annotation. With respect to contrast, however, the selection of annotatable elements is wider because a wider range of linguistic elements may be contrasted.

The annotation scheme must fulfill at least two, possibly conflicting, goals. On the one hand, the tag set must be large enough to capture the full range of the information that we need to answer research questions. On the other hand, the tags must be clearly defined to ensure a high degree of inter-annotator agreement. The IS annotation scheme tries to balance these two concerns. In addition the tag set should be applicable to all the languages in our corpus, which differ, among other things, by the fact that Greek has a definite article and other languages do not. Finally, we wanted the tag set to be used in combination with our morphological and syntactic annotation instead of duplicating the syntactic information in the IS annotation, with the associated risk of introducing inconsistencies.

7.1. *Information status*

Annotation of discourse accessibility goes back to Prince (1981), and the ideas behind most modern annotation schemes have roots in this paper. In developing our own scheme we evaluated Dipper *et al.* (2007), Nissim *et al.* (2004) and Riester and Lorenz (2009). Full compatibility with either of these schemes was not an important goal, as they are all applied to very different texts, which would limit the usefulness of comparing the data. We do not know about any attempt to tag accessibility on ancient texts,[24] so we decided to take an eclectic approach and develop our own scheme based on these works.

Both Nissim *et al.* (2004) and Dipper *et al.* (2007) are based on a fundamentally tripartite distinction into new/mediated/old (Nissim *et al.*, 2004) or new/accessible/given (Dipper *et al.*, 2007), but introduce different kinds of subdivisions of these major tags: for example, Nissim *et al.* (2004) use MEDIATED-PART for referents which are inferrable through part-whole relationships, but MEDIATED-EVENT when an entity can be inferred from a previous VP. By using such a hierarchical annotation scheme, it is possible to collapse some distinctions and get more reliable data (i.e. with higher inter-annotator agreement measures), so we kept this idea.

24. There also seem to be very few attempts to tag narrative texts: the schemes cited above are mainly applied to dialogues and news bulletins. And among the five schemes evaluated in Ritz *et al.* (2008), only one is applied to narrative texts.

The scheme of Riester and Lorenz (2009) diverges from the others in that it uses different categories for definite and indefinite NPs. For example, an indefinite NP which has not previously been mentioned and is not inferrable will be tagged as NEW, whereas a definite NP which has not previously been mentioned and is not somehow inferrable will be tagged as 'accessible-by-description.'

The approach of Riester and Lorenz (2009) cannot be directly applied to the PROIEL corpus, where most of the languages do not have definite articles. On the other hand, this scheme is the only one to be solidly grounded in linguistic theory, specifically Discourse Representation Theory (Kamp and Reyle, 1993), which includes a theory about different *contexts* where discourse referents may be identified by the hearer. We follow Riester and Lorenz (2009) in determining annotation status according to these contexts. This means that referents are placed on the following scale of discourse accessibility:

(12) OLD < ACC-SIT < ACC-INF < ACC-GEN < NEW

OLD referents are the ones that can be found in the preceding discourse context. The mid-part of the accessibility scale refers to *accessible* discourse referents. The category of 'accessible' is further subdivided as a) ACC-GEN (world knowledge, generally accessible), b) ACC-SIT (accessible from discourse situation), and c) ACC-INF (inferrable from preceding discourse).

(13) eipen de Iêsous pros tous paragenomenous pros auton arkhiereis
said PART Jesus-OLD to the arrived to him-OLD archpriests-NEW
kai stratêgous tou hierou kai presbuterous
and captains-NEW of-the temple-ACC-GEN and elders-NEW

'And Jesus said to the archpriests and captains of the temple and elders who had come to him.' (*Lk.* 22:52)

The pilot annotation for information status of discourse referents covers a total of 655 NPs. Several trial runs were made, followed by discussion of inconsistently tagged passages. In the final trial run, the tag set as a whole was applied more consistently by three annotators (kappa = 0.89 for the main tags OLD, ACC and NEW, kappa = 0.86 for the subdivisions of the ACC-tag).

When broken down on individual tags, the pilot is rather small, but some tendencies emerge. Among the individual tags, the OLD tag was applied most consistently: it accounted for 60% of cases of agreement between annotators (as compared to 52% of all tags) and NEW accounted for 20% (and 16% of all tags). This is easily explained by the fact that overt, local mention of a referent excludes the NEW option, as well as the different ACC tags. In addition, it is easy to check for the presence of an earlier mention of the referent, and old referents typically come in the form of (anaphoric) pronouns which encode their information status lexically. Thus, a high degree of consistency was to be expected.

