Annotating Speech, Attitude and Perception Reports

Corien Bary, Leopold Hess, Kees Thijs,
Radboud University Nijmegen
Department of Philosophy,
Theology and Religious Studies
Nijmegen, the Netherlands
l.hess,k.thijs,c.bary
@ftr.ru.nl

Peter Berck and Iris Hendrickx
Radboud University Nijmegen
Centre for Language & Speech Technology /
Centre for Language Studies,
Nijmegen, the Netherlands
p.berck,i.hendrickx
@let.ru.nl

Abstract

We present REPORTS, an annotation scheme for the annotation of speech, attitude and perception reports. The scheme makes it possible to annotate the various text elements involved in such reports (e.g. embedding entity, complement, complement head) and their relations in a uniform way, which in turn facilitates the automatic extraction of information on, for example, complementation and vocabulary distribution. We also present the Ancient Greek corpus RAG (Thucydides' *History of the Peloponnesian War*), to which we have applied this scheme using the annotation tool BRAT. We discuss some of the issues, both theoretical and practical, that we encountered, show how the corpus helps in answering specific questions about narrative perspective and the relation between report type and complement type, and conclude that REPORTS fitted in well with our needs.

1 Introduction

Both in our daily communication and in narratives we often refer to what other people said, thought and perceived. Take as an example (1) which has a speech, attitude and perception report in the first, second and third sentence, respectively:

(1) John came to Mary, bent on his knees, and asked her 'Will you marry me?' He was afraid that Mary didn't like him enough. He didn't look at her face.

Notice that not only does the type of the report differ (speech, attitude, perception), we also see different kinds of complements: a direct complement *'Will you marry me?'*, an indirect complement *that Mary didn't like him enough* and an NP complement *her face*. (Throughout this paper, by 'reports' we understand reports of speech acts, attitudes and perceptions - i.e., such things that can in principle have *propositional contents*, even if in a given case the report complement is only an NP. *John came to Mary* is not a report in this sense.)

The relation between the report type and the complement type (direct, indirect (further divided into e.g. complementizer + finite verb, participle, infinitive), NP) has been a major topic of research in semantics (Portner, 1992; Verspoor, 1990), syntax (Bresnan, 1970; Haumann, 1997), and language typology (Givón, 1980; Cristofaro, 2003; Cristofaro, 2008) alike.

A corpus annotated for speech, attitude and perception reports is a convenient tool to study this relation since it makes it possible to extract relevant information automatically. For a dead language like Ancient Greek - for which we developed our annotation scheme REPORTS in the first place - such a corpus is even more important, as the research is corpus-based by necessity.

In addition to the linguistic question of understanding the relation between report type and complement type, a corpus annotated for speech, attitude and perception reports is also of great use for questions of a more narratological nature. Narratology is the study of narratives, and one of the big topics here is that of *narrative perspective*, the phenomenon whereby literary texts often present events through the eyes of one of the characters in the story. Such a perspective emerges from the intricate interplay of all kinds of linguistic expressions, with an important role for speech, attitude and perception reports (whose thoughts and perceptions we read and what form they have).

In order to ultimately understand how narrative perspective works, a first step is a corpus annotated for speech, attitude and perception reports. Within the Perspective project,[1] we created such

[1] www.ru.nl/ncs/perspective

Proceedings of the 11th Linguistic Annotation Workshop, pages 46–56,
Valencia, Spain, April 3, 2017.

a corpus for Ancient Greek, RAG (Reports in Ancient Greek). Ancient Greek authors often create shifts to the perspectives of characters. Thucydides, for example, whose *History of the Peloponnesian War* (books 6 and 7) is the first corpus we annotated, was already in ancient times famous for this (Plutarch, *De Gloria* 3). How these authors achieved these perspective shifts is however an unsolved puzzle. We aim to shed light on these issues using the RAG corpus. Both the annotation guidelines and the annotated corpus are publicly available.[2]

RAG makes it possible to extract certain information about reports automatically, which will contribute to answering questions at both the linguistic and the narratological side. Although we developed the annotation scheme primarily with Ancient Greek in mind, we expect that it can be used for corpora in many other languages as well (see the evaluation below). A corpus annotated according to this scheme makes it for example easy to see which report types occur with which complement types. Here we distinguish not only between reports of speech, attitude and perception, but also annotate certain further subdivisions such as that between normal and manipulative speech reports (as in English *tell that* vs. *tell to*) which can be expected to be relevant.

