On Operationalizing Syntactic Complexity

Benedikt M. Szmrecsányi

University of Freiburg – Germany

benedikt.szmrecsanyi@uni-freiburg.de

Abstract

In the recent functional linguistic literature, the notion of syntactic complexity and similar concepts have received considerable attention. Yet, when trying to operationalize syntactic complexity as an independent variable in statistical research designs, it soon emerges that the notion is somewhat underdefined. I will report the results from an experiment comparing three measures of syntactic complexity — node counts, word counts, and a so-called 'Index of Syntactic Complexity' — with regard to their accuracy and applicability. While presupposing that node counts are cognitively the most 'real' measure, it turns out that the other two measures are near-perfect proxies of the former. I conclude that since node counts are in most cases unreasonably resource demanding to conduct, researchers can feel safe in using the measure that is most economically to conduct, word counts.

Keywords: stylometry, textual classification, text corpora and text encoding.

1. Introduction

What is meant by syntactic complexity (e.g. Ferreira, 1991; Givón, 1991), cognitive complexity (e.g. Mondorf, 2003; Givón, 1991; Rohdenburg, 1996), clause complexity (e.g. Kubon, 2001), linguistic complexity (e.g. Schleppegrell, 1992), structural complexity (e.g. Givón, 1991; Arnold *et al.*, 2000), or grammatical / syntactic weight (e.g. Wasow, 1997; Wasow and Arnold, 2003)[1] has received a good deal of scholarly attention, particularly — but not exclusively — by linguists working within functional frameworks. Yet, somewhat surprisingly, there is a dearth of precise definitions and convincing approaches to operationalize these concepts in a straightforward, objective, and non-intuitional way in empirical research designs. More often than not, the issue is avoided or side-stepped in that syntactically, cognitively, and/or structurally complex environments are sloppily 'defined' as contexts "that are for some reason more difficult, more complex, less entrenched, less frequent, less accessible or in any way cognitively more complex" (Mondorf, 2002: 252). Obviously, such definitions are not helpful when the task is to answer a variationist research question such as "How does the syntactic complexity of a slot's linguistic environment influence a speaker's decision to employ a pattern *x* rather than some other equivalent pattern *y*?". In this paper, therefore, I will report the results from an experiment comparing three measures of syntactic complexity — node counts, word counts, and a so-called 'Index of Syntactic Complexity' — with regard to their accuracy and applicability in empirical research designs.

---

[1] In this paper, all these terms will be considered synonymous and subsumed under 'syntactic complexity' since they all basically refer to syntactic structures which necessitate increased parsing and processing effort. Also note that due to space limitations, I will not be able to consider notions such as 'width' and 'depth' in relation to subordination.

2. Syntactic complexity of what?

A first question that has to be addressed is, what is the nature of the units of linguistic data to be compared? In most cases, a unit A of some sort will have to be compared to a unit B of some sort. To this end, units A and B must be roughly comparable in scope (for instance, it will hardly be operational to compare the syntactic complexity of a ten word utterance to the syntactic complexity of, say, a short story). In principle, then, it is possible to define scope in two different ways: (i) scope is defined in terms of pure length, duration, or size of a unit; (ii) scope is defined by appealing to notions independent of pure length, duration, or size. To illustrate, in approach (i) one could compare units of linguistic data which have the same number of words (for instance, syntactic complexity of one six word utterance is compared to syntactic complexity of another six word utterance); or which took speakers the same time to produce (for instance, a 10 second utterance is compared to another 10 second utterance); or that have the same size (for instance, one 100 byte chunk of text is compared to another 100 byte chunk of text). The main problem with such an approach is that appealing to length, duration, or size alone often makes impossible what is at stake here — using a structural measure of syntactic relationships in a unit of linguistic data. For illustration, consider (1) and (2):

(1)   I wasn't there cause I had to fill out all this.

(2)   I didn't do it, and the reason for this was that.

Both (1) and (2) have a length of 11 words, and are thus comparable in scope according to criterion (i). But while it is easy to see that (1) contains a main clause and a syntactically dependent adverbial clause of reason, (2) is a compound clause, as far as we can say with any reasonable degree of confidence. But the last word, *that*, might (e.g., *...the reason for this was that I was sick*) or might not (e.g., *the reason for that was that guy*) introduce a syntactically dependent complement clause. In other words, the measure of scope outlined here interferes with one's ability to assess syntactic complexity.

If, according to (ii), scope is defined by appealing to notions independent of pure length, one could compare units of linguistic data that have the same number of clauses, the same number of sentences, the same number of paragraphs, the same number of phrases, etc. To illustrate, if one were were to adopt the notion of 'sentence' as the basic unit of measure, one could compare (3) to (4), because both (3) and (4) constitute exactly one sentence:

(3)   I wasn't there cause I had to fill out all this.

(4)   I didn't do it, and the reason for that was that I was sick.

The major advantage of this approach, of course, is that notions such as 'clause', 'sentence', or 'paragraph' already have a requirement of meaningful, complete syntactic relationships built into them.