The relationship between new and accessible referents is more complicated: the referent is by definition not previously mentioned in the local context and there is thus no overt element to check for. This was reflected in the inter-annotator agreement values above.

In the test runs we made, the use of the ACC-GEN-tag proved to be the most problematic. The problems stem from the difficulty in making assumptions about general information available to discourse participants, in our case the original intended audience of the NT. The question of the geography of the Holy Land will illustrate this point. Words referring to geographical locations behave differently from words referring to e.g. actors in the narrative. Actors are typically introduced by means of presentational devices such as 'There was an X...' before their actions are described in more detail. Geographical locations, on the other hand, are usually not introduced in this way, but rather assumed to be identifiable to the reader (the definite article is frequent with geographical names in Greek). One solution would be to identify generally available referents through the fact that they may occur with the definite article on first mention. This would, however, have the adverse effect of making the data less useful for subsequent research on the behavior of the definite article because the presence of the article was part of the definition of the category. Instead, we wanted to look just at the nominal head when annotating for information status, ignoring definiteness.

The two remaining subdivisions of the accessible category were easier to apply. The tag ACC-SIT was used on referents accessible from the discourse situation. This tag is mostly used in sequences of direct speech, particularly on NPs which contain deictic expressions. The ACC-INF tag was used for referents that could plausibly be inferred from referents mentioned in the previous discourse.

New referents were generally easy to identify. A question arises, however, with respect to major participants who reappear at intervals in the narrative, such as the disciples of Jesus. When these participants reappear they do not occur in typical presentation constructions which signal new material, but are rather presented as generally known entities.

A separate problem relates to the use of direct speech within narrative. Does direct speech constitute a separate discourse universe, and how should we handle references going outside a direct speech context? The following examples illustrate a typical context:

(14) eteken ton huion autês
gave-birth the son-ACC-INF she
'...she gave birth to her son...' (*Lk.* 2:7)

(15) heurêsete brefos...
shall-find child-NEW
'You shall find a child...' (*Lk.* 2:12)

(16) aneuran tên te Mariam kai ton Iôsêf kai to brefos
found the and Mary-OLD and the Joseph-OLD and the child-OLD
'...they found Mary and Joseph and the child...' (*Lk.* 2:16)

The birth of Jesus is first described. Later an angel tells the shepherds that they will find a child, and finally the shepherds actually find Jesus. Within the direct speech of the angel, the child is introduced as an indefinite NP and tagged as NEW. Later, when the shepherds find Jesus, the same word is used, only now in definite form.

We have adopted the principle that direct speech forms its own discourse universe in the sense that referents which have previously been mentioned in the narrative may be considered NEW within passages of direct speech.

7.2. *Referential distance*

At an early stage, we experimented with distinguishing between *active* and *inactive* old referents, depending on whether the referent had been mentioned in the previous syntactic unit. Since this was in effect a distance measure, it was decided to abolish this distinction in favour of anaphoric links between anaphoric expressions and their referents, although we still keep the OLD-INACT for items that are further away than the maximum allowable length of anaphoric links.[25]

The major reason for this was that anaphoric links would provide more exact data on the distribution of anaphors. While the dichotomy associated with the earlier distinction between 'active' and 'inactive' old referents only made reference to the immediately preceding syntactic unit, the distance between anaphoric expression and referent may now be measured exactly in words or sentences and the inactive tag can be reserved for long-distance anaphoric relations. A sample text passage with anaphoric links is shown in examples (17-20).

The anaphoric links are strictly local: when a referent is taken up by means of several anaphoric expressions, the last anaphor refers to the immediately preceding anaphoric expression rather than directly to the referent itself. Thus, we build anaphoric chains which at some point terminate in the referent binding the anaphor(s). This allows us to measure the complexity of the anaphoric chains as well as the absolute distance of any anaphoric expression from its binder.

