Treebanks in historical linguistic research

Dag Haug*

1 Introduction

Historical linguistics has always been based on corpora. By definition, it deals with stages of language that are no longer spoken. Therefore the historical linguist is deprived of other data sources such as native speaker intuitions (unless reported in the historical record), or psycholinguistic experiments. All we have is the historical record. This applies both to diachronic historical linguistics, and to the activity that necessarily precedes it, the synchronic analysis of older language stages – there is nothing but text. Being limited to a single type of data we must make the most of it.

Corpus linguistics is often seen as a strictly empirical approach to linguistics, both by proponents who see a virtue in this and by critics who deride the empiricist stance. But corpus linguists are not limited to a pure empirical approach, and in practice, they rarely adopt one. For there is no observation without theory. All we have, literally speaking, is ink on parchment (or papyrus, or paper). Everything else is interpretation and hence informed by theory.

Many linguists are happy to leave the most basic issues to the philologists, who read and collate the manuscripts and produce authoritative editions of the texts. We will do the same here and assume that the texts that make up the corpus on which we base our linguistic studies are given once and for all.[1] Even so, the linguistic analysis presupposes all kinds of theoretical considerations. We group word forms into lexemes, analyze them morphologically, determine their syntactic

*I thank Lieven Danckaert, Rosemarie Lühr and Carlotta Viti for useful comments on a previous version of this paper.

[1] In practice the question is not so easy, in particular when we deal with rare or deviant forms. In such cases, there is often feedback from the linguistic analysis to the establishment of the correct text.

function and interpret them. It is worth keeping in mind that *all* these categorizations reflect linguistic theory. It is often only when syntax makes its appearance that linguists with an empiricist bent conclude that we deal with theoretical linguistics, as if morphology was inherently less theoretical. I believe it is not, but it is probably true that for most languages, there is less controversy around the morphological analysis than around the syntax. For that reason, we need to be more – not less – aware of issues of syntactic theory, and enrich our corpora with theoretically informed syntactic annotation to create *treebanks*, which can help decide the theoretical questions.

The structure of the paper is as follows. In section 2 we describe the problems of traditional corpus-based research. In section 3, we describe how corpora with sophisticated linguistic annotation help overcome these challenges. Section 4 we illustrate this with a case study. Finally, section 5 draws some conclusions. Throughout the paper, the focus is on methodology. Moreover, I am only concerned with the synchronic analysis of historical language stages. I believe this is a crucial prerequisite for good diachronic work. For a good overview of how computational methods can help us in the analysis of change itself, see Niyogi (2002).

2 Traditional corpora

The traditional corpora are merely collections of text in the form that the editor left them, as books in the old days, or often as computer text files today. It is always present, either explicitly or implicitly, as the data material on which a study is based. A traditional corpus of this kind requires much manual work. If you want to study, for example, the meaning of aspect in Herodotus, there is little you can do but to read through the entire text. If you have an electronic version of your corpus, such as the Thesaurus Linguae Graecae, there is still little else you can do, for there is no easy way to retrieve all verbal forms from a pure text corpus, let alone sort them according to their aspectual properties.

This feature is to some extent an advantage, for it requires the linguist to engage extensively with the primary material and form ideas about how, say, word order in the New Testament works. There is probably no substitute for this approach when it comes to hypothesis formation. In practice, of course, hypotheses are often formed through a less systematic process than a thorough corpus examination, but it will always reflect the researcher's knowledge of what the primary material looks like.

	Luke + Acts		Luke only	
	Rife(1933)	Davison(1989)	Rife(1933)	Kirk(2012)
VSO	15	20	9	14
SVO	50	56	19	13
SOV	9	8	8	5
VOS	3	4	2	3
OVS	6	6	1	1
OSV	1	1	0	1

Table 1: Word order in Luke/Acts according to various authors

But while traditional corpora are fine for hypothesis *formation*, they have severe limitations in hypothesis *testing*. This is because assumptions are not made explicit. In some cases, even the most basic assumptions are left implicit and the edition that was used is not mentioned. But in most cases the implicit assumptions relate not to such basic matters, but to issues that influence the way researchers categorize their material.