3. Previous approaches to assess syntactic complexity

Previous research has used a couple approaches to assess syntactic complexity (see, e.g., Wasow, 1997 on their advantages and disadvantages with regard to heavy noun phrase shift). They all basically take what was outlined above as approach (ii), i.e. they define the boundaries of what they measure by appealing to notions independent from length.

3.1. *Length as a proxy for syntactic complexity (number of words dominated)*

Arnold *et al.* (2000) and Hawkins (1990), for instance, measured syntactic weight of constituents as the difference in length (in words) between these constituents. Using length —
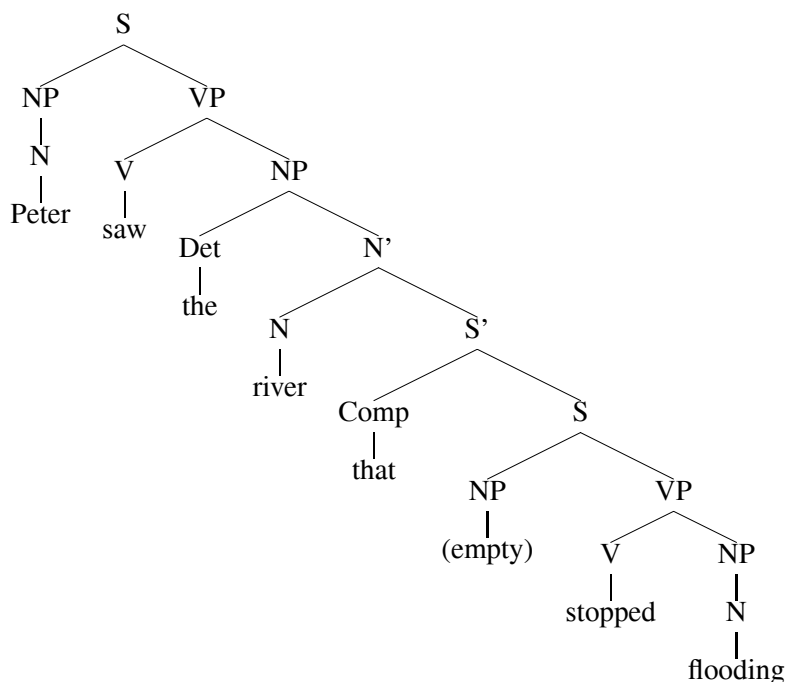
number of words, syllables, intonation units, etc. — as a proxy for syntactic complexity has the obvious advantage that this is a straightforward method which does not even necessarily involve manual coding. It is also probably the most time-honored proxy for complexity. Behagel (1909), in his seminal essay on what is today called 'the principle of end weight', already implicitly referred to it as a measure of weight. A drawback is that increased length does not necessarily always translate into increased syntactic complexity. For instance, concerning end weight, Chomsky (1975: 477) has remarked that "it is interesting to note that it is apparently not the length in words of the object that determines the naturalness of the transformation (i.e., heavy constituents shifted to final position, BS), but, rather, in some sense, its complexity."

If length is taken to mean the number of words dominated by a syntactic unit, then, this unit's syntactic complexity will be directly proportional to the number of words it contains. In what follows, the measure will be conceptualized by taking sentences as the basic unit and then determining sentence length (henceforth: SL). Thus, a sentence containing, for instance, 6 words will receive a syntactic complexity score of 6, while a sentence containing 9 words will receive a syntactic complexity score of 9.

3.2. *Counting nodes (number of nodes dominated)*

Johnson (1966), Ferreira (1991) and Rickford *et al.* (1995) counted nodes to determine syntactic complexity of phrases and sentences; Hawkins (1994) also suggested this method. In this approach, the more phrasal nodes a unit dominates, the more complex it is. Presupposing some notion of formal complexity, counting the number of nodes dominated is conceptually certainly the most direct and intuitively the most appropriate way to assess syntactic complexity. This is because the method reflects how the human parser is supposed to work. Thus, it is the structural measure which is psychologically most real. Note that syntactic structure — and complexity thereof — is not only a formal phenomenon, but also a a psychological one in that speakers must construct and produce material that conforms to the syntactic rules of their language. This is obviously more resource demanding for syntactically complex sentences than for syntactically simple sentences (for a discussion from a psycholinguistic perspective, see, for instance, Bock, 1982; Bock and Kroch, 1989; Ferreira, 1991; for a linguistic opinion, see Hawkins, 1990).

From a practical viewpoint, though, counting nodes almost always involves manual coding. This is the reason why the method is more popular in experimental research designs where only a very limited number of contexts has to be coded, in contrast to corpus-based designs where the number of contexts to code is potentially open-ended. If the method is adopted anyway, the idea is to measure the number of nodes in a unit of data using a parse tree.

```
                          S
                    ╱          ╲
                  NP            VP
                  │          ╱     ╲
                  N         V        NP
                  │         │     ╱    ╲
                Peter      saw  Det      N'
                            │    │    ╱    ╲
                           the  N      S'
                                │   ╱     ╲
                              river Comp    S
                                    │    ╱   ╲
                                   that  NP    VP
                                         │   ╱   ╲
                                      (empty) V    NP
                                              │     │
                                           stopped  N
                                                    │
                                                 flooding
```

To illustrate, the above tree diagram represents the sentence *Peter saw the river that stopped flooding*, which dominates 16 nodes (excluding the clause node S and the lexical items themselves) and therefore receives a syntactic complexity score of 16.