8. Semantic tags: animacy annotation

An advantage of having a publicly available text corpus is that the research based on the corpus is in principle replicable. To gain further from this advantage, the PROIEL corpus contains an annotation layer mainly intended for semantic tagging, as mentioned in section 2.2. In actual fact, however, this layer can be used for any

25. Since major referents in the narrative recur, it was necessary to specify the maximal distance between an anaphor and its potential antecedent. It was decided to put this limit at 13 sentences, and using OLD-INACT when the distance between it and a potential antecedent exceeded the maximal referential distance. Further experiments are clearly needed in order to ascertain the optimal limit.

(17) egeneto Iôannes ho baptizôn en têi erêmôi kai kêrussôn baptisma
was John.NEW ART baptist in ART desert.OLD and preaching baptism.NEW
metanoias eis afesin hamartiôn.
repentance for forgiveness of sins.ACC.GEN.

'John the Baptist appeared in the wilderness preaching a baptism of repentance for the forgiveness of sins.'

(18) kai exeporeueto pros auton pasa hê Ioudaia khôra kai hoi
and travel out to him.OLD all ART Judaean country.ACC.GEN. and ART
Hierosolumeitai pantes
Jerusalemite.ACC.GEN. all

'And all the country of Judea was going out to him, and all the people of Jerusalem.'

(19) kai PRO-SUBJ ebaptizonto hup' autou en tôi Iordanêi potamôi
and they.OLD were baptized by him.OLD in ART Jordan.ACC.GEN. river
exomologoumenoi tas hamartias autôn.
confessing ART sins.OLD of them.OLD'

'And they were being baptized by him in the Jordan River, confessing their sins.'

(20) kai ên ho Iôannês endedumenos trikhas kamêlou ...
and was ART John.OLD clothed hairs.NEW of camel

'And John was clothed with camel's hairs...' (*Mk.* 1:4–6)

user-defined tagging at lemma or token level, and is thus also a flexible tool for storing and making accessible the data work of individual scholars working on specific subjects. Thus, more fine-grained and specialized analyses can be tested and re-used by other scholars. We have conducted several tagging experiments, e.g. for aktionsart, event time and adjective class. All Greek noun lemmata are now tagged for animacy. The current section describes the principles for animacy tagging.

Animacy is a semantic category that is highly relevant to many of the central research questions in the PROIEL project. Not only is animacy actually emerging as a grammatical category in OCS (see section 4.2.1), but it is also likely to affect the choice of argument realization in the languages where the category is not grammaticalized, and is important to the question of topichood.

The pervasiveness of animacy effects in language is well known from typological studies, and there are many versions of implicational animacy hierarchies around in the literature. A common representation is that of the extended animacy hierarchy as found in Dixon (1979, p. 85): first/second person pronouns < third person pronoun < proper names < human common noun < nonhuman animate common noun < inanimate common noun. We may note, as does Croft (2003, p. 130), that this hierarchy is in fact very simple when it comes to animacy proper (human < nonhuman animate < inanimate), but includes morphological features such as part of speech and noun class. In the PROIEL application we can get at this information via the morphological

Animacy tag	Brief description
HUMAN	things that look and act like humans, including deities and spirits
ORG	a collectivity of humans with some degree of group identity
ANIMAL	non-human animates
CONCRETE	'prototypical' concrete objects or substances, excluding intangibles
VEH	vehicles
NONCONC	anything that is not prototypically concrete but clearly inanimate, e.g. events
PLACE	nominals that will normally serve as locations for human actions
TIME	expressions referring to periods of time

Table 3. *Tags for animacy annotation*

tagging. Including it in the actual animacy tags is thus neither necessary nor desirable, since such duplication of information may cause inconsistencies.

Animacy annotation is not very common in linguistic corpora, but to the extent that it is done, different degrees of granularity are chosen. Sometimes only the simple dichotomy human:non-human is tagged, as in the Swedish treebank Talbanken05 (Øvrelid, 2009). In the Potsdam SFB632 annotation guidelines, a four-way distinction is employed, between human animates, non-human animates, inanimates and inanimates with human-like properties (Dipper *et al.*, 2007, section 8). There are also corpora with very detailed animacy annotation, such as the 20-way distinction maintained in the Russian National Corpus.[26]

We have chosen a middle way, using a slightly simplified version of the annotation scheme of Zaenen *et al.* (2004); see table 3. We deem this scheme to be sufficiently granular to give interesting results, but at the same time simple enough to make the annotation process fairly straightforward. In particular, we expect there to be interesting differences between concrete and non-concrete inanimates. We also consider it an advantage to be able to access temporal and locative adverbials by way of the animacy tags.

