A Source of Variation: A Corpus-Based Study of the Choice between ἀπό and ἐκ in the NT Greek Gospels

Peter de Swart,^a Hanne M. Eckhoff,^b and Olga Thomason^c

a) Radboud University Nijmegen

p.deswart@let.ru.nl

b) University of Oslo

h.m.eckhoff@ifikk.uio.no

c) University of Georgia

olgath@uga.edu

Abstract

Using a quantitative methodology based on extensively annotated corpus data from the PROIEL corpus, we examine the distribution of ἀπό and ἐκ in the NT Greek Gospels. The original semantic opposition between these two prepositions in terms of an ablative-elative distinction started fading during the historical development of Greek and has been argued to be already much weaker at the time of the New Testament. To explore this we generate a semantic map without semantic pre-analysis on the basis of four parallel language samples. We then use statistical techniques to interpret this map. We find that there is still a fairly clean separation between ἐκ and ἀπό largely based on semantic role. However, ἀπό is quite frequently used in elative contexts. A lexical analysis clarifies that the use of ἀπό in this environment amounts to the preposition specialising with certain lexical items, some of them with variable interpretations, as seen in the case of toponyms.

Keywords

source encoding; prepositions (ἀπό; ἐκ); NT Greek Gospels; semantic variation; semantic maps

1. Introduction

Cases and adpositions are the main grammatical means employed by languages to encode spatial relations between entities, often referred to as Figure and Ground (Talmy 2000). Studies of the spatial systems of languages have identified two recurrent dimensions underlying them (e.g. Jackendoff 1983, Jackendoff 1990; see Lestrade 2010 chapter 3 for a recent overview and discussion). The first is configuration (or Place) and is concerned with the geometrical relation between Figure and Ground, e.g. the Figure can be under the Ground or on it or in it, etc. The second dimension, direction (or Path), refers to the movement of the Figure with respect to this configuration. Three directional

primitives can be distinguished (e.g. Jackendoff 1983, Jackendoff 1990; again see Lestrade 2010 for a recent discussion): (i) *Location*: the Figure is stationary relative to the configuration; (ii) *Goal*: the Figure moves to (and ends up at) configuration; and (iii) *Source*: the Figure moves away from the configuration.

In this paper we focus on the encoding of source expressions in the New Testament (NT) Greek Gospels. In particular, we will be concerned with the choice between the two prepositions *ἀπό* and *ἐκ*, both of which have source meaning as their core semantic value. Originally, these two prepositions had clearly distinct meanings which can be characterized in terms of the configurational relation between Figure and Ground: where *ἐκ* indicated that the source motion started from within the Ground ('out of'), *ἀπό* was underspecified for such contact between Figure and Ground and is often translated as 'away from'. This configurational opposition started to fade during the historical development of Greek and has been argued to be already much weaker at the time of writing of the New Testament. The following two parallel examples from the episode of Jesus being baptized by John in the river Jordan, as described in the Gospels of Mark and Matthew, serve to illustrate this:

- (1) καὶ εὐθύς ἀναβαίνων ἐκ τοῦ ὕδατος εἶδεν σχιζομένους
and straight go.up.PART.PRES EK DET.GEN water.GEN see.AOR.3SG split.PART
τοὺς οὐρανοὺς ...
DET.ACC heavens.ACC
- 'And straightway coming up out of the water he saw the heavens opened ...' (Mark 1:10)
- (2) βαπτισθεὶς δὲ ὁ Ἰησοῦς εὐθύς ἀνέβη ἀπὸ τοῦ
baptize.PART.AOR ADV ART.NOM Jesus straight go.up.AOR.3SG APO DET.GEN
ὑδατος
water.GEN
- 'And Jesus when he was baptized went up straight out of the water' (Matthew 3:16)

The two Gospel authors use very similar wordings to describe this scene. Both use a form of the verb *ἀναβαίνω* 'go.up' which takes τοῦ ὕδατος 'the water.GEN' as its prepositional complement. They differ, however, in the preposition used: Mark uses *ἐκ* and Matthew uses *ἀπό*. These examples suggest that in NT Greek the two prepositions represent closely related alternatives whose meanings substantially overlap. If this is indeed the case, we are faced with the question *what determines the choice between ἀπό and ἐκ*. Does it represent an instance of random variation or of conditioned variation? In the latter case, does the variation only reflect authorial differences, as could be argued in the examples above, or can we determine other factors which underlie it? Of course, in order to answer these questions we first have to determine to what degree there is actual variation.

Accordingly, the main objective of this paper is to get a firm handle on the variation in the choice between ἀπό and ἐκ in NT Greek. A better understanding of this variation will lead to a clearer view on the semantic similarities and differences between these two prepositions.

Our second objective is of a methodological nature. Research into older stages of languages such as Ancient Greek is corpus-based by necessity. Nevertheless, the usage of corpus-based methodologies and standards have not caught on in the research community at large. The main reason for this most probably lies in the absence of large-scale annotated corpora. For New Testament Greek, the situation has dramatically changed since the recent, ongoing development of the PROIEL corpus at the University of Oslo, Norway (Haug, Eckhoff, Majer, and Welo 2009). Such a large-scale annotated corpus ensures accountable and replicable research. Below we will show how corpus data in combination with analytical tools from corpus linguistics can shed light on the choice between ἀπό and ἐκ. More specifically, we want to investigate how the recently proposed methodology of *probabilistic semantic maps* (Wälchli 2010) can be applied to a small language sample.

The paper is structured as follows: in the next section we briefly review the literature on the semantics of ἀπό and ἐκ and the associated historical development. Section 3 forms the main part of the paper, as it presents the data set used and two ways of analyzing the variation in it: *probabilistic semantic maps* and *random forests*. In section 4 we will combine our quantitative approach with a more qualitative one by looking at the lexical profiles of the two prepositions, i.e. the types of complements they appear with. Finally, section 5 presents our conclusions.

2. The Expression of Source in Greek

In this section we discuss the semantics of ἀπό and ἐκ and their historical development based on two prominent works on prepositions in Ancient Greek (Luraghi 2003; Bortone 2010).

