Non-projectivity in the Ancient Greek Dependency Treebank

Francesco Mambrini
The Center for Hellenic Studies
Washington, DC
fmambrini@chs.harvard.edu

Marco Passarotti
Università Cattolica del Sacro Cuore
Milano, Italy
marco.passarotti@unicatt.it

Abstract

In this paper, we provide a quantitative analysis of non-projective constructions attested in the Ancient Greek Dependency Treebank (AGDT). We consider the different types of formal constraints and metrics that have become standardized in the literature on non-projectivity (planarity, well-nestedness, gap-degree, edge-degree). We also discuss some of the linguistic factors that cause non-projective edges in Ancient Greek. Our results confirm the remarkable extension of non-projectivity in the AGDT, both in terms of quantitative incidence of non-projective nodes and for their complexity, which is not paralleled by the corpora of modern languages considered in the literature. At the same time, the usefulness of other constraint (especially well-nestedness) is confirmed by our researches.

1 Introduction

The "free" word-order of Ancient Greek (AG) is a notorious problem for philologists and linguists. In spite of several studies devoted to the subject, the tendencies that govern the disposition of words and constituents in the sentence still lack a comprehensive explanation. Strictly connected to the word-order issue is the relevant amount of discontinuous constituents, which even casual readers of AG texts can experience[1].

The dependency-based treebanks of Classical languages (AG and Latin) that have been recently made available enable us to reconsider this long debate in the light of the abundant work on non-projective structures in dependency trees. Non-projectivity (see 2 for a formal definition) is a key issue in dependency grammar, both from the formal point of view and from a more descriptive linguistic perspective. From the standpoint of natural language processing, non-projectivity is also known to affect the efficiency of dependency parsers.

In a first attempt to improve parsing performances on AG, Mambrini and Passarotti (2012) reported that the amount of non-projective arcs occurring in the available treebanks of Classical languages is significantly higher than that attested in the corpora of modern languages used for CoNLL-X (Buchholz and Marsi, 2006, 155, tab. 1) and CoNLL 2007 shared tasks (Nivre et al., 2007, 920, tab. 1). Furthermore, the non-projective rate in the Ancient Greek Dependency Treebank is higher than in Classical and Medieval Latin (Passarotti and Ruffolo, 2010, 920, tab. 1).

In this paper, we want to discuss this claim in depth and substantiate it by applying to AG data the standard metrics for the different kinds of non-projective constructions established in the literature.

The paper is organized as follows. Section 2 provides a definition of the formal constraints considered and of the metrics that will be used: non-projectivity, planarity, well-nestedness, on the one hand, and gap-degree and edge-degree on the other. Section 3 introduces the corpus that will be tested, the Ancient Greek Dependency Treebank (AGDT).

Section 4 presents the evidence provided by the data. In 4.1 we report the results for the different constraints and metrics defined in section 2. Results for the distribution of non-projectivity in the different genres of the corpus are given and commented in 4.2.

In section 5, we discuss some of the linguistic issues that cause non-projectivity. Finally, section 6 reports our conclusions and sketches possible directions for additional research.

[1] On AG word-order see more recently Dik (1995; 2007), with bibliography of previous studies. On discontinuous structures see Devine and Stephens (2000).

Proceedings of the Second International Conference on Dependency Linguistics (DepLing 2013), pages 177–186,
Prague, August 27–30, 2013. © 2013 Charles University in Prague, Matfyzpress, Prague, Czech Republic

2 Non-projectivity

A dependency tree is a rooted tree where the nodes represent the words of a sentence, the edges represent the syntactic dependencies and the linear order of the nodes stands for the sequence of words.

According to the so-called ‘treeness constraint’ (Debusmann and Kuhlmann, 2010), a dependency tree requires (a) that no word should depend on itself, not even transitively (i.e. the tree must be acyclic), (b) that each word should have at most one governor, and (c) that a dependency analysis should cover all the words in the sentence.

If a node j depends on a node i, j is called a ‘child’ of i, and, symmetrically, i is the ‘parent’ of j (we write $i \rightarrow j$). On the other hand, we write $i \leftrightarrow j$ whenever the edge is considered regardless of the direction of the relation (i can be either the parent or the child of j). If i precedes j in the word order, i lies to the left of j (we write $i < j$); conversely, j lies to the right of i ($j > i$). The set of all nodes that can be reached from a given node i by following a directed path of zero or more edges is called the set of ‘descendants’ of i. A subtree of a tree T at a node i is the restriction of T (nodes and edges) to the descendants of i.

