Computational Linguistics and Intellectual Technologies:
Proceedings of the International Conference "Dialogue 2020"

Moscow, June 17–20, 2020

PRESENTING *HODEL*—A NEW RESOURCE FOR RESEARCH ON HOMERIC GREEK VERBS

Zanchi C.
University of Pavia

Luraghi S.
University of Pavia and National Research University Higher School of Economics (Moscow)

DOI: 10.28995/2075-7182-2020-19-1188-1200

1. Introduction[1]

In this paper, we present a new resource that has been created at the University of Pavia for the study of Homeric Greek verbs and their dependents: the ***Ho**meric Greek **De**pendency **L**exicon* (*HoDeL*). This resource is based on AGLDT 2.0 (**A**ncient **G**reek and **L**atin **D**ependency **T**reebank 2.0), which in its earliest version (AGLDT 1.0) is the first treebank for Ancient Greek and Latin.

HoDeL allows searching for verbs along with their dependents that are annotated as being included in the verbal valency in AGLDT 2.0. In the next sections, we discuss technical issues concerning the construction of *HoDeL* and the queries used to build it (Section 2), theoretical issues regarding the annotation scheme of AGLDT 2.0 (Section 3), and practical issues emerging from usage of the resource (Section 4).

[1] This research was carried out in the framework of the project *Dipartimenti di Eccellenza* 2018–2022 (Ministry of University and Research) and was funded by the Italian Ministry of Education and Research grant n. 20159M7X5P Prin 2015 *Transitivity and argument structure in flux*. The revised version of the paper was prepared as part of the Basic Research Program at the National Research University Higher School of Economics (HSE University). We thank three anonymous reviewers for helpful comments on an earlier version of this paper. This chapter results from joint work of the two authors. For academic purposes, Chiara Zanchi is responsible of Sections 2 and 3, and Silvia Luraghi of Section 4.

2. Technical issues

HoDeL is closely connected with the Homeric texts treebanked at AGLDT 2.0. It has been automatically induced from the syntactic (or 'analytical') annotation layer of AGLDT 2.0. **Figure 1a** shows the syntactic tree annotated at the analytical layer of *Iliad* 1.1–7 in (1).

(1)

mênin	*áeide*	*theà*	*Pēlēïádeō*
wrath.ACC	sing.IMPV.PRS.3SG	goddess.VOC	Peleus'_son.GEN
Akhilêos /	*oulomḗnēn,*	*hḕ*	*murí'*
Achilles.GEN	destructive.ACC	REL.NOM	countless.ACC.PL
Akhaioîs	*álge*	*éthēke, /*	*pollàs*
Achean.DAT.PL	woe.ACC.PL	put.AOR.3SG	many.ACC.PL
d'	*iphthímous*	*psukhàs*	*Áïdi*
PTC	valiant.ACC.PL	soul.ACC.PL	Hades.DAT
proḗapsen /	*hērṓōn*	*autoùs*	*dè*
send.AOR.3SG	hero.GEN.PL	3PL.ACC	PTC
helṓria	*teûkhe*	*kúnessin /*	*oiōnoîsí*
spoil.ACC.PL	make.IMPF.3SG	dog.DAT.PL	bird.DAT.PL
te	*pâsi,*	*Diòs*	*d'*
and	all.DAT.PL	Zeus.GEN	PTC
eteleíeto	*boulḗ, /*	*ex*	*hoû*
fulfil.IMPF.3SG.M/P	plan.NOM	from	REL.GEN
dḕ	*tà*	*prôta*	*diastḗtēn*
PTC	DEM.ACC.PL	first.ACC.PL	part.AOR.3DU
erísante /	*Atreḗdēs*	*ánax*	*andrôn*
fight.PTCP.AOR.NOM.DU	Atreus'_son.NOM	king.NOM	men.GEN
kaì	*dîos*	*Akhilleús*	
and	divine.NOM	Achilles.NOM	

'The wrath sing, goddess, of Peleus' son, Achilles, that destructive wrath which brought countless woes upon the Achaeans, and sent forth to Hades many valiant souls of heroes, and made them themselves spoil for dogs and every bird; thus the plan of Zeus came to fulfillment, from the time when first they parted in strife Atreus' son, king of men, and brilliant Achilles.' (*Il.* 1.1–7)

From this layer, *HoDeL* automatically extracts all Homeric finite and non-finite verb forms, along with their dependents labeled as SB (Subject, e.g. ***Luke** eats an apple*), OCOMP (Object Complement, e.g. *Luke makes Claire **happy***), PNOM (Predicate Nominal, e.g. *Luke is **happy***), and OBJ (Object, e.g. *Luke eats **an apple**, Luke gives **an apple to Claire**, Luke talks **to Claire***). The last label comprises all verbal arguments except subjects and arguments labeled as OCOMP and PNOM (i.e. accusative, dative, genitive nouns or pronouns, prepositional phrases, infinitive verbs, accusative+infinitive constructions, relative clauses, and other types of subordinate clauses).