Homeric Greek had essentially two grammatical means to express spatial notions, ignoring spatial adverbs for the moment. It could either use bare case forms or combine them with a preposition. The Greek oblique (non-nominative) cases showed a close correspondence with spatial notions. Thus, whereas the accusative case was used to express goal, the genitive was associated with source. The two main prepositions to express source meaning, which were unsurprisingly combined with the genitive case, were ἀπό and ἐκ. Our focus in this paper is on the distribution of these two prepositions in the NT Greek of the Gospels. A third preposition associated with source, παρά, is mostly left untouched due to its relatively low frequency in the NT Gospels.

The prepositions ἀπό and ἐκ were both used to express source meaning, but at the same time they displayed subtle semantic differences. ἀπό had an ablative semantic value ('away from') and ἐκ an elative one ('out of') (Luraghi 2003; Bortone 2010). According to Bortone (2010:164): "Greek ἀπό indicated position or distancing movement from the outer surface of the reference object (away from, cf. German cognate *ab-*), while ἐκ expressed a distancing motion from the inside of a three-dimensional volume (out of)". Likewise, Luraghi (2003:118) claims that ἐκ implied contact between Figure and Ground, whereas ἀπό was underspecified for this feature. The range of possibilities was saturated by παρά which implies absence of contact between Figure and Ground. ἀπό, thus, can be said to be less specific in meaning than ἐκ. In addition to their uses to indicate source proper, ἀπό and ἐκ were also used "for the reinforcement of partitive expressions (also rendered by the plain genitive)" (Bortone 2010:143).

Luraghi (2003) claims that the meanings of ἀπό and ἐκ are clearly distinct in Homer. This holds not only for their purely spatial uses, with ἐκ reserved for Grounds that can be conceived of as containers, but also in their meaning extensions outside the spatial domain (see Luraghi 2003, sections 3.2 and 3.4 for discussion and examples). Later on, starting with Herodotus, the meanings of the two prepositions start to converge and, although both follow a different path of semantic extension, "they ended up with much the same meaning" (Luraghi 2003:130). Bortone (2010) notes a similar change as part of a more general tendency where fine semantic differences between many pairs of spatial prepositions fade. He also claims that the distinction between elative and ablative meaning starts to fade already in Classical Greek. He notes that "[t]he confusion between ἐκ and ἀπό occurs within the same idiolect and with the same referent—even in adjacent paragraphs" (Bortone 2010:164, for examples from Thucydides). In particular, Bortone claims that ἐκ is losing functions to ἀπό and that the latter is more and more used as the former. These changes are part of a more general expansion of ἀπό to cover also other meanings. According to Bortone, this convergence between the two prepositions does not show up numerically in Biblical Greek, with ἐκ still being the more frequently used preposition, although papyri texts do show quantitative evidence for it.

In the next section we zoom in on the use of ἀπό and ἐκ in the Gospels of the Greek New Testament. This allows us to get a more detailed view on the variation in the choice of the prepositions.

3. ἀπό and ἐκ in the Gospels

Our data set comes from the PROIEL corpus, a parallel corpus of the Greek original of the New Testament and its translations into Latin, Old Church Slavonic, Gothic and Classical Armenian, developed by the members of the project

Figure 1. Distribution of ἀπό and ἐκ over semantic roles per Gospel.

“Pragmatic Resources in Old Indo-European Languages” at the University of Oslo, Norway.¹ The corpus contains detailed annotation on many levels: morphology, syntax, semantics and information structure. The texts are also automatically aligned at sentence and token level (the latter with a success rate of about 95 %, for more details on the annotation and alignment see Haug, Eckhoff, Majer, and Welo 2009). Of special relevance to the purposes of the present article is the fact that all Greek prepositional phrases (PPs) are individually tagged for semantic role, based on the classification in Thomason (2006). The text of the NT used in the corpus is Tischendorf’s critical edition of the New Testament (8th edition), which had been digitized by Ulrik Sandborg-Petersen (see Haug et al. 2009 for further discussion of the editions used in the corpus).

From the PROIEL corpus we extracted all occurrences of Greek PPs in the four Gospels involving an instance of ἐκ or ἀπό followed by a dependent with a case specified, thus excluding indeclinables and adverbs. This resulted in a total number of 682 PPs, 286 of which occurred with ἀπό and 396 with ἐκ.² Figure 1 summarizes the distribution of ἀπό and ἐκ over semantic roles per Gospel author.

¹) The corpus can be accessed at: <http://foni.uio.no:3000/>.

²) Inclusion of indeclinables and adverbs would have resulted in a slightly larger data set with 320 occurrences of ἀπό and 400 of ἐκ.

The figure shows that the authors differ in the frequency with which they describe the different types of meanings, with John's preference for elative meanings standing out. Zooming in on the expression of ablative meanings we see that ἀπό is still the predominant choice for all authors. Likewise, when we consider the expression of elative meanings, ἐκ comes out as the preferred expression. We do find more variation with elative expressions, ranging from John's preponderance of ἐκ to Matthew's strong preference for ἀπό with Luke and Mark sitting somewhere in between. This suggests that whereas ἀπό is still firmly reigning in the ablative domain, at the same time it is used in the elative domain traditionally associated with ἐκ.

3.1. *A Semantic Map of ἀπό and ἐκ*

In order to get a better view of the variation in the data, we exploited the cross-linguistic, parallel nature of the PROIEL corpus to automatically generate a semantic map of ἀπό and ἐκ based on a four-language sample, following work by Wälchli (2010). Semantic maps are used to visualize the internal structure of a semantic domain. This is achieved by placing concepts that are closely related closer on the map and ones that are less related further apart. Relatedness is determined by investigating the coding of meaning concepts in one or more languages. If a language uses the same encoding for two concepts, they are assumed to be semantically similar and hence will appear close together on the map (cf. Haspelmath 1997, Haspelmath 2003 for a general introduction to the methodology, see also Luraghi 2003:16–17 specifically for Ancient Greek). Semantic maps are in wide use in typology (e.g. Haspelmath 1997, Haspelmath 2003), but have also been employed for smaller-scale cross-linguistic work (Clancy 2006) and for diachronic investigations (Luraghi 2003).

Traditionally, semantic maps are drawn manually. This approach brings along several disadvantages. Such maps may, for instance, show the hand of the researcher as he makes analytical choices. Furthermore, such maps often give equal status to frequent and very infrequent patterns, in this way losing a distinction in generality between patterns. Finally, distance on hand-drawn maps is not interpretable and hence cannot be translated into semantic similarity. Recently, alternative ways of drawing have become available in which maps are generated computationally on the basis of a large database using a visualization tool such as multidimensional scaling (see below; Croft and Poole 2008, Wälchli 2010).