The condition of **projectivity**, which was formally defined by Marcus (1965), requires each dependency subtree to cover a contiguous region of the sentence: a word and its transitive dependents must span a contiguous sequence in the linear order. We may define the constraint of projectivity with the following formula (Havelka, 2007, 609):

$$i \rightarrow j \;\&\; v \in (i, j) \Longrightarrow v \in Subtree_i$$

which must be read in this way: let i be the parent of j ($i \rightarrow j$); if a node v lies between i and j in the linear order of the sentence, then v belongs to the subtree of i. If this condition does not hold, then the edge is non-projective and v is said to be in a **gap** ($v \in Gap_{i \leftrightarrow j}$). Example 1 illustrates this construction with a simplified version of a sentence from the AGDT (the first sentence of the *Iliad*), which is also represented in fig. 1[2]. The edges *mēnin-Achilēos* and *mēnin-ouloménēn* are non-projective, since the nodes of *áeide* and *theá* are in a gap in both cases.

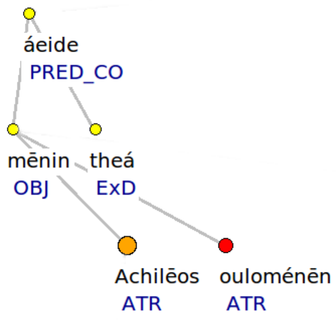

Figure 1: Non-projective edges: a simplified tree from the AGDT

(1) *mēnin* *áeide* *theá*
wrath.FEM.ACC sing Goddess.VOC
Achilēos *ouloménēn*
of-Achilles.GEN accursed.FEM.ACC
Sing, oh Goddess, the wrath of Achilles, the accursed wrath.
(simplified version of *Iliad* 1.1)

Non-projectivity, which was postulated by Marcus (1965) for the purposes of machine translation and language generation, is too strong a constraint for natural languages: a non-negligible number of constructions attested for many languages does not satisfy the condition. Several relaxations to the definition were subsequently introduced in order to better account for the linguistic data.

The condition of **planarity** involves two edges $i_1 \leftrightarrow j_1$ and $i_2 \leftrightarrow j_2$ and disallows any overlapping between them. Two edges are said to overlap if, for example, $i_1 > i_2 > j_1 > j_2$ or $i_1 < i_2 < j_1 < j_2$. Therefore, following Havelka (2007), a tree is non-planar if there are at least two edges $i_1 \leftrightarrow j_1$ and $i_2 \leftrightarrow j_2$ that meet the following condition:

$$i_1 \in (i_2, j_2) \;\&\; i_2 \in (i_1, j_1)$$

Example 2 (fig. 2) presents two non-planar edges from a tree of the AGDT.

(2) *mýri’* *Achaióis*
countless.NEUT.PL to-Achaeans.DAT
álge’ *éthēke*
grieves.NEUT.PL (it)-caused.3rd.SG
(which) inflicted countless grieves to the Achaeans
(*Iliad* 1.2)

[2] The Greek words and lemmata in the AGDT are written in Greek characters, transcribed according to an ASCII-based convention known as “Beta Code” (TLG, 2010). All the trees reported in this paper are transliterated in Latin script by the authors for ease of reading.

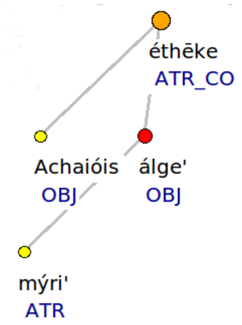

Figure 2: Non-planar (well-nested) edges

In the tree reported in fig. 2, the edges *éthēke* $\rightarrow$ *Achaióis* and *álge'* $\rightarrow$ *mýri'* are non-planar, because *éthēke* $>$ *álge'* $>$ *Achaióis* $>$ *mýri'*.

Well-nestedness introduces a further relaxation to projectivity. A (sub)tree is said to be well-nested if, for each pair of overlapping disjoint edges, the source node of one of the edges is a descendant of the source node of the other; conversely (Havelka, 2007), the (sub)tree is ill-nested if, for two edges $i_1 \leftrightarrow j_1$ and $i_2 \leftrightarrow j_2$:

$$i_1 \in Gap_{i_2 \leftrightarrow j_2} \;\&\; i_2 \in Gap_{i_1 \leftrightarrow j_1}$$

One may note that the two edges in fig. 2 are well-nested, since *álge'* is the child of *éthēke*.

In addition to these constraints, two metrics have become standard measures for non-projectivity.

Given a non-projective edge, **edge-degree** represents the number of nodes that are in the gap. This metric was introduced by Nivre (2006), but was named *edge-degree* by Kuhlmann and Nivre (2006) and it corresponds to *component degree* in Havelka (2007). The edge-degree can be estimated either by counting the edges that match the definition in a treebank, or by using the tree as a basis, the edge-degree of a tree T being equal to the highest edge-degree among the edges of T.