In addition to extracting information about the combinations of report types and complement types, RAG (and other corpora that use the scheme) also makes it possible to search for certain words in report complements only. An interesting class here is for example that of attitudinal particles such as Ancient Greek δή, μήν, and που. These small words express the attitude of the actual speaker towards his utterance (e.g. (un)certainty, confirming expectations, countering the assumptions of the addressee (Thijs, 2017)). In reports, however, they can also be anchored to the reported speaker. Both in light of a better understanding of these notoriously elusive words themselves (e.g. which layer of meaning they target), and in light of their role in creating a certain narrative perspective (whose point of view they express) (Eckardt, 2012), the behavior of particles in reports deserves special attention (Döring, 2013), the study of which is strongly facilitated by a corpus annotated for reports.

In parallel with RAG, we developed an Ancient Greek lemmatizer, GLEM, and POS tagger. This combination increases the possibilities for data extraction considerably, as we will see. An interesting application of the lemmatizer for the narratological question lies in determining the vocabulary distribution of a certain text. Are there for example significant differences between the words the narrator uses when speaking for himself and those in the complements of other people's speeches, attitudes and perceptions (and how does this differ for the different kinds of reports and complements)? The lemmatizer makes it possible to extract this information at the level of the lemma rather than the individual word form. If we apply the scheme REPORTS to other authors, we can also study differences between authors in this respect, for example, whether Herodotus has a stronger distinction between vocabularies, while in Thucydides this is more blurred. This could then explain why it is especially Thucydides that stands out as the author who creates especially sophisticated narrative effects.

A characteristic feature of Ancient Greek speech reports is that they are often quite long. Even indirect reports seem to easily extend to several sentences (rather than just clauses). RAG is also useful for a better linguistic understanding of these constructions. We can for example search for clitic words that are taken to come exclusively at the peninitial position within a sentence, e.g. connective particles such as γάρ (Goldstein, 2016), to see whether it is indeed justified to speak about 'complements' consisting of more than one sentence (and hence, in the case of infinitive complements, of sentences without a finite verb!).

In this paper we discuss related annotation work (section 2), and describe the annotation tool BRAT which we used (section 3) and our annotation scheme REPORTS (section 4). In section 4 we also discuss some choices we made regarding the implementation of REPORTS in our corpus RAG. The corpus is further described in section 5, where we also discuss the application of the lemmatizer and POS tagger and present the results of a small experiment testing inter-annotator agreement. We evaluate BRAT and REPORTS in section 6, including a discussion of the extendability of REPORTS to other languages. Section 7 concludes with final remarks.

[2] https://github.com/GreekPerspective

2 Related work

Previous attempts at corpus annotation for related topics include the annotation of committed belief for English (Diab et al., 2009) and the annotation of direct and indirect speech in Portuguese (Freitas et al., 2016). Our project differs from the former in its focus on complementation (rather than information retrieval) and from the latter in its broader scope (reports in general rather than only speech).

Also related are the annotation schemes for modality such as (McShane et al., 2004; Hendrickx et al., 2012). These schemes aim to grasp the attitude of the actual speaker towards the proposition and label such attitudes as for example belief or obligation. In contrast to modality annotation, which focuses on the attitude of the actual speaker, we are interested in speech, attitude and perception acriptions in general, including ascriptions to other people than the actual speaker. Another difference is our focus on the linguistic constructions used. In that respect our scheme also differs from (Wiebe et al., 2005), which, like RAG, annotates what we call reports, but without differentiating between e.g. different kinds of complements.

3 BRAT rapid annotation tool

BRAT is an open source web-based tool for text annotation (Stenetorp et al., 2012)[3] and is an extension of *stav*, a visualization tool that was designed initially for complex semantic annotations for information extraction in the bio-medical domain including entities, events and their relations (Ohta et al., 2012; Neves et al., 2012). BRAT has been used in many different linguistic annotation projects that require complex annotation such as ellipsis (Anand and McCloskey, 2015), co-reference resolution (Kilicoglu and Demner-Fushman, 2016), and syntactic chunks (Savkov et al., 2016).

As BRAT uses a server-based web interface, annotators can access it in a web browser on their own computer without the need for further installation of software. All annotations are conveniently stored on the server.

We considered several other possible annotation tools for our project, such as MMAX2 (Müller and Strube, 2006), GATE Teamware (Bontcheva et al., 2013) and Arethusa[4]. The main reasons for selecting BRAT as tool for the implementation of our annotation scheme were its web interface and its flexibility: BRAT accommodates the annotation of discontinuous spans as one entity and supports different types of relations and attributes.

Furthermore, BRAT offers a simple search interface and contains a tool for comparison of different versions of annotations on the same source text. BRAT also includes conversion scripts to convert several input formats such as the CoNNL shared task format, MALT XML[5] for parsing and the BIO format (Ramshaw and Marcus, 1995).

BRAT stores the annotation in a rather simple plain text standoff format that is merely a list of character spans and their assigned labels and relations, but that can easily be converted to other formats for further exploitation or search. We plan to port the annotated corpus to the ANNIS search tool (Krause and Zeldes, 2016) in a later stage to carry out more complex search queries.