In our opinion, a computational approach lacks the disadvantages mentioned above: it efficiently handles large data sets and allows for analysis with minimal preconditioning of the data (as opposed to the classical semantic maps

where the linguistic comparison is based on primitive categories which are set by a researcher). Moreover, multidimensional scaling visualizes similar meanings as clusters and semantic difference as distance: the more similar two concepts are in a sample, the closer they are on the map. This is not possible in a manual approach. Within the computational approach adopted here, semantic categories are not imposed on the data before the analysis, they emerge from the interpretation of the analysis itself: “Instead of assuming abstract functional domains, concrete instantiations of particular functions are considered (contextually embedded situations) as they are determined by given contexts. Functional domains will emerge in the analysis as clusters of situations if there is evidence for them in the cross-linguistic data set” (Wälchli 2010: 331). Due to the fact that computational maps are based on large datasets they are less likely to obscure frequency effects.

In this paper we follow the approach outlined in Wälchli (2010) which relies crucially on the notion of ‘similarity semantics’, as do semantic map approaches in general (see also Cysouw 2010 for discussion). The assumption is that the more similar two meanings are, the more likely they are to be expressed by the same form (in our case the preposition used) in any language. Thus, we expect linguistic expressions from many languages to cluster around the same functions. By visualizing such relationships between languages, semantic maps can “model general trends in the semantic organization of categories” (Wälchli 2010: 331).

As Wälchli (2010) demonstrates, the similarity of linguistic expressions can be calculated by the extent to which they occur in the same situations in a text (in his case, motion events in the Gospel of Mark). In our case a situation is instantiated by an occurrence of ἀπό or ἐκ in the Greek Gospels. In order to build a similarity matrix we enforced a restriction on our data set of ἀπό and ἐκ PPs in the Gospels described above. Each Greek PP should have a corresponding expression (of any form) in Latin, Gothic and Old Church Slavonic (OCS) which is token aligned with either the Greek P or its dependent. This restriction entails that we consider a much wider range of linguistic entities in the translation languages than in the original Greek. To include as much as possible, we also check for alignments with the complement of a Greek P if there is no alignment with the P itself (see Eckhoff, Thomason, and de Swart 2011 for more discussion on this issue). Another implication of this restriction is that it excludes all situations where one or more of the languages has no aligned translation of the Greek PP. Mostly, these are instances of lacunae in the other texts (large portions of the NT are missing in the OCS and the Gothic versions). There are also passages where a Greek PP is simply not translated or a rendition for various reasons could not be aligned with either a Greek P or its dependent, e.g. when the meaning of a Greek PP is included in the meaning

of a verb in the translation. The resulting dataset contains 358 examples (148 occurrences of ἀπό, 210 of ἐκ).

This reduced dataset was used to automatically construct a similarity matrix (see Wälchli 2010 for details) by first comparing the encoding of each of the 358 situations to all other situations in the database, i.e. comparing situation 1 to situations 2–358, situation 2 to situations 1,3–358, etc. This resulted in a similarity measure for each situation pair in each language. Next, the similarity measures for each situation pair were summed and divided by four (the number of languages) resulting in an overall similarity measure ranging from 0 (full similarity) to 1 (no similarity). Repeating this for each situation pair resulted in a 358-by-358 similarity matrix. This matrix was then visualized using multidimensional scaling (MDS) based on the `cmdscale()` function in the statistical software package R (R Development Core Team 2009). MDS is a multivariate statistical technique (see e.g. Borg and Groenen 2005 for a general introduction, and Baayen 2008 for linguistic applications) that takes as its input a set of (dis)similarities (our similarity matrix) and maps them into corresponding distances of an n -dimensional space. That is, the result is a set of points which can be plotted: a semantic map. The MDS solution tries to match the distance between two points as closely as possible with their (dis)similarity. The interpretation of the dimensions needed to represent the data is not given by the algorithm but has to be determined by the researcher himself.

Let's step back for a moment from the technical details to see what we would expect such a semantic map to show us. As stated above, semantically similar concepts will appear close together on the map. Thus, if ἀπό and ἐκ represent two semantically distinct concepts, they will appear as two clearly separated clusters on the map. On the other hand, if the meanings of the two prepositions overlap completely, we expect a less structured map on which ἀπό and ἐκ freely mix. Of course, any situation in between is possible as well.

Figure 2 shows a graphical representation of the first two dimensions of the MDS representation based on the similarity matrix described above. Each symbol on the graph represents a textual instance of ἀπό (A) or ἐκ (E) in one of the Gospels. ἀπό and ἐκ each take up two well-defined regions on the map. The two regions can almost be perfectly separated by the vertical axis. This map strongly suggests that the two prepositions represents two distinct concepts. Moreover, it suggests that the difference between these two concepts can be almost perfectly captured in one dimension, cf. the almost complete separation of the map on Dimension 1 (the horizontal axis).

Before we turn to the interpretation of this dimension, we should devote some discussion to the translational nature of our data set, for one could hypothesize that the structure in the map is due to strict translational equivalents of the Greek prepositions in the three other languages (see e.g. Luraghi and

Figure 2. Dimensions 1 and 2 of the MDS representation of the distribution of ἀπό and ἐκ. Each symbol on the map represents an instance of ἀπό (A) or ἐκ (E) in one of the Gospels.

Cuzzolin 2007 for a discussion of the effects of translation; see also Klein 1992 who discusses the effect of translation on prepositional usage in the Gothic translation of the Greek NT). For instance, if we would find that the Latin translation uses *ex* exclusively for ἐκ and *ab* for ἀπό, this would explain the distribution found in Figure 2. Inspection of the translations used in the three languages shows that this hypothesis remains unsupported. In Latin, we find 18 different ways of translating ἀπό and ἐκ of which *ab*, *ex* but also *de* are most frequently used. However, whereas *ex* is indeed almost exclusively used to translate ἐκ (84 out of 87 occurrences), both *de* (80/108) and *ab* (31/133) show much more variation. OCS presents 13 different ways of translating the Greek prepositions, but in this language we find one preposition *ot* to be responsible for the majority of translations of both ἀπό (112 times) and ἐκ (144 times). Finally, Gothic uses 16 different translations for ἀπό and ἐκ. This language is special compared to the other ones as it divides most of the translational burden over six different expressions. Four of these are mainly used to translate ἀπό: *fairra* (12 out of 12 occurrences), *faura* (10/10), *fram* (27/33), and *af* (60/77). Two others, bare genitive case (46/52) and *us* (132/152), are mainly used for ἐκ. The observed variation in translations within and between these languages strongly

suggests that translation is not the main factor determining the structure on the map. A very similar picture emerges from Eckhoff, Thomason, and de Swart (2011) who study source encoding in all four languages mentioned. They show that PP usage appears to be remarkably independent in all three translation languages. A full exploration and argumentation of the influence of translation on automatically generated semantic maps has to be postponed to future work.

