

# The Component Unit

## Introducing a Novel Unit of Syntactic Analysis

**Timothy Osborne**  
Zhejiang University  
Hangzhou  
China  
tjo3ya@yahoo.com

**Ruo Chen Niu**  
Zhejiang University  
Hangzhou  
China  
niuruochen@126.com

### Abstract

This contribution introduces a novel unit of syntactic analysis, which is called the *component*. The validity and utility of the component unit are established in terms of *chunking*. When informants organize the words of sentences into groups, they are creating *chunks*, and these chunks then qualify as components in dependency syntax. By acknowledging the nature of chunking and the component unit, it is possible to cast light on controversial aspects of dependency hierarchies. In particular, the component unit, informant data, and the reasoning based on these provide an argument in favor of the traditional DG assumptions about hierarchical status of many function words (auxiliary verbs, prepositions, subordinators, etc.), and in so doing, they contradict the Universal Dependencies (UD) annotation scheme. The data discussed here are from English, but the methodology and reasoning employed are easily extendable to other languages.

## 1 Introduction

The purpose of this manuscript is to probe the extent to which dependency syntax provides a basis for discerning how words are grouped together into units of meaning. The words that constitute sentences are of course not arranged arbitrarily, but rather they are grouped in such a manner that phrases and clauses can be acknowledged. According to the principle of *projectivity* (Hays, 1964; Gaifman, 1965; Robinson, 1970; Melčuk, 1988), dependents should be grouped together with their head, as opposed to together with one or more words that do not include their head. For instance, given a three-word

string such as *walk really fast*, a straightforward assumption is that the adverb *really* modifies the adverb *fast* and should hence be grouped together with *fast* before being grouped with *walk*. We therefore have *walk [really fast]*, not *[walk really] fast*.

While this analysis of *walk really fast* is not controversial, there are other cases where intuition about how the words should be grouped is not as clear. For instance, should an auxiliary verb be grouped first with the subject or with what follows it, e.g. *[I am] having lunch* vs. *I [am having lunch]*. Most phrase structure grammars (PSGs) would of course prefer the latter analysis. However, what does dependency syntax say about such examples? A DG analysis that subordinates both the subject *I* and the auxiliary verb *am* to the content verb *having* also predicts that the latter analysis, i.e. *I [am having lunch]*, should, for a reason discussed below, be preferred, whereas the alternative DG analysis, which positions both the subject *I* and the light verb *having* as immediate dependents of the finite auxiliary *am* predicts that neither one of the two groupings shown should be significantly preferred.

This manuscript makes and defends three major claims concerning the issue just sketched:

### Claim 1

Exactly how speakers of a language organize the words of sentences into groups can be determined by simple chunking data collected from informants.

### Claim 2

There is a novel unit of dependency syntax that helps predict how informants will chunk sentences. This unit is the *component*.

### Claim 3

One can use the component unit as a basis for motivating analyses of sentence structure. One can thus resolve areas of debate about the best hierarchical analysis.

Returning to the phrase *walk really fast*, informants can, for example, be prompted to divide the phrase into two chunks. The prediction in this area is that a significant majority of them will prefer to chunk the phrase as in (1a) rather than as in (1b):

- (1) a. walk | really fast  
b. walk really | fast.

The same experiment can be conducted on the sentence *I am having lunch*, whereby there are three potential responses:

- (2) a. I | am having lunch.  
b. I am | having lunch.  
c. I am having | lunch.

If a large majority of informants chunk the sentence as in (2a), one could then conclude that the auxiliary verb *am* can be grouped together with *having lunch* to the exclusion of *I*. If, in contrast, a large majority opts for the analysis in (2b), then one could conclude that *am* can be grouped with *I* to the exclusion of *having lunch*. If the sentence is chunked as in (2c) or there is a more even distribution of informant choices across (2a–c), then it is more difficult to acknowledge a clear grouping of the words in the sentence.

The component unit is the means by which the chunks just indicated in (1–2) can be interpreted. Our hypothesis is that informants prefer to chunk sentences in such a manner that the resulting chunks are components, whereby a component is *a word or a combination of words that form a string and are linked together by dependencies*.<sup>1</sup> This manuscript employs the component as the basis for shedding light on areas in which there is some disagreement among dependency grammarians about the best hierarchical analysis. In particular, it scrutinizes aspects of the Universal Dependencies (UD) annotation scheme.

<sup>1</sup> Another, more principled definition of the component unit is given in the next section.

## 2 Units of structure

The current DG is like many other DGs in understanding dependency as a one-to-one mapping of words to nodes and vice versa (e.g. Mel'čuk and Pertsov, 1987: 48, 57–8; Kahane, 1996: 45; Schubert, 1987: 78–86, 129; Engel, 1994: 25, 28; Bröker, 2003: 297; Hudson, 2007: 183). In addition, the current DG assumes trees and is monos-tratal in syntax, which means linear order (pre-cedence) and hierarchical order (dominance) are both primitive – as opposed to just hierarchical order being primitive and linear order being derived from hierarchical order. What this means is that the dependency trees assumed here always encode actual word order.