It is easy to illustrate this problem. Table 1 shows the findings of three different scholars on the word order in declarative main clauses in the Gospel of Luke, or the Gospel of Luke + the Acts (which were both written by Luke).

We see that the numbers differ between the scholars, even though they all claim to report the same, reasonably basic fact: word order in declarative main clauses where there is a nominal (as opposed to pronominal or zero) exponent of both the subject and the object argument. Such raw frequency facts are in themselves perhaps not very interesting, but they are input to other, higher-level debates, such as what (if any) the basic word order is in NT Greek, or to what extent NT Greek is influenced by Semitic.[2] How are we to settle such questions if we cannot even agree on the raw facts?

It is instructive to reflect on how the different authors arrived at their numbers, to draw some methodological lessons. Of course, some of the differences could be due to clerical errors. This in itself bears a methodological lesson: while we cannot eliminate the possibility of clerical errors, we can reduce their impact by making sure our study is replicable by other scholars.

Rife gives the least detail about how his investigation was performed. He only states the criteria for inclusion in the data set: “The investigation was limited to

[2] The basic word order in Hebrew is VSO, and many scholars think that influence from Hebrew led to increased frequency of VSO in Greek.

main declarative clauses where both subject and object are substantives." (Rife, 1933, p. 250) There is no indication of what edition he used. Essentially, we can trust his numbers or not, but there is little else we can use his data for, except (if we trust them) cite them as evidence for some hypothesis. It should be said that Rife follows a tradition which was very common at the time, and perhaps understandable at a time when the only available scholarly medium was paper. He could have told us what edition he used, and he could even have told us exactly which passages in the NT that he counted as cases of one or the other word order. But even if he had done so, we could do little except reread and recount if we wanted to replicate his study. And given that he is not very explicit about his criteria, it is not actually clear that we would be replicating his study.

Davison (1989) belongs to another world, that of budding philological use of computers. His study was based on the so-called Analytical Greek New Testament developed by Barbara and Timothy Friberg. This is a digital text of Aland, Black, Martini, Metzger and Wikgren's 1975 Bible edition published by the United Bible Societies (3rd edition), where each word has been tagged with detailed morphological information.[3]

As Davison acknowledges, such a text offers only limited help for the study of word order. He wrote a program to locate "clauses . . . which contained at least one nominative noun, one accusative verb and one indicate verb . . . Verbs normally followed by a genitive or a dative were traced using a concordance" (Davison, 1989, p. 24-25). (Incidentally, the last sentence gives a possible explanation of why Davison's numbers are mostly higher than Rife's - it is possible that the latter only counted accusative objects.) This material was then examined manually.

If we would want to replicate Davison's studies, there would be several possible sources of errors. First, his program located *clauses* with some specific properties, but the original text does not mark clause boundaries, so we would have to guess how his program approximated a definition of clausehood. (Perhaps using punctuation?) Second, the original text does not mark subjecthood and objecthood, so we would have to guess how Davison decided on this. In practice, subjecthood is relatively uncontroversial in Ancient Greek, but objecthood less so, especially when we consider genitive or dative objects. In sum, though Davison was explicit about his source, he was not explicit about his criteria. So again, if we wanted to replicate his study, we would have to reread and recount.

Even if Davison had made his notions of clausehood, subjecthood and objecthood explicit, we would still be left in the blue, for there are many more assump-

[3] A printed version was published as Friberg and Friberg (1981).

tions that must be made explicit. This is shown clearly in Kirk (2012), which is a linguistically much more sophisticated study of NT word order. She mentions the following criteria

- The clause contains at least an S(ubject), V(erb) and O(bject)
- The clause is continuous
- S and O are not embedded in a participial clause
- The verb assigns accusative, genitive, or dative to an argument that is a patient or theme
- The V consists of one word (no periphrastic forms, modal embeddings or light verbs)
- S and O are determiner phrases (this includes nominalizations) or quantifier phrases, and not clausal
- S and O are continuous strings

The point here is not to discuss the adequacy of these restrictions, but to illustrate how a deceptively simple task such as determining the frequencies of various word orders requires a large number of underlying assumptions. This not something we can get around – we could lift the restriction that the clause must be continuous, for example, but we would still make an assumption, only a different one. Which set of assumptions is preferable is a question for linguistic theory, but here we are only concerned with methodology. In that respect, the key feature, as we saw, is replicability. Could we redo the study with the same assumptions and receive the same answer?