All dependents mentioned above may either be direct child nodes of a finite/non-finite verb form, or be attached to the verbal head via one of the so-called bridge nodes, such as AUXP (preposition), AUXC (conjunction), COORD (coordinator, including coordinative conjunctions and particles), APOS (apposing elements, such

as commas). For example, the phrase *kúnessin oiōnoîsí te pâsi* 'for dogs and all birds' in Figure 1b is tagged with the bridge COORD (the coordinative particle *te*) and contains two objects, coordinated and hence tagged as OBJ_CO.[2]

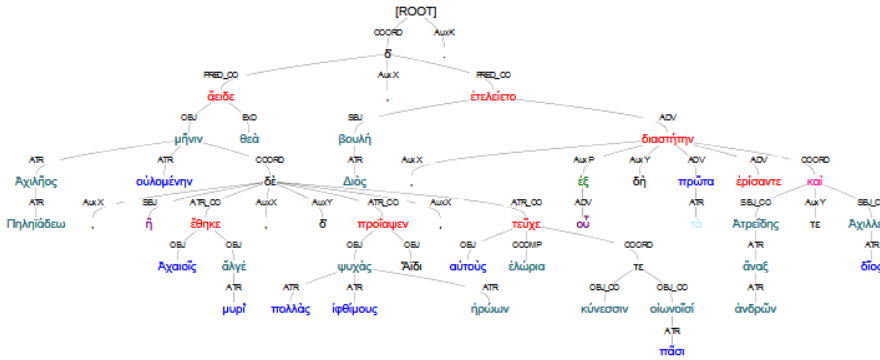

Figure 1a. *Iliad* 1.1–7 Syntactic annotation at the analytical layer

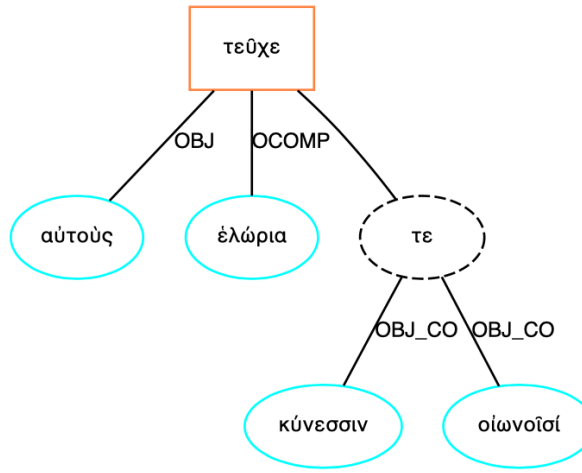

Figure 1b. An example of bridge (sub-tree from Figure 1a)

Bridge nodes can come in a sequence as in Figure 2. The verb *bátēn* 'they two walked' (AOR.3DU from *baínō*) takes an AUXP node *diá* 'through' that in turn functions as a head of two OBJ_CO, *éntea* 'arms' (ACC.PL) and *haíma* 'blood' (ACC).

[2] InfPs and other verb forms may be labeled as verbal head nodes and child simultaneously. This overlapping is due to the fact that a verb may be a head of a clause that in turn functions as an argument of another verb.