4 REPORTS: an annotation scheme for speech, attitude and perception reports

4.1 The scheme

The annotation scheme REPORTS consists of entities, events and attributes of both.

Entities are (possibly discontinuous) spans of text. Let's start with two central ones, the **attitude/ speech/ perception embedding entity**, like *confessed* in (2), and the report **complement**, here *that he was in love*.

(2) John confessed that he was in love.

The attitude/speech/perception embedding entity is most typically a verb form, as in (2), but may also be a noun phrase (e.g. *the hope that*).[6] The embedding entity and the complement stand in the two-place relation **report**, which we implemented as an event in BRAT.

Because this complement is internally complex in some cases – consisting of a series of connected complement clauses – we use as a third entity the **complement chunk**. Chunks are all of the individual complement clauses that are syntactically dependent upon one and the same embedding entity. In (3) we have one complement *that he was in love and had not slept for three nights*, which

[3] http://brat.nlplab.org

[4] http://www.perseids.org/tools/arethusa/app/#/

[5] https://stp.lingfil.uu.se/~nivre/research/treebank.xsd.txt

[6] Hence the term *embedding entity*, rather than just *verb*.

consists of two chunks *that he was in love* and *and had not slept for three nights*:

(3) John confessed that he was in love and had not slept for three nights.

The complement chunks stand in the **chunk-of** relation to the complement they are part of. Complement chunks have a **head**, the final entity we annotate. Heads are always verbs. It is the verb that is directly dependent on the embedding entity and can be either a finite verb, an infinitive or a participle, depending on the specific subordinating construction used. In (3) the heads are *was* and *had*. As one would expect they stand in the **head-of** relation to the chunk. Table 1 lists all the entities and events.

The table also shows the attributes assigned within each class. The attributes of the embedding entities concern its semantic type. Within the class of speech report we distinguish (i) normal speech, involving neutral expressions such as *say*, *answer*, *report*; (ii) manner of speech, which are restricted to entities that refer to the physical properties of the speech act (e.g. *scream*, *cry*, *whisper*); (iii) manipulative speech, which is reserved for speech entities that are meant to lead to future actions of the addressee, such as *order/persuade/beg someone to*. The attitude embedding entities (which cover a broadly construed range of propositional attitudes) are further subdivided into (i) knowledge (e.g. *know*, *understand that*), (ii) belief (e.g. *think*, *believe*, *assume that*), (iii) voluntative (e.g. *want*, *intend*, *hope*, *fear to*) and (iv) other (mostly emotional attitudes such as *be ashamed*, *be grieved*, *rejoice*). Entities of perception (e.g. *see*, *hear*) do not have a further subdivision.

The complement type is also specified by means of an attribute. Here, there are five options: (i) direct, (ii) indirect, (iii) mixed, (iv) NP and (v) preposed NP. The mixed category is used for those cases where a combination of direct and indirect speech is used – embedding constructions in Ancient Greek sometimes shift or slip from one construction into the other (Maier, 2015). The NP-category covers instances of complements which do not have a propositional (clausal) character, but only consist of an NP-object. An English example would be *he expects a Spartan victory*.

The category of preposed NPs is typical of Ancient Greek. In case of finite complement clauses, a constituent that semantically belongs to this complement is sometimes placed in a position preceding the complementizer, i.e. syntactically outside of the complement clause. This happens for reasons of information structure – Ancient Greek is a discourse-configurational language, in which word order is determined mainly by information-structural concepts like topic and focus (Allan, 2014; Goldstein, 2016). It may even happen that this constituent is syntactically marked as a main clause argument – this phenomenon is called prolepsis in the literature (Panhuis, 1984). As a whole, constructions like these are annotated as containing two complements – a preposed NP and an indirect one – as well as two report relations.

Let's consider some Ancient Greek examples from RAG.

(4) οἱ δὲ ἄλλοι ἐψηφίσαντό
the.NOM PRT others.NOM vote.PST.3PL
τε ξυμμαχίαν τοῖς Ἀθηναίοις
PRT alliance.ACC the.DAT Athenians.DAT
καὶ τό ἄλλο στράτευμα
and the.ACC other.ACC army.ACC
ἐκέλευον ἐκ Ῥηγίου κομίζειν
invite.PST.3PL from Rhegium.GEN fetch.INF
'the others voted for an alliance with the Athenians and invited them to fetch the rest of their army from Rhegium.' (Thuc. 6.51.2)

Figure 1 shows the visualization of (4) (with some context) as it is annotated in BRAT. Here, we have a manipulative speech verb (ἐκέλευον) that governs a discontinuous infinitival complement that consists of one chunk (τὸ ἄλλο στράτευμα ... ἐκ Ῥηγίου κομίζειν); its head is the infinitive κομίζειν.

