Proceedings of 1st Workshop on Language Technologies for Historical and Ancient Languages, pages 59–67
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

Automatic Semantic Role Labeling in Ancient Greek Using Distributional Semantic Modeling

Alek Keersmaekers
KU Leuven/Research Foundation - Flanders
Blijde Inkomststraat 21, 3000 Leuven
alek.keersmaekers@kuleuven.be

Abstract

This paper describes a first attempt to automatic semantic role labeling in Ancient Greek, using a supervised machine learning approach. A Random Forest classifier is trained on a small semantically annotated corpus of Ancient Greek, annotated with a large amount of linguistic features, including form of the construction, morphology, part-of-speech, lemmas, animacy, syntax and distributional vectors of Greek words. These vectors turned out to be more important in the model than any other features, likely because they are well suited to handle a low amount of training examples. Overall labeling accuracy was 0.757, with large differences with respect to the specific role that was labeled and with respect to text genre. Some ways to further improve these results include expanding the amount of training examples, improving the quality of the distributional vectors and increasing the consistency of the syntactic annotation.

Keywords: Semantic Role Labeling, Ancient Greek, distributional semantics

1. Introduction

In the last couple of years there has been a large wave of projects aiming to make the large and diachronically diverse corpus of Ancient Greek linguistically searchable. Some large treebanking projects include the Ancient Greek Dependency Treebanks (Bamman, Mambrini, and Crane, 2009), the PROIEL Treebank (Haug and Jøhndal, 2008), the Gorman Trees (Gorman, 2019) and the Pedalion Treebanks (Keersmaekers et al., 2019). Altogether (also including some smaller projects) the Greek treebank material already contains more than 1.3 million tokens – and it is still growing – offering a solid basis for corpus-linguistic research. There have also been recent efforts to automatically annotate an even larger body of text using natural language processing techniques: see Celano (2017) and Vatri and McGillivray (2018) for the literary corpus and Keersmaekers (2019) for the papyrus corpus. However, despite this large amount of *morphologically* and *syntactically* annotated data, *semantic* annotation for Ancient Greek is far more limited. A label such as “ADV” (adverbial) in the Ancient Greek Dependency Treebanks, for instance, refers to a large category of adverbials that do not necessarily have much in common: e.g. expressions of time, manner, place, cause, goal, and so on. While there have been some smaller scale initiatives for semantic role annotation in Greek, these only amount to about 12500 tokens (see section 2). This can be explained by the fact that manual annotation is a time-intensive task. Therefore this paper will present a first attempt to automatic semantic role labeling in Ancient Greek, using a supervised machine learning approach.

This paper is structured as follows: after introducing the data used for this project (section 2), section 3 will describe the methodology. Section 4 will give a detailed overview and analysis of the results, which are summarized in section 5.

2. The data

Devising a definite list of semantic roles for Ancient Greek is not a trivial task. Looking at semantic annotation projects of modern languages, we can also see a wild amount of variation in the number of roles that are annotated, ranging from the 24 roles of *VerbNet* (Kipper Schuler, 2005) to the more than 2500 roles of *FrameNet* (Baker, Fillmore, and Lowe, 1998). Obviously learning 2500 semantic roles is not feasible in a machine learning context (and even the 39 roles in the *Ancient Greek Dependency Treebanks* are a little on the high side considering the amount of training data we have, see below). Therefore I decided to make use of the roles of the *Pedalion* project (Van Hal and Anné, 2017). These are based on semantic roles that are commonly distinguished both in cross-linguistic typological frameworks and in the Greek linguistic tradition (in particular Crespo, Conti, and Maquieira 2003, although their list is more fine-grained). The 29 Pedalion roles I used for this project (see table 1) are a reasonable enough amount to be automatically learned through machine learning, and they are also specifically relevant for Ancient Greek, in the sense that no role of this list is expressed by the exact same set of formal means as any other role: e.g. while both an instrument and a cause can be expressed with the dative in Greek, a cause can also be expressed by the preposition ἕνεκα (*héneka*: “because of”) with the genitive while an instrument cannot.

For this task I limited myself to nouns and other nominalized constructions, prepositional groups and adverbs, depending on a verb. I excluded a number of constructions from the data (on a rule-based basis), either due to a lack of semantic annotation in the data I used (see below) or because they did not express any of the semantic roles listed in table 4 (e.g. appositions): nominatives, vocatives, accusatives when used as an object, infinitive and participial clauses (they are still included when nominalized with an article, see e.g. sentence 1 below), and words with a syntactic relation other than ADV (adverbial), OBJ (complement) or PNOM (predicate nominal).[1] ADV is used for optional modifiers (e.g. “**Yesterday** I gave him a book”), while OBJ is used for obligatory arguments of non-copula verbs (e.g. “Yesterday I gave **him** a book”) and PNOM for obligatory arguments of copula verbs (e.g. “I was **in Rome**”).

[1] While I am planning to include nominatives and accusatives in future versions of the labeler, this was not possible at this moment because none of the projects I included annotated them.

I took semantically annotated data from the following sources:

(1) The **Ancient Greek Dependency Treebanks** (AGDT) (Bamman, Mambrini, and Crane 2009), which has semantic data from the *Bibliotheca* of Pseudo-Apollodorus, Aesop's *Fables* and the Homeric *Hymn to Demeter* (1119 semantically annotated tokens in total).[2] The annotation scheme is described in Celano and Crane (2015): since it was more fine-grained (39 unique roles) than the one this project uses, some of their categories needed to be reduced (e.g. "relation", "connection", "respect" and "topic" to "respect"). Additionally, there are two other projects that are not included in the AGDT but use the same annotation scheme: a treebank of **Apthonius' *Progymnasmata*** (Yordanova, 2018, 752 tokens in total) and of the **Parian Marble** (Berti, 2016, annotated by Giuseppe G. A. Celano, 61 tokens in total).

(2) The **Harrington Trees** (Harrington, 2018), consisting of *Susanna* from the Old Testament, the first part of Lucian's *True Histories* and the *Life of Aesop* (Vita G): in total 1118 semantically annotated tokens. While their annotation scheme is quite compatible with the Pedalion scheme, their role set is a little smaller (22 unique roles), so I manually checked their data and disambiguated some roles (in particular "extent", "orientation" and "indirect object"). Syntactically its annotation scheme does not make a distinction between obligatory (OBJ) and non-obligatory (ADV) modifiers, so they were also disambiguated manually.