On the contrary, **gap-degree** is not based on a single edge, but rather on the 'projection' (or on the 'blocks') of a node, which is defined as the longest non-empty sequence of nodes that goes down to a terminal node in a chain succession from father to child[3]. A gap (or interval) in the projection is a discontinuity such that, given a node j_k in the sequence, $j_k - j_{k+1} > 1$. The gap-degree corresponds to the number of gaps in the sequence (Kuhlmann and Nivre, 2006), while the gap-degree of a tree is equal to the highest gap-degree for each sequence in the tree[4].

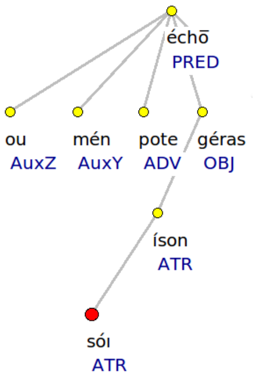

Figure 3: Gap-degree = 2

(3) *ou mén sói pote íson*
not PRTCL to-you.DAT ever equal.ACC
échō géras
(I)-have.1st.SG gift.ACC
Never do I get a gift that matches yours
(*Iliad* 1.163)

In fig. 3, which represents the tree of example 3, the segment *géras-íson-sói* is interrupted twice, namely by *échō* and *pote*. The words in the gaps do not form a single continuous interval in the linear order: the segment has therefore gap-degree = 2.

3 The corpus

The Ancient Greek Dependency Treebank (Bamman et al., 2009) is a dependency-based treebank of Greek literary texts of the Archaic and Classical age published by the Perseus Digital Library[5].

In its theoretical framework and guidelines, the AGDT is inspired by the analytical layer of the Prague Dependency Treebank of Czech (Böhmová et al., 2001). Currently, the last published version of the AGDT (1.7) includes 354,529 tokens. The collection is constituted by unabridged works that

[3] A projection of a node may consist of one or more 'blocks', i.e. maximal, non-empty intervals of descendants.

[4] **Interval-degree** is an edge-based version of the gap-degree (Havelka, 2007). Given a non-projective edge $i \rightarrow j$ with $v_{1,n}$ in the gap, the interval-degree corresponds to the number of intervals in the sequence $v_{1,n}$.

[5] Perseus Digital Library: `http://www.perseus.tufts.edu/hopper/`.

belong to three literary genres: epic poetry (the *Iliad*, the *Odyssey*, and the complete works of Hesiod), tragedy (the complete work of Aeschylus and five plays of Sophocles), philosophical prose (the *Euthyphro* of Plato). Chronologically, the texts range from the 8th to the 4th Century BCE. The composition of the AGDT 1.7 is resumed in table 1[6].

Author/Work	**Genre**	**Date**	**Tokens**
Iliad	Epic	8th(?)	128,102
Odyssey	Epic	8th(?)	104,467
Hesiod	Epic	7th(?)	18,881
Aeschylus	Drama	5th	48,261
Sophocles	Drama	5th	48,721
Plato	Prose	4th	6,097
Total			354,529

Table 1: AGDT 1.7

4 Results

4.1 Constraints and measures in the AGDT

Table 2 reports the number and percentage of the trees that do not respect the constraints of projectivity, planarity, and well-nestedness in the AGDT 1.7. The Ancient Greek data are compared with those of six other languages from the CoNLL-X shared task that display the highest rate of non-projective constructions (Havelka, 2007)[7]. The languages are sorted according to the percentage of non-projective trees in decreasing order.

As it may be seen, AG shows a remarkably high rate of non-projective trees in comparison with the other languages. Non-projective edges are found in almost three out of every four sentences of the AGDT, a distribution that nearly reverses that of German, Czech or Slovene. The abundance of non-projective constructions in the AGDT stands out even more clearly when one considers the rate of non-projective edges instead of non-projective trees. The numbers are reported in table 3; the comparative data are again taken from Havelka (2007).

Language	**Tot. edges**	**Non-proj. edges**	
		No.	%
A. Greek	301848	45731	15.15
Dutch	179063	10566	5.90
German	660394	15844	2.40
Czech	1105437	23570	2.13
Slovene	25777	550	2.13
Portuguese	197607	2702	1.37

Table 3: Non-projective edges in AG and other languages

Although AG is exceptional for the incidence of non-projective edges, if one considers the different conditions of relaxation of projectivity, AG data seem to reflect the same tendencies already observed in other languages (Kuhlmann and Nivre, 2006; Havelka, 2007). Non-planarity does not prove to mark a significant relaxation. On the contrary, well-nestedness is a very effective constraint in AG too. Although the absolute rate is higher than in the other languages, the number of ill-nested trees is considerably smaller.