The semantic map in Figure 2 suggests that the semantic distinction between ἀπό and ἐκ can almost be reduced to one single dimension.³ That is, one can almost draw a line completely separating the Es from the As.⁴ However, as stated above, the MDS solution does not come with an interpretation of its dimensions. The MDS algorithm computes the mathematically most sound solution not hindered by any linguistic knowledge. As a result, a dimension can correspond to both a well-known linguistic feature, but also to any ((non)-linear) combination of features. This means that each dimension requires interpretation by the researcher. One way of approaching this is by overlaying the map with features known to be relevant for the phenomenon.

Figure 3 shows such an overlay for the semantic role expressed by a PP. These roles are taken from the annotation provided in the PROIEL corpus.

The figure is split into four different maps each representing a different semantic role: ablative (top left), elative (top right), partitive (bottom left), and other roles (bottom right). As in Figure 2 above, each symbol on a map represents an instance of ἀπό (A) or ἐκ (E) in one of the Gospels. If Dimension 1, which almost perfectly separates occurrences of ἀπό (A) or ἐκ (E), corresponds with a partition in semantic roles, we would expect that one subset of roles occurs in one region on the map and the other roles in some other region. This amounts to saying that one set of semantic roles is represented by only Es and the other set by only As. As Figure 3 shows, this is not the case. In all four panels we find that the respective semantic role is expressed by both prepositions and hence occurs on both sides of the map. The partitive/non-partitive distinction comes closest to such a complete separation. Nevertheless, it is clearly the case that roles have a stronger association with specific parts of the map. Thus, elative situations are mostly, but not exclusively, found on the left side of the map and

³) Inspection of various goodness of fit measures (cmdscale's GOF, relative values of eigenvalues, difference between original distance matrix and matrix returned by the MDS algorithm) all indicate that more than one dimension has to be taken into account to come to a faithful representation of the data. These measures suggest the optimal number of dimensions to be four.

⁴) Incidentally, this line almost coincides with the vertical axis. Note that the vertical axis is only drawn as a visual aid, no theoretical significance is ascribed to it.

Figure 3. Semantic role overlay on the MDS representation of Figure 2. The overlay is split into four different roles: elative (top left), ablative (top right), partitive (bottom left), and other roles (bottom right). Each symbol on a map represents an instance of ἀπό (A) or ἐκ (E) in one of the Gospels.

the converse holds for ablative situations. This suggests that semantic role may be an important part of the interpretation of Dimension 1.

Consider Figure 4 as a potential alternative interpretation of the splitting criterion. This figure represents an overlay of the animacy features of the PP complements extracted from the corpus. The overlay is again split into four panels, each representing a degree of animacy (roughly following Zaenen, Carletta, Garretson, Bresnan, Koontz-Garboden, Nikitina, O'Connor, and Wasow 2004): animates (top left), concrete inanimates (top right), non-concrete inanimates (bottom left), and place (bottom right). As before, each symbol on a map represents an instance of ἀπό (A) or ἐκ (E) in one of the Gospels.

Figure 4. Animacy overlay on the MDS representation of Figure 2. The overlay is split into four degrees of animacy: animates (top left), non-concrete inanimates (top right), concrete inanimates (bottom left), and place (bottom right). Each symbol on a map represents an instance of ἀπό (A) or ἐκ (E) in one of the Gospels.

Figure 4 shows that animacy is most likely less relevant for accounting for Dimension 1. Neither of the 4 degrees shows commitment to one of the two sides of the map. Thus, neither the semantic role expressed by the PP nor the animacy of its complement can serve as a full explanation of Dimension 1 of the semantic map. In the next section we discuss a more systematic way of finding such an interpretation.

3.2. Recursive Partitioning of the Data: A Random Forest

In order to get a better view of the motivations underlying to choice for ἀπό or ἐκ in the Gospels, we have selected eight factors for inclusion in a multifactorial

analysis. These factors represent features of the dependent of the preposition, its animacy, definiteness, number and part of speech, features of the prepositional phrase, its semantic role, its syntactic function in the clause and the gospel in which it occurs, and a verbal feature (presence of a prefix). Table 1 provides an overview of the factors and their levels. Each factor is based on annotations taken from the PROIEL corpus and each item in our data set is annotated for all factors. To avoid data sparseness we have reduced the number of levels for some factors by grouping levels with low numbers together under the heading ‘other’. The next step is to find out if and in what way these factors are related to the prepositional choice under investigation. As it is not possible for the human eye and mind to keep track of the influences of and interactions between these eight different factors, we will employ automated statistical procedures to do so for us.

FACTOR	DESCRIPTION	# LEVELS	LEVEL DESCRIPTION
Animacy	animacy of the dependent	5	animate (human, animal, organisation); concrete inanimate; non-concrete inanimate; place; time.
ART	presence of article on dependent	2	yes; no.
Author	author of the Gospel	4	John; Luke; Mark; Matthew.
Number	number of the dependent	2	singular; plural.
POS	part of speech of the dependent	6	Adjective (Adj); Common noun (CN); Proper noun (PN); Personal pronoun (Ppers); Reflexive pronoun (Pref); other.
Prefix	type of prefix of the main verb of the clause of the PP	3	apo; ek; absent (no prefix or other than apo/ek).
SemRole	semantic role of the PP	4	ablative; elative; partitive; other.
SyntRel	syntactic relation of the PP	4	adverbial (adv); verbal argument (obl); adnominal partitive modifier (part); other.

Table 1. Descriptions of the factors used in the random forest analysis and their levels. All annotations are taken from the PROIEL corpus.