Given these assumptions about the nature of dependency syntax, key units of syntax can be defined as follows:

### String

A word or a combination of words that are continuous with respect to precedence

### Catena

A word or a combination of words that are continuous with respect to dominance

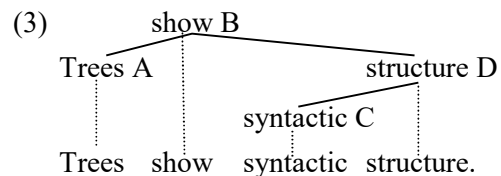
### Component

A word or a combination of words that are continuous with respect to both precedence and dominance

### Constituent

A component that is a complete subtree

These units are illustrated using the following dependency tree:



The capital letters abbreviate the words. All the distinct strings, catenae, components, and constituents in (3) are listed next:

### 10 distinct strings in (3)

A, B, C, D, AB, BC, CD, ABC, BCD, and ABCD

### 10 distinct catenae in (3)

A, B, C, D, AB, BD, CD, ABD, BCD, and ABCD

### 8 distinct components in (3)

A, B, C, D, AB, CD, BCD, and ABCD

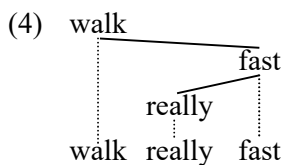
### 4 distinct constituents in (3)

A, C, CD, and ABCD

Of these four units, the focus below is on the component. The reason the other three are presented here together with the component is the desire to increase understanding of the one through comparison with the other three.

Most theories of syntax acknowledge strings, and the validity of the catena unit as just defined has been thoroughly established in a series of articles (e.g. O’Grady, 1998; Osborne et al., 2012). The constituent is generally viewed as a unit of phrase structure grammar. However, some DGs have also acknowledged constituents as just defined over dependency structures (e.g. Hudson, 1984: 92; Starosta, 1988: 105; Hellwig, 2003: 603; Anderson, 2011: 92).

While the component has been acknowledged in the DG literature (Osborne and Groß, 2016: 117), it has not been the focus of particular research efforts until now. It is therefore necessary to establish a solid understanding of this unit. To do this here now, the two examples discussed in the introduction above are examined more carefully. The first example:

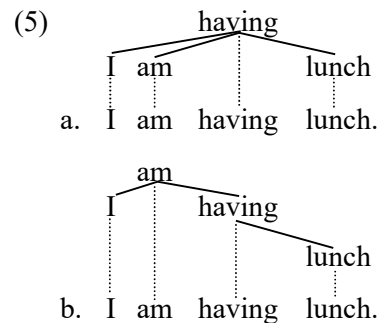


This hierarchical analysis is, as stated above, not controversial. Each individual word is a component by definition. The word combinations that are strings and consist of two words are of particular interest in this case, since predictions made about chunking apply directly to them. There are two two-word strings: *walk really* and *really fast*. The former of these is not a component according to the hierarchy in (4), whereas the latter is.

The prediction concerning chunking, then, is that informants will prefer to chunk this phrase in a manner that the two resulting chunks are component strings, as opposed to one of them being a

non-component string. In other words, informants will NOT chunk this phrase as *walk really | fast* because the chunk *walk really* would not be a component. They will instead chunk the phrase as *walk | really fast*, because the chunk *really fast* is a component (and so is the one-word string *walk*, of course).<sup>2</sup>

Turning to the second example, i.e. *I am having lunch*, there are two conceivable structural analyses that DGs are likely to pursue:



The analysis in (5b) has a long tradition in DG, reaching back to Franz Kern (1883, 1884). This tradition positions the finite verb as the clause root and then subordinates the subject to the finite verb. The type of analysis in (5a) has recently gained many adherents; it is the one advocated by the Universal Dependencies (UD) annotation scheme (e.g. de Marneffe et al., 2014).<sup>3</sup> This scheme systematically subordinates function words such as the auxiliary *am*, to the content words with which they co-occur.

The account of chunking in terms of components predicts that if the hierarchical analysis in (5a) is correct, then informants will prefer to chunk the sentence as *I | am having lunch* because the chunk *am having lunch* would then be a component; they would not chunk the sentence as *I am | having lunch*, because according to the hierarchy in (5a), *I am* would not be a component. The hierarchical analysis in (5b), in contrast, predicts that informants will chunk the sentence

<sup>2</sup> In our original rounds of data collection, we did not test the phrase *walk really fast*. In a follow-up round of data collection, however, we did test it. The informant responses strongly verified expectation:

- (i) walk | really fast – 30 responses
- (ii) walk really | fast – 1 response

<sup>3</sup> At the time of writing this manuscript (April 2017), an over view of the Universal Dependencies project and of its annotation scheme were available at the following web address: <http://universaldependencies.org/>.

as *I | am having lunch* or *I am | having lunch* or *I am having | lunch* because in all three cases, each of the chunks shown would be a component.