AG deviates from the trend observed in other languages also for the complexity of non-projective structures, as measured by both gap- and edge-degree. The observations reported in table 4 highlight the main differences with the other languages studied[8].

The fact that in AG the percentages of gap-degree 0 and 1 appear to be almost inverted in comparison with those of the other languages is not surprising, given the general proportion of non-projective trees in the languages discussed. In all the languages but AG, a threshold set to gap-degree = 1 is a very strong constraint, which allows to account for more than 99% of the total of trees[9]. In AG, instead, the number of trees with gap-degree = 3 is still a non-negligible fraction (more than 6% of the total).

The rate of trees with edge-degree ≥ 1 is significantly higher in AG than in the other languages too. While in other treebanks an edge-degree = 2 is already sufficient to cover more than 99% of the

[6] The dates reported in the table refer to the century BCE. In some cases (like for the Homeric poems and the works of Hesiod), even this general chronology is very hypothetical.

[7] Note that in the CoNLL data format, each sentence is provided with a technical root node placed before the sentence (i.e. the root is the leftmost node in dependency trees for left-to-right languages). All the dependency analyses are attached directly, or indirectly to the root node. As edges from technical roots may introduce non-planarity, we disregard all such edges when counting trees conforming to the planarity constraint. Since the same is done by Havelka (2007, p. 21), this makes our data comparable with those reported there.

[8] The data for Danish and Czech are taken from Kuhlmann and Nivre (2006). Hindi and Urdu are two of the Indian languages studied by Bhat and Sharma (2012), whence the numbers were taken.

[9] Urdu with, 98.43%, is only a limited exception.

Language	Tot. trees	Non-proj.		Non-plan.		Ill-nested	
		trees	%	trees	%	trees	%
Ancient Greek	24825	18568	74.80	15334	61.77	656	2.64
Dutch	13349	4865	36.44	4115	30.83	15	0.11
German	39216	10883	27.75	10865	27.71	416	1.06
Czech	72703	16831	23.15	13783	18.96	79	0.11
Slovene	1534	340	22.16	283	18.45	3	0.20
Portuguese	9071	1718	18.94	1713	18.88	7	0.08

Table 2: Non-projective, non-planar, ill-nested trees in AG and other languages

Language	Trees	Gap-degree (%)					Edge-degree (%)				
		gd0	gd1	gd2	gd3	gd4	ed0	ed1	ed2	ed3	ed4
A. Greek	24825	25.20	68.33	6.17	0.28	0.02	25.20	43.73	14.15	7.07	3.88
Danish	4393	84.95	14.89	0.16	-	-	84.95	13.29	1.32	0.39	0.05
Czech	73088	76.85	22.72	0.42	0.01	<0.01	76.85	22.69	0.35	0.09	0.01
Hindi	20497	85.14	14.56	0.28	0.02	na	85.14	14.24	0.45	0.11	0.03
Urdu	3192	77.85	20.58	1.31	0.12	na	77.85	19.20	1.97	0.56	0.22

Table 4: Gap-degree and edge-degree

trees, it is only at edge-degree = 7 (0.84%) that the frequency of the AGDT trees drops under 1% of the corpus[10].

To sum up, we can conclude that, whereas well-nestedness appears to be an effective constraint in AG too, the thresholds of gap-degree 1 and edge-degree 2 in the AGDT do not have the same impact as that observed in other treebanks.

4.2 The role of genre

Genre difference is known to have a strong effect on the performances of dependency parsers for AG texts (Mambrini and Passarotti, 2012). It is important, therefore, to observe if the values reported above are at variance in the different genres included in the AGDT.

The frequencies of non-projective constructions in each of the three genres of the AGDT are reported in table 5. The most relevant fact is the difference between the poetic genres (epic and drama) on the one hand, and the philosophical dialogue in prose on the other. This difference can be appreciated especially when one looks at the rate on non-projective edges, which does not vary significantly between epic and drama (respectively, 15.30% and 15.09%), but it is quite different if prose is concerned (9.81%). Plato's *Euthyphro* (the sole prose work included in the corpus at the moment) is the only text where the number of non-projective trees is less than 50%, although the incidence of non-projectivity is still sensibly higher than in corpora of modern languages, as reported above.

These distributions may lead to the claim that the high number of discontinuous constituents is due to poetic style and, possibly, to the metrical constraints that operate in poetic language[11]. Unfortunately, no conclusions can be drawn on this matter. Limited as they are to one single author and text, the presently available data for prose language can hardly point to more than a working hypothesis for future research. It will be possible to test this hypothesis as soon as new texts of the same genre and/or author will be added to the corpus.