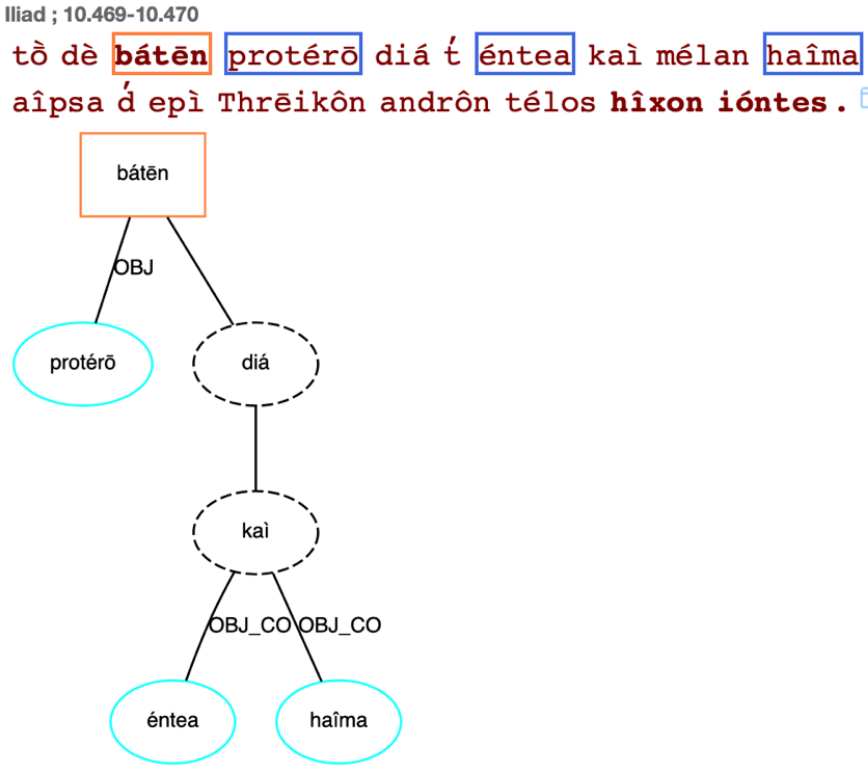

Figure 2. A sequence of bridge nodes

Extracted dependents are included in the verbal valency according to the guidelines of AGLDT 2.0. We did not extract dependents that are tagged as ADV (adverbials, which provide the event with background information), ATR (NP modifiers) and ATV/ATVV (non-governed complements, i.e. predicative noun phrases/adjectives which may morphologically agree with their head noun, but qualify the whole event denoted by the verb) that the AGLDT 2.0 guidelines do not consider as belonging to the verbal valency. Argumental dependents have been extracted using a series of SQL queries.

Extracted data was then recorded in a spreadsheet, which works as a backend for users' interface. The original query algorithm and its implementation were conceived to build a valency lexicon of Thomas Aquinas' texts treebanked at the IT-TB (*Index Thomisticus TreeBank*). The resulting lexicon *IT-VaLex* is documented in [McGillivray & Passarotti 2009]. While building *HoDeL* such queries have been adapted to the AGLDT 2.0 tagset.

An earlier version of *HoDeL* was released in 2016 [Zanchi et al. 2018] and was based on AGDT 1.0. Transliteration of the Greek script and links to English translation (**Section 4**) have also been added to the new version.

3. Theoretical issues

According to the guidelines of the analytical layer of AGLDT 2.0, the dependents extracted with our query correspond to verbal arguments, and extraction of these dependents from the Homeric poems treebanked at AGLDT 2.0 is expected to yield a corpus-driven valency lexicon for Homeric Greek. A similar procedure was presented as being able to extract, from AGLDT 1.0, "the first corpus-driven valency lexicon for Ancient Greek verbs." [McGillivray & Vatri 2015: 105].[3] However, things are not so simple, and such valency lexica can be said to be corpus-driven only in the sense that they are automatically induced from a corpus, which is an obvious misconception of what a corpus-driven approach is (see e.g. [Biber 2009: 12–17]).

In fact, queries based on annotation schemas, such as ours, can only retrieve whatever has been annotated according to a given annotation schema which, for obvious reasons, cannot be theory-free [Tognini Bonelli & Sinclair 2006: 214], [Haug 2015: 193]. In our case, verbal arguments extracted are those that were considered such in the creation of AGLDT 2.0.

The theory of valency underlying AGLDT 2.0 (as well as that of the *Prague Dependency Treebank* PDT guidelines on which AGLDT 2.0 heavily relies) turns out to be idiosyncratic and often inconsistent.[4] Examples include the following:

(i) Passive agents are annotated as OBJ (passive voice is generally acknowledged as a valency-decreasing strategy, by removing agents from argument structure; [Siewierska 2005], this issue is discussed in [Zanchi et al 2018: 230–231]).

(ii) The notion of obligatoriness is not elaborated in the guidelines, resulting e.g. in a varying annotation of Beneficiary and Instrumental dative dependents. This last issue results in inconsistencies: compare (2), in which the instrumental phrase *ophthalmoîsi teoîsin* 'with your eyes' is tagged as OBJ, with (3) in which a similar phrase, *ophthalmoîsi* 'with (my) eyes' is tagged as ADV (see also (4)).

(2)

eí	*pou*	*ópōpas*	***ophthalmoîsi***	***teoîsin***	*ē*
if	anywhere	see.PRF.2SG	eye.DAT.PL	your.DAT.PL	or
állou	*mûthon*	*ákousas*			
another.GEN	word.ACC	hear.AOR.2SG			

'(For this now I would come to your knees, if you wanted to tell me his horrible death,) if you had ever seen it **with your eyes** or (if) you had heard the word of another man.' (*Od.* 3.92–94)

---

[3] To our knowledge the complete valency lexicon presented in [McGillivray & Vatri 2015] has never been made available online.

[4] A distinction is usually made between syntactic valency, i.e. the number of slots opened by a verb that need to be filled with constituents governed by the verb to produce a well-formed sentence, and the wider notion of semantic valency, i.e. the number of participants evoked by the situation denoted by the verb and encoded by both governed and non-governed constituents [Luraghi & Parodi 2008: 196]. Here we refer to syntactic valency, as do the PDT guidelines available at https://ufal.mff.cuni.cz/pdt2.0/doc/manuals/en/a-layer/html/ch03.html (with explicit reference to governed constituents) and hence the AGLDT 2.0 guidelines. See [Zanchi et al 2018: 229–234] for further discussion of issues arising from the PDT view of verbal valency and its output in the AGLDT annotation scheme.