The choice between $\acute{\alpha}\pi\acute{o}$ and $\acute{\epsilon}\kappa$ can be seen as an instance of a structural alternation not uncommon in linguistic research. In recent years much corpus-based research into such structural alternations (e.g. the dative alternation) has seen the light which tries to uncover their determinants by means of statistical modeling, in particular logistic regression. Logistic regression provides a natural method to determine the relevance and importance of (sets of) predictor variables for linguistic phenomena involving a binary choice (see Baayen 2008 for a linguistic introduction). The method also has its limitations, for instance, in cases of complete separation, where one predictor or a linear combination of predictors perfectly predicts the target value (cf. Hosmer and Lemeshow 1989:141), as Figure 2 suggests is the case in our data. In addition, such mod-

els may run into estimation problems when the number of predictors is large in comparison to the number of data points or in case of categorical predictors with many levels.

In order to avoid these potential problems in our dataset we will employ the non-parametric alternative of classification trees to get a better grip on the conditioning factors in the choice between $\acute{\alpha}\pi\acute{o}$ and $\acute{\epsilon}\chi$. Classification trees assess the importance of predictor variables through a process known as *recursive partitioning*. They start out with the full data set and all relevant predictor variables, which together can be visualized as a rectangular space. The algorithm then splits this space into two parts which in their turn can be split again, and so on until some stopping criterion (e.g. impurity reduction or number of observations) is reached. Each partition is created such that similar outcomes, in our case the use of either $\acute{\alpha}\pi\acute{o}$ or $\acute{\epsilon}\chi$, are grouped together as much as possible. Each partition is associated with one out of two mutually exclusive set of levels of one of the predictor variables. Examples of such sets in the case of the semantic roles discussed above are {ablative} vs. {elative,partitive,other} or {ablative,partitive} vs. {elative,other} or any of the other remaining combinations. This process of recursive partitioning results in a classification tree (see the bottom panels of Figures 6–8 for examples) in which partitions are visualized as binary branching nodes. For each terminal node in the tree, the model takes the value of the majority of the outcomes in that partition as the predicted outcome for each observation in the partition. Thus, if in our case $\acute{\alpha}\pi\acute{o}$ is the most frequently used preposition in a partition, then the model will predict for all data points in that partition $\acute{\alpha}\pi\acute{o}$ as the encoding used (even though in reality $\acute{\epsilon}\chi$ may be used). By comparing the predicted outcome with the actual outcome one can assess the goodness of fit of such a classification tree and hence the relevance of the predictors included in the model.

Classification trees lack the mentioned problems of logistic regression models, but they do face a problem of their own, which is the potential instability of the predicted outcomes. Due to the way a tree is grown, partitions made earlier on influence partitions made later on, and hence the outcome might be heavily influenced by small changes in the input data. In order to overcome this instability, one can grow a set of classification trees resulting in what is known as a *random forest*. The following discussion of random forests is mainly based on Strobl, Malley, and Tutz (2009), to which we refer the interested reader for a highly accessible, in-depth discussion of this method.

In a random forest predicted outcomes are based on the averaged outcomes of multiple classification trees resulting in much more stable prediction accuracies (see Strobl et al. 2009 for more advantages of random forests over single trees). A random forest is random in two ways: (i) each tree is grown on the basis of random sampling from the full data set, either by selecting a subsam-

ple or a same-size sample with replacement; (ii) each split in a tree is based on a subset of the predictor variables, in this way allowing for the detection of low-level influences of factors which were otherwise obscured by stronger competing factors. This also makes random forests ideal for data sets with collinearity of factors (see Shih and Grafmiller 2011 for distinguishing the influence of four operationalizations of grammatical weight by means of random forests).

In contrast to single classification trees, a random forest is much harder to visualize by a tree structure. Instead, the relevance of the different factors is represented by their variable importance measure which can be presented graphically (see Figure 5). Variable importance is computed by randomly permuting the values of the predictor variables, in this way breaking the original relation with the outcome. Prediction accuracies are then compared. The rationale behind this is that if the predictor is strongly associated with the outcome, permutation will result in a decrease in prediction accuracy. On the other hand, if a predictor is only weakly associated with the outcome, permutation will have only a minor effect on the prediction accuracy of the model.

Figure 5 shows the conditional variable importance of the eight factors introduced above (cf. Table 1) in a random forest of 1000 trees grown on the full data set ($n = 682$) and selecting four variables per split.⁵

This figure shows that the semantic role expressed by the PP is by far the most important factor in predicting the preposition used. The author of the Gospel comes second. This suggests that there may be much variation between the authors in the selection of factors influencing their choice between $\acute{\alpha}\pi\acute{o}$ and $\acute{\epsilon}\kappa$. The presence and type of prefix and the part of speech of the PP-complement come third and fourth, followed by the syntactic relation of the PP and the animacy of its complement, which both show relatively little influence. Finally, the presence of an article does not seem to affect the outcome, nor does the number of the PP-complement, which indeed did not reach significance. The prediction accuracy of this full model reaches 87.5% and hence presents a substantial increase in comparison to a baseline model which always predicts the most frequent occurring preposition as its outcome (in this dataset $\acute{\epsilon}\kappa$ is used 58.1% of the times).

⁵) All analyses were run with the functions `cforest()` and `varimp()` of the `party` package (Hothorn, Hornik, Strobl, and Zeileis 2010) freely available within the statistical software package R (R Development Core Team 2009).

Figure 5. Variable importance of the random forest classifier of the full dataset ($n = 682$). The classifier was constructed on the basis of 1000 trees and selection of 4 variables per split. Number comes out as a non-significant factor. The prediction accuracy of the model was 87.5% compared to a baseline of 58.1% ($C = 0.9498$, $D_{xy} = 0.8996$). The model misclassified 50 occurrences of $\acute{\alpha}\pi\acute{o}$ and 35 occurrences of $\acute{\epsilon}\chi$.

As ‘author’ emerges as a highly ranked factor in the full model, we decided to grow a random forest for each Gospel separately in order to examine potential variation between authors. The results of these analyses are summarized in Figures 6–9. The top panels provide a visualization of the variable importance for each author and the bottom panels show a single classification tree to give an impression of the behavior of the factors in each Gospel.⁶ Below we will discuss some of the main trends emerging from this analysis.

These figures show that for each of the Gospel authors the semantic role expressed by the PP is the single most important factor in choosing between $\acute{\alpha}\pi\acute{o}$ and $\acute{\epsilon}\chi$. The individual classification trees do seem to suggest a difference in sensitivity to the different semantic roles, as John and Luke split the data differently than Mark and Matthew. This could be explained by the relatively higher frequency of the partitive construction with the first two authors.

