Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA 2022), pages 73–83
Language Resources and Evaluation Conference (LREC 2022), Marseille, 25 June 2022
© European Language Resources Association (ELRA), licensed under CC-BY-NC 4.0

In Search of the Flocks: How to Perform Onomasiological Queries in an Ancient Greek Corpus?

Alek Keersmaekers, Toon Van Hal
University of Leuven
Blijde-Inkomststraat 21, 3000 Leuven, Belgium
{alek.keersmaekers; toon.vanhal}@kuleuven.be

Abstract

This paper explores the possibilities of onomasiologically querying corpus data of Ancient Greek. The significance of the onomasiological approach has been highlighted in recent studies, yet the possibilities of performing 'word-finding' investigations into corpus data have not been dealt with in depth. The case study chosen focuses on collective nouns denoting animate groups (such as *flocks* of people, *herds* of cattle). By relying on a large automatically annotated corpus of Ancient Greek and on token-based vector information, a longlist of collective nouns was compiled through morpho-syntactic extraction and successive clustering procedures. After reducing this longlist to a shortlist, the results obtained are evaluated. In general, we find that πλῆθος can be considered to be the default collective noun of both humans and animals, becoming especially prominent during the Hellenistic period. In addition, specific tendencies in the use of collective nouns are discerned for specific semantic classes (e.g. gods and insects) and over time. Throughout the paper, special attention is paid to methodological issues related to onomasiologically searching.

Keywords: onomasiology, data querying, collective nouns, Ancient Greek

1. Introduction

This paper explores the possibilities of onomasiologically querying corpus data of Ancient Greek. The significance of the onomasiological approach has been highlighted in recent studies, yet the possibilities of performing 'word-finding' investigations into corpus data have not been dealt with in depth. English has a wide range of words denoting groups of animals or people, such as a "pack of dogs", "a school of fish" and "a gang of bandits". This paper aims to explore how similar collective nouns can be detected in the Ancient Greek corpus by adopting an onomasiological approach to the data.

The paper is organized as follows. A survey of the state of the field (Section 2) precedes an outline of our strategies adopted to tracing collective nouns in Greek (Section 3). Section 4 analyzes various groups of animate entities in Ancient Greek by means of corpus data and discusses onomasiological change in Ancient Greek. In the concluding part (Section 5), alternative approaches and further avenues are discussed. The case studied in this paper has identifiable morpho-syntactic characteristics (see Section 3.2), but in the future it should be also made possible to find words expressing a certain concept for which the availability of syntactical and morphological annotation is not helpful.

2. State of the field

2.1 Onomasiological searching

Corpus-based research is usually based on a (set of) predefined term(s), of which the meaning is traced. In addition to this semasiological or 'sense-finding' approach, it is also conceivable to take a certain meaning (concept or notion) as a starting point, and examine which terms are used to shape this meaning in a corpus. In recent decades, linguists have strongly emphasized the importance and relevance of such an onomasiological or 'word-finding' approach (see e.g. Grzega; 2002; Geeraerts, 2009; Fernández-Domínguez, 2019), and more recently there have been increasing advocates of the onomasiological approach among conceptual historians too (see e.g. Müller & Schmieder, 2016; Cananau, 2019). For obvious reasons, querying corpora with a semasiological, word-based approach is much easier than meaning-based onomasiological queries, because unlike a meaning a term is a tangible starting point. In a methodological survey paper published against the backdrop of a corpus-based computational historical semantics project, Bernhard Jussen and Gregor Rohmann mention the onomasiological approach, yet the case-studies they present are semasiological in nature (Jussen & Rohmann, 2015). While there has been some research on querying onomasiological dictionaries (Kipfer, 1986; Sierra, 2008; Moerdijk et al., 2008 on the development of 'semagrams'), the literature on how to onomasiologically querying corpora is limited (see McGillivray, 2020; see also Kutuzov, 2020). In general, onomasiological search strategies generally boil down to making use of annotations that approximate the concept under investigation as much as possible, the results of which are complemented through bottom-up approaches (see e.g. Goossens, 2013). Hence, this presupposes the presence of an annotated corpus, which is a demanding and time-consuming investment, if such a corpus is not yet available (see Mehl, 2016: 50; 92; Atallah et al., 2018). The type of annotations required depends on the onomasiological task at stake. For certain tasks, part-of-speech tags can be helpful, while for other tasks more detailed morphological, syntactic, semantic and/or pragmatic information is needed.

2.2 Collective nouns

Words as 'flock' and 'herd' are styled quantifying collectives and collective nouns by Biber et al. (2003: 61-62). The terms have been criticized for being too vague (see the references in Dedè, 2012). Some scholars have treated collective nouns as classifiers (or 'classifier constructions', cf. Lehrer, 1986). Aikhenvald (2000: 115-116) however argues why such terms do not meet the criteria of genuine

classifiers. Zwarts (2020) distinguishes 'crowd' and 'club' words. The first type of collective nouns has its starting point in a number of *individuals*, which are spatially so closely associated to each other that a line can be drawn to establish the collective (dynamic **a** → **b** in Fig. 1). Conversely, club words have their starting point in the *whole*, which is open to individual members (dynamic **c** → **b** in Fig. 1.).

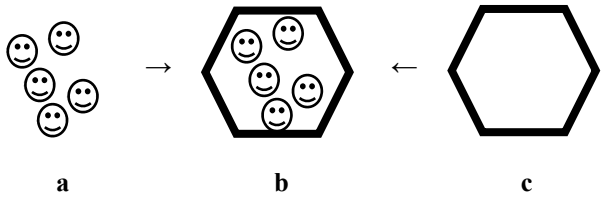

Figure 1: The distinction between crowd words (a → b) and club words (c → b) after Zwarts (2020: 539)

Collective nouns have attracted much scholarly attention for their behavior in subject agreement: in the singular, they typically refer to more than one entity, while they can be combined both with plural and singular verbs (see Birkenes & Sommer, 2014). In the Anglo-Saxon tradition, collective nouns are even defined on the basis of this concord criterium, whereas in continental research strands semantic criteria prevail (cf. Joosten et al., 2007: 88).

Many modern European languages have a proliferation of, often highly specialized or idiosyncratic, animal collective nouns (such as "a murder of crows" and "a rout of wolves" in English) — a phenomenon having its roots in middle age hunting practice and Books of Courtesy (Rhodes, 2014). To the best of our knowledge, there has been so far no systematic research into the range of collective nouns in Ancient Greek, even though the 'oldest grammar of the West' by (Pseudo-)Dionysius Thrax already defines the collective noun (περιληπτικὸν) as τὸ τῷ ἑνικῷ ἀριθμῷ πλῆθος σημαῖνον ("signifying a multitude in the singular number"), offering the examples of δῆμος, χορός and ὄχλος (see Swiggers & Wouters, 1998). In most grammars, collective nouns are mainly discussed against the background of subject agreement, with only a few examples offered (see e.g. Kühner-Gerth, 1966: §359, more extensive treatment in Viteau, 1896: 103-11). The research undertaken by Birkenes & Sommer (2014) is limited to a very small number of collective nouns in Ancient Greek. A number of contributions aim to prove that the etymology of certain Greek words suggests a past of a collective noun (see e.g. Leroy, 1956; Kaczyńska, 2019), while others examine specific terms or a limited set of collective nouns, mostly in contexts other than linguistics (see e.g. Dieckhoff, 2018). The following section explores therefore how one can computationally trace the equivalents used in Ancient Greek to express such notions.