Kirk is admirably concrete in stating her assumptions in a separate appendix (Kirk, 2012, p. 259-268), but even here some questions remain. For example, her criterion that something is an object is that it bears a patient or theme thematic role. These notions are not themselves entirely clear. But Kirk does include a list of the verbs that were included on this criterion, so we could replicate her study.

Another question looms, however. For if we want not only to replicate the study, but assess the correctness of its assumptions, we would need to know not only what was included, but what was excluded. What are the verbs that govern dative or genitive arguments that are not patients or themes? Or on a different criterion – how much does the picture change if we do include discontinuous clauses? And by the way, how many clauses were in fact excluded on the basis of discontinuity?

Asking such questions is also an important part of hypothesis testing. Kirk excludes discontinuous clauses because their structure is poorly understood. It could conceivably be different from that of continuous clauses, with the consequence that word order follows entirely different principles. But it could also follow the same principles. In any case, we would like to know how much material was excluded from the study on this criterion. This is not strictly about replicability of the study, but about controlling that the data was not reported in a selective manner.

There is only one possible way to achieve that, namely for researchers to share all the data underlying their research. This means not only the data that was subjected to analysis, but also the data that was potentially relevant but eventually discarded, such as discontinuous clauses. Kirk does not do this, and for good reason: it is not something that can easily be done in a book format.

To sum up, the reason we still – despite Rife (1933); Davison (1989); Kirk (2012) (and several others) – do not know the frequencies of the various word orders in the New Testament is that the question is much more complicated than its simple appearance betrays. To answer it we need to know not just what the frequency is under some specific set of assumptions, but what it is under all possible sets of (reasonable) assumptions. Mathematically, we need to describe a function from possible assumptions to frequency distributions.

There is limited possibility to do this in the book/article format. But a culture of secrecy and lack of willingness to share data also inhibit progress. For example, the Fribergs' Analytical Greek New Testament, on which Davison's study was based, is not freely available, but must be purchased. It is entirely reasonable that researchers want something back from the time, energy and funding they put into creating such a resource, but the unfortunate result is that the work cannot be modified and restributed. If for example Davison had added his assumptions about clausehood, subjecthood and objecthood to the source files, he could not have distributed this data, since it would also give away the Fribergs' original work.

The result is that researchers spend a large amount of time redoing work that others have done before them. Although Rife, Davison and Kirk all had different assumptions and therefore arrived at different results, it is clear that there is a core of data that were counted by all of them independently. Moreover, although we could see each separate study as a partial description of the function from assumptions to word order frequency distributions, there is no obvious way to put together their respective answers to an overall picture, because their source data is not available. The answer to all this, I claim, is to create structured corpora that are annotated for all the assumptions that go into the research based on them. In the domain of syntactic research, this means *treebanks*, i.e. a corpus that is annotated with syntactic structures, and in most cases also morphological information and lemmatization.

3 Treebanks in historical linguistic research

The PROIEL project developed a parsed corpus of the Greek New Testament as well as several of the early translations (Haug and Jøhndal, 2008; Haug et al., 2009). As our example study will be drawn from the PROIEL corpus, we will start by describing the information in the corpus.

The edition used in the corpus is Tischendorf (1869–1872). This is a relatively old edition, which was chosen because it is in the public domain and could be freely distributed. It is also a well-known and respected edition, although obviously not up to date on e.g. the last century's papyrus finding. The text had already been tagged and lemmatized by Ulrik Sandborg-Petersen.[4] We used this morphological analysis as the basis for our own, but the whole text was gone through anew, first by annotators, and then by reviewers correcting the annotator's work.

The morphology of Greek is well understood and not really controversial, so it is safe to say that the annotation in most cases reflect a scholarly consensus. Sandborg-Petersen's tagging also had a high quality, so the changes we made to the original annotation were mostly not error corrections, but the application of more fine-grained distinctions in some domains. For example, the traditional analysis that Sandborg-Petersen followed does not distinguish between subordinating and coordinating conjunctions, or in modern terms complementizers/subjunctions and conjunctions. But this distinction is crucial in a corpus with syntactic annotation.