(3) The **Pedalion Treebanks** (Keersmaekers et al., 2019), annotated by a group of people involved at the University of Leuven in the annotation scheme described in this paper (syntactically, they are annotated in the same way as the AGDT). This is the largest amount of data this project uses (9446 semantically annotated tokens, or 76% of the total) and contains a wide range of classical and post-classical authors.

In total this data includes 12496 tokens of 29 roles, as described in table 4 at the end of this paper.

3. Methodology

Next, I used this dataset of 12496 annotated roles as training data for a supervised machine learning system. Traditionally, automated approaches typically make use of formal features such as part-of-speech tags and morphology, syntactic labels, lemmas and sometimes encyclopedic knowledge such as lists of named entities (see e.g. Gildea and Jurafsky, 2002; Màrquez et al., 2008; Palmer, Gildea, and Xue, 2010), essentially excluding semantic information. This seems counter-intuitive, but was necessary at the time due to a lack of good methods to represent lexical semantics computationally. Recently, however, due to the rise of so-called distributional semantic models (or "vector space models") and word embeddings, it has become possible to computationally represent the meaning of a word as a vector, with words that are similar in meaning also having mathematically similar vectors. This methodology has been highly successful for several natural language processing tasks, including semantic role labeling (e.g. Zhou and Xu, 2015; He et al., 2017; Marcheggiani and Titov, 2017).

Therefore one of the crucial features used for this task was a distributional vector of both the verb and the argument that bears the semantic relationship to the verb. The method of computing these distributional vectors is explained in more detail in Keersmaekers and Speelman (to be submitted). In short, they are calculated by computing association values (with the PPMI "positive pointwise mutual information" measure) of a given target lemma with its context elements, based on a large (37 million tokens) automatically parsed corpus of Ancient Greek (see Turney and Pantel, 2010 for a more detailed explanation of this methodology). These context elements are lemmas with which the target lemma has a dependency relationship (either its head or its child).[3] Next, these vectors are smoothed and their dimensionality is reduced by a technique called latent semantic analysis (LSA). This technique (using so-called Singular Value Decomposition) enables us to retrieve vectors with a lower dimensionality, where the individual elements do not directly correspond to individual contexts but the 'latent meaning'[4] contained in several context elements (see Deerwester et al., 1990 for more detail). Experimentally I found that reducing the vector to only 50 latent dimensions was sufficient for this task, with no significant improvements by increasing the number of dimensions.[5]

Apart from the distributional vector of both the verb and its argument, the following additional features were included:

- The form of the construction, subdivided into three features: the preposition (or lack thereof), the case form of its dependent word and a feature that combines both; e.g. for ἀπό+genitive (*apó*: "from") these features would be {ἀπό,genitive,ἀπό+genitive}. Combinations that did occur less than 10 times were set to "OTHER" (179 in total).
- The lemma of both the verb and its argument. For verbs or arguments that occurred less than 50 times, the value of this feature was set to "OTHER". Only 26 argument lemmas and 25 verb lemmas occurred more than 50 times; however, altogether these lemmas account for 34% of all tokens for the arguments and 34% of all tokens for the verbs as well.

[2] While the AGDT treebank is also available in the Universal Dependencies project, I used their original version (in the style of the Prague Dependency Treebank) to ensure compatibility with the other projects included.

[3] This is the *DepHeadChild* model in the Keersmaekers and Speelman (to be submitted) paper.

[4] This "latent meaning" simply refers to the fact that several context features tend to be highly correlated: e.g. a word such as ἐξέρχομαι (*exérkhomai*) and ἀπέρχομαι (*apérkhomai*) "go away" would typically be used with similar nouns. These "latent meanings" can therefore be seen as generalizations over several correlated features.

[5] I used the function *svds* from the *R* package *RSpectra* (Qiu et al., 2019).

- The syntactic relation between verb and argument, which was either “OBJ” (complement), “ADV” (adverbial) or “PNOM” (predicate nominal).
- Animacy data, taken from an animacy lexicon coming from several sources: the PROIEL project (Haug and Jøhndal, 2008) as well as data annotated at the University of Leuven (see Keersmaekers and Speelman, to be submitted). It categorizes nouns into the following groups: *animal*, *concrete object*, *non-concrete object*, *group*, *person*, *place* and *time*. For 5249 (42%) arguments a label from this category could be assigned; the others were set to “unknown”.
- The part-of-speech of the argument to the verb: *adjective, article, demonstrative pronoun, indefinite pronoun, infinitive, interrogative pronoun, noun, numeral, participle, personal pronoun and relative pronoun*.
- Morphological features of the argument and of the verb: *gender* and *number* for the argument and *number, tense, mood* and *voice* for the verb.

I trained a *Random Forest* classifier on this data, using *R* (R Core Team 2019) package *randomForest* (Breiman et al., 2018), building 500 classification trees[6] – this classifier turned out to perform better than any other machine learning model I tested. The results were evaluated using 10-fold cross-validation (i.e. by dividing the data in 10 roughly equally sized parts as test data, and training 10 models on each of the other 9/10 of the data).

4. Results and analysis

Overall labeling accuracy was 0.757, or 9460/12496 roles correctly labeled.[7] However, there were large differences among specific roles, as visualized in table 1. These results are calculated by summing up the errors for each of the 10 test folds.