Our second example is more complicated. The annotations in BRAT are shown in Figure 2, again with some context.

(5) [A ship went from Sicily to the Peloponnesus with ambassadors,]

οἵπερ τά τε σφέτερα
who.REL.NOM the.ACC PRT own.affairs.ACC
φράσουσιν ὅτι ἐν ἐλπίσιν εἰσὶ
tell.FUT.3PL that in hopes.DAT be.PRS.3PL
καὶ τὸν ἐκεῖ πόλεμον ἔτι μᾶλλον
and the.ACC there war.ACC even more
ἐποτρυνοῦσι γίγνεσθαι
incite.PRS.3PL become.INF
'who should tell that their own affairs were hopeful, and should incite [the Peloponnesians] to prosecute the war there even more actively.' (Thuc. 7.25.1)

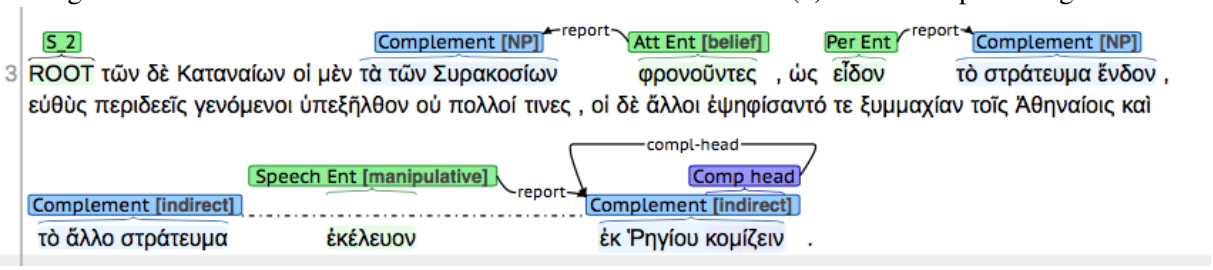
Entities		
embedding entity[a]	speech	normal
		manner of speech
		manipulative
	attitude	knowledge
		belief
		voluntative
		other
	perception	
complement	direct	
	indirect	
	mixed	
	noun phrase	
	preposed noun phrase	
complement chunk		
head of complement chunk	finite, not optative	
	finite, optative	
	infinitive	
	participle	

Events (relations)
report
chunk-of
head-of

Table 1: RAG's entities, events and their attributes

[a] In the implementation in BRAT there actually is no such entity as an underspecified embedding entity, instead we go straight to the speech, attitude and perception embedded entities. The reason is that BRAT does not allow attributes of attributes, which we would otherwise need for the attributes normal etc.

Figure 1: visualization of annotation in BRAT of the sentence in (4) with some preceding context

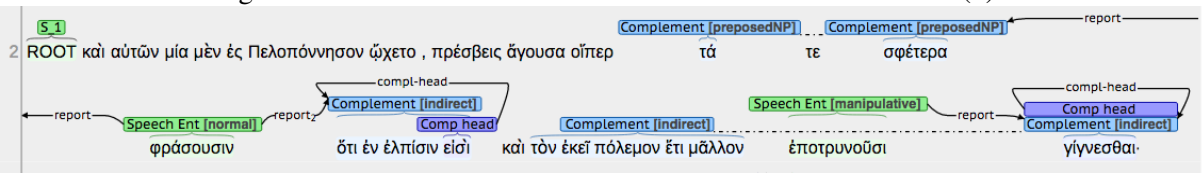


Here τά σφέτερα is a preposed NP, the complement clause being marked by the complementizer ὅτι 'that'. In the second part of the example, however, we find an infinitive construction instead of a finite clause with a complementizer and the whole complement clause is annotated as one (discontinuous) complement span again, like in (4).

4.2 Choices that we made

This basic scheme can be felicitously used for a great deal of the actual data in our corpus, but we also encountered on the one hand practical and on the other hand more complex issues that asked for additional annotation rules. The issues were discussed in the test phase and the rules were spelled out in an elaborate annotation manual. Some examples of the choices we made are the following.

Figure 2: visualization of annotation in BRAT of the sentence in (5)



NPs We only made annotations when a complement is explicitly present. In other words, speech or attitude verbs used in an absolute sense (*he spoke for a long time*) are left out. We did include, however, NP-complements that have a prepositional form, as in περὶ τῆς ἀρχῆς εἰπεῖν (*to speak about the empire*) or ἐς Συρακοσίους δέος (*fear of the Syracusans*). With regard to NP-complements in general, we excluded instances of indefinite and demonstrative pronominal NP-objects (e.g. *he expects something/this*), since they are not interesting for our present research goals due to their lack of meaningful content.

chunks and heads As follows from the definition of a chunk as a subordinated clause, we did not annotate chunks in the case of NP and direct complements (nor heads, since a head is always head of a chunk).

attributes of the head We did not make manual annotation for the attributes of the complement head, i.e. whether it is an indicative, optative, infinitive or participle form. Instead, we used the output from the independently-trained POS tagger (see section 5).