Such cases are few, however, and the morphological analysis is mostly uncontroversial. The same cannot be said about the syntax. As we already saw, Greek has a relatively free word order, of the kind that has triggered much discussion in the theoretical linguistic literature. It is enough in this context to point to the seminal works of Hale (1983) on non-configurational languages and Kiss (1995) on discourse configurational languages. The analysis of such languages in terms

[4] Sandborg-Petersen's files are available on https://github.com/morphgnt/tischendorf

of phrase structure is still a very controversial matter.

This prompts some cautionary remarks. Although as I will argue, there is much added value in organizing linguistic categorization in treebanks, it is also clear that in some sense, what you get out of it is determined by what you put into it. A treebank does not in itself define the *actual* assumptions of research based on it, but it defines the set of *possible* assumptions that a researcher can make in using it, for the assumptions must be framed in terms of the source annotation.[5]

For example, if the annotation is structured around a configurational analysis, it is hard to use the corpus to question the basic phrase-structure analysis itself. Let us assume for a moment that the corpus takes a strict configurational approach along the lines of early Chomskyan syntax and defines the subject as the NP c-commanding VP, perhaps from specIP, and the object as the sister to V^0. We can now recast the question of word order frequecy as a question about the frequency of right- and left-branching in IP and VP. But we cannot easily question the existence of a VP in the first place.[6]

The existence of a VP in free word order languages is indeed controversial and we believe it is best not to preempt a conclusion. There are several ways to avoid this. Phrase structure based corpora, such as the family of corpora from the Linguistic Data Consorium at UPenn, use a much flatter phrase structure than any practicioners of theoretical phrase structure grammars assume and thereby avoid many contentious decisions. The other option, which was taken in the PROIEL corpus, is to use a dependency-based analysis, where grammatical relations, such as subject, object, and adverbial, are taken as primitive.

There are numerous practical advantages of using a dependency-based analysis, especially for free word order languages. For example, when faced with a discontinuous NP where, say, a modifier has been separated from its head, a phrase structure analysis is forced to make a number of non-trivial decisions about how to analyze the structure. These make the annotation more error-prone and theory-dependent. By contrast, a dependency analysis will simply mark the modifier as a dependent of its head, in spite of the word order. This does not free us from thinking about why the modifier ended up where it did. But we do not have to encode hypotheses about this in the annotation and therefore, as we will see, the dependency analysis is actually better suited for research on constituency.

[5] We ignore the possibility of doing theoretically motivated transformations of the source data, which can be a powerful technique in corpus linguistics.

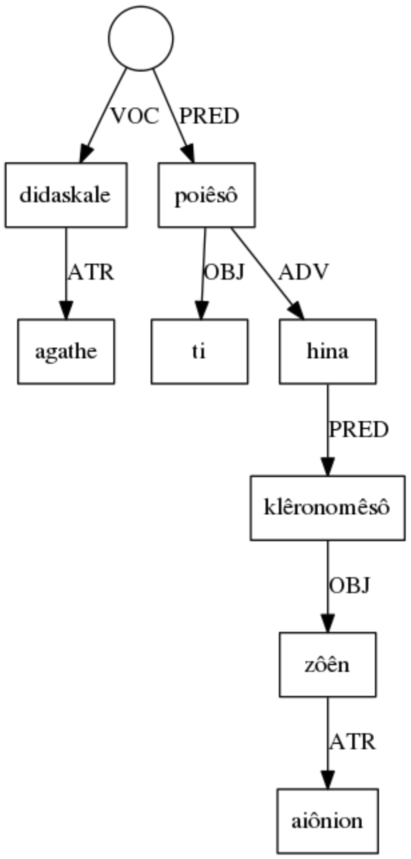
[6] On the other hand, the *annotation* of the treebank could be seen as a hypothesis test, for if the analysis is wrong, it would be impossible to apply it consistently in annotation. But we focus here on the use rather than the creation of treebanks.