The informant responses we have collected resolve this issue and others. The hierarchical analysis in (5b), which corresponds to the more traditional stance towards the hierarchical status of auxiliary verbs, receives support. Auxiliary verbs are heads over the content verbs with which they co-occur.

### 3 Methodology

Two rounds of handouts were designed to obtain data that reveal how speakers chunk sentences. The instructions at the beginning of each handout provided an introduction to the chunking concept as well as illustrations of how a sentence might be divided into chunks. The handout then prompted the informants to chunk a number of sentences.

The first round of data collection, i.e. the pilot test, consisted of ten English sentences that varied in length and type. The handout was arranged in such a way that sentences of the same type and in the same length were randomly scattered. Participants were invited to divide the sentences into three chunks by using two dividers “|”.

The second round of data collection via a handout obtained participants’ responses to sentences of which the hierarchical structure is under debate. It consisted of two parts: part one was composed of five sentences containing auxiliary and content verbs, where informants were asked to divide the sentence into two chunks by inserting only one divider “|”; part two had fifteen sentences concerning controversial issues, such as the status of auxiliary verbs, the status of prepositions, and the status of object predicatives. Informants were invited to divide each sentence into three chunks.

All the informants involved in the surveys were undergraduate students learning English at a major university in China.<sup>4</sup> Their level of Eng-

lish was evaluated as intermediate to advanced, CET3 (College English Test Band 3). The simple sentences in each handout were easy for them to read and understand.

All the responses obtained from the informants were recorded using Microsoft Office Excel 2007. Exactly how informants divided each sentence and how many informants did so in that way, i.e. the tokens, were recorded below each sentence. Handouts containing responses that did not follow the requirements were excluded from recording. The number of handouts recorded for the pilot test and the second round was 46 (two excluded) and 43 (one excluded), respectively.

## 4 Discussion of results

### 4.1 Auxiliary verbs

As stated above, there are two competing analyses within DG regarding the status of auxiliary verbs. There is the traditional analysis that is assumed in DG frameworks such as Lexicase Grammar (Starosta, 1988), Word Grammar (Hudson, 1990, 2007) and Meaning-Text Theory (Mel’čuk, 1988), and in numerous prominent DG works such as Kunze (1975), Schubert (1987), Heringer (1996), Eroms (2000). The central status of the finite verb, which is an auxiliary verb if an auxiliary verb is present, reaches back to the earliest works in DG, namely to the treatises of Franz Kern (e.g. 1883, 1884) – Kern emphasized time and again the central role that the finite verb plays as the sentence root. The competing analysis is more recent; it is associated mainly with the annotation scheme of Universal Dependencies (UD) – see footnote 3.

Of the 26 initial sentences we tested on informants, 15 of them contained an auxiliary verb. The tendency in this area is that informants prefer to chunk the sentence immediately before the auxiliary verb if the subject is a noun (phrase) or immediately after the auxiliary verb if the subject

---

were simple sentences of English of the sort that certainly none of the informants had difficulty reading and understanding. The second is that we did a smaller, follow-up round of data collection from native informants, testing most of the key sentences presented in this manuscript. With one exception, the results we obtained from the native informants were similar to the results obtained from the much larger number of Chinese informants. This issue is acknowledged and discussed briefly in the concluding section.

---

<sup>4</sup> Since we were testing English sentences, native speakers of English would have been preferred as informants, of course. We unfortunately did not have access to large numbers of English native speakers at this stage of our project. Two important factors moderate this weakness in the informant responses. The first is that the sentences we tested

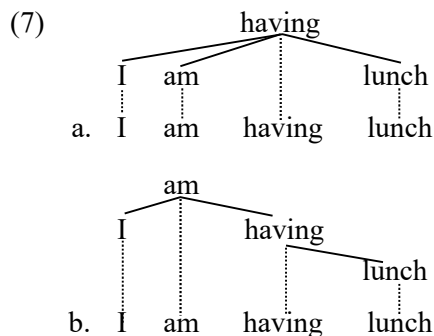
is a pronoun. This variation is best accommodated on the structural analysis illustrated above with (5b), where the finite auxiliary verb is the sentence root. If the finite auxiliary verb is the sentence root, both strings – the string consisting of the subject and the finite auxiliary as well as the string consisting of the finite auxiliary and everything following the finite auxiliary – qualify as components.