⁶ The classification trees were grown with the function `ctree()` of the `party` package (Hothorn et al. 2010) freely available within the statistical software package R (R Development Core Team 2009).

Figure 6. *Top panel:* Variable importance of the random forest classifier of the Gospel of John ($n = 196$). The classifier was constructed on the basis of 1000 trees and selection of 4 variables per split. Prefix comes out as a non-significant factor. The prediction accuracy of the model was 95.4 % compared to a baseline of 82.7 % ($C = 0.9683$, $Dxy = 0.9366$). The model misclassified 8 occurrences of ἀπό and 1 occurrence of ἐκ. *Bottom panel:* Single classification tree of the Gospel of John.

Figure 7. *Top panel:* Variable importance of the random forest classifier of the Gospel of Luke (n = 196). The classifier was constructed on the basis of 1000 trees and selection of 4 variables per split. Prefix, Article and Number come out as non-significant factors. The prediction accuracy of the model was 82.7% compared to a baseline of 57.1% (C = 0.9153, Dxy = 0.8306). The model misclassified 19 occurrences of ἀπό and 15 occurrences of ἐκ. *Bottom panel:* Single classification tree of the Gospel of Luke.

Figure 8. *Top panel:* Variable importance of the random forest classifier of the Gospel of Mark (n = 106). The classifier was constructed on the basis of 1000 trees and selection of 4 variables per split. The prediction accuracy of the model was 82.1% compared to a baseline of 61.3% ($C = 0.9589$, $Dxy = 0.9178$). The model misclassified 16 occurrences of ἀπό and 13 occurrences of ἐκ. *Bottom panel:* Single classification tree of the Gospel of Mark.

Figure 9. *Top panel:* Variable importance of the random forest classifier of the Gospel of Matthew (n = 184). The classifier was constructed on the basis of 1000 trees and selection of 4 variables per split. The prediction accuracy of the model was 84.2% compared to a baseline of 53.3% ($C = 0.9292$, $Dxy = 0.8583$). The model misclassified 15 occurrences of ἀπό and 14 occurrences of ἐκ. *Bottom panel:* Single classification tree of the Gospel of Matthew.

Inspection of the ranking of the remaining factors shows that Matthew stands out in comparison to the other three authors. We have already seen above (cf. Figure 1) that he is the most advanced in using ἀπό for relative meanings. The high ranking of the factor Prefix is related to the presence of the prefix ἐκ- on a verb and suggests that this acts as a counterforce, i.e. some sort of prepositional concord: 15 occurrences of this prefix are in an elative context,

the preposition ἐκ is used 14 times in such passages, as well. To compare, the three occurrences of an ἐκ-prefix in ablative contexts all take ἀπό.

The three other authors have a higher ranking of POS, the part of speech of the PP-complement, in common. The classification trees of John and Mark suggest that the influence of this factor is due to the association between proper names and reflexive pronouns with ἀπό in elative contexts, thus also underlining the authors' overall resemblance resulting from the random forest analyses. The relevance of this split within the factor POS is reinforced by the fact that it also emerges in a classification tree of the full data set (not shown here). A potential interpretation is that it provides a window on ἀπό's original entrance into the elative domain. In John, most cases of elative contexts involve reflexive pronouns, and these may have used causal agents as a model, as they also appear with ἀπό.

Mark, by contrast, uses ἀπό in elative contexts only with place names. This may be explained by the fact that toponyms are inherently flexible in interpretation, fluctuating between a rigid container reading and a more vague region reading. The latter reading makes them a suitable candidate for ἀπό in general and this again could have functioned as a model for its use also in elative contexts.

In sum, this section has provided a detailed quantitative view of the choice between ἀπό and ἐκ in the Gospels. It has shown that there can still be drawn a rather clean separation between the two prepositions, and that this is mainly due to the near-monopoly of ἀπό in the ablative domain. Elative contexts, by contrast, are not dominated by ἐκ alone, but also frequently encoded by ἀπό. This extended use of the latter preposition is fully in line with the general observations on the expansion of its use in Luraghi (2003) and Bortone (2010), discussed in section 1 above. It is important to stress that our results cannot be taken as direct evidence for an expansion of ἀπό. It may well be that application of our methods to older stages of the language yields results different from those reported in the literature based on more traditional research.

In addition, we have seen how the different Gospel authors determine their choice for the prepositions on the basis of (partially) different factors. In particular, Matthew stands out from the other three authors by the most frequent use of ἀπό in elative contexts and by the factors that guide his choice in general.

4. Lexical Profiles of ἀπό and ἐκ

Thus far we have fully concentrated on the quantitative aspects of the choice in prepositional encoding. In this section we combine our quantitative approach with a more qualitative one by investigating the *lexical profiles* of ἐκ and ἀπό, i.e.

Figure 10. Lexical profiles of ἀπό (left) and ἐκ (right). Each graph also shows the ten most frequent lemmata (in their nominative citation form) for each preposition followed by the number of occurrences.

the lexemes used as their complements. This may reveal low-level differences and similarities between the two prepositions.

Figure 10 gives an overview of the lexical distribution of complements of both ἀπό and ἐκ by plotting each unique lexeme occurring as its complement against its frequency in our data set. The results are two graphs with shapes well known from lexical frequency distribution: a few items with a high number of occurrences followed by the majority of items with a low number of occurrences. In addition these graphs also list the most frequently occurring lemmata for each preposition. Unsurprisingly, the versatile pronoun αὐτός is the most frequent complement with both ἀπό and ἐκ. Closer inspection of the occurrences of this pronoun reveals an asymmetry in its use: αὐτός occurs with ἐκ mostly in the plural (29/38), whereas the singular is more frequent with ἀπό (19/34). This difference is explained by the high number of partitive expressions with ἐκ that require divisibility of their complements.

Partitivity is also the explanation for the regular occurrence of the 1st person plural pronoun ὑμεῖς with ἐκ, whereas ἀπό goes with first and second person singular pronouns. A further characteristic of ἐκ is its occurrence with the location-denoting complements δεξιός ('right') and εὐώνυμος ('left'). ἀπό, by contrast, has a preference for complements denoting time (ἀρχή ('beginning'), but also ὥρα ('moment, hour')). The frequent combination of ἀπό with the noun ζύμη ('leaven') represents an idiomatic phrase often accompanied by the verb προσέχω ('beware of').