	Precision	**Recall**	**F1**
agent (364)	0.875	0.712	0.785
beneficiary (715)	0.649	0.691	0.669
cause (753)	0.728	0.681	0.704
companion (424)	0.870	0.682	0.765
comparison (198)	0.882	0.455	0.600
condition (5)	(never used)	0.000	0.000
degree (295)	0.745	0.793	0.768
direction (1006)	0.809	0.874	0.840
duration (221)	0.821	0.665	0.735
experiencer (259)	0.742	0.444	0.556
extent of space (67)	0.917	0.164	0.278
frequency (78)	0.704	0.487	0.576
goal (282)	0.696	0.422	0.525
instrument (507)	0.628	0.673	0.650
intermediary (16)	1.000	0.688	0.815
location (1436)	0.702	0.808	0.752
manner (1596)	0.745	0.809	0.775
material (22)	1.000	0.727	0.842
modality (17)	0.385	0.294	0.333
possessor (127)	0.781	0.701	0.739
property (6)	0.000	0.000	0.000
recipient (1289)	0.879	0.942	0.909
respect (800)	0.708	0.733	0.720
result (15)	0.667	0.133	0.222
source (803)	0.724	0.885	0.797
time (943)	0.805	0.752	0.777
time frame (45)	0.786	0.489	0.603

Table 1: Precision, recall and F1 scores for each semantic role (number of instances between brackets)

In general low recall scores for a specific role can be explained by a lack of training examples: roles that had very little training data such as condition (only 5 instances), property (6 instances) and result (15 instances) expectedly had very low recall scores (0 for condition and property, and 0.133 for result). Figure 1 plots the recall score of each role as a function of the (logarithmically scaled) token frequency of the role in the training data, showing that the amount of training examples is one of the main factors explaining the performance of each role. Figure 2 shows a confusion matrix detailing how often each role (“Reference”) got labeled as another role (“Prediction”).

Next, we can estimate the effect of each variable by testing how well the classifier performs when leaving certain variables out of the model.[8] As can be inferred from table 2, there were only two features that had a substantial effect on the overall model accuracy: the word vectors (-8% accuracy when left out) and the syntactic label (-2.4% accuracy when left out). Lemmas, morphology, animacy and part-of-speech were less essential, as the accuracy decreases less than half a percentage point when either of them (or all of them) is left out. Probably the information that is contained in the lemma, animacy and part-of-speech features is already largely contained in the word vectors,

[6] This is the default setting for the *randomForest* package, but this amount can be decreased to as low as 250 without having a large negative effect on labeling accuracy (0.756, or -0.1%).

[7] While this set of roles is quite fine-grained, a reduction in the number of roles did not have a large effect on accuracy: when I merged some less frequent roles with more frequent ones (‘condition’ to ‘respect’, ‘extent of space’ to ‘location’, ‘frequency’ and ‘time frame’ to ‘time’, ‘intermediary’ and ‘value’ to ‘instrument’, ‘material’ to ‘source’, ‘modality’ to ‘manner’, ‘property’ to ‘possessor’ and ‘result’ to ‘goal’, reducing the amount of roles to 19 from 29), accuracy only increased with 1.1% point (0.768). This is probably because these roles, while semantically quite similar, typically use other formal means in Greek to express them (e.g. ‘time frame’ is typically expressed by the genitive, but ‘time’ by the dative).

[8] I did not test leaving out the three variables indicating the form of the construction since I considered them essential for the classification task. The variable importances calculated by the random forest also indicate that these variables are by far the most important (in the order “combined preposition/case” > ”preposition” > ”case”). While including a feature “combined preposition/case” might seem superfluous, considering that the regression trees are able to model the interaction between them natively, when it is excluded there is a relatively big drop in accuracy, from 0.757 to 0.726 (-3.1%). Presumably due to the low amount of training data and the large feature space, the data often gets partitioned into too small groups during the construction of the tree so that this interaction effect is not modelled (see also Gries, 2019, who argues that adding such combined features in a Random Forest can be beneficiary for regression as well).

while most morphological features are not that important for semantic role labeling. [9]

	Accuracy
Overal accuracy	0.757
Excluding word vectors	0.677 (-8.0%)
Excluding syntactic label	0.734 (-2.3%)
Excluding lemmas	0.759 (+0.2%)
Excluding morphology	0.754 (-0.3%)
Excluding animacy class	0.758 (+0.1%)
Excluding part-of-speech	0.756 (-0.1%)
Excluding lemmas, morphology, animacy class and part-of-speech	0.754 (-0.3%)

Table 2: Accuracy when leaving out certain features

As for part-of-speech differences, interrogative pronouns (accuracy 0.893; however, 3/4 of examples are the form τί *tí* "why"), adverbs (0.822) and personal pronouns (0.807) did particularly well, while relative pronouns (0.528), articles (0.616), numerals (0.629, but only 35 examples) and infinitives (0.667) did rather badly. The results of relative pronouns are not particularly surprising, since they are inherently anaphoric: therefore it would likely be better to model them by the vector of their antecedent (which is directly retrievable from the syntactic tree) rather than the "meaningless" vector of the lemma ὅς (*hós*: "who, which"). As for infinitives, the issue might be that they are modelled with the same vectors as nouns, while their usage is quite different: in sentence (1), for instance, whether the lemma of the infinitive is θολόω (*tholóō*: "disturb") or any other lemma is irrelevant, and the causative meaning is instead inferred from the verb ἐμέμφετο (*emémpheto*: he reproached) combined with the ἐπί + dative (*épi*: "because of") infinitive construction (in the future it might therefore be better to model infinitive arguments with a singular vector generalizing over all occurrences of an infinitive). Similarly, articles are modelled with the vector of the lemma ὁ (*ho*: "the"), which covers all usages of this lemma, while the (dominant) attributive usage is quite different from its pronominal usage (as a verbal argument): therefore restricting the vector of ὁ to pronominal uses might also help performance.

(1) *ἐμέμφετο* ***ἐπὶ*** ***τῷ***
emémpheto ***epí*** ***tõi***
reproach.3SG.IMPF **for** **the.DAT**
τὸν ***ποταμὸν*** ***θολοῦν***
tón ***potamón*** ***tholoũn***
the.ACC **river.ACC** **disturb.INF.PR**
He reproached [him] **for disturbing the river**

Finally, there were some genre differences, as can be seen in table 3.

	Accuracy
Religion	0.838 (932/1112)
Documentary	0.809 (1332/1646)
History	0.765 (1439/1881)
Drama	0.751 (1091/1453)
Narrative	0.751 (2019/2689)
Rhetorical	0.723 (1086/1503)
Philosophy	0.714 (1076/1506)
Epic and lyric poetry	0.687 (485/706)

Table 3: Accuracy per genre

Unsurprisingly, the texts that did well are quite repetitive in nature, have a large amount of training examples and use an everyday, non-abstract language: religious and documentary texts. On the other side of the spectrum are poetic texts, which often express their semantic roles with other formal means than prose texts (which are the majority of the training data), and philosophical and rhetorical texts, which use relatively abstract language (see also below).