(3) *kapnòn* *d'* *enì* *méssei* *édrakon* ***ophthalmoîsi*** *dià*
smoke.ACC PTC in midst.DAT see.AOR.1SG eye.DAT.PL through
drumà *puknà* *kaì* *húlen*
bush.ACC.PL thick.ACC.PL and wood.ACC
'In the midst (of the island) I saw **with (my) eyes** smoke through the thick bush and the wood.' (*Od.* 10.196–197)

Moreover, the analytical layer of AGLDT 2.0 does not include any nodes for referential null arguments. This has the effect of omitting all non-overtly expressed subjects, which are numerous in a pro-drop language such as Ancient Greek. As noted in [Zanchi et al 2018: 235] "though being part of argument structure, subjects cannot always be automatically retrieved ... This has the unwelcome consequence of treating in the same way ... zero-valent impersonal verbs, such as 'rain', and monovalent verbs with omitted subjects."

In addition, referential direct objects are also omitted under certain syntactic and pragmatic conditions [Luraghi 2003]; [Keydana & Luraghi 2012]. In much the same way as null subjects, null referential direct objects fill in slots of the verbal valency, and the fact that they are not annotated in AGLDT 2.0 results in an incomplete account of the valency of specific verbs. An example is the null object of the verb *ídon* 'they saw' in (4), which is coreferential with *erōidion* 'a heron' in the preceding sentence as shown in **Figures 3** and **4**.

(4) *toîsi* *dè* *dexiòn* *hêken*
DEM.DAT.PL PTC right send.AOR.3SG
erōidiòn *eggùs* *hodoîo* *Pallàs*
heron.ACC near road.GEN Pallas.NOM
Athēnaíē *toì* *d'* *ouk*
Athena.NOM DEM.NOM.PL PTC NEG
ídon *ophthalmoîsi* *núkta* *di'*
see.AOR.3PL eye.DAT.PL night.ACC through
orphnaíēn, *allà* *klágxantos* *ákousan*
dark.ACC but cry.PTCP.AOR.GEN heard.AOR.3PL
'Pallas Athena sent them a heron on their right next to the road. They did not see (**it**) with (their) eyes in the dark night, but heard (**it**) crying.' (*Il.* 10.274–276)

The passage in (4) contains two sentences. In the former, the verb *hêken* '(she) sent' (from *híēmi* 'send') takes two arguments, the direct object *erōidiòn* 'a heron' and the indirect object *toîsi* '(to) them', both correctly tagged as OBJ. In the latter sentence, the verb *ídon* 'they saw' (from *eîdon* 'see') is annotated as not taking any argument (the fact that *núkta* 'night' is annotated as OBJ is an error discussed in **Section 4**). Still, it shares the direct object of *hêken*. Things are even more complicated for the verb *ákousan* 'they heard', which similarly shares the same direct object. In this case, the verb does not take any OBJ in the annotation, but it features as dependent the predicative participle *klágxantos* 'crying', which refers to the omitted direct object. Notably, the verb *akoúō* 'hear' takes the genitive, contrary to *híēmi* 'send' and *eîdon* 'see', which take the accusative. The assumed case of the omitted object is reflected in predicative participle, which is inflected in the genitive, hence agreeing with a null constituent.

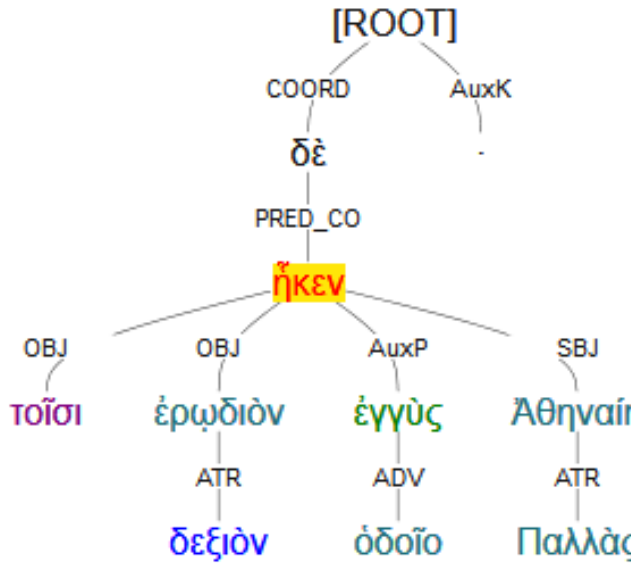

Figure 3. *Iliad* 10.274

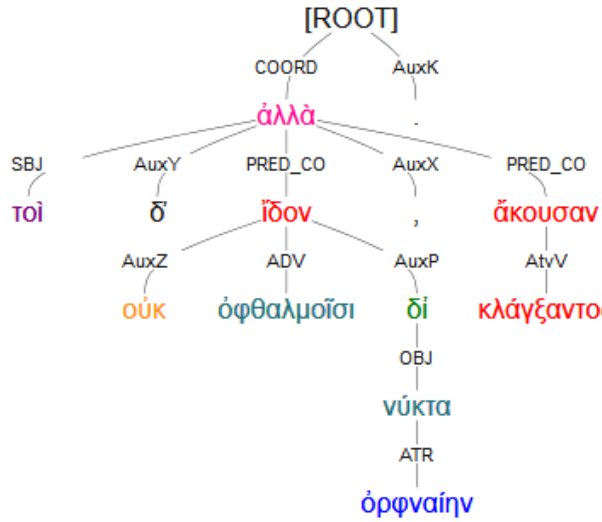

Figure 4. *Iliad* 10.275–276

Finally, the usage of the PNOM label is rather inconsistent in the treebank: for example, in a sentence such as *Luke is* ***in the kitchen***, the prepositional phrase is occasionally tagged as PNOM, despite the verb 'to be' functioning as an existential/locative verb and not as a copula in this context (hence it should be tagged as OBJ). An example is (5), annotated as in Figure 5.