Figure 11 restricts our attention to the words with a clear lexical content by showing the noun lemmata that occur at least five times in our data set.

Figure 11. Frequency of occurrence of noun lemmata with *ἀπό* plotted against frequency of occurrence with *ἐκ*. The following four coordinates each have two nouns with equal frequency counts: (0,5) *πατήρ*, *χειρ*, (1,4) *ὄφθαλμός*, *ὕδωρ*, (2,4) *γῆ*, *πνεῦμα*, and (4,6) *μνημεῖον*, *πόλις*.

The graph shows the frequency of occurrence with *ἀπό* on the horizontal axis and the frequency with *ἐκ* on the vertical axis. The absence of occurrences of the nouns *μαθητής* ('disciple') and *Φαρισαῖος* ('Pharisee') with *ἀπό* and their frequent use with *ἐκ* again is due to partitivity. *ἀπό*, by contrast, matches up more frequently with geographical names, be it names of cities (*Ἱεροσόλυμα* ('Jerusalem')) or regions (*Γαλιλαία* ('Galilee')), also *ἀνατολή* ('the east')). This preference for *ἀπό* is substantially reduced when the more general term *πόλις* ('city') is used. In similar contexts, this noun is found with both prepositions:

- (3) ἐξερχόμενοι ἀπὸ τῆς πόλεως ἐκεῖνης καὶ τὸν κονιορτὸν ἀπὸ
 go.out.PART.PRES APO DET.GEN city.GEN DEM.GEN ADV DET.ACC dust.ACC APO
 τῶν ποδῶν ὑμῶν ἀποτινάσσετε
 DET.GEN.PL feet.GEN you.PL.GEN shake.off.IMP.2PL
 'when you go out of that city shake the dust of your feet' (Luke 9:5)
- (4) ἐξῆλθον ἐκ τῆς πόλεως καὶ ἦρχοντο πρὸς αὐτόν.
 go.out.AOR.3PL EK DET.GEN city.GEN CONJ come.IMP.3PL to him.ACC
 'Then they went out of the city and came to him.' (John 4:30)

In the following example from Luke we find *ἀπό* with the region and *ἐκ* with *πόλις*. The respective choice of the two prepositions is most probably due to the fact that we progress from a non-specific to a more specific entity, which is enforced by the parallel structure found with the goal NPs. *ἐκ*, then, is better suited to be used with the more specific, container-like *πόλις*.

- (5) Ἀνέβη δὲ καὶ Ἰωσήφ ἀπὸ τῆς Γαλιλαίας ἐκ πόλεως Ναζαρέθ εἰς
 go.up.AOR.3SG ADV ADV Joseph APO DET.GEN Galilee EK city.GEN Nazareth to
 τὴν Ἰουδαίαν εἰς πόλιν Δαυείδ
 DET.ACC Judaea to city David

‘And Joseph also went from Galilee out of the city of Nazareth into Judaea to the city of David’
 (Luke 2:4)

The place-denoting noun *οὐρανός* (‘heaven’) shows a strong preference for *ἐκ*. This is partly due to the high frequency of this noun in John (15 out of 36 occurrences), who has a tendency to talk about things coming out of heaven and going out of the world (all occurrences of *κόσμος* (‘world’) are with John). As shown above, John has a strong preference to use *ἐκ* and this is sustained with *οὐρανός* for which he uses *ἐκ* 14/15 times. Also all his uses of *κόσμος* take *ἐκ*. But also in almost identical linguistic contexts we find variation, even with this author:

- (6) ὅτι καταβέβηκα ἀπὸ τοῦ οὐρανοῦ
 CONJ go.down.PF.1SG APO DET.GEN heaven.GEN
 ‘For I came down from heaven ...’
 (John 6:38)
- (7) ὅτι ἐκ τοῦ οὐρανοῦ καταβέβηκα
 CONJ EK DET.GEN heaven.GEN go.down.PF.1SG
 ‘that I came down from heaven’
 (John 6:42)

Similar cases of within-author variation for *οὐρανός* can be found when it modifies the noun *σημεῖον* (‘sign’):

- (8) φόβητρά τε καὶ σημεῖα ἀπ’ οὐρανοῦ μεγάλα ἔσται.
 fear.NOM.PL CONJ CONJ sign.NOM.PL APO heaven.GEN great.NOM.PL be.FUT.3SG
 ‘fearful sights and great signs shall there be from heaven.’
 (Luke 21:11)
- (9) ἕτεροι δὲ πειράζοντες σημεῖον ἐξ οὐρανοῦ ἐζητοῦν παρ’ αὐτοῦ.
 others PRT try.PART.PRES sign EK heaven.GEN seek.IMP.3PL PARA him.GEN
 ‘And others tempting him sought of him a sign from heaven.’
 (Luke 11:16)

In contrast to *οὐρανός* and *κόσμος* and also *στόμα* (‘mouth’), all frequent with *ἐκ*, the noun *σταυρός* (‘cross’) is very hard to conceptualize as a container. As a result, use of *ἐκ* would be unexpected, and indeed it occurs four times only with *ἀπό* (‘coming down from the cross’). This example points towards *ἐκ* not being used in the ablative domain so much.

Consider the noun *ἄνθρωπος* (‘man’) as our final example. This noun provides a valid illustration of the overlap between *ἀπό* and *ἐκ* as it occurs with both

prepositions in very similar contexts involving motion verbs. Similar observations hold for the noun θεός ('god'). The following two examples illustrate this for the verb ἐξέρχομαι ('go out'):

- (10) Ὅταν τὸ ἀκάθαρτον πνεῦμα ἐξέλθῃ ἀπὸ τοῦ
 when DET.NOM unclean.NOM spirit.NOM go.out.SUBJ.AOR.3SG APO DET.GEN
 ἀνθρώπου, ...
 man.GEN
 'When the unclean spirit is gone out of a man, ...' (Luke 11:24)
- (11) ἔλεγεν γὰρ αὐτῷ, Ἵξελθε τὸ πνεῦμα τὸ
 say.IMP.3SG ADV him.DAT go.out.IMPERA.AOR.2SG DET.VOC spirit.VOC DET.VOC
 ἀκάθαρτον ἐκ τοῦ ἀνθρώπου.
 unclean.VOC EK DET.GEN man.GEN
 'For he said to him: "Unclean spirit, come out of the man"' (Mark 5:8)

In sum, the more qualitative inspection of the lexical profiles of ἀπό and ἐκ confirms the results from the quantitative analysis in the previous section. On the one hand, we find overlap in the lexical complements of the two prepositions, with which they appear in (almost) identical contexts. On the other hand, we also find nouns with a strong preference for one of the two prepositions. In the latter cases, this preference can often be traced back to the semantic nature of the noun and the original semantic profile of the preposition.