Moving towards a more detailed analysis of the results, the following will give a short overview of the specific problems associated with some roles that turned out to be especially problematic. As for **condition**, **property**, **result** and **modality**, which all had recall scores of less than 0.3, there are simply not enough training tokens in the data to make any conclusions about the performance of these roles (5, 6, 15 and 17 respectively). **Intermediary** and **material** did perform relatively well, on the other hand (recall of 0.688 and 0.727), even though they do not have that many training examples either (16 and 22 respectively). However, they are rather uniformly represented in the training data: each example of "intermediary" that was classified correctly was encoded by διά + genitive (*diá:* "through") and had either the verb γράφω (*gráphō*: "write"), κομίζω (*komízō*: "bring") or πέμπω (*pémpō*: "send") with it, while every single example of "material" that was classified correctly was a genitive object of either πίμπλημι (*pímplēmi*) or ἐμπίμπλημι (*empímplēmi*) "fill". Because of this large level of uniformity, their relatively high performance with respect to their token frequency is not particularly surprising.

Extent of space, on the other hand, did quite bad even when its frequency of 67 training examples is taken into account, as can be seen on figure 1. From the confusion matrix in figure 2, we can see that it was, unsurprisingly, most commonly misclassified as "location" (almost half of all cases) and, to a much lower extent, "direction" and "cause". One of the difficulties is that most expressions that can be used to express this role can also express a location: e.g. διά with the genitive (*diá*: "through"), ἐπί with the accusative (*epí* "at, to"), κατά with the accusative (*kata*: "along") and so on (sometimes this role was also misclassified as "location" in the data, which obviously did not help the learning or evaluation process). As an additional difficulty, the lemmas used with this role do not

[9] In the variable importances, *gender* and *number* of the argument of the verb were considered to be the most important, while in particular *person*, *number* and *voice* of the verb ranked lower than any other feature (including any of the 100 vector elements). As for *voice* of the verb, this can probably be explained because I did not label subjects, making the number of roles where this would be a factor relatively limited (mainly "agent" and possibly "experiencer").

substantially differ from the lemmas typically used for the role “location” (e.g. lemmas such as ἀγορά *agorá* “market”, γῆ *gẽ* “land” etc.). Instead it is typically an interaction of the meaning of the verb and the form of the construction that determines that the semantic role should not be “location” but “extent of space”, which is likely too difficult to learn with the limited amount of training examples for this role. Similar problems arise for the roles **time frame** and **frequency**, which are often expressed with the same argument lemmas as “time” and therefore are often confused with this role: however, the degree of confusion is less than with “extent of space”, likely because the formal means to express these roles are quite different from the ones used to express “time” (e.g. time frame is mostly expressed with the genitive, while time is rarely so; frequency uses several adverbs such as πολλάκις *pollákis* “frequently”, δίς *dís* “twice” etc. that can only express this role). More training examples would probably be beneficial in these cases: while **source** and **direction**, for instance, are also often used with the same arguments as “location”, their recall scores are quite high, likely because they have many training examples to learn from (803 and 1006 respectively).

Moving to the more frequent roles, there were three roles in particular that received a wrong classification quite frequently even with a relatively high amount of training examples: **comparison**, **experiencer** and **goal**. As for **comparison**, one problem is that there are a wide range of formal means to express this role: 21 in total, which is on the high side, considering that the median role only has 12 formal means and that there is only an average amount of training examples for this role (198 in total). Another problem is that unlike for roles such as “time” and “location”, the argument of the verb can be almost any lemma (and, when it is used in an adverbial relationship, the verb itself as well): if we look at sentence 2, for instance, neither the verb ἔχω (*ékhō:* “have”) nor the noun ἄνθρωπος (*ánthrōpos*: “human”) is particularly useful to identify the role of ἀντί (*antí*: “instead”): instead ἀντί functions more as a “mediator” between κυνοκέφαλος (*kunoképhalos*: “baboon”) and ἄνθρωπος. Involving not only the verb but also its dependents would help in this case, but since the comparative construction can refer to any element in the sentence this problem is rather complicated (and might be more appropriate to solve at the parsing stage).

(2)			
	τίς	*αὐτὸν*	*θελήσει*
	tís	*autón*	*thelḗsei*
	who.NOM	he.ACC	want.3SG.FUT
	ἀγοράσαι	*καὶ*	*κυνοκέφαλον*
	agorásai	*kaí*	*kunoképhalon*
	buy.INF.AOR	and	baboon.ACC
	ἀντὶ	***ἀνθρώπου***	***ἔχειν;***
	antí	***anthrṓpou***	***ékhein*?**
	instead.of	**human.GEN**	**have.INF.PR**

Who will want to buy him and have a baboon **instead of a human**?

The **experiencer** role is most often confused with the beneficiary/maleficiary role. This happens in particular when this role receives the label ADV “adverbial” (recall 0.173) rather than OBJ “complement” (recall 0.817). In this case both “beneficiary” and “experiencer” refer to a person who is affected in some way by the action of the main verb, and the difference between being advantaged or disadvantaged by an action and being affected by it is often only subtle (and sometimes also inconsistently annotated). In sentence 3, for instance, σοί (*soí* “for you”) has been labeled as an experiencer, but might also be considered a beneficiary: “the rest is according to your wishes **for your benefit**”. In general verbs that denote an action that have clear results (e.g. ποιέω *poiéō* “make”, παρασκευάζω *paraskeuázō* “prepare” etc.) would be more likely to have a beneficiary rather than an experiencer adverbial, but more training data is likely needed to learn this subtle difference.

(3)				
	εἰ	*(...)*	*τὰ*	*λοιπά*
	ei		*tá*	*loipá*
	if		the.ACC.PL	rest.ACC.PL
	σοί	*ἐστιν*	*κατὰ*	*γνώμην,*
	soí	*estin*	*katá*	*gnṓmēn,*
	you.DAT	be.3SG	according	will.ACC
	ἔχοι	*ἂν*	*καλῶς*	
	ékhoi	*án*	*kalõs*	
	have.3SG.PR.OPT	PTC	good	

If (…) the rest is according to your wishes **for you**, it would be good.