5 Discussion: linguistic causes of non-projective edges in the AGDT

In this section we discuss some of the linguistic causes of non-projective edges in the AGDT. We first analyze one specific kind of nodes in the gap (the clitics: 5.1); then we will focus on those non-projective edges in the AGDT that are governed by a verb or by a noun, also in the light of the typologies studied for Czech (Hajičová et al., 2004).

[10] The maximum edge-degree found in the AGDT is the abnormal value of 45, which however results from an annotation error.

[11] Sensible remarks (with minimal bibliography) about the question of linguistic and metrical constraints on word-order in AG tragedy can be read in Dik (2007, 3, 168-224).

Measures and	**Epic**	**Drama**	**Prose**
constraints	T = 16359	T = 8040	T = 426
	E = 217539	E = 79162	E = 5147
non-proj trees (%)	82.25	60.95	49.77
non-proj edges (%)	15.30	15.09	9.81
non-planar (%)	66.67	52.70	44.84
ill-nested (%)	2.50	3.00	1.41
gap-deg = 0 (%)	17.75	39.05	50.23
gap-deg = 1 (%)	76.27	53.43	44.37
gap-deg = 2 (%)	5.78	7.00	5.40
edge-deg = 0 (%)	17.75	39.05	50.23
edge-deg = 1 (%)	50.21	31.64	23.00
edge-deg = 2 (%)	14.38	6.85	9.62
edge-deg = 3 (%)	7.51	3.09	4.69

Table 5: Measures and constraints in the AGDT, grouped by genre (T = tot. trees, E = tot. edges)

5.1 The role of clitics

As it is partly the case with the Czech conjunction *-li* (Hajičová et al., 2004), clitics are likely to have a strong role in non-projective structures of the AGDT. It is well known that in AG clitics tend to stick to a fixed position in the sentence or clause, in accordance to the so-called "Wackernagel's law", which is common to many Indo-European languages (Wackernagel, 1892; Ruijgh, 1990). The words that belong to the class of *postpositives* (i.e. the clitics and a few other words that cannot occupy the clause-initial position) tend to be placed in second position, even when this collocation breaks a syntactic constituent. In example 4, the coordinating particle *d'* (*dé*) is placed in second position and separates one of the two (non-coordinated) attributes (*pollás*) from the rest of the noun phrase (*iphthímous psychás*).

(4) *pollás* ***d'*** *iphthímous*
many.ACC.FEM **and** strong.ACC.FEM
psychás *Háidi* *proíapsen*
souls.ACC.FEM to-Hades (it)-sent

and (it) sent forth to Hades many valiant souls (*Iliad* 1.1)

This situation is normal also with the enclitic *te* (*and*, = Latin *-que*), which is regularly placed after the first word of coordinated clauses. Whenever the word that precedes *te* has one or more right descendants in the dependency tree, *te* comes to be in a gap.

Some facts hint that this tendency of clitics is a relevant issue for non-projectivity: we have resumed them in table 6. The first five most frequent words recurring in gaps belong all to the class of postpositives; in total, postpositives account for about 40% (39.16%) of the nodes attested in gaps[12]. As for the most frequent syntactic labels, coordinating conjunctions (COORD) and sentence adverbials (AuxY), which are the two typical functions of postpositives, are again the two groups ranking in the first and second place. Finally, second position (i.e. the one which is usually occupied by postpositives) is by far the most

		% nodes in gap
Lemmata	dé	25.85
	te	4.87
	mén	3.17
	gár	2.09
	án	0.69
Syntactic relations	COORD	22.78
	AuxY	15.35
	PRED	12.57
	ADV	9.47
	OBJ	6.91
Positions	2	18.63
	5	7.62
	4	7.00

Table 6: Most frequent lemmata, syntactic relations and positions in the sentences for words in a gap in the AGDT

[12] A list of postpositives can be found in Dik (1995, 32). Note that we left personal pronouns out of our analysis.

often attested for nodes in a gap[13].

5.2 Verb-headed and noun-headed non-projective edges in the AGDT

In this section we will focus on those non-projective edges that are governed by either a noun/pronoun or a verb, as they cover more than 76% of all the non-projective edges in the AGDT.