To make this point concrete, the results we obtained for the example sentence discussed above, i.e. *I am having lunch*, are presented next. When informants were asked to divide this sentence into two chunks, the following results obtained:

- (6) a. I am | having lunch. – 26 responses
- b. I | am having lunch. – 10 responses
- c. I am having | lunch. – 7 responses

These data reveal three things about how the words are organized into groups. The first is that they refute the initial binary division of the clause associated with most PSGs. Phrase structure syntax typically divides the clause into a subject NP and a predicate VP. If that division were real, the expectation would have been for a greater number of informants to chunk the sentence as in (6b). The fact that a significant majority of informants chose to chunk the sentence as in (6a) refutes the NP-VP division of most PSGs.

The second thing that the data in (6a–c) reveal is that the string *I am* is likely a component. This then refutes the UD analysis of auxiliary verbs. The two competing structural analyses are repeated here as (7a–b):



do, however, is reveal that prosodic factors influence which particular components will be chosen as chunks.

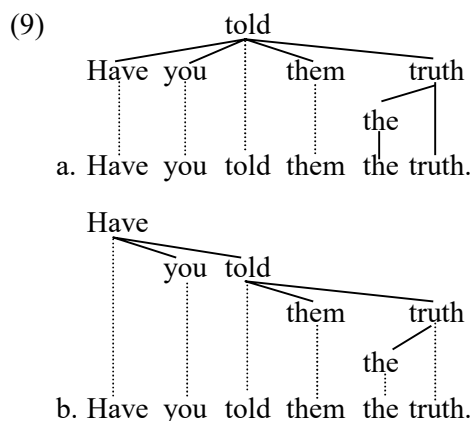
## 4.2 Subject-auxiliary inversion

Four of the sentences tested contained subject-auxiliary inversion. The responses we received in this area reveal that informants are reluctant to chunk between the subject and auxiliary verb. This reluctance again supports the traditional analysis which maintains a direct dependency between the subject and finite verb.

The four sentences we tested containing subject-auxiliary inversion are listed next: *Have you told them the truth?*, *Why did he quickly leave?*, *Did you send it out?*, and *Where did you go?*. The results for the first of these four sentences are provided here for discussion. The informants were invited to divide the sentence into three chunks. We received the following responses:

- (8) a. Have you | told them | the truth? – 39  
 b. Have you told | them | the truth? – 4  
 c. Have | you told them | the truth? – 2  
 d. Have you | told | them the truth? – 1

The two relevant and competing structural analyses of this sentence are as follows:



The analysis given as (9a) is that of UD; both the subject *you* and the auxiliary *have* appear as a dependent of the content verb *told*. The more traditional analysis is given as (9b); the finite verb, which is the auxiliary verb, is the root of the sentence there.

The fact that a large majority of the informants, 39 of 46, chose to chunk the sentence as in (8a) supports the traditional analysis given as (9b) over the UD analysis given as (9a). This conclusion follows from the status of the string *Have*

*you* as a non-component in (9a), but as a component in (9b). Observe also that each of the five chunks indicated in (8a) and (8b) is a component.

Worth considering in this area is that only 3 of the 46 informants chunked the sentence in a manner that was inconsistent with the traditional analysis given as (9b). The chunk *you told them* in (8c) is not a component on the analysis in (9b), and the chunk *them the truth* in (8d) is also not a component on the analysis in (9a) and (9b). Anomalous responses like these were not unusual. For most of the sentence we tested, there was a small minority of informants that chunked the sentence at hand in a manner that contradicted the traditional analysis. It was usually the case, however, that a large majority of informants chunked the sentence at hand in a manner that contradicted the UD annotation scheme.

The results for the other three sentences containing subject-auxiliary inversion were similar. The results we obtained were more consistent with an analysis that takes the subject and auxiliary verb as forming a component than with one where the two do not form a component.

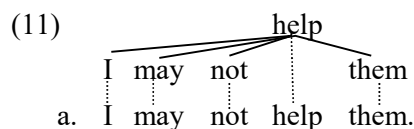
## 4.3 Sentence negation

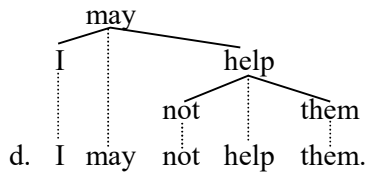
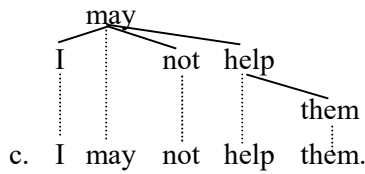
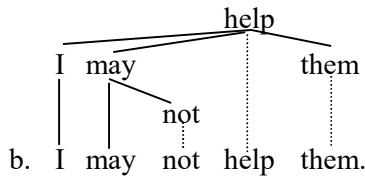
We tested two sentences containing an auxiliary verb and the standard clausal negation *not*. The results we obtained again support the traditional hierarchical analysis of auxiliary verbs over the UD approach. Further, the results we obtained also support an analysis that positions the negation *not* as a postdependent of the auxiliary verb.