Finally, as for **goal**, its large amount of confusion with roles such as “cause” or “respect” is not very surprising, as they are expressed by similar argument lemmas. However, the role is also frequently confused with roles such as “direction” and “location” (to a lesser extent). While the same formal means are often used to express goals and directions (e.g. εἰς/κατά/ἐπί/πρός + accusative), one would expect directions to be used predominantly with concrete objects and goals with non-concrete objects. However, in general non-concrete objects do perform quite badly: their accuracy is only 0.655, as opposed to 0.744 for all nouns in general. This might suggest that these nouns are not that well modelled by their distributional vector (which we also found to some extent in Keersmaekers and Speelman to be submitted), although other explanations (e.g. non-concrete objects typically receiving roles that are harder to model in general) are also possible. Other than that, there was also a large influence of the syntactic label: the recall of goals that had the label ADV was 0.493 while it was only 0.111 for the label OBJ – and 35/48 of the goals that were misclassified as direction had the label “OBJ”: this is consistent with the fact that goals predominantly have the ADV label (80%) while directions predominantly have OBJ (83%), and some of the goals that were classified as OBJ were in fact misclassifications.

5. Conclusion

This paper has described a first approach to automatic semantic role labeling for Ancient Greek, using a Random Forest classifier trained with a diverse range of features.

While the amount of training data was relatively low (only about 12500 tokens for 29 roles), the model was still able to receive a classification accuracy of about 76%. The most helpful features were distributional semantic vectors, created on a large corpus of 37 million tokens, while other features (lemmas, morphology, animacy label, part-of-speech) did not contribute as much. Probably it is exactly this small amount of training samples that explains why these vectors are so important: since there are a large amount of lemmas in the training data (about 2700 argument lemmas and 1900 verb lemmas), the model is able to reduce this variation by assigning similar vectors to semantically similar lemmas. The distinctions that features such as morphology are able to make (e.g. the role *agent* as expressed by *ὑπό hupó* “by” with the genitive is rare with active verbs) might be too subtle, on the other hand, to be statistically picked up by the model with the relatively low training examples we have, and therefore these features would perhaps be more helpful when there is more data to learn from.

An in-depth error analysis reveals a number of ways for further improvement. First of all, the most important step would be expanding the amount of training data, since there is an obvious correlation between the amount of training examples and the performance of each role. Secondly, while the distributional semantic approach works well for most words, some categories (e.g. relative pronouns, infinitives) are not modelled that well and might require a special treatment. Thirdly, non-concrete words turned out to be particularly problematic, and need to be investigated in more detail (particularly by examining if their meaning is modelled well by their semantic vector). Finally, the syntactic relation (adverbial or complement) was also relatively influential in the model, and some wrongly classified instances had in fact received the wrong syntactic label. Therefore improving the syntactic data with regards to this distinction would also likely improve results, especially when moving from manually disambiguated syntactic data (as used in this paper) to automatically parsed data.

The semantic role labeling system used in this paper, as well as the training data on which the system was trained (including all modifications of existing treebanks) is available on GitHub.[10] Hopefully this will encourage corpus annotators to add a semantic layer to their project (since there is already an automatically annotated basis to start from), so that their data can also be integrated in the system and results can be further improved.

6. Abbreviations used in interlinear glosses

ACC	accusative
AOR	aorist
DAT	dative
FUT	future
GEN	genitive
IMPF	imperfect
INF	infinitive
NOM	nominative
OPT	optative
PL	plural
PR	present
PTC	particle
SG	singular

7. Acknowledgements

The present study was supported by the Research Foundation Flanders (FWO) [grant number: 1162017N]. Some preliminary work was carried out by Colin Swaelens (whose master thesis I supervised). I would like to thank my PhD supervisors (Toon Van Hal and Dirk Speelman), as well as three anonymous reviewers for their detailed and helpful feedback.

8. Bibliographical references

Baker, C.F., Fillmore, C.J. and Lowe, J.B. (1998). The Berkeley FrameNet Project. In Proceedings of the 17th International Conference on Computational Linguistics Volume 1, pages 86–90, Montreal, Quebec, Canada, august. Association for Computational Linguistics.

Bamman, D., Mambrini, F. and Crane, G. (2009). An Ownership Model of Annotation: The Ancient Greek Dependency Treebank. In Marco Passarotti, Adam Przepiórkowski, Savina Raynaud, Frank Van Eynde, editors, Proceedings of the Eighth International Workshop on Treebanks and Linguistic Theories (TLT 8), pages 5–15, Milan, Italy, december. EDUCatt.

Berti, M. (2016). The Digital Marmor Parium. Presented at the Epigraphy Edit-a-thon. Editing Chronological and Geographic Data in Ancient Inscriptions, Leipzig.

Breiman, L., Cutler, A., Liaw, A., and Wiener, M. (2018). randomForest: Breiman and Cutler's Random Forests for Classification and Regression. https://CRAN.R-project.org/package=randomForest.

Celano, G.G.A. and Crane, G. (2015). Semantic Role Annotation in the Ancient Greek Dependency Treebank. In Marcus Dickson, Erhard Hinrichs, Agnieszka Patejuk, Adam Przepiórkowski, editors, Proceedings of the Fourteenth International Workshop on Treebanks and Linguistic Theories (TLT14), pages 26–34, Warsaw, Poland, december. Institute of Computer Science, Polish Academy of Sciences.

Crespo, E., Conti, L., and Maquieira, H. (2003). Sintaxis Del Griego Clásico. Madrid: Gredos.

Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., and Harshman, R. (1990). Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science*, 41(6):391–407.

Gildea, D. and Jurafsky, D. (2002). Automatic Labeling of Semantic Roles. *Computational Linguistics*, 28(3):245–288.

Gries, S. (2019). On Classification Trees and Random Forests in Corpus Linguistics: Some Words of Caution and Suggestions for Improvement. *Corpus Linguistics and Linguistic Theory*.

Haug, D.T.T. and Jøhndal, M. (2008). Creating a Parallel Treebank of the Old Indo-European Bible Translations. In Caroline Sporleder and Kiril Ribarov (Conference Chairs), et al., editors, Proceedings of the Second Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2008), pages 27–34, Marrakech,

[10] https://github.com/alekkeersmaekers/PRL

Morocco, may. European Language Resource Association (ELRA).

He, L., Lee, K., Lewis, M., and Zettlemoyer, L. (2017). Deep Semantic Role Labeling: What Works and What's next. In Regina Barzilay, Min-Yen Kan, editors, Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 473–483, Vancouver, Canada, july-august. Association for Computational Linguistics.