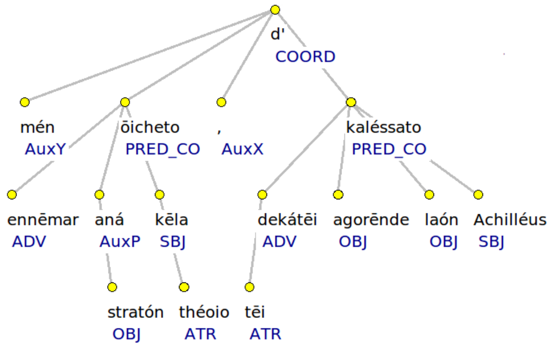

Figure 4: *Iliad* 1.53-4

Verb-headed edges: in the non-projective structures of the AGDT, verbal complementations precedes their verbal head in the 67.93% of the cases. Both arguments (subjects, predicatives and direct/indirect objects) and adjuncts (several kinds of adverbial modifications) are very frequently involved in such a movement of the complements to the left, which often results in non-projective constructions.

In Czech, this situation is produced notably by contrastive contextually bound elements moved towards the beginning of the clause. As the AGDT does not feature annotation of information structure yet, it is not easy for us to evaluate its quantitative impact in data, but a reading of the examples that can be extracted from the treebank suggests that the same tendency is at work also in AG[14].

In the two coordinated clauses of example 5 (fig. 4), for instance, the temporal complementations create a contrastive frame for the action (*ennēmar... tēi dekátēi*). In order to highlight their function, both complements are moved to the first position of their respective clauses. This movement causes non-projectivity, as the nodes for *mén* and *d'* result to be in a gap.

Another phenomenon that may generate non-projective constructions in AG is the raising of complementations of infinitive verbs that are moved to the left outside the subordinate clause. The importance of this pattern can be seen by measuring the distribution of verbal mood in the non-projective arcs. Indicative dominates in general (75% vs 14% of infintives), but if one considers the subset of cases with complement–head order and with a gap wider than one single clitic, then the rate of infinitives increases considerably (50% vs 42% of indicatives).

(5) *ennēmar* *mén*
for-nine-days on-the-one-hand
aná *stratón ōicheto kēla*
throughout camp went arrows
théoio, *tēi* *dekátēi d'*
of-the-god, (on-)the tenth while
agorēnde *kaléssato laón* *Achilléus*
to-assembly called army Achilles

For nine days the arrows of the god swept throughout the camp; on the tenth, Achilles called the army to the assembly (*Iliad* 1.53-4)

(6) *hoi mén* *epépleon* *hygrá*
they on-the-one-hand sailed-over watery
kéleutha, *laoús* *d'*
paths, men.ACC while
Atreídēs *apolymáinesthai*
son-of-Atreus.NOM to-purify.INF.REFL
ánōgen
commanded

So they sailed over the watery paths, while the son of Atreus commanded the men to purify themselves (*Iliad* 1.312-2, slightly simplified)

Example 6 (fig. 5) shows a combination of the two aforementioned phenomena: *laoús*, subject of the infinitive *apolymáinesthai*, is moved to the left of the main verb of the clause, outside the subordinate governed by the infinitive. At the same time, one may note the structure of the sentence, where two clauses are contrasted. In the first part, the departure of a small embassy of twenty selected

[13] The value of this observation is very limited. There are a number of cases (starting with those where a particle follows a coordinating conjunction and has the second coordinated clause in its scope) where the "second position" of clitics does not correspond to the second word of a sentence. However, it seems significant, also in light of the other observation reported above, that rank n. 2 scores so highly.

[14] It is also known that topic elements in AG tend to be placed at the beginning of the clause (Dik, 1995; Matić, 2003).

Greek soldiers (whose preparation is the subject of the preceding lines) is mentioned. In the second, the events at the Greek camp and the actions of the main army that remains at Troy are narrated. When the whole army is mentioned again, thus, the contrastive contextually bound word (*laoús*) is raised in prominent position and this movement causes non-projectivity, since the subject of the governing clause (*Atreídēs*) comes to be in a gap[15].

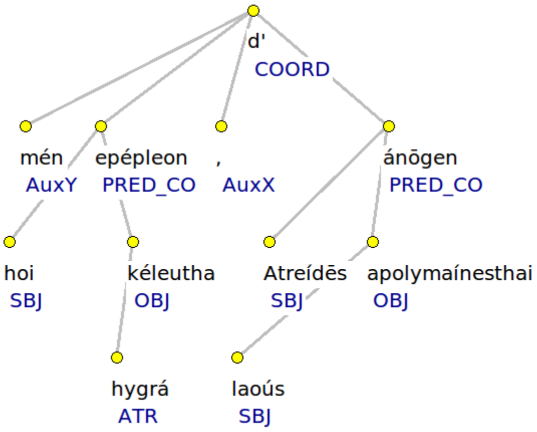

Figure 5: *Iliad* 1.312-3 (simplified)

Noun-headed edges: nouns can govern attributes (in the form of adjectives, nouns in genitives, or relative clauses), predicatives and valency complements (especially for deverbal nouns). In the case of nouns, the preference toward the complement–head order is less marked than with verbs (57% complement–head vs 43 head–complement%).