Eventually we want all nouns, pronouns, substantivized adjectives and participles to be tagged for animacy in each project language, as well as denominal adjectives in OCS. The tags are token-level tags, pertaining to the animacy of each token's referent, but in order to make the annotation process maximally efficient, we do as much as possible at lemma level. We then go on to adjust the annotations at token level. As a first step, all Greek noun lemmata were tagged for animacy by one project participant. The tags were then reviewed by another project member. The lemma-level annotation gave quick and high-quality results. Most of the annotations would also be valid at token level. We found that most of the problems encountered in the annotation process were due to lemmata that were used with different animacy status in different contexts.

26. See their guidelines for semantic annotation at `http://ruscorpora.ru/en/corpora-sem.html`.

A case in point is the lemma *kardia* 'heart', which was annotated as CONCRETE by the annotator, as this is the default choice for body parts. However, the reviewer found that none of the tokens of this lemma referred to physical hearts, but rather to people's minds and opinions. The annotation was therefore changed to NONCONC at lemma level.

Given the strong tendency for nouns to be translated into nouns,[27] we found that the animacy annotation could be transferred to the other languages via the token alignments.[28] We performed a test transfer to OCS: We found all OCS nouns and adjectives that were token aligned with Greek nouns. Each OCS token may be token aligned with more than one Greek token, and may thus potentially be associated with more than one animacy tag. For each OCS token, we therefore selected the most frequently occurring animacy tag within the set of aligned Greek tokens, and transferred that tag to the OCS token. The results were good: over 95% of the OCS annotations were correct.

As an illustration of errors we got in the transfer process, consider the lemma *kъnigy* 'book, writing', which was asigned the tag NONCONC. The reason was that *kъnigy* did not get its animacy tag from *biblion* 'document, book,' as one could perhaps expect, but rather from *graphê* 'writing', with which it is aligned much more frequently. Since *graphê* most commonly refers to laws and prophecies rather than to actual objects of writing, it bears the tag NONCONC. However, although *kъnigy* is more often aligned with *graphê* than with *biblion*, it still most often refers to concrete objects of writing. Clearly, then, the lemma-level annotation must be checked and adjusted at token level. This is of course particularly important after automatic tag transfers from the Greek to another language, but it is also necessary in the Greek.

Currently only nouns have animacy tags. In a further stage we can extend the annotation to at least pronouns by using the anaphoric links in the information structure annotation (see section 7.2). All members of an anaphoric chain must necessarily have the same animacy status.

9. Conclusion

The PROIEL database of NT translations provides multi-layered annotation and alignment possibilities. This information may be combined to identify complex interactions between morphology, syntax, information structure and semantics. While the immediate goal of the project is to investigate how the elements of the grammatical systems of the languages involved are used in expressing pragmatic notions, the database will function as a flexible tool for scholars working on very different linguistic issues, as well as, to some extent, for non-linguists.

27. 93% even in OCS, where nouns in the genitive are regularly translated into denominal adjectives, 97–98% in Latin and Gothic.

28. Even without doing this, the animacy tags would be accessible (with a certain margin of error) for all the languages by way of the token alignments with the Greek.

	Reviewed	%	Annotated	%	Unannotated	%	Total
Greek	23120	16.8	91805	66.7	22788	16.5	137713
Latin	25445	20.2	81861	65.0	18606	14.8	125912
OCS	10297	17.6	46485	79.6	1578	2.7	58360
Gothic	8190	14.5	46536	82.6	1580	2.8	56306

Table 4. *Progress in the annotation of the NT (by tokens)*

In the process of annotation and review, we have encountered various challenges which have all served to clarify problematic issues and pointed the way toward their solution. As a result, we are able to provide increasingly stable and sophisticated analyses of the corpus data. This will continue as the revision progresses.

The annotation of the NT translations is almost complete (see table 4). The review process has also reached a mature state where the analyses of frequently occurring constructions are fixed, but work remains to be done on reviewing the entire corpus and developing analyses of less frequent phenomena.

The reviewed part of the corpus is openly available, and we encourage others to use it. The choice of an open-source architecture for the PROIEL application means that our work can be re-used by others for related purposes, avoiding duplication of effort.