UID As mentioned in the introduction, Ancient Greek reports, even indirect ones, can be very long. Quite frequently we find what is called Unembedded Indirect Discourse (Bary and Maier, 2014), as in (6).

(6) [A general sends messengers to his allies,]

ὅπως μὴ διαφρήσωσι
in.order.that not let.through.SBJV.3PL
τοὺς πολεμίους ἀλλὰ
the.ACC enemies.ACC but
ξυστραφέντες κωλύσωσι
combine.PTCP.PASS prevent.SBJV.3PL
διελθεῖν. ἄλλῃ γὰρ αὐτοὺς οὐδὲ
pass.INF elsewhere for them.ACC not.even
πειράσειν.
try.INF.FUT
'in order that they would not let the enemies through, but would combine themselves and prevent them from passing; for [he said] elsewhere they would not even attempt it.'
(Thuc. 7.32.1)

UID has a form that is usually associated with a dependent construction (infinitive or the Ancient Greek reportative mood called the optative), but in cases like the second sentence in (6) there is no embedding verb it is syntactically dependent on. As the clause with the infinitive or optative expresses the content of the report, we do annotate it as a complement (although the term complement may be misleading in this case).

parenthetical reports We made a different choice in the case of parenthetical report constructions, *Xerxes builds a bridge, as it is said* (ὡς λέγεται in Greek). Although here we do annotate the parenthetical verbs (since they have an important narrative function – the narrator attributes a thought or story to someone other than himself), we do not annotate the main clause *Xerxes builds a bridge* as a complement because there is no report morphology (infinitive or optative). In such cases the boundaries of the complement are also often very vague. Thus, while UID is annotated as a report complement without an embedding entity (or report relation), a parenthetical verb is annotated as an embedding entity without a complement.

defaults for ambiguous cases Some of the embedding entities have multiple meanings, which belong to different semantic categories in our classification of attributes. In some of these cases the choice of the attribute depends on the construction used for the complement clause. Just as English *tell*, mentioned in the introduction, εἶπον with a bare infinitive means *tell someone to* and is classified as a manipulative speech verb, whereas εἶπον with a subordinated *that*-clause means *say that* and belongs to the category of normal speech. In case of speech verbs governing an accusative constituent and an infinitive, however, there may still be an ambiguity in interpretation between the so-called Accusative plus Infinitive-construction (*He told that Xerxes builds a bridge*), where the ac-

cusative constituent functions exclusively as the subject of the complement infinitive clause, and a construction with an accusative object and a bare infinitive (*He told Xerxes to build a bridge*) (cf. (Rijksbaron, 2002)). Usually a decision can be easily made by looking at the surrounding context (as is the case in (4) above).

In other cases, the semantics of embedding entities is truly ambiguous between two categories, irrespective of the complement construction. Perception verbs like *see*, for instance, can easily mean *understand* or *know*. For verbs like these, we have made default classification rules such as the following: 'a perception entity is annotated as such by default; only if an interpretation of direct physical perception is not possible in the given discourse context it is annotated as an attitude knowledge entity.' Moreover, a list was made of all embedding entities and their (default) classification.

personal passive report constructions If the embedding entity is a passive verb of speech or thought, as in *Xerxes is said to build a bridge*, its subject is coreferential with the subject of the complement clause. (This is the so-called Nominative plus Infinitive construction (Rijksbaron, 2002)). What is reported here, of course, is the fact that Xerxes builds a bridge. However, we have decided not to include the subject constituent within the annotated complement in these cases, mainly to warrant consistency with other constructions with coreferential subjects for which it is more natural to exclude the subject from the complement (as in *Xerxes promised to build a bridge/that he would build a bridge*). There is a similar rule for constructions like δοκεῖ μοι 'X seems to me' and φαίνομαι'to appear'.

complement boundaries In complex, multi-clause report complements, which are not rare in Ancient Greek, it is sometimes difficult to tell which parts actually belong to the report and which are interjections by the reporting speaker. As a default rule, we only treat material within the span of a report complement as an interjection (i.e. not annotate it as part of the complement) if it is a syntactically independent clause. Thus, for instance, relative clauses in non-final positions always belong to the span of the complement.

These and similar choices that we made in the progress of fine-tuning our annotation were motivated primarily by practical considerations, but they already led to a better conceptualization of some substantial questions, such as complement boundaries or relevant kinds of syntactic and semantic ambiguities.

5 RAG: a Greek corpus annotated for reports

So far we have annotated Thucydides' *History of the Peloponnesian War*, books 6 and 7, which consists of 807 sentences and 30891 words. In addition to Thucydides, we are also currently working on Herodotus' *Histories*.