The two sentences containing *not* that we tested were *Jill did not laugh* and *I may not help them*. The informants were invited to divide these sentences into three chunks. The results we obtained for the latter sentence were as follows:

- (10) a. I | may not | help them. – 27  
 b. I may not | help | them. – 7  
 c. I may | not | help them. – 6  
 d. I | may not help | them. – 5  
 e. I may | not help | them. – 4

Four potential structural analyses of this sentence are as follows:





The UD annotation scheme would likely pursue the analysis in (11a) or (11b), whereas more traditional assumptions would be along the lines of (11c) or (11d). Given the component unit and chunking data, it is possible to discern which of the four analyses is the best.

The chunking in (10a) and (10b) reveal first and foremost that *may not* and *I may not* should have component status. Since the analysis in (11c) is the only one of the four that grants both of these strings component status, it is preferable. Observe as well that the chunks indicated in (10c) and (10d) are also all components on the analysis in (11c). Only the chunking in (10e), which was produced by just four informants, contradicts the hierarchical analysis given as (10c), because the chunk *not help* in (10e) is not a component in (11c).

Concerning the other sentence containing *not* that we tested, i.e. *Jill did not laugh*, the results we obtained were as follows:

- (12) a. Jill | did not | laugh. – 44  
 b. Jill did | not | laugh. – 1  
 c. Jill | did | not laugh. – 1

These results are uninteresting insofar they do not clearly support one analysis over another, for if the negation *not* here is interpreted as a postdependent of the auxiliary verb *did*, similar to the analyses shown in both (11b) and (11c), then *did not* is a component on both accounts, the UD account and the traditional account.

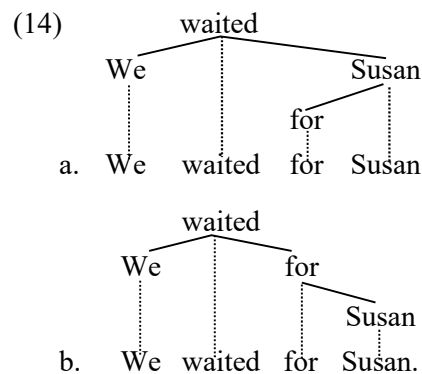
#### 4.4 Prepositions

Most DGs acknowledge prepositional phrases, that is, they view prepositions as heads over the nouns with which they co-occur. The UD annotation scheme, in contrast, positions prepositions as dependents of the nouns with which they co-occur. To shed light on these alternative analyses of prepositions, we included sentences containing prepositional phrases in our test sentences. The informant responses we obtained again support the traditional analysis over the UD approach.

Six of the sentences we tested contained a prepositional phrase. These six sentences are listed next: *Friends of mine are arriving now*, *I am in the classroom*, *One of the people protested*, *We are looking out for the teacher*, *He sleeps on his bed*, *We waited for Susan*. When invited to divide the last of these sentences into three chunks, the informants responded as follows:

- (13) a. We | waited for | Susan. – 32  
 b. We | waited | for Susan. – 6  
 c. We waited | for | Susan. – 5

The two relevant and competing hierarchical analyses of this sentence are given next:



The UD analysis is shown as (14a), and the more traditional analysis as (14b). The difference lies with the hierarchical position of the preposition.

The preferred way to chunk the sentence supports the traditional analysis. A large majority of informants, 32 of them, chunked the sentence in such a manner that *waited for* appears as a chunk. Since *waited for* is not a component on the UD analysis in (14a) but is a component on the traditional analysis in (14b), the traditional analysis is again more consistent with predictions based upon the component unit.

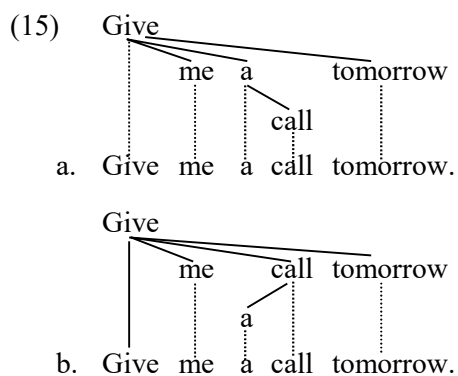
The results for the other five sentences con-

taining a preposition were similar. While there were a few anomalies, the informants by and large chunked the sentences in ways that support the existence of prepositional phrases. Note that an important caveat concerning the data in (13) is mentioned in the conclusion below.