On the downside, dependency grammar (DG) is not well-developed as a linguistic theory. There is an emerging branch of theoretical dependency grammar, spurred by its success in computational approaches, as evidenced by the recent conference series on dependency linguistics (Depling). And there have been some applications to ancient languages such as Latin (Happ, 1976; Kienpointner, 2010). But there is little that would satisfy the more theoretically oriented linguist. However, ideas from dependency grammar have been very influential in some linguistic theories: in particular, the functional structures of Lexical-Functional Grammar (Kaplan and Bresnan, 1982; Dalrymple, 2001; Bresnan, 2001) in fact encode dependencies and use grammatical relations as primitives to do so. Moreover, it makes the explicit claim that many facts about universal grammar are better captured in terms of these grammatical relations than in terms of phrase structuree. The annotation scheme of PROIEL is motivated by this theoretical framework – especially in the points where it deviates from strict DG.

DG has become a de facto standard in computational work, to a large extent because of its simplicity. On standard DG assumptions, a sentence is analysed as a set of asymmetric head-dependent relations between the words of the sentence, such that the relations form a tree rooted in a designated node which dominates the main predicate as well as any material that does not belong to the sentence, such as vocatives, parenthetical predications etc. (2) gives a sample analysis of the sentence in (1).

(1) didaskale agathe ti
teacher.SG.M.VOC good.SG.M.VOC what.SG.N.ACC
poiêsô hina zôên aiônion
do.1.SG.PFV.PST.SBJV.ACT that life.SG.F.ACC eternal.SG.F.ACC
klêronomêsô
inherit.1.SG.PFV.PST.SBJV.ACT
‘Good teacher, what should I do to inherit eternal life? (Mark 10.17)

(2)



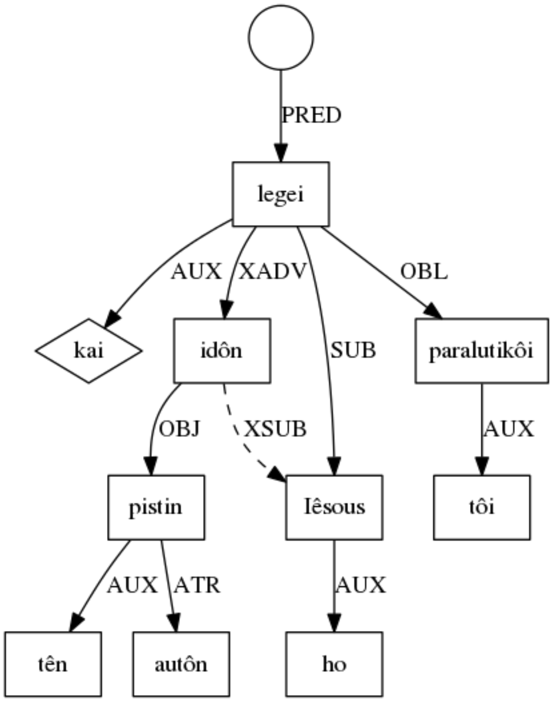
The arrows symbolize head-dependent relations, and are labelled with grammatical relations such as VOCative, ATtRibute, PREDicate and OBJect.

However, there are well-known linguistic structures that violate the underlying (and greatly simplifying) assumption that a word has a unique head. Control structures are a case in point. We will illustrate this with adjunct control. It is quite common in Greek that an action that is somehow related to that of the main verb is expressed with a so-called predicate participle, whose subject is an argument shared with the matrix verb. In such cases, the shared argument has two heads, rather than a single one.[7] An example is shown in (3), together with the PROIEL corpus analysis in (4).

[7] There are various technical ways of overcoming this conclusion, for example by introducing empty categories such as PRO. But these too violate the basic assumptions of dependency grammar.

(3) kai idôn ho Iêsous
and see.SG.PFV.PST.PTCP.ACT.M.NOM the.SG.M.NOM Jesus.SG.M.NOM
tên pistin autôn legei
the.SG.F.ACC faith.SG.F.ACC him.3.PL.M.GEN say.3.SG.PRES.ACT
tôi paralutikôi
the.SG.M.DAT paralytic.SG.M.DAT
'Seeing their faith, Jesus said to the paralytic.'(Mark 2.5)

(4)



We see that in this analysis, *Iêsous* has two heads. It is the ordinary SUBject of the matrix verb *legei*, and the external subject (XSUB) of the participle *idôn*.