The Thucydides digital text is provided by the Perseus Digital Library (Crane, 2016). As it was in betacode we converted it into unicode (utf8) using a converter created by the Classical Language Toolkit (Johnson and others, 2016).[7]

As mentioned before, we combine the manual reports annotation with automated POS-tagging and lemmatization (Bary et al., 2017), which we developed independently and which is open source.

The POS tagger made life easier for the annotator. We only annotated what is the head of the complement chunk and let it to the POS tagger to decide automatically whether this head is e.g. an infinitive, participle or finite verb and if finite, whether it has indicative mood or for example the reportative optative mood.

The lemmatizer enables us to discover whether a specific verb (e.g. all forms of λέγω 'to say') occurs with, say, a complement which contains the particle μήν or a complement with an oblique optative, without having to specify the (first, second, third person etc) forms of the verb manually.

For Herodotus, we can also adapt the manual annotations (including syntactic dependencies) made in the PROIEL project (Haug and Jøhndal, 2008; Haug et al., 2009),[8] whose text we use.

All of the corpus has been annotated by two annotators (PhD students with MA in Classics) working independently. An inventory of differences has been made for every chapter by a student assistant (partly extracted from BRAT automatically using the built-in comparison tool). All the errors and differences were then reviewed by

[7] https://github.com/cltk/cltk/blob/master/cltk/corpus/greek/beta_to_unicode.py

[8] http://www.hf.uio.no/ifikk/english/research/projects/proiel/

the annotators (the most difficult issues were discussed in project meetings) to arrive at a common and final version. Most often differences between annotators concerned two types of issues, where clear-cut criteria are impossible to define: categorization of embedding verbs and syntactic structure ambiguities. The former issue involved verbs which could, depending on interpretation, be annotated with two or more different attributes. For example, ἐλπίζω may be a 'voluntative' verb ('to hope') or a 'belief' verb ('to expect'), (cf. discussion of εἶπον above); some verbs are ambiguous between factive ('knowledge' attitude entity category) and non-factive ('belief' category) senses etc. Even with the use of the more specific rules in the manual, different readings were often possible. The latter issue involved many kinds of ambiguities, most typically concerning relation between the complement clause and other subordinate and coordinate clauses. For example, a final relative clause whose antecedent is within the scope of the complement may, depending on interpretation, belong to the complement as well (its content is part of the content of the reported speech act or attitude) or be an external comment. (Purpose and conditional clauses give rise to similar issues.)

A small selection of the results are listed in Table 2. Here we see for example that γάρ, which is taken to come exclusively at the second position within a sentence, quite frequently occurs within a non-direct complement, suggesting that in these cases we have to do with main clauses rather than dependent clauses. Likewise we can easily search for the particle δή within complements to investigate whose perspective it expresses.

Inter annotator agreement

We performed a small experiment to measure the inter annotator agreement for labeling the main labels in this annotation task. We compared the span annotations of the following sample: book one of Thucydides, chapters 138-146, which contain 1932 words and 56 sentences. We counted the main labels (complement, complement chunk, head of chunk, attitude, speech and perception entities). We wielded a strict form of agreement: both the span length and span labels had to match to count as agreement. One annotator labeled 192 spans while the other labeled 182 spans leading to an inter annotator agreement of 83.4% mutual F-score (Hripcsak and Rothschild, 2005).

# embedding entities	670
speech	189
attitude	441
perception	40
# complements	702
indirect	543
direct	15
NP	138
preposed NP	19
with speech embedding entities	186
with attitude embedding entities	460
with perception embedding entities	39
unembedded	17
# δή/δὴ in non-direct complements	10
# multisentential complements	9
# γάρ/γὰρ after sentence-boundary in non-direct complements	12
total # words in complements[a]	17.836
average # of words per complement	25.41
indirect	14.25
direct	630.60
NP	4.09
preposed NP	3.89

Table 2: Some numbers for RAG

[a] Embedded ones counted twice.

6 Evaluation

In this section we evaluate both the BRAT tool and the REPORTS scheme with respect to their convenience and usefulness.

BRAT is a convenient annotation tool, offering perspicuous visualization and easy to use without any prior training or IT skills (although such skills are needed, of course, to set up an annotation scheme in BRAT). It does not even require typing any commands - after selecting a span of text, a window opens from which the annotation can be chosen with a click of the mouse. However, it has its limitations. The following remarks can be seen as suggestions for future versions or extensions of the program.

For example, with complex annotations involving multiple entities and relations (where often one report is embedded in another) the visualization ceases to be easily legible. In this respect, it seems that an annotation scheme of the complexity of

REPORTS reaches the limits of BRAT's usefulness. Also, since it is currently impossible to assign attributes to attributes, we could not have speech, attitude and perception as attributes of the category embedding entities (see the footnote in Table 1). As a result we need to query the conjunction of speech, attitude and perception entities if we want to draw conclusions about this class in general.