(5) *laoì* *d'* ***ein*** ***agorêi*** *ésan* *athróoi*
men.NOM.PL PTC in assembly.DAT be.AOR.3PL in_crowds.NOM.PL
'The folk, gathered together, was in the place of assembly.' (*Il.* 18.497)

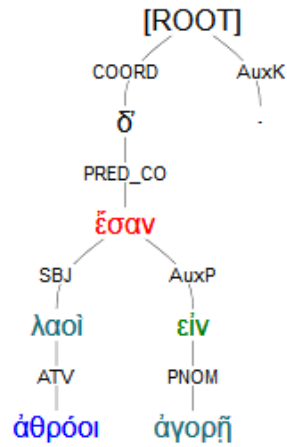

Figure 5. *Iliad* 18.497

4. Practical issues

In this section, the practical use of *HoDeL* is first illustrated by showing some simple queries. Three types of practical issues that may affect the user's experience when searching *HoDeL* are then addressed.

HoDeL homepage shows a list of Homeric verbs alphabetically ordered. After each lemma its frequency is provided (Figure 6).

HoDeL by default also shows that, according to AGLDT 2.0, the Homeric poems contain 2,482 verbal lemmas for a total of 40,693 occurrences. Verbal lemmas function as heads of 4,219 dependent lemmas that are annotated as having the argumental functions of SB, OBJ, PNOM, and OCOMP (cf. Section 3). The total number of occurrences of such dependent lemmas is 49,137. As users add filters to their queries, *HoDeL* always provides these and other frequency counts.

Users can also visualize all lemmas that depend on Homeric verbs, as shown in Figure 7. A number of verbs, such as *aeírō* 'lift', *hairéō* 'take', etc., appear in this list: these verbs function as main verbs in dependent SB or OBJ clauses (cf. fn. 2).

In lists of Figures 6 and 7, each verbal or dependent lemma is in turn clickable. By clicking on a verb—say, *akoúō* 'hear'—users get (i) all its forms contained in the Homeric poems, (ii) their ordered contexts of occurrences (automatically chunked by an algorithm that exploits punctuation marks), (iii) syntactic subtrees representing the queried verb and its dependents tagged as SB, OBJ, PNOM, OCOMP (Figure 8).

By clicking on a dependent lemma, users get the list of verbs that take that dependent. In turn, by clicking on each of the resulting verbs, users obtain the relative contexts and subtrees for that verb.