3. Identifying collective nouns in Ancient Greek

3.1 Starting points of the research

As pointed out in 2.1, onomasiological queries highly benefit from corpus annotations. This is especially true for Ancient Greek, a language with a highly flexible word order and complex inflectional morphology, which reduces the effectiveness of strictly form-based (as opposed to lemma-, morphology- and syntax-based) queries. For Ancient Greek, the most well-known corpora are the Greek treebanks (several annotators, consisting of dependency trees with syntactic, morphological and lemma annotation: see Celano, 2019 and Keersmaekers et al., 2019), which are manually annotated but not extremely large (1.5M tokens), and the Diorisis corpus (a corpus that is annotated for lemmas and morphology, cf. Vatri and McGillivray, 2018), which is relatively sizable (10.2M tokens) but is automatically annotated and does not contain syntactic information. We therefore made use of the (so far unreleased) GLAUx corpus (Keersmaekers, 2021), a corpus containing literary (8th century BC-3th century AD) and documentary texts (3th century BC-8th century AD) automatically annotated for lemmas, morphology and syntax (28.8M tokens): see Keersmaekers (2021) for an evaluation of the quality of the annotation, which was high enough not to provide any substantial obstacles for the research described below.

Although some steps for semantic annotation of Greek have been taken (see Celano & Crane, 2015 and Keersmaekers, 2020 for semantic role annotation; Bizonni et al., 2014 and Biagetti et al., 2021 for Ancient Greek WordNet), so far no large-scale semantically annotated corpus resource for Ancient Greek with the level of granularity that is necessary for the research described in this paper has been created. We therefore made use of a bottom-up approach that has become highly popular in recent years to represent semantics computationally, the so-called 'distributional' approach to semantics, where meaning is represented by vectors of real numbers (with semantically similar words or constructions receiving mathematically similar vectors). These vectors are based on the context patterns of words in large text corpora (see Erk, 2012; Lenci, 2018 for more detail). Distributional semantic methods have been applied to Ancient Greek by Boschetti (2010), Rodda, Senaldi & Lenci (2017), Rodda, Probert & McGillivray (2019), Keersmaekers (2020), Keersmaekers & Van Hal (2021), and Perrone et al. (2021). For this paper we use the implementation of Keersmaekers & Van Hal (2021), which calculates word vectors on the basis of PPMI-scaled syntactic dependency-based co-occurrence counts in the GLAUx corpus, with an SVD-based dimension reduction to 100 latent dimensions.

3.2 Morpho-syntactic extraction

In Greek, collective nouns are syntactically well-defined, since they are usually accompanied by a so-called partitive genitive (Benvenuto, 2013). Based on the GLAUx corpus, we could extract all constructions of type 'noun + animate entity in the genitive plural, having 'attribute' as its syntactic feature'. The animacy was determined via supervised machine learning techniques, training a deep learning model[1] on data annotated for the animacy class of the lemma as the dependent variable and a 100-dimensional word vector of the lemma (as described in 3.1) as the independent variable(s) (see Keersmaekers, 2020: 103-116). Our training data was an animacy lexicon containing

[1] As implemented in R package *h2o* (LeDell et al., 2022), using a multi-layer feedforward artificial neural network trained with stochastic gradient descent using back-propagation.

486 animate and 2650 inanimate entities; animate entities yielded precision of 0.941 and recall of 0.914 – an estimation via 10-fold cross validation on the training data. On this basis, 1991 lemmas were labeled as animate, which allowed us to identify possible collective nouns.

Our approach is not infallible: in addition to possible errors in the automatically annotated data, we should note that there are a number of alternative constructions that can express collective nouns and that are not included in the extracted data. For example, the genitive can sometimes be replaced with an adjective (e.g. the LSJ dictionary of Ancient Greek, Jones et al., 1996, cites μελισσαῖος οὐλαμός "a swarm of bees", with the adjective μελισσαῖος 'consisting of bees'). Some collective nouns have no (need for) further attributive specification – especially if the animal is already lexicalized in the collective noun itself (e.g. βουκόλιον "a group of cows", συβόσιον "a group of pigs", αἰπόλιον "a group of goats"). Obviously, plural morphology might also be used to indicate a group of animate entities (e.g. simply αἶγες 'goats' instead of αἰπόλιον αἰγῶν 'a flock of goats'). Finally, constructions with a genitive in the singular are conceivable (e.g. 'a swarm of vermin' in English). Of the extracted lemmas (5488), only lemmas with a frequency of ≥ 5 (frequency of the lemma accompanied by an animate genitive plural) were retained (1266 in total). These lemmas thus count as potential collective nouns, out of which we attempted to identify the real collective nouns using several computational techniques.

3.3 Visualization and clustering techniques

The query defined in section 3.2 likely has a high recall, since we expect most collective nouns to occur in the construction defined there, even though there are some other ways to express groups of animate entities as discussed above. However, its precision is rather low, since many nouns occurring in the noun + animate genitive plural construction are not collective nouns: this construction admits many more types of nouns such as body parts (e.g. 'the legs of the horses'), possession relations (e.g. 'the money of the men') and so on. To retrieve collective nouns from this large set (1266 nouns), we used a variety of dimension reduction and clustering techniques to find structure in our dataset, as well as lexicographical data (the LSJ dictionary, Jones et al., 1996) and corpus examples from the GLAUx corpus (in case of doubt) to identify collective nouns in these structured datapoints. The dimension reduction and clustering techniques were applied to the cosine distances between the nouns in our dataset, which mathematically represent the 'semantic distances' between the nouns (see Erk, 2012: 636-637).

As a first step, we made use of t-SNE (t-distributed stochastic neighbor embedding, Van den Maaten and Hinton, 2008), a dimension reduction technique that allows us to represent high dimensional data (a 1266x1266 matrix representing the cosine distances between nouns) in a low-dimensional (in our case two-dimensional) space, with words that are similar in meaning occurring close to each other on the tsne-map.[2] This enables us to find structure in the data and identify which words are worth looking at to retrieve collective nouns. For instance, the cluster on the bottom right of Fig. 2 (in dark yellow) contains words that clearly refer to body parts (e.g. μῦς 'muscle', γαστήρ 'belly', θρίξ 'hair'). It is unlikely that a collective noun would occur in such a cluster, so these words can safely be discarded after identifying the thematic coherence of the cluster. Instead, on the bottom/center-left of the plot there are several clusters that clearly contain many collective nouns: military units (in red: e.g. ἴλη, λόχος, οὐλαμός), words referring to herding (in dark blue, with several words that mean 'a flock or herd', such as αἰπόλιον, ἀγέλη, πῶυ, but also some non-collective nouns such as νομεύς 'herdsman') and a small cluster of words referring to groups in general (in yellow: πλῆθος, ὄχλος, πληθύς, ὅμιλος, ἑσμός, σμῆνος); additionally, a little more doubtful are the clusters in pink (generally containing words related to transport such as ἅμαξα 'wagon', φορτίον 'load' and ἵππος 'horse' but also some collective nouns such as συνωρίς 'pair of horses' and κτῆνος 'beast', but also 'flock'), dark green (mainly poetic words referring to family such as φῦλον 'tribe', but maybe also 'swarm', γένεθλον 'family', but also unrelated poetic words such as σημάντωρ 'leader') and light blue (two words referring to the action of collecting or coming together but maybe also to a collection or group, viz. ἄθροισμα and συνδρομή). After identifying these clusters, we used dictionaries and corpus data to check whether each word occurring in these clusters is actually a collective noun.

Next, we used two cluster techniques that are prevalent in corpus linguistics to identify additional nouns that we may have missed with the t-SNE analysis, viz. partitioning around medoids (PAM) and hierarchical agglomerative clustering (AGNES).[3] The former technique divides the data into a predetermined number (*k*) of clusters. After experimenting with the values for k, in the end we settled for a small number of *k*=20 clusters. The latter technique hierarchically clusters all nouns into a tree, with similar words occurring in the same 'branches' of the tree – a subpart of the tree, containing many collective nouns, is shown in Fig. 3. As with the t-SNE analysis, we analyzed the thematic coherence of each cluster that was formed (in the case of PAM, simply each of the 20 clusters; in the case of AGNES, branches of the tree occurring roughly at the same height), and looked into more detail at the more 'promising' clusters containing many collective nouns. These techniques allowed us to identify some additional collective nouns that we had previously missed: these were especially words in the festive or public domain including θίασος, χορός and σύλλογος, along with some words thematically related to the words we previously found such as σύστημα (a military unit, or also a group in general) and νέφος (literally 'cloud', but also a group of people or animals).