Deleting and correcting complex annotations is not straightforward. Crossing and overlapping spans frequently give rise to errors, which are then impossible to repair from the level of BRAT's interface and require manual access to source files.

A useful function would be that of creating different annotation layers that could be switched on and off in the visual representation - which is possible in e.g. MMAX2 (Müller and Strube, 2006) and would be helpful in this project to use for the annotations of POS and lemma information.

Finally, it would have been convenient if it had been possible to formulate default features, such as the attribute 'normal' in the case of speech embedding entities.

As for the annotation scheme itself, it involves a relatively small number of entities, relations and attributes, but its application is not straightforward and it necessitated the creation of additional documents (described above): a manual containing explicit rules for annotation and a list of embedding verbs in the different categories. Both documents have been extended and amended in the course of work on the annotation. Annotators also required time to get accustomed to the scheme. Nonetheless, after the initial period it was possible to achieve a good level of inter-annotator agreement, as shown by the experiment mentioned above.

The annotation scheme is easily extendable to other languages which share the same typology of complements (direct vs. indirect vs. NP) in speech, attitude and perception reports (that includes at least all major European languages). The categorization of embedding verbs should be universally applicable. Some simplifications are possible in many languages, e.g. removing the category of preposed NP complements or the additional layer of complement chunks (which may not be as useful for many languages as it is for Ancient Greek, where complements often contain several clauses of different types). For modern literary languages it would probably be necessary to create a category for Free Indirect Discourse (but perhaps this would not require more than adding a new attribute of complement entities - the scheme already supports unembedded reports). More substantial changes would be needed for languages which have different typologies of reports (e.g. with no strict distinction between direct and indirect reports) or use other constructions besides embedding verbs to convey reports (e.g. evidential morphemes).

7 Conclusion

BRAT, despite some limitations, is a useful annotation tool that made it possible to implement an annotation scheme which covers all the categories and distinctions that we had wanted to include.

Our annotation scheme REPORTS serves its purpose well, as it makes it possible to easily extract from the corpus information that is relevant to a variety of research questions, concerning e.g. relations between semantics of embedding entities and syntax of complement clauses, factive presuppositions, distribution of vocabulary (including special subsets such as discourse particles, evaluative expressions, deictic elements) in different types of report complements and outside of them, narrative perspective and focalization etc. Both the corpus and the annotation scheme, which are made publicly available, can therefore be a valuable resource for both linguists and literary scholars.

Acknowledgments

This research is supported by the EU under FP7, ERC Starting Grant 338421-Perspective. We thank Anke Kersten, Celine Teeuwen and Delano van Luik for their help.

References

Rutger Allan. 2014. Changing the topic: Topic position in Ancient Greek word order. *Mnemosyne*, 67(2):181–213.

Pranav Anand and Jim McCloskey. 2015. Annotating the implicit content of sluices. In *The 9th Linguistic Annotation Workshop held in conjuncion with NAACL 2015*, pages 178–187.

Corien Bary and Emar Maier. 2014. Unembedded Indirect Discourse. *Proceedings of Sinn und Bedeutung*, 18:77–94.

Corien Bary, Peter Berck, and Iris Hendrickx. 2017. A memory-based lemmatizer for Ancient Greek. Manuscript.

Kalina Bontcheva, Hamish Cunningham, Ian Roberts, Angus Roberts, Valentin Tablan, Niraj Aswani, and Genevieve Gorrell. 2013. Gate teamware: a web-based, collaborative text annotation framework. *Language Resources and Evaluation*, 47(4):1007–1029.

Joan W. Bresnan. 1970. *On Complementizers: Toward a Syntactic Theory of Complement Types*. Springer, Dordrecht.

Gregory R. Crane. 2016. Perseus Digital Library. `http://www.perseus.tufts.edu`. [Online; accessed Dec 16, 2016].

Sonia Cristofaro. 2003. *Subordination*. Oxford University Press, Oxford.

Sonia Cristofaro. 2008. A constructionist approach to complementation: Evidence from Ancient Greek. *Linguistics*, 46(3):571–606.

Mona T. Diab, Lori S. Levin, Teruko Mitamura, Owen Rambow, Vinodkumar Prabhakaran, and Weiwei Guo. 2009. Committed Belief Annotation and Tagging. In *Third Linguistic Annotation Workshop*, pages 68–73, Singapore. The Association for Computer Linguistics.

Sophia Döring. 2013. Modal particles and context shift. In *Beyond expressives: Explorations in use-conditional meaning*, pages 95–123, Leiden. Brill.

Regine Eckardt. 2012. Particles as speaker indexicals in Free Indirect Discourse. *Sprache und Datenverarbeitung*, 36(1):1–21.