#### 4.5 Determiners

The status of determiners has been controversial since the term *determiner phrase* (DP) first became established in the mid 1980s (e.g. Abney 1987). While the dominant view among DGs was and still is that determiners are dependents of their nouns, there have been exceptions. Most notably, Richard Hudson has argued in a number of works (e.g. 1984: 90–2, 1990: 268–276), that determiners are heads over their nouns. The component unit and chunking tasks can be brought to bear on this issue. The results we have obtained support the traditional NP analysis of nominal groups over the DP analysis.

Of the sentences we tested, eight of them contained a determiner, e.g. *Give me a call tomorrow*. Concerning the nominal group *a call* in this example, the two competing views about the hierarchical nature of nominal groups are present in the following analyses of the sentence:



The DP analysis of *a call* shown in (15a) predicts that some informants would choose to chunk between *a* and *call*, since the structure in (15a) shows *Give me a* as a component. The NP analysis of *a call* shown in (15b), in contrast, predicts that informants will not chunk between *a* and *call*, because on that account, *Give me a* and *me a* would not be components.

The informant responses in this area were mostly consistent. With only 13 exceptions (among hundreds of responses), the informants chunked the eight sentences containing deter-

miners in such a manner that the determiner was grouped together with the following noun. For instance, sentence (15) was chunked as follows:

- (16) a. Give me | a call | tomorrow. – 44  
 b. Give | me | a call tomorrow. – 2

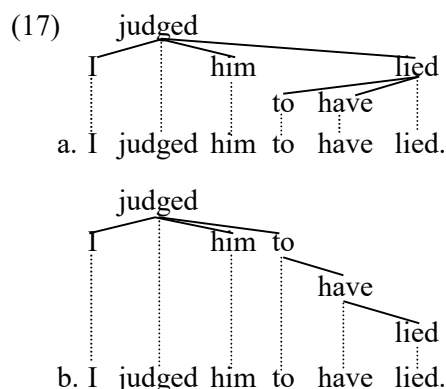
Not one of the informants who chunked this sentence chose to chunk between *a* and *call*. The three chunks shown in (16a) are components. The latter chunk in (16b), i.e. *a call tomorrow*, is the exception, since it is not a component in (15b) (and 15a).

The conclusion concerning determiners is therefore that informants prefer to group determiners together with the nouns that follow them. This fact supports the traditional NP analysis of nominal groups over the DP analysis.

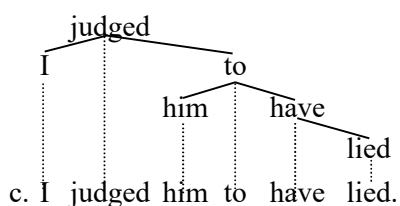
#### 4.6 Object predicatives (“small clauses”)

The hierarchical status of object predicative expressions, e.g. *I judged him to have lied*, has been a source of much debate among syntacticians. A ternary-branching analysis has been in competition with a strictly binary branching analysis. From the DG point of view, there are two conceivable analyses of these predicatives. The component unit and chunking task can be brought to bear on this issue. They reveal that the ternary-branching analysis should be preferred.

We tested four sentences that contained object predicatives: *I judged him to have lied*, *My parents expect me to become a doctor*, *We believe Sam to be upset*, and *They want you to go home*. Three possible structural analyses of the first of these four sentences are given next:







The UD approach is likely to pursue (17a). A “small clause” analysis would be along the lines of (17c). Analysis (17b) can be viewed as the traditional analysis of object predicatives in DG (see Kunze, 1975: 111–2; Schubert, 1987: 94–6; and Heringer, 1996: 76–7).

The informant responses we received for this sentence are listed next:

- (18) a. I | judged him | to have lied. – 25  
 b. I judged him | to | have lied. – 4  
 c. I | judged him to | have lied. – 4  
 d. I judged | him | to have lied. – 3  
 e. I judged | him to | have lied. – 3  
 f. I | judged | him to have lied. – 3  
 g. I judged | him to have | lied. – 1

The fact that a majority of informants preferred to chunk this sentence immediately after the object *him* provides guidance about the structure. The small clause analysis (17c) can be immediately rejected because it does not grant the chunk *judged him* component status. Choosing between (17a) and (17b) is more difficult based on the informant responses. The informants that chunked the sentence as in (18c) did, however, provide some guidance insofar as the chunk *judged him to* is a component in (17b) but not in (17a).

Other considerations allow the approach to more confidently choose between (17a) and (17b). The discussion of auxiliary verbs above in Sections 4.1, 4.2, and 4.3 established that auxiliary verbs are plausibly viewed as heads over the content verbs with which they co-occur. This fact hence points to (17b) as the best analysis, since (17b) does, but (17a) does not, position the auxiliary verb *have* as head over *lied*.

The informant responses we received for the other four sentences that contained an object predicative further support the conclusion. The traditional, ternary-branching analysis of object predicatives, as in (17b), is well motivated based upon the reasoning from chunking and the component unit.