[2] We made use of the R package *Rtsne* (Krijthe and Van der Maaten, 2018). We used a perplexity of 5, theta of 0.0 and 5 iterations.

[3] As implemented in R package *cluster* (Maechler et al., 2022). We used out-of-the-box settings.

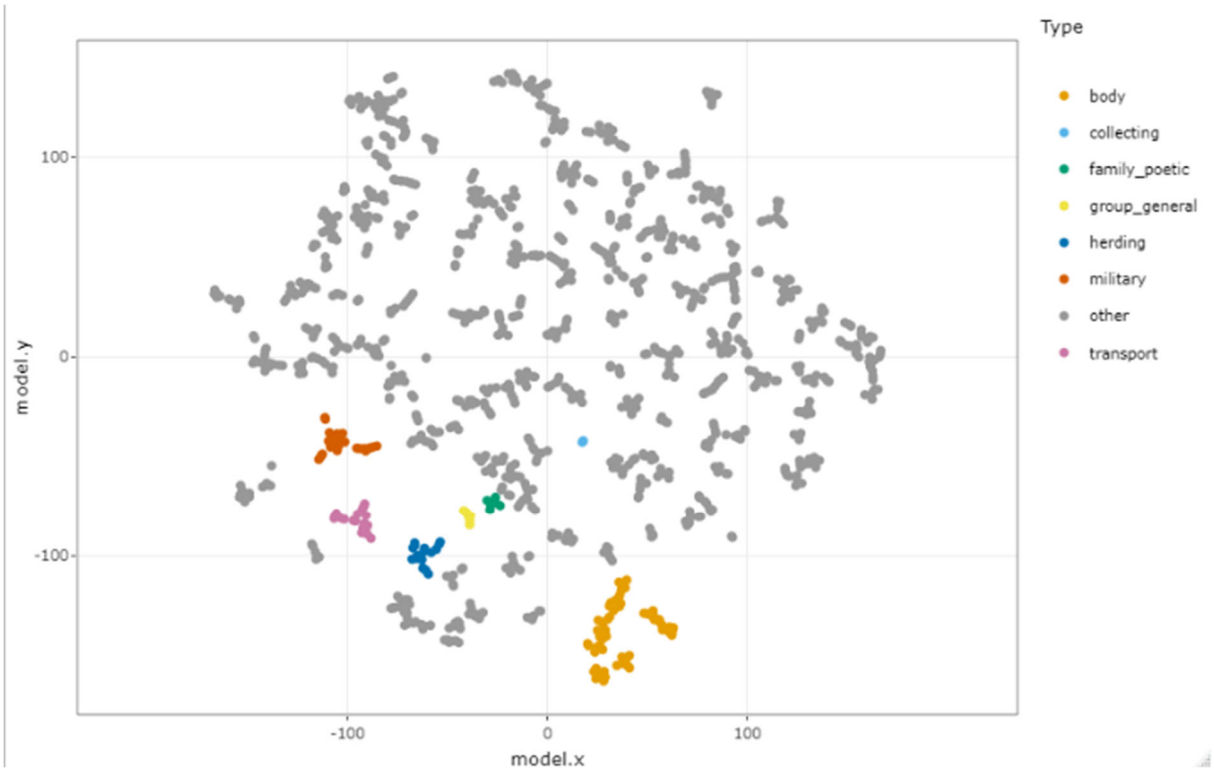

Figure 2: Visualization of the t-SNE embeddings

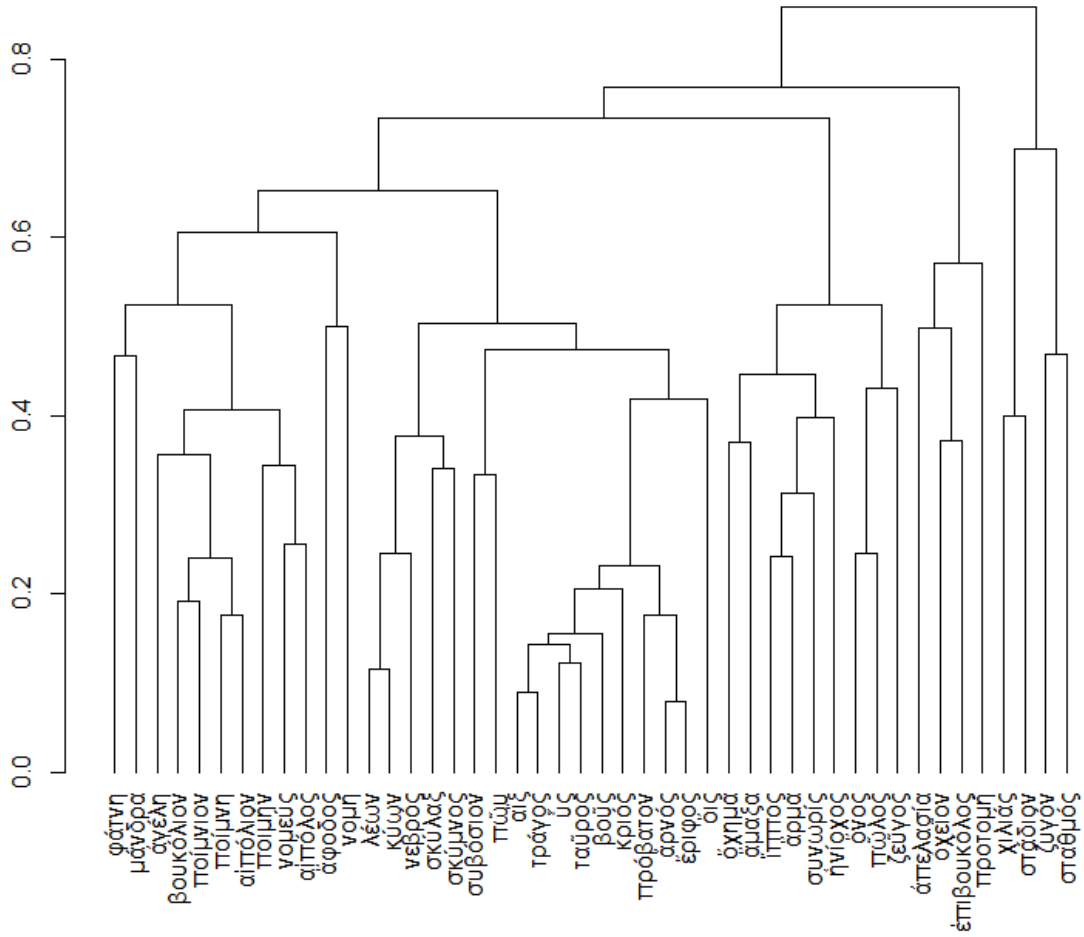

Figure 3: Subpart of the tree of hierarchical agglomerative clustering (AGNES)

4. Results and discussion

4.1 Discussion of shortlist in general

In total, we have traced 40 collective nouns (see Table 1). The dataset on which this paper is based is available through a csv-file.