The head-noun of a non-projective edge can be the salient element moved toward the beginning of the sentence. This is the case in example 1 (fig. 1), where the noun *mēnis* ("wrath"), which introduces the main subject of the whole poem, is the focus of the sentence. This word is placed in first position, before the invocation to the Muse, and detached from the possessive genitive ("of Achilles"), of the epithet ("accursed") and of the long series of relative clauses that further specify the noun (not reported in the example above). The left-movement of the noun that isolates one of the key-themes in the poem occurs in the first sentence of each of the epic texts of the AGDT, with the only exception of the *Shield of Herakles*.

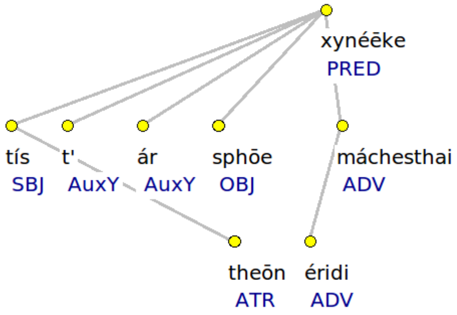

Figure 6: *Iliad* 1.8

Another case where we can observe the isolation of one focus element at the beginning of the sentence is with the interrogative pronoun *tís* ("who/which?"). Often, this left-movement separates the pronoun from the determiners that further specify it; in the case of example 7 (fig. 6) the pronoun is separated from the partitive genitive (*tís*. . . *theōn*) by two particles (*t'* and *ár*) and by the direct object of the verb (*sphōe*)[16].

(7) *tís t' ár sphōe*
who and then them-two.DU.ACC
theōn éridi xynéēke
of-the-gods with-strife pitted
máchesthai?
to-fight?
Who was it of the gods who pitted the two against each other so that they contended in strife? (*Iliad* 1.8)

Predicative adjectives, which specify the manner of the action expressed by the verb but agree with a nominal head, are very frequent in AG. They are syntactically dependent on the agreeing noun, but they modify semantically the verb as well. This sort of "double gravity" is a potential source of non-projective constructions. Often, as it is the case with *autómatos* ("of his initiative, unbidden") in example 8 (fig. 7), predicative adjectives convey the most salient information in the sentence and are therefore attracted toward a pre-eminent position.

[15] Note that even in the first clause the contrastive contextually bound subject (*hoi*) and the verb (*epépleon*) are non-contiguous, with the particle *mén* placed in the gap.

[16] The fact that the identity of the god is the focus of the question is evinced from the answers that is given (as nominal sentence) in the line that follows: "the son of Leto and Zeus. For he, angered against the king" etc.

(8) *autómatos dé hoi ēlthe boēn*
unbidden.NOM and to-him came cry
agathós Menélaos.
good Menelaos.NOM?
And Menelaos, good at the war-cry, came to him unbidden (*Iliad* 2.408)

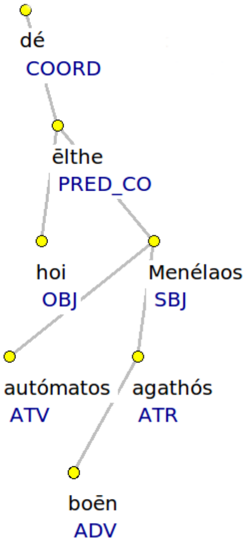

Figure 7: *Iliad* 2.408

6 Conclusions and future work

Our survey has confirmed the remarkable extension of non-projectivity in the Ancient Greek Dependency Treebank (1.7). The AGDT stands out for the relevant amount of both non-projective trees and edges, which are unmatched by the rate of discontinuous structures known from dependency treebanks of other languages and for the complexity of these constructions.

The edge-degree and gap-degree measures of non-projective trees from the AGDT are equally unmatched. In particular, the non-neglectable rates of trees with gap-degree ≥ 2 (which include more than 6% of the sentences in the AGDT) contradicts the assumptions that were inferable from other languages. On the other hand, in spite of these peculiarities, AG data confirm other conclusions that were drawn in previous literature, especially about the efficacy of the well-nestedness constraint.

The peculiar nature of these results may partially depend on the genres represented in the corpus, more than 98% of which is taken from poetic texts. The only prose work that is included in the collection shows a lesser degree of non-projective trees and edges, without conforming, however, to the rates known from other languages.