Cláudia Freitas, Bianca Freitas, and Diana Santos. 2016. Quemdisse? Reported speech in Portuguese. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 4410–4416, Paris. European Language Resources Association (ELRA).

Talmy Givón. 1980. The binding hierarchy and the typology of complements. *Studies in Language*, 4(3):333–377.

David Goldstein. 2016. *Classical Greek Syntax: Wackernagel's Law in Herodotus*. Brill, Leiden.

Dag Haug and Marius Jøhndal. 2008. Creating a parallel treebank of the old Indo-European Bible translations. In *Proceedings of the Language Technology for Cultural Heritage Data Workshop (LaTeCH 2008), Marrakech, Morocco, 1st June 2008*, pages 27–34.

Dag Haug, Marius Jøhndal, Hanne Eckhoff, Eirik Welo, Mari Hertzenberg, and Angelika Müth. 2009. Computational and linguistic issues in designing a syntactically annotated parallel corpus of Indo-European languages. *Traitement Automatique des Langues (TAL)*, 50(2):17–45.

Dagmar Haumann. 1997. *The Syntax of Subordination*. Max Niemeyer, Tübingen.

Iris Hendrickx, Amália Mendes, and Silvia Mencarelli. 2012. Modality in Text: a Proposal for Corpus Annotation. In *LREC'2012 – Eighth International Conference on Language Resources and Evaluation*, pages 1805–1812, Istanbul, Turkey, May. European Language Resources Association (ELRA).

George Hripcsak and Adam S. Rothschild. 2005. Agreement, the f-measure, and reliability in information retrieval. *Journal of the American Medical Informatics Association (JAMIA)*, 12(3):296–298.

Kyle P. Johnson et al. 2016. CLTK: The Classical Languages Toolkit. `https://github.com/cltk/cltk`. [Online; accessed Nov 12, 2016].

Halil Kilicoglu and Dina Demner-Fushman. 2016. Bio-SCoRes: A Smorgasbord Architecture for Coreference Resolution in Biomedical Text. *PLoS ONE*, 11(3).

Thomas Krause and Amir Zeldes. 2016. Annis3: A new architecture for generic corpus query and visualization. *Digital Scholarship in the Humanities*, 31(1):118–139.

Emar Maier. 2015. Reported speech in the transition from orality to literacy. *Glotta: Zeitschrift für griechische und lateinische Sprache*, 91E(1):152–170.

Marjorie McShane, Sergei Nirenburg, and Ron Zacharski. 2004. Mood and modality: out of theory and into the fray. *Natural Language Engineering*, 10(01):57–89.

Christoph Müller and Michael Strube. 2006. Multi-level annotation of linguistic data with MMAX2. In Sabine Braun, Kurt Kohn, and Joybrato Mukherjee, editors, *Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods*, pages 197–214. Peter Lang, Frankfurt a.M.

Mariana Neves, Alexander Damaschun, Andreas Kurtz, and Ulf Leser. 2012. Annotating and evaluating text for stem cell research. In *Third Workshop on Building and Evaluation Resources for Biomedical Text Mining (BioTxtM 2012) at Language Resources and Evaluation (LREC)*.

Tomoko Ohta, Sampo Pyysalo, Jun'ichi Tsujii, and Sophia Ananiadou. 2012. Open-domain anatomical entity mention detection. In *Proceedings of the Workshop on Detecting Structure in Scholarly Discourse*, pages 27–36. Association for Computational Linguistics.

Dirk Panhuis. 1984. Prolepsis in Greek as a discourse strategy. *Glotta: Zeitschrift für griechische und lateinische Sprache*, 62:26–39.

Paul Portner. 1992. Situation theory and the semantics of propositional expressions. Ph.D. Dissertation, University of Massachusetts, Amherst.

L.A. Ramshaw and M.P. Marcus. 1995. Text chunking using transformation-based learning. In *Proceedings of the Third Workshop on Very Large Corpora*, pages 82–94.

Albert Rijksbaron. 2002. *The Syntax and Semantics of the Verb in Classical Greek: An Introduction*. Gieben, Amsterdam.

A. Savkov, J. Carroll, R. Koeling, and J. Cassell. 2016. Annotating patient clinical records with syntactic chunks and named entities: the Harvey Corpus. *Language Resources and Evaluation*, 50:523–548.

Pontus Stenetorp, Sampo Pyysalo, Goran Topic, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. BRAT: a web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107.

Kees Thijs. 2017. The Attic particle μήν: intersubjectivity, contrast and polysemy. *Journal of Greek Linguistics*.

Marjolijn Verspoor. 1990. Semantic criteria in English complement selection. Ph.D. Dissertation, Rijksuniversiteit Leiden.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language resources and evaluation*, 39(2):165–210.