5 collective nouns appeared more than 100 times in the data. At the top of the list, without a doubt, is the word πλῆθος (token frequency 998), which can be regarded as the default collective noun for animate referents. There is no semantic class in which this collective noun does not occur. In some cases, πληθύς is used too, which is according to most dictionaries merely an Ionian variant (this, however, should be checked against the data). πλῆθος is followed by ἀγέλη (215), φῦλον (129), τάξις (125) and χορός (109). Some words are exclusively used as collective nouns for animals, such as πῶυ (10), αἰπόλιον (8), βουκόλιον (7) and συβόσιον (5), while words that occur only with human referents are more numerous, viz. λόχος (37); τάγμα (34); σύνταγμα (22); σύλλογος (12); θίασος (10); οὐλαμός (10); ἄθροισμα (6).

	total		animal		human	
πλῆθος	998	40%	331	41%	667	40%
ἀγέλη	215	9%	184	23%	31	2%
φῦλον	129	5%	37	5%	92	6%
τάξις	125	5%	8	1%	117	7%
χορός	109	4%	8	1%	101	6%
ὄχλος	84	3%	1	0%	83	5%
ἴλη	74	3%	5	1%	69	4%
ὅμιλος	70	3%	4	0%	66	4%
ζεῦγος	68	3%	60	7%	8	0%
λόχος	37	1%	0	0%	37	2%
συναγωγή	37	1%	3	0%	34	2%
νομή	36	1%	31	4%	5	0%
φάλαγξ	35	1%	2	0%	33	2%
σύνοδος	34	1%	2	0%	32	2%
τάγμα	34	1%	0	0%	34	2%
σύστημα	32	1%	4	0%	28	2%
στρατιά	30	1%	1	0%	29	2%
στρατός	28	1%	2	0%	26	2%
πληθύς	27	1%	6	1%	21	1%
στῖφος	25	1%	1	0%	24	1%
σύνταγμα	22	1%	0	0%	22	1%
σμῆνος	22	1%	20	2%	2	0%
ἐκκλησία	20	1%	1	0%	19	1%
συνέδριον	20	1%	1	0%	19	1%
ἑσμός	17	1%	15	2%	2	0%
νέφος	16	1%	13	2%	3	0%
σπεῖρα	13	1%	3	0%	10	1%
ἄγημα	12	0%	2	0%	10	1%
συνωρίς	12	0%	11	1%	1	0%
σύλλογος	12	0%	0	0%	12	1%
ποίμνιον	12	0%	9	1%	3	0%
ποίμνη	11	0%	9	1%	2	0%
θίασος	10	0%	0	0%	10	1%
πῶυ	10	0%	10	1%	0	0%
οὐλαμός	10	0%	0	0%	10	1%
αἰπόλιον	8	0%	8	1%	0	0%
κτῆνος	8	0%	7	1%	1	0%
βουκόλιον	7	0%	7	1%	0	0%
ἄθροισμα	6	0%	0	0%	6	0%
συβόσιον	5	0%	5	1%	0	0%

Table 1: Collective nouns denoting animals and humans

It is important to point out which collective nouns are not included in our data, and why. The word δῆμος, also cited by Dionysius Thrax, was not clearly identified in the cluster techniques applied. Related to δῆμος are words like λεώς and ἔθνος, all of which first of all refer to a 'people' or 'tribe' rather than to a 'group'. This however implies that Homeric collocations such as ἔθνεα [...] μελισσάων ("clouds of bees", Il. 2.87-89) are not captured in the data.[4] Some words designating 'flock' or 'group', such as βοτά and κῶμος mentioned in the Woodhouse English-Greek dictionary (Woodhouse, 1987), turn out to be very infrequent in our data. In addition, we have deliberately excluded a fairly long list of words which did turn up via our methodology, but (a) where inspection of the examples showed that only a minority of cases could count as a collective noun or (b) where its status as a collective noun is more doubtful. The cases in question are the following: ἀγών; ἀποσκευή; βόσκημα; βουλή; γένεθλον; γέννα; δεκάς; διατριβή; δικαστήριον; διλοχία; ἑορτή; ἱππαρχία; κατάλογος; λεία; μόρα; ὁμιλία; πομπή; πρόβατον; συνουσία; συσσίτιον; σχολή; χιλιάς; χρῆμα. A case of (a) is συνουσία 'company, intercourse': although there are some corpus examples which may allow for a collective reading, in the vast majority of cases it rather means 'the state of being together' (or 'sexual intercourse'); a case of (b) is δικαστήριον 'court', which is a club word (like many other words in this list) referring to a group of judges. Although we included several club words in our shortlist, we excluded those where people assemble for a highly specialized purpose, in this case making a judicial decision. It should be emphasized that by eliminating these words (as well as the δῆμος and equivalents mentioned above) we have eliminated some clear examples of collective nouns. However, we felt that the inclusion of these words would give way to considerable noise in the data. Conversely, it should be noted that not all instances included in the shortlist unambiguously refer to a collective noun. This is certainly the case for a word like τάξις, which is very polysemous (e.g. also 'order', 'class', 'rank' etc.). Due to the scale of our undertaking, it was infeasible to inspect the data token-wise. We are however aware that our type-based approach is vulnerable to noise.

4.2 Classes of animals and humans

Next, we divided the lemmas of animals and people in a number of subgroups. These subgroups were semi-automatically created: through hierarchical clustering (AGNES) of the vectors of these lemmas, we first checked which of them were highly semantically related and created subgroups on this basis, but the final groups were created with a high degree of human control (e.g. if the cluster algorithm would cluster a specific fish with a bird together, we would put this fish in the group 'water animals and fish' rather than 'birds'). Fig. 4 shows the frequency of collective nouns with the most frequent groups of animals,

[4] The data in Johnston (2019) suggests that collective nouns are rare in Homeric similes.

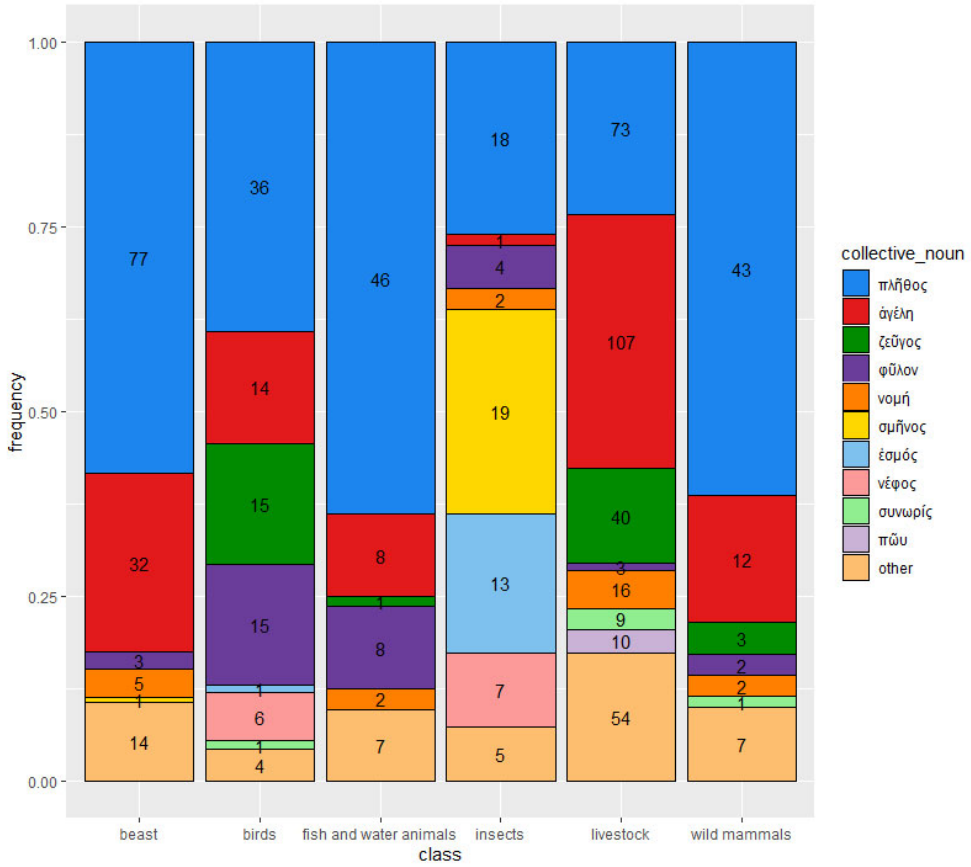

Figure 4: Variation in the presence of animal collective nouns according to semantic subgroups