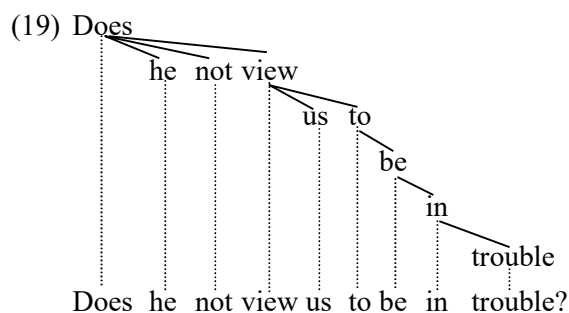
## 5 Overall relevance of component unit

Given the aspects of sentence structure established in the previous sections, it has become possible to put everything together in order to arrive at a motivated analysis of the overall structures discussed. We calculated the overall component to non-component ratio given traditional structures as opposed to structures corresponding to the UD annotation scheme. These overall numbers provide a cumulative argument in favor of the traditional structures.

The following points were established above:

1. Auxiliary verbs are heads over content verbs.
2. The nature of subject-auxiliary inversion further supports the stance that auxiliary verbs are heads over content verbs.
3. The sentence negation *not* is typically a postdependent of the auxiliary verb that precedes it.
4. Prepositional phrases exist, that is, the preposition is the head of the phrase it introduces.
5. The traditional NP analysis of nominal groups is preferable over the DP analysis.
6. Object predicatives are best analyzed with a ternary-branching structure that positions the object as an immediate dependent of the matrix verb.

To illustrate all of these points in one structure, we offer the following example:



If one produces hierarchical analyses of this sort for all 25 of the original sentences we tested on informants, one can then check to see how many of the chunks (consisting of two or more words) produced by informants were and were

not components.<sup>5</sup> By doing so, one arrives at an overall number that can be used to summarize the validity of varying assumptions about syntactic structure (e.g. traditional analysis vs. UD).

We obtained a total of 2,252 chunks from informants that consisted of two or more words. Among these 2,252 chunks, 2,124 of them were identified as components on traditional assumptions about sentence structure (as illustrated with 19). Thus, the ratio of component strings to total strings (abbreviated as  $R_C$ ) reached 94%. In contrast, a smaller number of these chunks, i.e. 1,632 of them, were components on competing UD assumptions about sentence structure, rendering  $R_C$  at 72%.

The results just mentioned are given visually with the following pie charts:

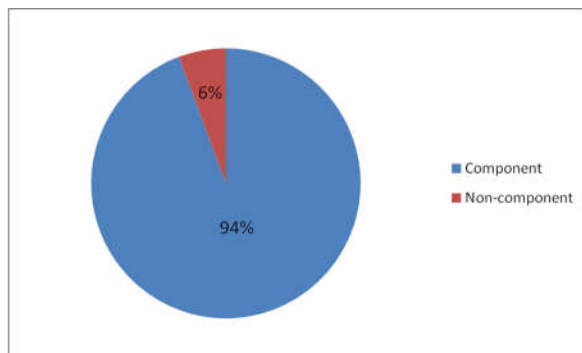


Figure 1. Component to non-component chunks on traditional analysis;  $R_C = 94\%$

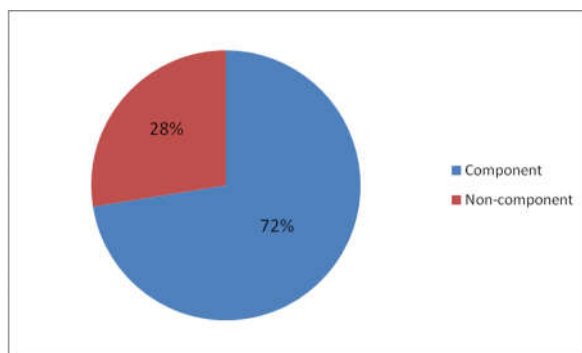


Figure 2. Component to non-component chunks on UD annotation scheme;  $R_C = 72\%$

The higher number of component chunks on the traditional analysis supports traditional DG

<sup>5</sup> The qualification “consisting of two or more words” is important. Individual words are always components by definition (regardless of the hierarchical analysis assumed). They were therefore excluded from the calculations in order to more strongly draw out the contrasts in the numbers across the two competing hierarchical analyses.

assumptions about the structure of sentences in English over the UD annotation scheme.

## 6 Concluding comments

This manuscript has employed the component unit, which is a novel unit of dependency syntax, to shed light on aspects of sentence structure in English. It should be evident that the simple methodology and reasoning employed can be easily extended to other languages. As long as one has access to a significant number of speakers of the language under investigation, the relevant data can be easily collected and analyzed to resolve issues about the hierarchical structures of that language. Indeed, the methodology and reasoning we have employed in this study are currently being extended to Chinese to resolve issues about the hierarchical analysis of Chinese sentences.