It would be futile to deny that representations such as in (2) and (4) are based on linguistic theory. However, they are reasonably theory neutral in that they are based on concepts that are recognized in most linguistic theories. It is useful in this respect to compare dependency trees with phrase structure trees, which are a very different linguistic structure, but still incorporate many similar concepts.[8]

[8] We chose to illustrate 'theory-neutrality' by comparing the annotation with what one expects from a phrase structure analysis, since this is a very different framework. Other formalisms, and traditional grammatical analysis, are much closer to DG.

For example, the notion of heads and dependents, while not native to phrase structure representations, have been introduced via X′ theory and the notion of government (Chomsky, 1986, p. 8-9). Multiple headedness, as in (4), are captured either via movement (permitting a single constituent to be appear in several positions in a structure), coindexed empty categories or even multidominance in recent minimalist work (see e.g. Citko (2011)). Finally, grammatical relations, though not viewed as primitive, are often defined in terms of phrase structure configurations. So although configurational theories will not accept e.g. the SUB relation that appears in (4) as a theoretical primitive, it will reconstruct the same notion in terms of phrase structure configurations, in a way that makes it possible to account for the same phenomena related to case, agreement, binding etc.

Representations such as (2) and (4) therefore contain structures that most linguistic theories will acknowledge in some way or another.[9] In addition to the structure as represented in the tree, the corpus also contains information about the linear order of the words in the sentence. However, one thing that is *not* represented in the corpus, is a notion of constituency. And that is precisely why a dependency corpus is ideal for investigating constituency.

We will show this in a case study, but before we do that, let us reflect on the nature of the relationship between corpus research and linguistic theory. We observed in the introduction that when we do historical linguistics, there are no native speakers' intuitions to be had and corpus data is all we have got. This is most worrysome if your theoretical persuasions tend towards mentalism, but even if not, the absence of negative data in corpora is a problem.

Chomsky (1957), for example, famously argued that since neither (5) nor (6) had ever occurred in an English discourse, a statistical model of grammaticality would rule both out as equally remote from English, even if (5) is grammatical and (6) is not.

(5) Colorless green ideas sleep furiously.

(6) *Furiously sleep ideas green colorless.

But Chomsky's argument is correct only if the statistical model assigns zero probability to all unseen events. A model that does so badly overfits the training data. Not all statistical models are like this, however. As shown in Pereira (2000), a simple probabilistic model trained on newspaper text can in fact estimate that the probabilities of (5) and (6) differ by five orders of magnitude. The theoretical up-

[9] But it is not certain - in fact highly improbable - that all linguists working on Greek will agree with the actual application of these structural concepts to the data in the treebank.

shot of this is that by careful statistical analysis, we *are able* to provide something approaching negative evidence, if the data are sufficient. It is true that statistics will never provide us with categorical judgments, but then speakers' judgments are also not as categorical as linguists' report of them.

4 Case study

Let us now look at how we can use the PROIEL treebank to examine how constituency works in Greek.

First, observe that in (3), *Iêsous*, the subject of the main clause, appears embedded in the participle clause of which it is also the subject. This is not visible in the dependency tree, since such trees by their nature abstract away from the relationship between grammatical relations and word order. But the word order in (3) raises deep questions about the configurational structure and how the two subject positions are linked. Concerns about this were the reason why Kirk (2012) removed such sentences (with transitive matrix verbs) from her data, so solving the question also has repercussions for the study of word order overall.

There are (at least) three hypotheses we may want to entertain about the structure of (3).

1. The main clause verb has an empty subject (Ancient Greek is a "prodrop language") which refers anaphorically to *Iêsous*

2. Ancient Greek word order is free not just inside clauses but even across clauses (at least participial ones)

3. *Iêsous* appears in the participle clause but controls the subject position in the matrix

The first hypothesis seems the simplest one, but it can safely be rejected because the nominative case on *Iêsous* comes from the matrix clause.[10] We are therefore left with the two hypotheses 2 and 3.