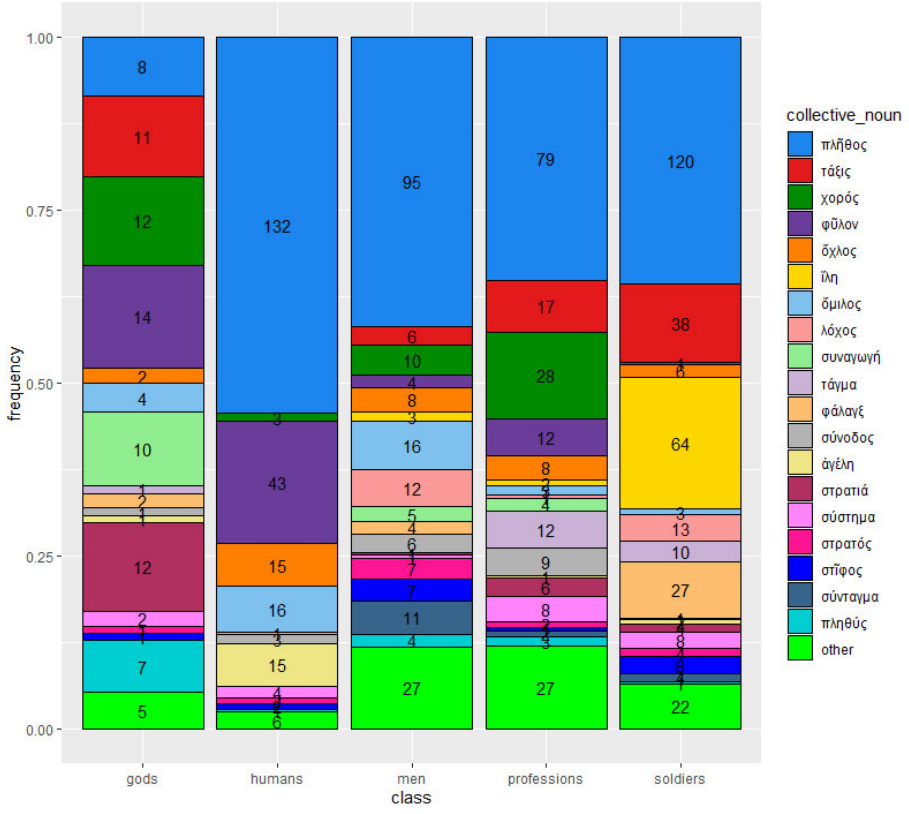

Figure 5: Variation in the presence of human collective nouns according to semantic subgroups

viz. animals in general, birds, water animals and fish, insects, livestock and wild mammals. The default word πλῆθος is used in all categories, although remarkably less frequently in the case of livestock, insects and, to a lesser extent, birds. It is also notable that ἀγέλη (often translated as 'flock') is used in every subgroup, and not exclusively for livestock. In the category of insects, the subgroup where ἀγέλη is underrepresented, one can notice the use of a number of specific collective nouns that are almost exclusively used for insects, viz. ἑσμός, σμῆνος and νέφος, the latter of which is also used for birds (see also 4.3). This specialization seems to be rather atypical: φῦλον, νομή and, to a lesser extent, ζεῦγος are animal collectives that can be used for almost any subgroup. One should also notice the high degree of 'other' with livestock: besides the 'default' options of ἀγέλη and πλῆθος, Ancient Greek has a large number of specialized words for livestock (e.g. βουκόλιον 'group of cows')."

The humans can also be divided into a number of subgroups. Fig. 5 makes a distinction between the most frequent subgroups, viz. gods, humans in general, men, professions and soldiers. The data shows that in case of the gods certain collective nouns, viz. χορός and φῦλον, outnumber the 'default' use of πλῆθος. Among the military category one finds the most specialized collective nouns, such as τάξις, ἴλη, λόχος and φάλαγξ. Strikingly, στρατιά and στρατός ('army') are hardly represented in this category. In general, it is noticeable that there are plenty combinatory possibilities.

4.3 Degree of specialization

This leads us to the question of to what extent there are exclusive combinations in Greek, showing one-to-one correspondences between a specific collective noun and a specific animate type, like the English 'murder of crows'. Table 2 shows the top results of a Pointwise Mutual Information (PMI) calculation, a measure of association showing whether two variables co-occur more frequently than expected based on their individual frequencies (see Gries 2010: 275-277 for more detail). We have only included collocational combinations that occur at least five times. The results show that in some cases there is a clear etymological connection between the collective noun and the species at stake (συβόσιον, αἰπόλιον, βουκόλιον), thus logically excluding alternative combinations (such as βουκόλιον and αἴξ). Excluding these words and the Homeric word πῶυ, which is exclusively used for ὄις 'sheep', it seems that especially specific insects (μέλισσα 'bee'; ἀκρίς 'grasshopper', the same goes for less frequent insects such as κηφήν 'drone' and σφήξ 'wasp') are combined with specific collective nouns (νέφος; ἑσμός; σμῆνος), which are rarely used for animals other than insects (except νέφος which is also often combined with birds). A group of pigeons (τρυγών or περιστερά) is mostly referred to as ζεῦγος, likely indicating a duo.

Collective		Child		Collocation	PMI
συβόσιον	5	ὗς	7	5	8.5
πῶυ	10	ὄις	12	10	7.7
αἰπόλιον	8	αἴξ	26	8	6.6
νέφος	16	ἀκρίς	12	5	6.0
τάγμα	34	λοχαγός	7	5	5.7
ἑσμός	17	μέλισσα	35	12	5.7
σμῆνος	22	μέλισσα	35	14	5.5
συναγωγή	37	υἱός	16	9	5.2
ποίμνιον	12	πρόβατον	44	8	5.2
ζεῦγος	68	τρυγών	5	5	5.2
ποίμνη	11	πρόβατον	44	7	5.2
βουκόλιον	7	βοῦς	83	7	4.9
ὅμιλος	70	μνηστήρ	6	5	4.9
φάλαγξ	35	ὁπλίτης	54	22	4.9
ὄχλος	90	οἰκότριψ	6	6	4.8

Table 2: Strongest PMI associations between collective nouns and the genitives occurring with them

Closer inspection reveals that some of the exclusive correspondences in Table 2 might be somewhat deceptive, for example, because all the attestations come from one author. This is the case for the association between ὄχλος and οἰκότριψ, which seems to be a personal style characteristic of Origenes.

4.4 Diachronic developments

In the previous sections, we mapped the onomasiology of Ancient Greek collective nouns in a static way. However, this onomasiology is obviously prone to semantic change, i.e. the terms used to express groups of animate entities change over time. This section will consider how computational methods can shed light on this onomasiological change. To this aim, we have divided the data into archaic (8th-6th century BC), classical (5th-4th century BC), Hellenistic (3rd-1st century BC) and Roman eras (1st-4th century AD) (see Fig. 6 and Fig. 7). However, caution is advised here: for instance, we have almost exclusively epic texts from the archaic period, so that developments between the archaic and the classical period may be explained in terms of genre rather than in diachronic terms. In the archaic period, the number of data points is very limited. For the classical period the data for the animals are also rather limited, so that the transition between the Hellenistic and Roman period especially lends itself for a study of diachronic developments.

The main evolution that can be traced with respect to the animal collectives (cf. Fig. 6) is the prominence of πλῆθος in the Hellenistic period, which clearly decreases in the Roman period. There is no clear challenger; rather, there seems to be a diversification in general, with, for example, a more frequent use of ἀγέλη, φῦλον and ζεῦγος (even though πλῆθος and ζεῦγος are likely not simply interchangeable). In a few cases we can also observe a tendency to specialization: it is especially in the Roman period that words for insects are associated with specific collective nouns, namely νέφος; ἑσμός; σμῆνος, whereas in the Hellenistic period πλῆθος is still predominating here (see Table 3).