The principle objection that can be raised against the message delivered in this contribution concerns the informants – see note 4. Native speakers of Chinese may chunk English sentences differently than native speakers of English. In an effort to address this objection, a follow-up round of data collection was conducted on 13 native speakers of English. Most of the sentences presented and discussed above were tested. The results obtained matched those of the Chinese informants, with one exception.

The exception is present in the following numbers:

- |                             |     |
|-----------------------------|-----|
| f. We   waited   for Susan. | – 9 |
| We   waited for   Susan.    | – 3 |
| We waited   for   Susan.    | – 1 |

Comparing these numbers with the numbers for sentence (13) above, there is an obvious difference. These numbers from native informants are more congruent with the UD analysis, which subordinates the preposition *for* to the proper noun *Susan*.

Caution is therefore warranted concerning the greater conclusion. Solid claims about syntactic structure of English sentences will become possible only after the project has been extended to include data from large numbers of native informants.

## Acknowledgement

The research presented in this article was funded by the Ministry of Education of the People's Republic of China, Grant # 15YJA74001.

## References

- Paul Steven Abney. 1987. The English noun phrase in its sentential aspect. Doctoral dissertation, MIT, Cambridge, MA.
- John Anderson. 2011. *The substance of language volume I: The domain of syntax*. Oxford University Press.
- Norbert Bröker. 2003. Formal foundations of dependency grammar. In Vilmos Ágel et al. (eds.), *Dependency and valency: An international handbook of contemporary research*, vol. 1, 294–310. Walter de Gruyter, Berlin.
- Ulrich Engel. 1994. *Syntax der deutschen Gegenwartssprache*, 3<sup>rd</sup> ed. Erich Schmidt Verlag, Berlin.
- Hans-Werner Eroms. 2000. *Syntax der deutschen Sprache*. Walter de Gruyter, Berlin.
- Chaim Gaifman. 1965. Dependency systems and phrase-structure systems. *Information and Control* 8, 304–337.
- Thomas Groß and Timothy Osborne. 2009. Toward a practical dependency grammar theory of discontinuities. *SKY Journal of Linguistics* 22, 43–90.
- David Hays. 1964. Dependency theory: A formalism and some observations. *Language* 40, 511–525.
- Peter Hellwig. 2003. Dependency Unification Grammar. *Dependency and valency: An international handbook of contemporary research*, ed. by Vilmos Ágel et al., 593–635. Walter de Gruyter, Berlin.
- Hans Jürgen Heringer. 1996. *Deutsche Syntax: Dependentiell*. Stauffenburg, Tübingen.
- Richard Hudson. 1984. *Word Grammar*. Basil Blackwell, Oxford.
- Richard Hudson. 1990. *An English Word Grammar*. Basil Blackwell, Oxford.
- Richard Hudson. 2007. *Language networks: The new Word Grammar*. Oxford University Press.
- Sylvain Kahane. 1996. If HPSG were a dependency grammar... *Actes de TALN*, Marseille, 22–24, 45–49.
- Franz Kern. 1883. *Zur Methodik des deutschen Unterrichts*. Nicolaische Verlags-Buchhandlung, Berlin.
- Franz Kern. 1884. *Grundriss der Deutschen Satzlehre*. Nicolaische Verlagsbuchhandlung, Berlin.
- Jürgen Kunze. 1975. *Abhängigkeitsgrammatik*. Series: *studia grammatica XII*. Akademie-Verlag, Berlin.
- Marie-Catherine de Marneffe, Timothy Dozat, Natalia Silvaire, Katrin Haverinen, Filip Ginter, Joakim Nivre, Christopher D. Manning. 2014. Universal Stanford Dependencies: A cross-linguistic typology. *LREC* 14.
- Igor Mel'čuk. 1988. *Dependency syntax: Theory and practice*. Albany: State University of New York Press.
- Igor Mel'čuk and Nikolai Pertsov. 1987. *Surface syntax of English: A formal model with the Meaning-Text Framework*. Benjamins, Amsterdam.
- William O'Grady 1998. The syntax of idioms. *Natural Language and Linguistic Theory* 16, 79–312
- Timothy Osborne and Thomas Groß. 2016. The do-so-diagnostic: Against finite VPs and for flat non-finite VPs. *Folia Linguistica* 50, 1, 97-35
- Timothy Osborne, Michael Putnam, and Thomas Groß. 2012. Catenae: Introducing a novel unit of syntactic analysis. *Syntax* 15, 354-396.
- Jane Robinson. 1970. Dependency structures and transformational rules. *Language* 46, 259–285.
- Klaus Schubert. 1987. *Metataxis: Contrastive dependency syntax for machine translation*. Foris Publications, Dordrecht.
- Stanley Starosta. 1988. *The case for Lexicase: An outline of Lexicase grammatical theory*. Pinter Publishers, London.