Keersmaekers, A. (2019). Creating a Richly Annotated Corpus of Papyrological Greek: The Possibilities of Natural Language Processing Approaches to a Highly Inflected Historical Language. *Digital Scholarship in the Humanities*.

Keersmaekers, A., Mercelis, W., Swaelens, C., and Van Hal, T. (2019). Creating, Enriching and Valorizing Treebanks of Ancient Greek. In Marie Candito, Kilian Evang, Stephan Oepen, Djamé Seddah, editors, Proceedings of the 18th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2019), pages 109–117, Paris, France, august. Association for Computational Linguistics.

Keersmaekers, A. and Speelman, D. (to be submitted). Applying Distributional Semantic Models to a Historical Corpus of a Highly Inflected Language: The Case of Ancient Greek.

Kipper Schuler, K. (2005). VerbNet: A Broad-Coverage, Comprehensive Verb Lexicon. Dissertation in Computer and Information Science. University of Pennsylvania.

Marcheggiani, D. and Titov, I. (2017). Encoding Sentences with Graph Convolutional Networks for Semantic Role Labeling. In Martha Palmer, Rebecca Hwa, Sebastian Riedel, editors, Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 1506–1515, Copenhagen, Denmark, september. Association for Computational Linguistics.

Màrquez, L., Carreras, X., Litkowski, K.C., and Stevenson, S. (2008). Semantic Role Labeling: An Introduction to the Special Issue. *Computational Linguistics*, 34(2):145–159.

Palmer, M., Gildea, D., and Xue, N. (2010). Semantic Role Labeling. Morgan & Claypool.

Qiu, Y., Mei, J., Guennebaud, G., and Niesen, J. (2019). RSpectra: Solvers for Large-Scale Eigenvalue and SVD Problems. https://CRAN.R-project.org/package=RSpectra.

R Core Team. (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing. https://www.R-project.org/

Turney, P.D. and Pantel, P. (2010). From Frequency to Meaning: Vector Space Models of Semantics. *Journal of Artificial Intelligence Research*, 37:141–188.

Van Hal, T. and Anné, Y. (2017). Reconciling the Dynamics of Language with a Grammar Handbook: The Ongoing Pedalion Grammar Project. *Digital Scholarship in the Humanities*, 32(2):448–454.

Vatri, A. and McGillivray, B. (2018). The Diorisis Ancient Greek Corpus. *Research Data Journal for the Humanities and Social Sciences*, 3(1):55–65.

Zhou, Y. and Xu, W. (2015). End-to-End Learning of Semantic Role Labeling Using Recurrent Neural Networks. In Chengqing Zong, Michael Strube, editors, Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 1127–1137, Beijing, China, july. Association for Computational Linguistics.

9. Language Resource References

Giuseppe G. A Celano. (2017). Lemmatized Ancient Greek Texts. 1.2.5. https://github.com/gcelano/LemmatizedAncientGreekXML.

Vanessa Gorman. (2019). Gorman Trees. 1.0.1. https://github.com/vgorman1/Greek-Dependency-Trees.

Matthew J. Harrington. (2018). Perseids Project - Treebanked Commentaries at Tufts University. https://perseids-project.github.io/harrington_trees/.

Polina Yordanova. (2018). Treebank of Aphtonius' Progymnasmata. https://github.com/polinayordanova/Treebank-of-Aphtonius-Progymnasmata.