Collective	Hellenistic	Roman
πλῆθος	9	8
ἀγέλη	0	1
φῦλον	0	3
νομή	0	1
σμῆνος	3	15
ἑσμός	1	10
νέφος	3	4
χορός	0	2

Table 3: Collective nouns used for insects in the Hellenistic and Roman period

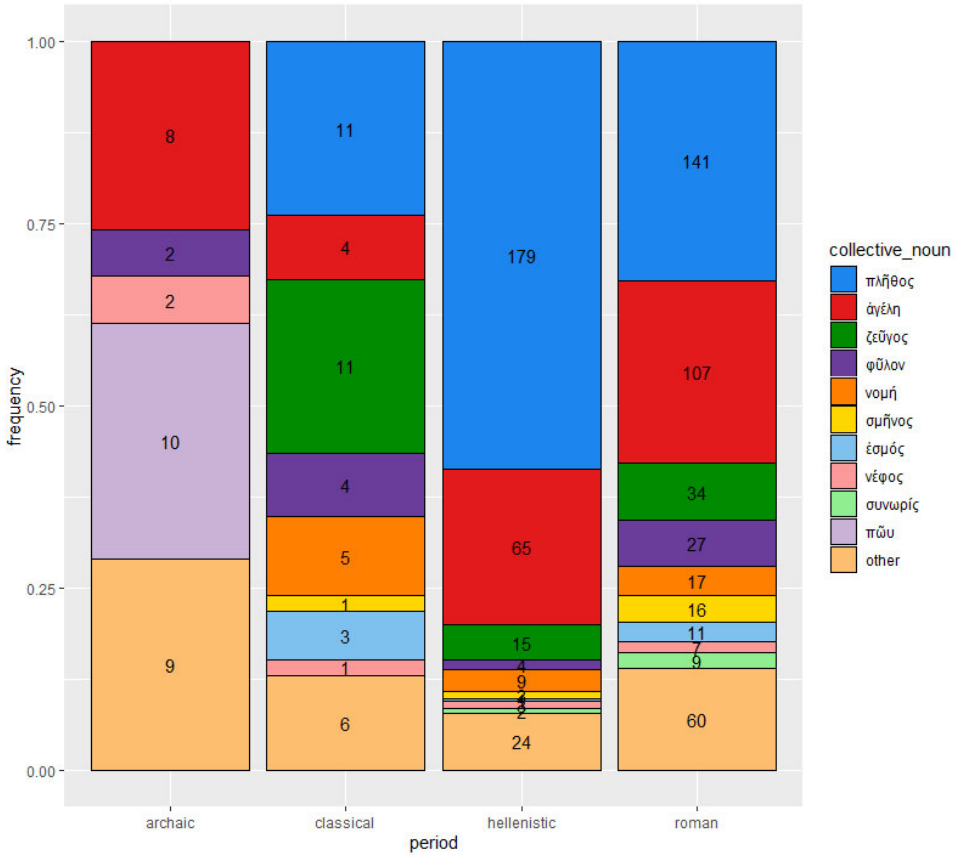

Figure 6: Evolution in the presence of animal collective nouns over time

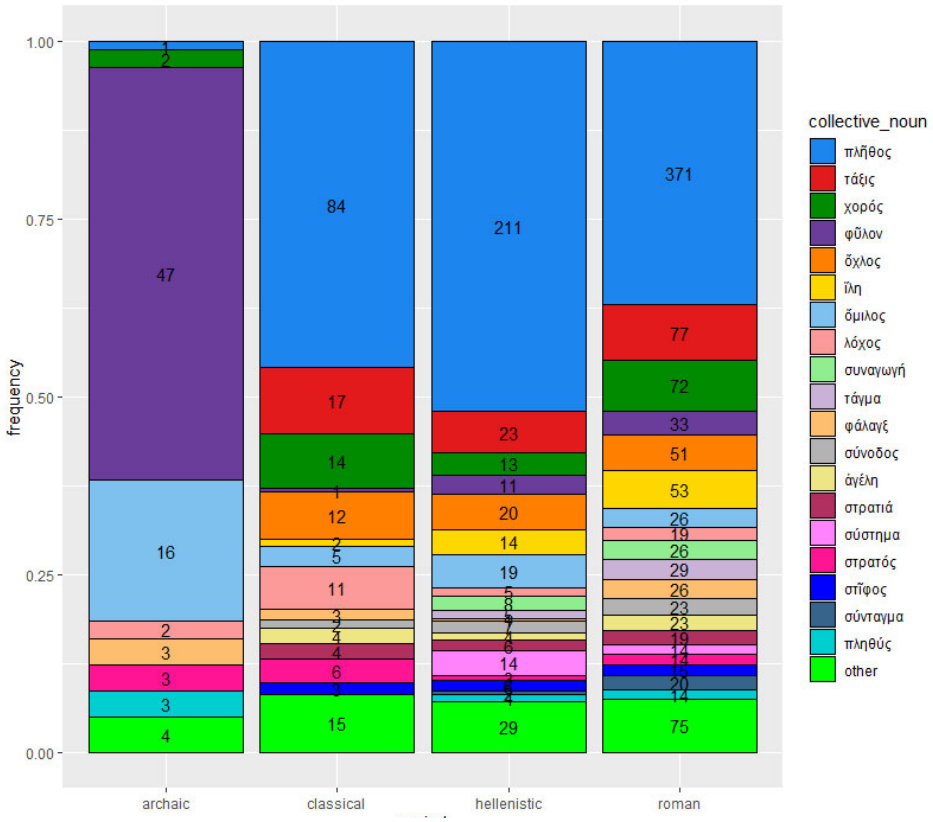

Figure 7: Evolution in the presence of human collective nouns over time

For the human collective nouns the diachronic changes are less clear (cf. Fig. 7). A number of archaic collective nouns are used much less frequently in later periods. A clear example is ὅμιλος, which in later periods is mainly taken up by a few authors (especially Philo Judaeus). Another example is φῦλον. Here again we observe the prominence of πλῆθος in the Hellenistic period, but the decline in the Roman period is less pronounced. What is particularly striking is the diachronic increase of ἴλη as a collective noun for soldiers in the Roman period (12 instances of ἴλη and 58 instances of πλῆθος in the Hellenistic period versus 52 instances of ἴλη and 53 instances of πλῆθος in the Roman period). In addition, there is a clear increase of χορός among certain 'professions' (2 instances of χορός and 19 instances of πλῆθος in the Hellenistic period versus 23 instances of χορός and 51 instances of πλῆθος in the Roman period). Inspecting the data, this is especially true when the profession has a 'didactive' or 'heralding' function, e.g. philosophers, teachers and prophets.

5. Conclusions and outlook

The syntactic/morphological-based extraction and clustering techniques have allowed us to detect a large number of collective nouns. Nevertheless, there are some important caveats. The quantitative methods used have enabled us to compile a longlist. A final manual selection, reducing the longlist to a shortlist, nevertheless remained necessary. This step involves a large degree of subjective decisions, many of which can be debated. In addition, we cannot evaluate which relevant words were not found ('recall'). Furthermore, polysemy causes any clustering technique to be problematic. The multidimensional nature of semantics implies that Ancient Greek equivalents for polysemous and idiosyncratic collocations (such as e.g. English 'murder of crows') will be difficult to identify. Some 'collective nouns' can also be frequently combined with inanimate entities (e.g. πλῆθος χρημάτων "a group or amount of money"). While these examples were filtered out during the animacy detection described in section 3.2 (i.e. we only included words with a sufficient number of animate genitive attributes in the cluster analysis in section 3.3, and similarly only analyzed words with such attributes during the corpus analysis described in 4), these contexts with inanimate entities were still included in the word vectors, and therefore might distort the results of the cluster analysis. In the future, word vectors modelling the meaning of a word in context rather than the general meaning of a word might allow for a higher degree of precision. The results could also be improved by means of an objective set of criteria whether or not a word can be considered a collective noun. Another difficulty resides in the data scarcity, which makes it very difficult to make statements about the significance of the connection between certain collective nouns and specific animals. By way of example, we see that for θύννος ('tuna'), attested thrice in the data, three different collective nouns are used: besides the generic πλῆθος, στρατός and ἴλη occur. Table 1 shows that both στρατός and ἴλη tend to be used as collective nouns of humans (especially in a military context; see 4.4). The question here is whether we are dealing with a fixed, conventional collective noun for tuna or a context-related metaphor. Obviously, close reading of the relevant passages may shed more light on the matter. For this particular case, it seems to be an occasional metaphor twice. However, if there would have been more data, it could be determined with more certainty to what extent the use of στρατός and ἴλη is rooted in context or convention. The same applies to many other lemmas, so that it is very difficult to make firm statements about which combinations were idiomatically acceptable in Greek.