Both of these hypotheses imply that the two subject positions are related via control. According to 2, *Iêsous* is structurally the subject of *legei* and controls the subject position of *idôn*. The surface order is the result of a relatively shallow word order rearrangment. According to 3, it is the other way around: *Iêsous* is

[10] In the interest of space, we forgo a demonstration of this fact. It is not controversial in Greek grammar.

	discontinuous	continuous
Finite clauses	0	9894
Infinitive clause	36	770
Complement ptcp. clause	6	205
Absolute ptcp. clause	0	167
Conjunct ptcp. clause (ext. subj.)	48	1260
Conjunct ptpc. clause (int. subj.)	0	1308

Table 2: Discontinuities in clausal categories in the NT

structurally the subject of *idôn* and controls the subject position of *legei*. The surface order reflects the structural relations.

The third hypothesis involves 'backward control' (from a structurally lower clause into a structurally higher on), which has important theoretical consequences for control theory (see for example Polinksy and Potsdam 2002a) that must be resolved before the analysis is viable. But here we focus on the interpretation of the Greek data.

Hypothesis 2 implies a much freer view of Greek word order than 3, and we can use a corpus to test whether these predictions are borne out. In particular, hypothesis 2 predicts that participial clauses can be discontinuous in the surface syntax. Refuting this prediction in in principle requires negative data, i.e. judgments that a surface discontinuous participial clause is ungrammatical. The corpus will not give us this directly, so we will instead test the prediction in two stages.

First, observe that on hypothesis 2, the fact that *Iêsous* is (directly or indirectly) the subject of *idôn* plays no role in licensing its surface position. So hypothesis 2 leads us to expect to find participial clauses that are interrupted by material that is functionally external to them. To check this we can investigate the surface continuity of participle clauses both on the assumption that their subject can be internal to the them and on the assumption that they cannot. The results are shown in Table 2, together with continuity data from other clausal categories.[11]

We see that out of 1308 participle clauses, 48 are discontinuous. But all the discontinuities disappear if we consider the subject as internal to the participle clause. In other words, all discontinuous participle clauses are of the type in (3). There is no independent motivation for the acceptability of discontinuous

[11] Note that a long-distance wh-extraction is not counted as a discontinuity, as it results from a well-defined syntactic process found also in languages with strict(er) word order.

participle clauses.

This is still a far cry from negative evidence, however. Although it is hard within hypothesis 2 to come up with a syntactic explanation of why it is always the participle's subject that intervenes, there could conceivably be a pragmatic explanation. For example, one could argue that *Iêsous* comes close to *idôn* because the two words 'belong together' (Behaghel's law, Behaghel 1932, p. 4–7). To test this hypothesis, we can compare controlled participial adjuncts to a group of clauses that show clear signs of discontinuity, namely subject control infinitives. In Table 3, discontinuous participle and control infintive clauses are categorized by the intruding material: is the clause's semantic subject (as in (3)), or the verb that governs the clause, or both, or entirely external material.

	subject	head	head-and-subject	external
Adjunct ptcp.	48	0	0	0
Control inf.	5	6	1	3

Table 3: Intrusion types in the Gospels (p=2.350e-08, Fisher's exact test)

We see that the distribution is very different. Control infinitives are mostly interrupted by their own head, according to the same pattern that we find in normal hyperbaton (for example the NP complement of a preposition split by the head P, or the object NP split by its governing verb), whereas participle clauses are only interrupted by their subjects. The pattern is statistically significant at a level strong enough to approach negative evidence: it is very likely that participle clauses interrupted by something other than their subject is ungrammatical in Ancient Greek. Theoretical considerations, which we skip over here, may then lead us to conclude that participle clauses are not at all discontinuous, because the subject is in fact internal to the clause, as in hypothesis 3.

We must end our brief investigation of Greek word order here. But I hope the case study shows the advantages of using corpora in this kind of research. In particular, we saw how a dependency treebank, which makes no assumptions about the constituency of clauses, can help us explore clausal structure. Moreover, the corpus analyses that we have seen are completely replicable, as everyone can download the underlying data and perform it themselves.

5 Conclusions and outlook

I hope the case study that we briefly went through in section 3 demonstrate the potential of treebanks for syntactic research on Greek and Latin. Although we considered a relatively specific phenomenon, the status of the subject in certain participle clauses, it is clear that the database can be used to address the more global question of Greek word order too: it does in fact contain information about word order in all clauses that it contains, and it is possible to access this information according to specific assumptions, to see for example, how the data changes if we consider discontinuous clauses in addition to the continuous ones, or perhaps include discontinuous subject and object noun phrases to the extent that can be ordered linearly.