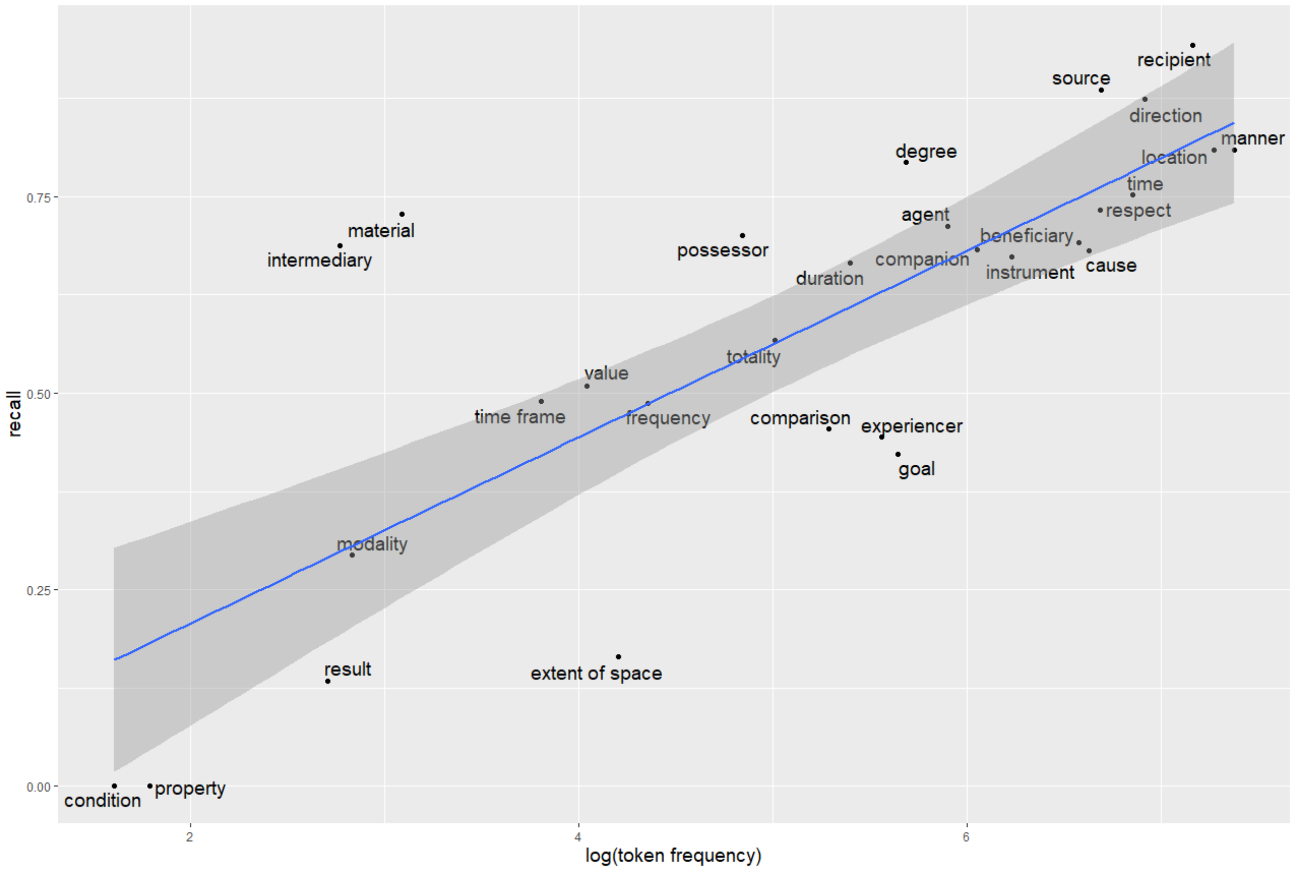
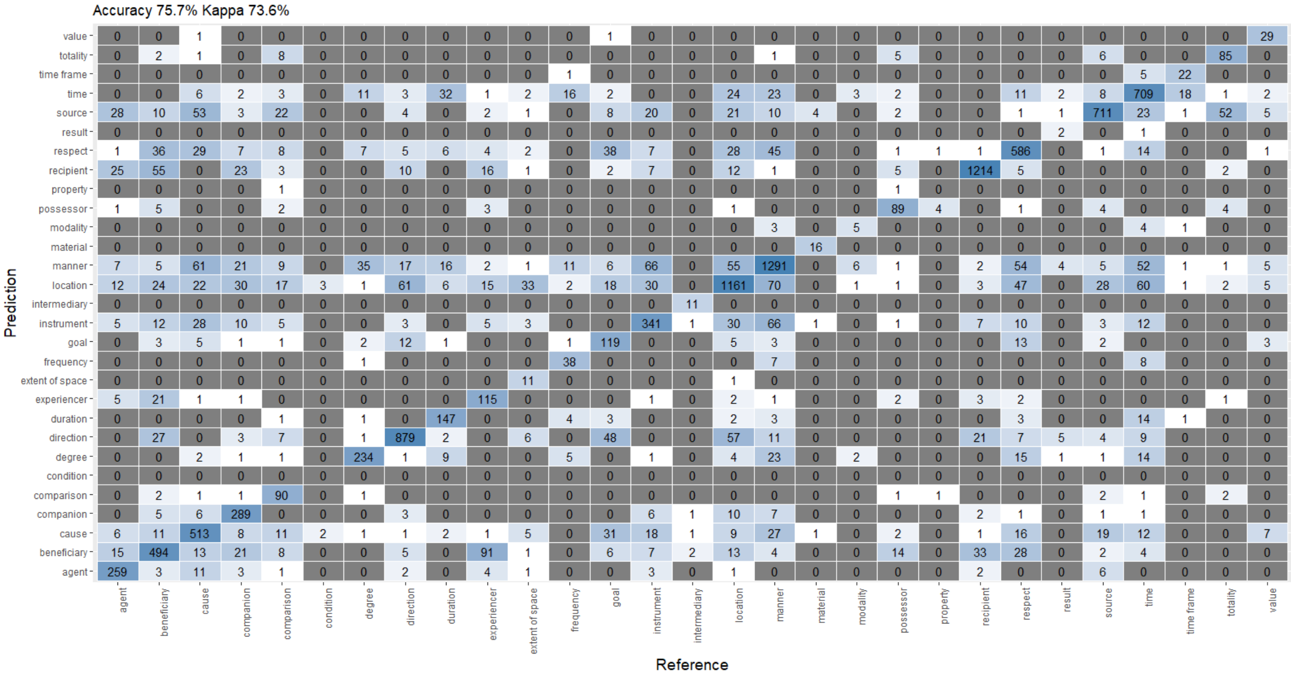


Figure 1: Recall scores for semantic roles as a function of their logarithmically scaled token frequency



Accuracy 75.7% Kappa 73.6%

Prediction

Reference

	agent	beneficiary	cause	companion	comparison	condition	degree	direction	duration	experiencer	extent of space	frequency	goal	instrument	intermediary	location	manner	material	modality	possessor	property	recipient	respect	result	source	time	time frame	totality	value
value	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	29
totality	0	2	1	0	8	0	0	0	0	0	0	0	0	0	0	0	1	0	0	5	0	0	0	0	6	0	0	85	0
time frame	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	5	22	0	0
time	0	0	6	2	3	0	11	3	32	1	2	16	2	0	0	24	23	0	3	2	0	0	11	2	8	709	18	1	2
source	28	10	53	3	22	0	0	4	0	2	1	0	8	20	0	21	10	4	0	2	0	0	1	1	711	23	1	52	5
result	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	1	0	0	0
respect	1	36	29	7	8	0	7	5	6	4	2	0	38	7	0	28	45	0	0	1	1	1	586	0	1	14	0	0	1
recipient	25	55	0	23	3	0	0	10	0	16	1	0	2	7	0	12	1	0	0	5	0	1214	5	0	0	0	0	2	0
property	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
possessor	1	5	0	0	2	0	0	0	0	3	0	0	0	0	0	1	0	0	0	89	4	0	1	0	4	0	0	4	0
modality	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	5	0	0	0	0	0	0	4	1	0	0
material	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	16	0	0	0	0	0	0	0	0	0	0	0
manner	7	5	61	21	9	0	35	17	16	2	1	11	6	66	0	55	1291	0	6	1	0	2	54	4	5	52	1	1	5
location	12	24	22	30	17	3	1	61	6	15	33	2	18	30	0	1161	70	0	1	1	0	3	47	0	28	60	1	2	5
intermediary	0	0	0	0	0	0	0	0	0	0	0	0	0	0	11	0	0	0	0	0	0	0	0	0	0	0	0	0	0
instrument	5	12	28	10	5	0	0	3	0	5	3	0	0	341	1	30	66	1	0	1	0	7	10	0	3	12	0	0	0
goal	0	3	5	1	1	0	2	12	1	0	0	1	119	0	0	5	3	0	0	0	0	0	13	0	2	0	0	0	3
frequency	0	0	0	0	0	0	1	0	0	0	0	38	0	0	0	0	7	0	0	0	0	0	0	0	0	8	0	0	0
extent of space	0	0	0	0	0	0	0	0	0	0	11	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
experiencer	5	21	1	1	0	0	0	0	0	115	0	0	0	1	0	2	1	0	0	2	0	3	2	0	0	0	0	1	0
duration	0	0	0	0	1	0	1	0	147	0	0	4	3	0	0	2	3	0	0	0	0	0	3	0	0	14	1	0	0
direction	0	27	0	3	7	0	1	879	2	0	6	0	48	0	0	57	11	0	0	0	0	21	7	5	4	9	0	0	0
degree	0	0	2	1	1	0	234	1	9	0	0	5	0	1	0	4	23	0	2	0	0	0	15	1	1	14	0	0	0
condition	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
comparison	0	2	1	1	90	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	2	1	0	2	0
companion	0	5	6	289	0	0	0	3	0	0	0	0	0	6	1	10	7	0	0	0	0	2	1	0	1	1	0	0	0
cause	6	11	513	8	11	2	1	1	2	1	5	0	31	18	1	9	27	1	0	2	0	1	16	0	19	12	0	0	7
beneficiary	15	494	13	21	8	0	0	5	0	91	1	0	6	7	2	13	4	0	0	14	0	33	28	0	2	4	0	0	0
agent	259	3	11	3	1	0	0	2	0	4	1	0	0	3	0	1	0	0	0	0	0	2	0	0	6	0	0	0	0