There are also alternative methods possible for onomasiological queries, including searching for English translations of the concept in question through lexica (e.g. the English-Greek dictionary by Woodhouse 1987, or reverse-searching the LSJ lexicon by Jones et al. 1996) or through parallel translations, as well as using Ancient Greek WordNets – a first Ancient Greek WordNet was created by Bizzoni et al. (2014), while recently a new attempt has been undertaken by Biagetti et al. (2021). Although we could not systematically compare these approaches to the one adopted in this paper due to time and space constraints, we will briefly address the advantages and disadvantages of both through a quick exploration. Searching the Woodhouse and LSJ lexica for words such as 'flock', 'herd', 'crowd' and 'group' returned many words listed in Table 1, but also missed some (e.g. neither lexicon included νέφος under an English lemma referring to a collective noun, for example, and Woodhouse expectedly does not contain Homeric words such as πῶυ or post-classical words such as ἴλη as it is limited to the Classical Attic dialect). On the other hand, they also include words missed by our computational approach, especially low-frequent ones that we filtered out in an initial step (see 3.2), e.g. κῶμος (only 3 occurrences with an animate genitive noun). Additionally, they also reveal some alternative constructions to express a group of living beings rather than the noun + genitive construction, e.g. adjective + noun constructions such as μελισσαῖος οὐλαμός (see section 3.2) or δρακονθόμιλος συνοικία "a swarm of dragons" (Woodhouse). However, a big limitation of this approach is that it simply shifts the burden of determining which on words or constructions can express a particular concept from one language (Ancient Greek) to another one (e.g. English). For instance, the word ἄθροισμα is defined, among other definitions, as 'aggregate' in the LSJ lexicon. While 'aggregate' is certainly a collective noun in English, one must take this English term into account as one of the many possibilities to express collective nouns in order to retrieve ἄθροισμα with a lexical-based method. While the Ancient Greek WordNets seem to be less vulnerable in this respect, as they encode semantic relations between words in the target language – in this case Greek – the WordNet designed by Bizzoni et al. (2014) was in fact based on automatic linking between Greek-English lexica and therefore prone to similar problems, while the Biagetti et al. (2021) WordNet is still in active development. All these human-curated resources are also highly dependent on human judgments and the data they have considered during their developments, while the automatic approach discussed in this paper can easily take the whole Greek corpus into account (although it is fair to say that the quality of the semantic methods is highly dependent on the frequency of specific genres in the input data, see also Perrone et al. 2019).

For this first exploration of onomasiologically searching, we have deliberately chosen a case with identifiable syntactic characteristics. The challenge for future research

consists in choosing less straightforward cases, where syntactic and morphological encoding is significantly less decisive. Without doubt, one of the greatest onomasiological challenges is to trace in the Ancient Greek corpus concepts that may be present but for which lexicalized words are missing (possible examples include modern concepts such as 'queer', 'fashion', etc.).

6. Acknowledgements

This research was made possible by FWO grant 3H200733: "Language and Ideas: Towards a New Computational and Corpus-Based Approach to Ancient Greek Semantics and the History of Ideas". We would like to thank three anonymous reviewers for their stimulating criticisms.

7. Bibliographical References

Aikhenvald, A.Y. (2000). Classifiers: a Typology of Noun Categorization Devices. Oxford: Oxford University Press.

Atallah, C., Bras, M., & Vieu, L. (2018). Exploring a Corpus Annotated in Causal Discourse Relations for the Study of Causal Lexical Clues. In *Final Action Conference TextLink 2018, Mar 2018, Toulouse, France*. Retrieved from https://hal.archives-ouvertes.fr/hal-02982984.

Benvenuto, M.C. (2013). Genitive. In Encyclopedia of Ancient Greek Language and Linguistics. Leiden: Brill. Retrieved from <http://referenceworks.brillonline.com/entries/encyclopedia-of-ancient-greek-language-and-linguistics/genitive-COM_00000140>.

Biagetti, E., Zanchi, C., & Short, W.M. (2021). Toward the Creation of WordNets for Ancient Indo-European Languages. In *Proceedings of the 11th Global Wordnet Conference*. University of South Africa (UNISA): Global Wordnet Association, pp. 258-266.

Biber, D., Conrad, S., & Leech, C. (2003). *Longman Student Grammar of Spoken and Written English*. Harlow: Longman.

Birkenes, M.B., & Sommer, F. (2014). The agreement of collective nouns in the history of Ancient Greek and German. In C. Gianollo, A. Jäger & D. Penka (Eds.), Language Change at the Syntax-Semantics Interface. Berlin: De Gruyter, pp. 183-221.

Bizzoni, Y., Boschetti, F., Diakoff, H., Del Gratta, R., Monachini, M., & Crane, G. (2014). The Making of Ancient Greek WordNet. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Reykjavik: European Language Resources Association (ELRA), pp. 1140-1147.

Boschetti, F. (2010). A Corpus-based Approach to Philological Issues. Unpublished PhD Thesis. University of Trento.

Cananau, I. (2019). Toward a Comparatist Horizon in Conceptual History. History of European Ideas, 45(1), pp. 117-120. https://doi.org/10.1080/01916599.2018.1493307

Celano, G.G. (2019). The Dependency Treebanks for Ancient Greek and Latin. In M. Berti (Ed.), Digital Classical Philology. Berlin & Boston: Walter de Gruyter, pp. 279-298.

Celano, G.G. A., & Crane, G. (2015). Semantic Role Annotation in the Ancient Greek Dependency Treebank. In M. Dickinson, E. Hinrichs, A. Patejuk, & A. Przepiórkowski (Eds.), *Proceedings of the Fourteenth International Workshop on Treebanks and Linguistic Theories (TLT14)*. Warsaw, pp. 26-34.

Dedè, F. (2012). Some Remarks on the Metalinguistic Usage of the Term 'Collective'. In *Proceedings of the First Workshop on the Metalanguage of Linguistics. Models and Applications. University of Udine – Lignano, March 2-3, 2012*. Roma: Il calamo, pp. 81-94.

Deroy, L. (1956). La valeur du suffixe préhellénique *-nth-* d'après quelques noms grecs en -νθος. Glotta, 35(3/4), pp. 171-195.

Dieckhoff, A. (2019). Peuples et populisme. Annuaire international de justice constitutionnelle, 34, pp. 691-698. https://doi.org/10.3406/aijc.2019.2719

Erk, K. (2012). Vector Space Models of Word Meaning and Phrase Meaning: A survey. Language and Linguistics Compass, 6(10), pp. 635-653.

Fernández-Domínguez, J. (2019). The Onomasiological Approach. In Oxford Research Encyclopedia of Linguistics. Oxford: Oxford University Press. https://doi.org/10.1093/acrefore/9780199384655.013.579.

Geeraerts, D. (2009). Theories of Lexical Semantics. Oxford: Oxford University Press.