5. Conclusions

In this article we took a quantitative approach to NT Greek prepositional semantics. Using extensively annotated corpus data from the PROIEL corpus, we explored the distribution of ἀπό and ἐκ by way of a semantic map generated without semantic pre-analysis, but rather on the basis of four parallel language samples. We then used statistical techniques to interpret this map. Our results offer a better understanding of these two prepositions in NT Greek: We find that there is still a fairly clean separation between ἐκ and ἀπό largely based on semantic role. However, the data also showed some semantic overlap between the two prepositions, in particular the use of ἀπό in the elative domain. A lexical analysis clarified some of the details of such uses of ἀπό, in particular it is clear that it amounts to the preposition specializing with certain lexical items, some of them with variable interpretations, as seen in the case of toponyms. We believe that other cases of structural alternations in Ancient Greek may benefit from similar treatment in light of the rapidly improving corpus resources now available.

Acknowledgements

The work on which this paper is based was started when Olga Thomason and Peter de Swart were visitors of the PROIEL project at the Department of Philosophy, Classics, History of Arts and Ideas at the University of Oslo. We would like to thank the editors of this special issue, two anonymous reviewers, Dag Haug and the other members of the PROIEL project, and the audience at ‘Ancient Greek and Semantic Theory’ for valuable feedback.

Peter de Swart received financial support from the Netherlands Organization for Scientific Research (NWO) (grant 275-89-003, The Status of Hierarchies in Language Production and Comprehension), which is hereby gratefully acknowledged.

References

- Baayen, R. Harald. 2008. *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge: Cambridge University Press.
- Borg, Ingwer and Patrick J.F. Groenen. 2005. *Modern Multidimensional Scaling, theory and applications, 2nd edition*. New York: Springer.
- Bortone, Pietro. 2010. *Greek prepositions: From antiquity to the present*. Oxford: Oxford University Press.
- Clancy, Steven J. 2006. The topology of Slavic case: Semantic maps and Multidimensional Scaling. *Glossos*, 7(1): 1–28.
- Croft, William and Keith T. Poole. 2008. Inferring universals from grammatical variation: Multidimensional Scaling for typological analysis. *Theoretical Linguistics* 34(1): 1–37.
- Cysouw, Michael. 2010. Semantic maps as metrics on meaning. *Linguistic Discovery* 8(1): 70–95.
- Eckhoff, Hanna M., Olga Thomason and Peter de Swart (2011). A map of the source domain: Evidence from old Indo-European parallel data. Unpublished manuscript Oslo/Georgia/Nijmegen.
- Haspelmath, Martin. 1997. *Indefinite pronouns*. Oxford: Oxford University Press.
- Haspelmath, Martin. 2003. The geometry of grammatical meaning: Semantic maps and cross-linguistic comparison. In: Michael Tomasello (Ed.), *The new psychology of language*, pp. 211–242. Mahwah, NJ: Lawrence Erlbaum.
- Haug, Dag Trygve Truslew, Hanne Eckhoff, Marek Majer, and Eirik Welo. 2009. Breaking down and putting back together: analysis and synthesis of New Testament Greek. *Journal of Greek Linguistics* 9: 56–92.
- Hosmer, David W. and Stanley Lemeshow. 1989. *Applied logistic regression*. New York: Wiley.
- Hothorn, Torsten, Kurt Hornik, Carolin Strobl, and Achim Zeileis. 2010. A laboratory for recursive partitioning. Party Package manual.
- Jackendoff, Ray. 1983. *Semantics and cognition*. Cambridge: the MIT Press.
- Jackendoff, Ray. 1990. *Semantic structures*. Cambridge: the MIT Press.
- Klein, Jared S. 1992. On the idiomatic nature of the Gothic New Testament: A comparative study of prepositional usage in Gothic and New Testament Greek. *Transactions of the Philological Society* 90: 1–80.
- Lestrade, Samuel. 2010. *The space of case*. Ph.D. thesis, Radboud University Nijmegen.
- Luraghi, Silvia. 2003. *On the meaning of prepositions and cases: The expression of semantic roles in Ancient Greek*. Amsterdam: John Benjamins.

- Luraghi, Silvia and Pierluigi Cuzzolin. 2007. Mediating culture through language: Contact-induced phenomena in the early translations of the Gospels. In Paolo Ramat and Elisa Roma (Eds.), *Europe and the Mediterranean as linguistic areas*, pp. 133–158. Amsterdam/Philadelphia: John Benjamins Publishing Company.
- R Development Core Team. 2009. *R: A language and environment for statistical computing*. Vienna, Austria. URL: <http://www.R-project.org>: R Foundation for Statistical Computing.
- Shih, Stephanie. and Jason Grafmiller. 2011. Weighing in on end weight. Paper presented at the LSA 2011 Annual Meeting.
- Strobl, Carolin, James Malley, and Gerhard Tutz. 2009. An introduction to recursive partitioning: Rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological Methods* 14(4): 323–348.
- Talmy, Leonard. 2000. *Towards a cognitive semantics*. Cambridge, MA: the MIT Press.
- Thomason, Olga. 2006. Prepositional Systems in Biblical Greek, Gothic, Classical Armenian, and Old Church Slavic. Ph.D. thesis, University of Georgia.
- von Tischendorf, Constantin. 1869–1872. *Novum Testamentum Graece*. Leipzig: Hinrich.
- Wälchli, Bernhard. 2010. Similarity semantics and building probabilistic semantic maps from parallel texts. *Linguistic Discovery* 8(1): 331–371.
- Zaenen, Annie, Jean Carletta, Gregory Garretson, Joan Bresnan, Andrew Koontz-Garboden, Tatiana Nikitina, M. Catherine O'Connor, and Tom Wasow. 2004. Animacy encoding in English: Why and how. In Bonnie Webber and Donna Byron (Eds.), *Proceedings of the 2004 ACL Workshop on Discourse Annotation*, pp. 118–125. Association for Computational Linguistics.