Figure 2: Confusion matrix of semantic roles

Role	**Example**
Agent (364 instances)	δύο δὲ παῖδες **ὑπὸ μητρὸς** τρεφόμενοι “Two children being raised **by their mother**”
Beneficiary/Maleficiary[11] (715 instances)	**ὑπὲρ τῆς πατρίδος** ἀποθανεῖν δυνήσομαι “I will be able to die **for my native land**”
Cause (753 instances)	ἐκπλαγῶν **διὰ τὸ παράδοξον τῆς ὄψεως** “Being struck **by the incredibility of the sight**”
Companion (424 instances)	τοῦτον **μετὰ Σιτάλκους** ἔπινον τὸν χρόνον “During that time I was drinking **with Sitalces**”
Comparison (198 instances)	πάντα ἐοικότες **ἀνθρώποις** πλὴν τῆς κόμης “Completely looking **like humans** except for their hair”
Condition (5 instances)	κελεύοντος **ἐπ’ αὐτοφώρῳ** τὸν μοιχὸν κτείνεσθαι “Commanding that an adulterer should be killed **in case he is caught**”
Degree (295 instances)	ξεῖνε **λίην** αὐχεῖς ἐπί γαστέρι “Stranger, you are boasting **too much** about your belly”
Direction (1006 instances)	**εἰς Θετταλίαν** αὐτοὺς ἀγαγὼν “Bringing them **to Thessaly**”
Duration (221 instances)	εὐφράνθη **ἐφ’ ἡμέρας τέσσαρες** “She was happy **for four days**”
Experiencer (259 instances)	σὺ δέ **μοι** δοκεῖς αἰτιᾶσθαι τὸν γάμον “You seem **to me** to defend marriage”
Extent of space (67 instances)	**διὰ Καϋστρίων πεδίων** ὁδοιπλανοῦντες “Wandering **through Castrian plains**”
Frequency (78 instances)	ἀποθνήσκομεν ὅτι οὐ βλέπομέν σε **καθ’ ἡμέραν** “We are dying because we do not see you **every day**”
Goal (282 instances)	ὥσπερ **ἐπὶ δεῖπνον** ἀποδεδημηκὼς εἰς Θετταλίαν “As if going to Thessaly **for a banquet**”
Instrument (507 instances)	**τοῖς δακτύλοις** τῶν ἑαυτοῦ βλεφάρων ἡπτόμην “I felt my own eyelids **with my fingers**”
Intermediary (16 instances)	ἔπεμψά σοι ἐπιστολὴν **διὰ τοῦ ἀρτοκόπου** “I’ve sent you a letter **by the baker**”
Location (1436 instances)	**ἐν Βυζαντίῳ** διατρίβειν δυναμένοις “Being able to stay **in Byzantium**”
Manner (1596 instances)	ἐάν τις τῷ **εὖ** λέγοντι μὴ πείθηται “If someone does not believe the person who speaks **well**”
Material/Content (22 instances)	ἔπλησεν τόν ἀσκόν **ὕδατος** “He filled the sack **with water**”
Modality (17 instances)	**ἴσως** οἶδας τί σοι ἔγραψα “**Perhaps** you know what I’ve written to you”
Possessor (127 instances)	ἔσται **τῇ Σαρρα** υἱός “**Sara** will have a son” (lit. “There will be a son **to Sara**”)
Property (6 instances)	ὅ ἦν **ἀγαθοῦ βασιλέως** “What is typical **of a good king**”
Recipient (1289 instances)	τὰ ἱμάτια αὐτοῦ ἔδωκεν **τῷ Αἰσώπῳ** “He gave **Aesop** his clothes”
Respect (800 instances)	μήτε ἀλγεῖν **κατὰ σῶμα** μήτε ταράττεσθαι **κατὰ ψυχήν** “Neither having pain **in the body** neither being disturbed **in the soul**”
Result (15 instances)	φαίνῃ **εἰς μανίαν** ἐμπεπτωκέναι “You seem to be fallen **into madness**”
Source (803 instances)	ῥίπτει δὲ αὐτὸν **ἐξ οὐρανοῦ** Ζεὺς “Zeus threw him **from Heaven**”
Time (943 instances)	**τετάρτῳ τε καί εἰκοστῷ τῆς βασιλείας ἔτει** νόσῳ διεφθάρη “He died from disease **in the twenty-fourth year of his reign**”
Time frame (45 instances)	μηδ’ εἰληφέναι μηθὲν **ἐνιαυτοῦ** “Not receiving anything **over the course of the year**”
Totality (150 instances)	ἐπιλαμβάνεται **τῆς χειρὸς αὐτῆς** “He took her **by the hand**”
Value (57 instances)	**ἑξήκοντα δηναρίων** τοῦτον ἠγόρακα “I’ve bought him **for sixty denarii**”

Table 4: Pedalion semantic roles

[11] I combined these two roles because they were not distinguished in the data, but since some prepositions (e.g. ὑπέρ + genitive) can only be used for a beneficiary, while others (e.g. κατά + genitive) only for a maleficiary, in the future it might be better to keep them apart.