Goossens, D. (2013). Assessing Corpus Search Methods in Onomasiological Investigations. In H. Hasselgård, J. Ebeling, S. Oksefjell Ebeling (Eds.), Corpus Perspectives on Patterns of Lexis. Amsterdam & Philadelphia: John Benjamins, pp. 271-292.

Gries, S. Th. 2010. Useful Statistics for Corpus Linguistics. In *A Mosaic of Corpus Linguistics: Selected Approaches*, In A. Sánchez & M. Almela (Eds.), Studien zur romanischen Sprachwissenschaft und interkulturellen Kommunikation. Frankfurt am Main: Peter Lang, pp. 269-291.

Grzega, J. (2002). Some Aspects of Modern Diachronic Onomasiology. *Linguistics*, *40*(5), pp. 1021-1045. https://doi.org/10.1515/ling.2002.035

Johnston, I. (2019). A List of Homeric (Epic) Similes from the Iliad and Odyssey. Retrieved from http://johnstoniatexts.x10host.com/homer/homericsimiles.html.

Jones, H.S., Henry George Liddell, MacKenzie, R., Scott, R., & Thompson, A.A. (1996). A Greek-English Lexicon (New ed. with new supplement). Oxford: Clarendon.

Joosten, F., Sutter, G.D., Drieghe, D., Grondelaers, S., Hartsuiker, R.J., & Speelman, D. (2007). *Dutch Collective Nouns and Conceptual Profiling*. Linguistics, *45*(1), pp. 85-132.

Jussen, B., & Rohmann, G. (2015). Historical Semantics in Medieval Studies: New Means and Approaches. Contributions to the History of Concepts, 10(2), pp. 1-6. https://doi.org/10.3167/choc.2015.100201

Kaczyńska, E. (2019). Laconian βοῦα 'Band of Boys' as a collective noun. Graeco-Latina Brunensia, (1), pp. 93-103. https://doi.org/10.5817/GLB2019-1-7

Keersmaekers, A. (2020). A Computational Approach to the Greek Papyri: Developing a Corpus to Study

Variation and Change in the Post-Classical Greek Complementation System. Unpublished PhD Dissertation. KU Leuven. Retrieved from https://lirias.kuleuven.be/retrieve/590983.

Keersmaekers, A. (2021). The GLAUx corpus: methodological issues in designing a long-term, diverse, multi-layered corpus of Ancient Greek. In *Proceedings of the 2nd International Workshop on Computational Approaches to Historical Language Change 2021*, pp. 39-50.

Keersmaekers, A., Mercelis, W., Swaelens, C., & Van Hal, T. (2019). Creating, Enriching and Valorizing Treebanks of Ancient Greek. In *Proceedings of the 18th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2019)*. Association for Computational Linguistics (ACL), pp. 109-117.

Keersmaekers, A., & Van Hal, T. (2021). A Corpus-Based Approach to Conceptual History of Ancient Greek. In G. Kristiansen, K. Franco, S. De Pascale, L. Rosseel, & W. Zhang (Eds.), Cognitive Sociolinguistics Revisited. Berlin & Boston: Walter de Gruyter, pp. 213-225.

Kipfer, B.A. (1986). Investigating an Onomasiological Approach to Dictionary Material. Dictionaries: Journal of the Dictionary Society of North America, *8*, pp. 55-64.

Krijthe, J., & van der Maaten, L. (2018). Rtsne: T-Distributed Stochastic Neighbor Embedding using a Barnes-Hut Implementation (Version 0.15). Retrieved from https://CRAN.R-project.org/package=Rtsne.

Kühner, R., & Gerth, B. (1966). Ausführliche Grammatik der griechischen Sprache. München: Hueber.

Kutuzov, A. (2020). Distributional Word Embeddings in Modeling Diachronic Semantic Change. Unpublished PhD-dissertation. Oslo University.

LeDell, E., Gill, N., Aiello, S., Fu, A., Candel, A., Click, C., … H2O.ai. (2022). h2o: R Interface for the 'H2O' Scalable Machine Learning Platform (Version 3.36.0.4). Retrieved from https://CRAN.R-project.org/package=h2o.

Lehrer, A. (1986). English Classifier Constructions. Lingua, 68(2-3), pp. 109-148.

Lenci, A. (2018). Distributional models of word meaning. Annual Review of Linguistics, 4, pp. 151-171.

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., Hornik, K., Studer, M., … Murphy, K. (2022). cluster: 'Finding Groups in Data': Cluster Analysis Extended Rousseeuw et al. (Version 2.1.3). Retrieved from https://CRAN.R-project.org/package=cluster.

McGillivray, B. (2020). Computational methods for semantic analysis of historical texts. In K. Schuster & S. Dunn (Eds.), Routledge International Handbook of Research Methods in Digital Humanities. London: Routledge, pp. 261-274.

Mehl, S. (2016). *Corpus Onomasiology: A study in World Englishes*. Unpublished PhD-dissertation: UCL London.

Meyer, P., & Tu, N.D.T. (2021). A Word Embedding Approach to Onomasiological Search in Multilingual Loanword Lexicography. In *Proceedings of eLex 2021*, pp. 78-91.

Moerdijk, F., Tiberius, C., & Niestadt, J. (2008). Accessing the ANW Dictionary. In *Coling 2008: Proceedings of the Workshop on Cognitive Aspects of the Lexicon (COGALEX 2008)*. Manchester: Coling 2008 Organizing Committee, pp. 18-24. Retrieved from https://aclanthology.org/W08-1903.

Müller, E., & Schmieder, F. (2016). Begriffsgeschichte und historische Semantik: ein kritisches Kompendium. Frankfurt am Main: Suhrkamp.

Perrone, V., Palma, M., Hengchen, S., Vatri, A., Smith, J.Q. & McGillivray, B. (2019). GASC: Genre-Aware Semantic Change for Ancient Greek. In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*. Florence: Association for Computational Linguistics, pp. 56-66.

Perrone, V., Hengchen, S., Palma, M., Vatri, A., Smith, J.Q., & McGillivray, B. (2021). Lexical Semantic Change for Ancient Greek and Latin. In N. Tahmasebi, L. Borin, A. Jatowt, Y. Xu, & S. Hengchen (Eds.), Computational Approaches to Semantic Change. Berlin: Language Science Press, pp. 287–310.

Rhodes, C. (2014). An Unkindness of Ravens: A Book of Collective Nouns. London: Michael O'Mara Books.

Rodda, M.A., Probert, P., & McGillivray, B. (2019). Vector space models of Ancient Greek word meaning, and a case study on Homer. TAL Traitement Automatique Des Langues, 60(3), pp. 63-87.

Rodda, M.A., Senaldi, M.S., & Lenci, A. (2016). Panta Rei: Tracking Semantic Change with Distributional Semantics in Ancient Greek. In *CLiC-It/EVALITA*.

Sierra, G. (2008). Natural Language Searching in Onomasiological Dictionaries. In *Coling 2008: Proceedings of the Workshop on Cognitive Aspects of the Lexicon (COGALEX 2008)*. Manchester: Coling 2008 Organizing Committee, 32-38. Retrieved from https://aclanthology.org/W08-1905.

Swiggers, P., & Wouters, A. (1998). De Tékhnē grammatikē van Dionysius Thrax: de oudste spraakkunst in het Westen. Leuven: Peeters.

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

Viteau, J. (1896). Étude sur le Grec du Nouveau Testament: comparé avec celui des septante. Sujet, complément et attribut. Paris: E. Bouillon.

Woodhouse, S.C. (1987). English-Greek dictionary: a vocabulary of the Attic language (Repr.). London: Routledge and Kegan Paul.

Zwarts, J. (2020). Contiguity and membership and the typology of collective nouns. In *Proceedings of Sinn Und Bedeutung*, *24*(2). Osnabrück : Osnabrück University, pp. 539-554.