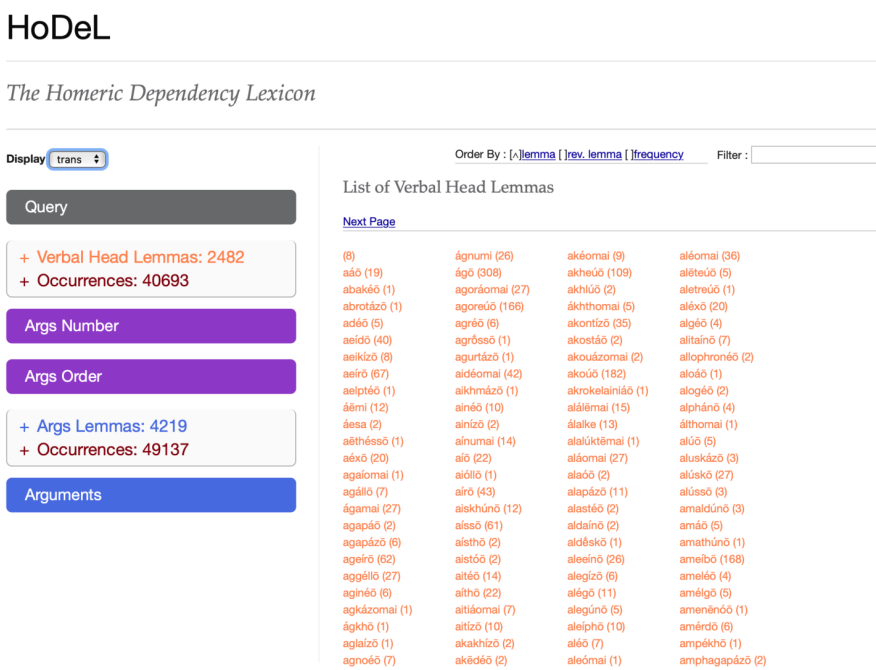

Figure 6. *HoDeL* homepage

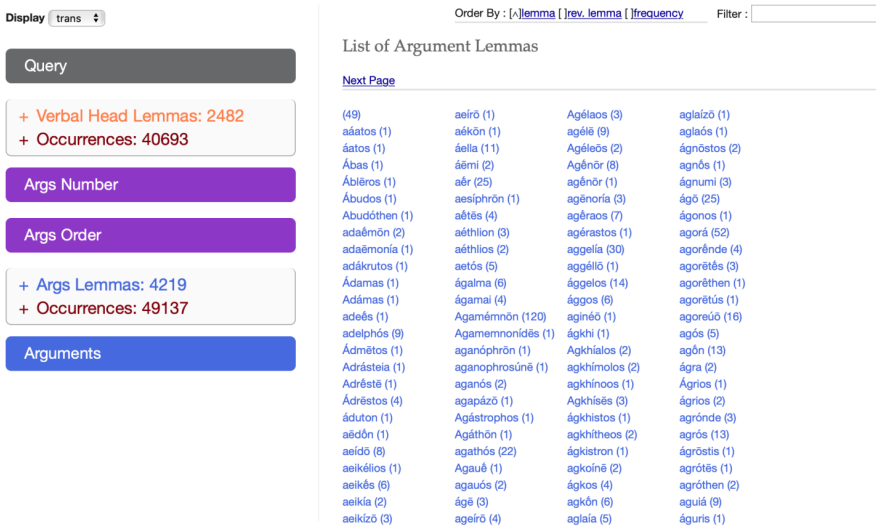

Figure 7. List of argument lemmas

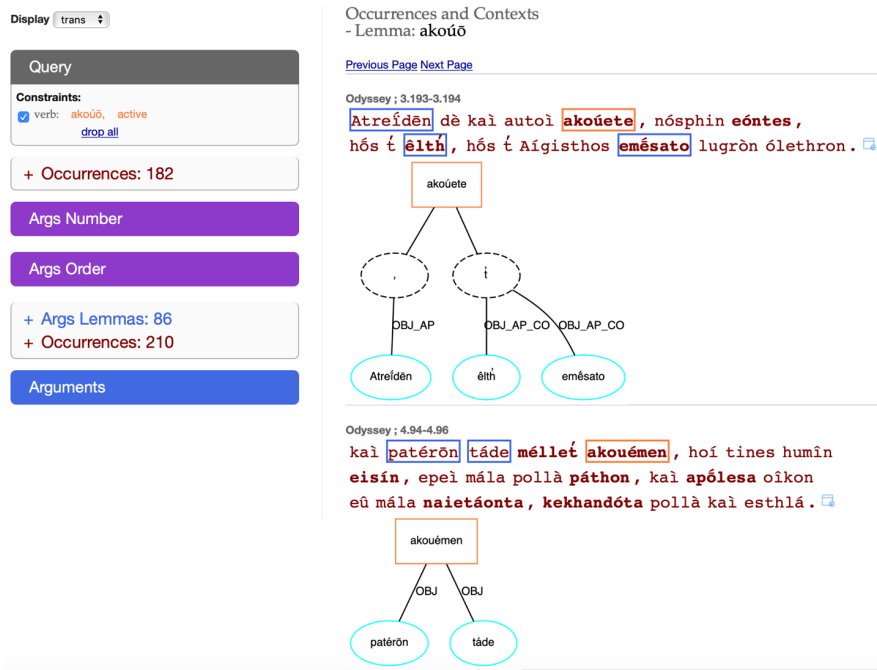

Figure 8. Searching for *akoúō* 'hear'

HoDeL also allows directly typing in verbal and dependent lemmas. By clicking on the box **Query**, the query window opens, in which the requested lemma—say *ágō* 'carry'—can be typed in Beta Code. *HoDeL* gives back all forms of *ágō* in the Homeric poems, their contexts, and the relevant syntactic subtrees (cf. Section 3, Figure 9).

If users point whatever word in the output contexts, they obtain morphological information as annotated in AGLDT 2.0 (Figure 10). Moreover, clicking on the blue folder after the Greek text, the English translation of the passage is given. Translations have been aligned automatically with the Greek text using an algorithm that exploits punctuation marks and text chunks contained in the text provided at the *Perseus Project*. The automatic alignment has been manually checked and, if necessary, modified according to the translation available at the *Chicago Homer* (see Figure 10).

By clicking on the box **Args Number**, users can see how many times a verb takes a certain number of arguments. Furthermore, users can get a list of functions that dependents of a verb carry out and their frequencies. *HoDeL* can also be used to obtain information regarding constituent order in Homeric Greek. In addition, users can filter outputs by argument relation and case/mood, clicking on the blue box **Arguments**. Additional filters are also incorporated in the **Query** box. Users can work on a single poem and search for verbs in a specific morphological voice. They can search for specific argument lemmas and filter their outputs based on features related to them (relation, case/mood, preposition, conjunction, position with respect to the verb). All these parameters can be combined with one another. *HoDeL* also allows users to search for more than one argument at one time.