At the same time, it is clear that word order in Greek is not just a matter of syntax. Information structure and discourse factors are certainly important in determining the order of words. The general nature of this influence has been worked out in several studies (see in particular Dik 1995; Matić 2003), but details are still unclear, in particular because information structural categories such as topic and focus are much more subjective than syntactic ones such as subject and object. Here too, a corpus approach would lead to better replicability. In the PROIEL corpus, we have been experimenting with such higher levels of annotation, and the entirety of the Greek gospels have been annotation for givenness. The results are encouraging and described in Haug et al. (2014). But the methodology lessons are essentially the same as described in the present article: Make your assumptions explicit and make your raw data public.

References

Behaghel, Otto. 1932. *Deutsche Syntax IV*. Heidelberg: Carl Winter.

Bresnan, Joan. 2001. *Lexical-Functional Syntax*. Oxford: Blackwell.

Chomsky, Noam. 1957. *Syntactic structures*. The Hague: Mouton.

Chomsky, Noam. 1986. *Barriers*. Cambridge, MA: MIT Press.

Citko, Barbara. 2011. Multidominance. In Cedric Boeckx (ed.), *The Oxford Handbook of linguistic minimalism*, pages 119–142, Oxford: Oxford University Press.

Dalrymple, Mary. 2001. *Lexical Functional Grammar*. San Diego: Academic Press.

Davison, M. E. 1989. New Testament Greek Word Order. *Literary and Linguistic Computing* 4, 19–28.

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

Friberg, Barbara and Friberg, Timothy. 1981. *Analytical Greek New Testament*. Grand Rapids, MI: Baker Book House.

Hale, Ken. 1983. Warlpiri and the grammar of non-configurational languages. *Natural Language & Linguistic Theory* 1(1), 5–47.

Happ, Heinz. 1976. *Grundfragen einer Dependenzgrammatik des Lateinischen*. Göttingen: Vandenhoeck & Ruprecht.

Haug, Dag, Eckhoff, Hanne and Welo, Eirik. 2014. The theoretical foundations of givenness annotation. In Kristin Bech and Kristine Eide (eds.), *Information Structure and Syntactic Change in Germanic and Romance Languages*, Amsterdam: Benjamins.

Haug, Dag and Jøhndal, Marius L. 2008. Creating a Parallel Treebank of the Old Indo-European Bible Translations. In *Language Resources and Evaluation*, Marrakech, Morocco.

Haug, Dag, Jøhndal, Marius L., Eckhoff, Hanne, Welo, Eirik, Hertzenberg, Mari and Müth, Angelika. 2009. Computational and Linguistic Issues in Designing a Syntactically Annotated Parallel Corpus of Indo-European Languages. *Traitement Automatique des Langues* 50, 17–45.

Kaplan, Ronald M. and Bresnan, Joan. 1982. Lexical-Functional Grammar: A Formal System for Grammatical Representation. In Joan Bresnan (ed.), *The Mental Representation of Grammatical Relations*, pages 173–281, MIT Press.

Kienpointner, Manfred. 2010. *Latein-Deutsch kontrastiv*. Tübingen: Julius Groos.

Kirk, Allison. 2012. *Word order and information structure in New Testament Greek*. Ph. D.thesis, Universiteit Leiden.

Kiss, Katalin É. 1995. *Discourse configurational languages*. Oxford: Oxford University Press.

Matić, Dejan. 2003. Topic, focus, and discourse structure. *Studies in Language* 27, 573–633.

Niyogi, Partha. 2002. The Computational Study of Diachronic Linguistics. In David Lightfoot (ed.), *Syntactic Effects of Morphological Change*, pages 351–365, Oxford: Oxford University Press.

Pereira, Fernando. 2000. Formal grammar and information theory: together again? *Philosophical Transactions of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences* 358, 1239–1253.

Polinksy, Maria and Potsdam, Eric. 2002a. Backward control. *Linguistic Inquiry* 33, 245–282.

Rife, J. Merle. 1933. The mechanics of translation Greek. *Journal of Biblical literature* 52, 244–252.

Tischendorf, Constantin von. 1869–1872. *Novum Testamentum Graece*. Leipzig: Hinrichs, 8. edition.