Since we have stable guidelines, trained annotators and mature annotation software, it is natural to consider further extensions to the corpus. A natural first step will be to include texts from the older, classical stages of Greek and Latin. But we would also like to broaden the scope of the project by including more languages. We believe the PROIEL treebank will be an important tool for the historical syntax of Indo-European languages, but to fully achieve this goal, it should eventually include data from all the major, old branches of Indo-European.

10. References

Brants T., Skut W., Uszkoreit H., "Syntactic annotation of a German newspaper corpus", *in* A. Abeillé (ed.), *Treebanks: Building and Using Parsed Corpora*, Kluwer, Dordrecht, p. 73-87, 2003.

Cejtlin R. M., Večerka R., Blagova E. (eds.), *Staroslavjanskij slovar'*, Russkij jazyk, Moscow, 1994.

Croft W., *Typology and Universals*, Cambridge University Press, Cambridge, 2003.

Cysouw M., Biemann C., Ongyrth M., "Using Strong's numbers in the Bible to test an automatic alignment of parallel texts", *Sprachtypologie und Universalienforschung*, vol. 60, p. 1-16, 2007.

Dipper S., Götze M., Skopeteas S., *Information Structure in Cross-Linguistic Corpora: Annotation Guidelines for Phonology, Morphology, Syntax, Semantics, and Information Structure*. 2007, available at `http://www.sfb632.uni-potsdam.de/~d1/sfb632_guidelines/`.

Dixon R., "Ergativity", *Language*, vol. 55, n° 1, p. 59-138, 1979.

Gale W. A., Church K. W., "A Program for Aligning Sentences in Bilingual Corpora", *Computational Linguistics*, vol. 19, n° 1, p. 75-102, 1993.

Haug D., Jøhndal M. L., "Creating a Parallel Treebank of the Old Indo-European Bible Translations", *Language Resources and Evaluation*, Marrakech, Morocco, 2008.

Hunt J. W., McIlroy M. D., An Algorithm for Differential File Comparison, Computing Science Technical Report n° 41, Bell Laboratories, 1976.

Jagić V., *Quattuor evangeliorum versionis palaeoslovenicae Codex Marianus glagoliticus*, Weidmann, Berlin, 1883.

Jøhndal M. L., Eckhoff H., Haug D., "Aligning Syntax in Early New Testament Texts: The PROIEL Corpus", *Zeitschrift für Slawistik*, forthcoming.

Kamp H., Reyle U., *From discourse to logic: introduction to model-theoretic semantics of natural language, formal logic and discourse representation theory*, Kluwer, 1993.

Klenin E., *Animacy in Russian: A New Interpretation*, Slavica, Columbus, Ohio, 1983.

Künzle B., *Das altarmenische Evangelium*, Peter Lang, Bern, 1984.

Nissim M., Dingare S., Carletta J., Steedman M., "An Annotation Scheme for Information Status in Dialogue", *Language Resources and Evaluation*, Lisbon, 2004.

Øvrelid L., "Empirical evaluations of animacy annotation", *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, 2009.

Prince E. F., "Toward a taxonomy of given-new information", *in* P. Cole (ed.), *Radical pragmatics*, Academic Press, New York, p. 223-255, 1981.

Riester A., Lorenz D., *Richtlinien zur Annotation von Informationsstatus (Gegebenheit) in Projekt A1, SFB 732*. 2009.

Ritz J., Dipper S., Götze M., "Annotation of Information Structure: An Evaluation Across Different Types of Texts", *Language Resources and Evaluation*, Marrakech, Morocco, 2008.

Rooth M., "A theory of focus interpretation", *Natural Language Semantics*, vol. 1, n° 1, p. 75-116, 1992.

Streitberg W., *Die gotische Bibel*, Carl Winter, Heidelberg, 1919.

Tischendorf C. v., *Novum Testamentum Graece*, 8th edn, Hinrichs, Leipzig, 1869–1872.

Zaenen A., Carletta J., Garretson G., Bresnan J., Koontz-Garboden A., Nikitina T., O'Connor M. C., Wasow T., "Animacy Encoding in English: why and how", *in* B. Webber, D. K. Byron (eds.), *ACL 2004 Workshop on Discourse Annotation*, Association for Computational Linguistics, Barcelona, Spain, p. 118-125, July, 2004.

Zôhrapean Y., *Astowacašownč' matean hin ew nor ktakaranac'*, Lazar, Venice, 1805.