We have also isolated a number of specific constructions and we have tried to highlight some linguistic factors that can bring about syntactic discontinuity. Section 5 does not want to be an exhaustive classification of the linguistic aspects that stand behind non-projectivity: further work is required. Especially, on account of the well known influence of topic-focus articulation on AG word-order, this research would greatly benefit from the interaction of layers of syntax, pragmatics and information structure in annotated data.

References

David Bamman, Francesco Mambrini, and Gregory Crane. 2009. An ownership model of annotation: The Ancient Greek Dependency Treebank. In *Proceedings of the Eighth International Workshop on Treebanks and Linguistic Theories (TLT 8)*, pages 5–15, Milan. EDUCatt.

Riyaz Ahmad Bhat and Dipti Misra Sharma. 2012. Non-projective structures in Indian language treebanks. In *Proceedings of the 11th Workshop on Treebanks and Linguistic Theories (TLT11)*, pages 25–30, Lisbon. Colibri.

Alena Böhmová, Jan Hajič, Eva Hajičová, and Barbora Hladká. 2001. The Prague Dependency Treebank: A three-level annotation scenario. In Anne Abeillé, editor, *Treebanks: Building and Using Syntactically Annotated Corpora*, pages 103–127. Kluwer, Boston.

S. Buchholz and E. Marsi. 2006. CoNLL-X shared task on multilingual dependency parsing. In *Proceedings of the Tenth Conference on Computational Natural Language Learning (CoNLL-X '06)*, pages 149–164, Stroudsburg, PA, USA. ACL.

Ralph Debusmann and Marco Kuhlmann. 2010. Dependency grammar: Classification and exploration. In Matthew W. Crocker and Jörg Siekmann, editors, *Resource-Adaptive Cognitive Processes*, pages 365–388. Springer, Berlin and Heidelberg.

Andrew Devine and Laurence Stephens. 2000. *Discontinuous Syntax: Hyperbaton in Greek*. Oxford University Press, Oxford.

Helma Dik. 1995. *Word Order in Ancient Greek: A pragmatic account of word order variation in Herodotus*. J.C. Gieben, Amsterdam.

Helma Dik. 2007. *Word Order in Greek Tragic Dialogue*. Oxford University Press, Oxford.

Eva Hajičová, Petr Sgall, and Daniel Zeman. 2004. Issues of projectivity in the Prague Dependency Treebank. *Prague Bulletin of Mathematical Linguistics*, 81:5—22.

Jiří Havelka. 2007. Beyond projectivity: Multilingual evaluation of constraints and measures on non-projective structures. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, volume 45, pages 608–615, Prague. ACL.

Marco Kuhlmann and Joakim Nivre. 2006. Mildly non-projective dependency structures. In *Proceedings of the COLING/ACL 2006. Main Conference Poster Sessions*, pages 507–514, Sidney, Australia. ACL.

Francesco Mambrini and Marco Passarotti. 2012. Will a parser overtake Achilles? First experiments on parsing the Ancient Greek Dependency Treebank. In *Proceedings of the 11th Workshop on Treebanks and Linguistic Theories (TLT11)*, pages 133–144, Lisbon. Colibri.

Solomon Marcus. 1965. Sur la notion de projectivité. *Mathematical Logic Quarterly*, 11(2):181–192.

Dejan Matić. 2003. Topic, focus, and discourse structure: Ancient Greek word order. *Studies in Language*, 27(3):573–633.

J. Nivre, J. Hall, S. Kübler, R. McDonald, J. Nilsson, S. Riedel, and D. Yuret. 2007. The CoNLL 2007 shared task on dependency parsing. In *Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL 2007*, pages 915–932, Prague, Czech Republic. ACL.

Joakim Nivre. 2006. Constraints on non-projective dependency parsing. In *Proceedings of EACL-06. Trento, Italy*, pages 73–80, Trento. ACL.

Marco Passarotti and Paolo Ruffolo. 2010. Parsing the Index Thomisticus Treebank. Some preliminary results. In *Latin Linguistics Today. Akten des 15. Internationalem Kolloquiums zur Lateinischen Linguistik*, volume 137, pages 714–725. Institut fur Sprachwissenschaft der Universität Innsbruck.

Cornelis J. Ruijgh. 1990. La place des enclitiques dans l'ordre des mots chez Homère d'après la loi de Wackernagel. In Heiner Eichner and Helmut Rix, editors, *Sprachwissenschaft und Philologie. Jacob Wackernagel und die Indogermanistik heute*, pages 213–33. Reichert, Wiesbaden.

TLG. 2010. The TLG beta code manual 2010. http://stephanus.tlg.uci.edu/encoding/BCM2010.pdf.

Jacob Wackernagel. 1892. Über ein Gesetz der indogermanischen Wortstellung. *Indogermanische Forschungen*, 1:333–436.