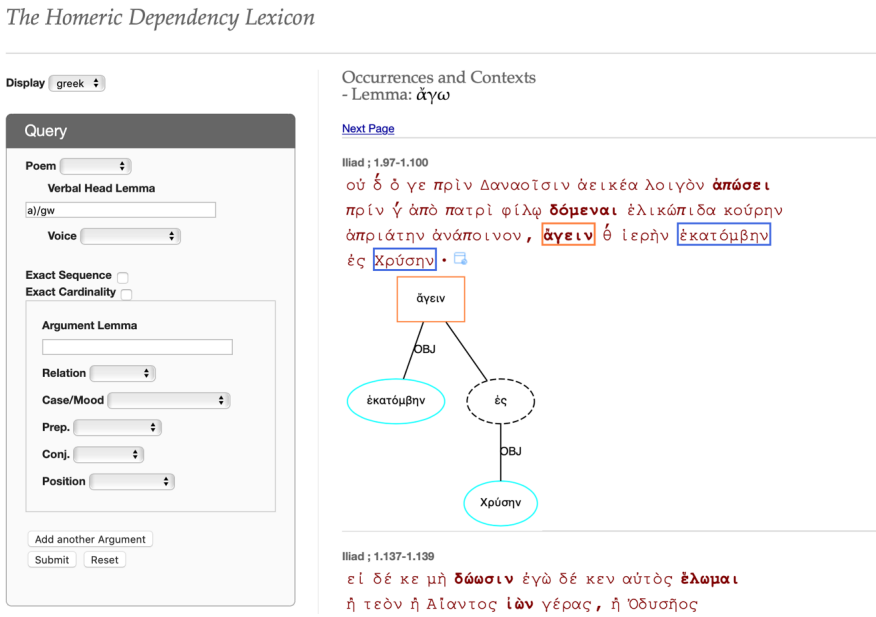

Figure 9. Searching for *ágō* 'lead'

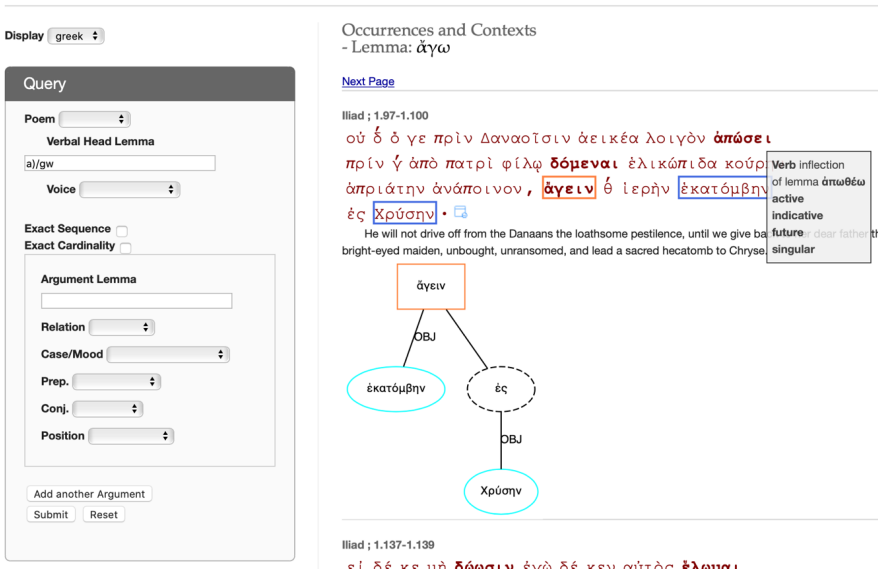

Figure 10. Visualizing morphological information and translation

Possible issues that may arise when using *HoDeL* are of three types. In the first place, as *HoDeL* takes its data directly from AGLDT 2.0, it also inherits a number

of annotation errors contained in the treebank. Some of them have been or are being fixed. For example, voice information has been manually re-annotated for all verbal forms, which are now tagged according to strict morphological criteria. Another example is constituted by errors in the syntactic tags that do not affect the structure of trees. In **Figure 3**, the prepositional phrase *di(à) núkta orphnaíēn* 'through the dark night' is correctly tagged as a child node of the verb *ídon* 'they saw', but it is incorrectly assigned the label OBJ, rather than ADV. These and similar corrections are currently being implemented.

A second error inherited from AGLDT 2.0 concerns the lemmatization of some verbs. For example, the verb *horáō* 'see' has a paradigm consisting of three stems, the present *horáō*, the aorist *eîdon*, and the perfect *ópopa*. They are lemmatized as two different entries, the former including forms from the stems *horáō* and *ópopa*, and the latter including forms from the stem *eîdon*. However, the prefixed verb *eis-oráō* 'behold', whose paradigm also features the same three stems, is lemmatized as a single entry. Fixing this inconsistency would be more complex, as it would imply restructuring the lemmatization system underlying the treebank.

A final issue is raised by peculiarities of the Homeric language, such as tmesis. The word 'tmesis', which literally means 'cut' refers to cases in which verbal prefixes are separated from the verbal stem that they semantically modify (see the exhaustive account in [Zanchi 2019]). An example is the verb *epeîmi* 'approach' in (6).

(6)	*núkta*	*phulassoménoisi*		*kakḗn*	*pedíon*
	night.ACC	keep_watch.PTCP.PRS.M/P.DAT		evil.ACC	plain.ACC
	=de	*gàr*	*aieì*	*tetráphath',*	*hoppót'*
	toward	PTC	ever	turn.PRF.M/P.3PL	when
	epì	*Trṓōn*	*aíoien*	***ióntōn***	
	PREV	Trojan.GEN.PL	hear.OPT.PRS.3PL	come.PTCP.GEN.PL	

'They kept watch through the evil night; for toward the plain were they ever turning if they might hear the Trojans coming on.' (*Il.*10.188–10.189)

In the example, the prefix *epí* is anticipated and occurs earlier in the sentence than the verb with which it belongs semantically. It precedes the main verb *aíoien* 'they might hear' and its object *Trṓōn* 'the Trojans. This pattern is typical of Homeric Greek, in which preverbs still preserved a partly free status. In AGLDT 2.0, the prefix is tagged as AUXZ, the label reserved for logical operators that are adverbs, such as those meaning 'not', 'even', and 'also'. This may be at odds with the function of verbal prefixes, which may change the meaning and thus argument structure of the base. As Homeric Greek is a free word order language, there is no straightforward way to account for such cases in the annotation system.

5. Conclusion

In this paper a new resource for research on Homeric Greek verbs has been presented. Besides discussing technical features and theoretical questions that lie at the foundation of this work, we have illustrated its practical usage and possible issues arising from it. More work is currently being done to improve the quality of the base data.

In spite of some remaining issues *HoDeL* is an interesting and useful resource for researchers as well as for student of Classics, as well as for any linguist interested in investigating verbs' argument structure and word order, on account of the richness of the data it contains and the friendliness of the interface. [5]

References

1. *AGLDT 2.0:* https://perseusdl.github.io/treebank_data.
2. *Biber, D.* (2009), Corpus-Based and Corpus-driven Analyses of Language Variation and Use, The Oxford Handbook of Linguistic Analysis, Oxford University Press, Oxford, pp.159–191.
3. *Chicago Homer:* https://homer.library.northwestern.edu.
4. *Haug D. T. T.* (2015), Treebanks in historical linguistics research, Perspectives on Historical Syntax, Benjamins, Amsterdam, pp. 187–202.
5. *HoDeL:* https://studiumanistici.unipv.it/hodel.
6. *IT-TB:* https://itreebank.marginalia.it/view/ittb.php.
7. *IT-VaLex:* https://itreebank.marginalia.it/itvalex.
8. *Keydana G., Luraghi S.* (2012), Definite referential null objects in Vedic Sanskrit and Ancient Greek, Acta Linguistica Hafniensia, Vol. 44/2, pp. 116–128.
9. *Luraghi S.* (2003), Definite referential null objects in Ancient Greek, Indogermanische Forschungen, Vol. 108, pp. 169–196.
10. *Luraghi S., Parodi, C.* (2008), Key terms in syntax and syntactic theory, Continuum, London.
11. *McGillivray B., Passarotti M.* (2009), The Development of the Index Thomisticus Treebank Valency Lexicon, Proceedings of LaTeCHSHELT& R Workshop 2009, Athens, pp. 43–50.
12. *McGillivray B., Vatri A.* (2015), Computational valency lexica for Latin and Greek in use: a case study of syntactic ambiguity, Journal of Latin Linguistics, Vol. 14/1, pp. 101–126.
13. *Prague Dependency Treebank 3.0:* https://ufal.mff.cuni.cz/pdt3.0.
14. *Perseus Project:* http://www.perseus.tufts.edu/hopper.
15. *Siewierska A.* (2005), Passive constructions, WALS Online, available at http://wals.info/chapter/107.
16. *Tognini Bonelli E., Sinclair J. M.* (2006), Corpora, Encyclopedia of Language and Linguistics, 2nd edition, vol. 3, Elsevier, Amsterdam, pp. 206–219.
17. *Zanchi C.* (2019), Multiple preverbs in ancient Indo-European languages. A comparative study on Vedic, Homeric Greek, Old Church Slavic, and Old Irish, Narr, Tübingen.
18. *Zanchi C., Sausa E., Luraghi S.* (2018), HoDeL, a Dependency Lexicon for Homeric Greek: Issues and Perspectives, Formal Representation and the Digital Humanities, Cambridge Scholars Publishing, Cambridge, pp. 221–246.

[5] With respect to other resources equally induced from freely available annotated data, such as Sketch Engine, our main focus here is not on detecting collocations based on linear order, but on showing verbal arguments irrespective of where they occur in the sentence.