3.3. *An index of syntactic complexity (number of high complexity indicators dominated)*

A final approach to assessing syntactic complexity is conceivable. It would involve establishing what I would like to call an 'Index of Syntactic Complexity' (henceforth: ISC). For Beaman (1984: 45), "syntactic complexity in language is related to the number, type, and depth of embedding in a text. Syntactically simple authors [...] rely more heavily on coordinated structures [...] Syntactically complex authors [...] use longer sentences and more subordinate clauses." In her study, Beaman determines syntactic complexity of spoken and written discourse by simply comparing the percentage of subordinate clause structures in both discourse types. In a similar vein, Givón (1991) appears to consider embedded and/or subordinate structures — as opposed to conjoined structures — indicative of marked, and hence cognitively more complex, categories. Givón (1991: 347) cites a sizable body of psycholinguistic studies which demonstrate that subordinate clause structures are more complex to process than conjoined main clause structures.

Following Givón's and Beaman's emphasis on embeddedness, then, syntactic complexity of a given context could be established by counting linguistic tokens that can be considered telltale signs of increased grammatical subordinateness and embeddedness, such as (i) subordinating conjunctions (henceforth: SUB) (for instance, *because, since, as, when, that*, etc.), and (ii) WH-pronouns (henceforth: WH) (*who, whose, whom, which*). In addition, tokens to be included in the index should also include (iii) verb forms (henceforth: VF), both finite and non-finite, and (iv) noun phrases (henceforth: NP).

Because subordinators and WH-pronouns are the most straightforward indicators of increased embeddedness — and thus of high complexity —, these features should be weighted more heavily than verb forms and noun phrases. I would, then, like to suggest the following formula — which, admittedly, is somewhat tentative and ad-hoc — to establish (ISC):

$$\text{ISC}(u) = 2 \times n(u, SUB) + 2 \times n(u, WH) + n(u, VF) + n(u, NP)$$

Let u be the the unit of linguistic data under analysis, let $\text{ISC}(u)$ be the ISC of the unit of linguistic data under analysis, and let n(u,SUB) be the number of occurrences of SUB in the unit of linguistic data under analysis, etc. According to this formula, ISC of a given unit of data is twice the number of occurrences of subordinating conjunctions and WH-pronouns plus the number of occurrences of verb forms and noun phrases in that unit. A structural measure such as ISC has the advantage that it is relatively easy to establish, especially when compared to counting nodes. If the data to be analyzed is part of speech (POS) tagged, it can be even done automatically by software. To illustrate ISC, consider (5):

> (5) The jury further said in term-end presentments that the City Executive Committee, which had over-all charge of the election, deserves the praise and thanks of the City of Atlanta for the manner in which the election was conducted.

This sentence would receive an ISC score of 21: it contains 1 subordinating conjunction (*that*), 2 WH-pronouns (*which, which*), 5 verb forms (*said, had, deserves, was, conducted*), and 10 NPs (*jury, term-end presentments, City Executive Committee, over-all charge, election, praise, thanks, City of Atlanta, manner, election*). Just for purposes of comparison, according to the SL measure, this sentence would receive a score of 38 (it has 38 words); according to the measure of how many nodes the sentence commands (henceforth: NODE), it would receive a score of 68 (it dominates 68 nodes, not including the clause node S and the lexical items themselves).

4. Comparing structural measures of syntactic complexity

4.1. *Method and data*

To systematically compare differences between the aforementioned approaches to measure syntactic complexity, each of the three measures was applied to the first 20 sentences of text A01 of the *Susanne Corpus* and the first 30 sentences of text T01 of the *Christine Corpus*, establishing syntactic complexity per sentence as analytical unit. The convenient advantage of the above corpora is that they are so-called treebanks, i.e. they are tagged for both POS and syntax. The Susanne corpus comprises a subset of the *Brown Corpus of American English* and is a corpus of written English. The Christine Corpus comprises a subset of the spoken-demographic section of the *British National Corpus* and is a corpus of spoken English (see http://www.cogs.susx.ac.uk/users/geoffs/SueDoc.html and http://www.cogs.susx.ac.uk/users/geoffs/ChrisDoc.html, respectively, for more information on these corpora). Both corpora are syntactically annotated using a phrase structure approach to English grammar.

4.2. *Results*

Numerical structural complexity scores returned by different structural measures are not cardinally comparable. They are comparable in an ordinal fashion only, i.e. how they rank different sentences with regard to syntactic complexity. This is what is indicated in brackets in table 1. When a sentence occupies rank 1, this means that there is no sentence that is syntactically more complex in the dataset, while the lowest rank indicates that there is no sentence which is syntactically less complex in the dataset. Note that because two or more sentences may receive identical syntactic complexity scores, more than one sentence may occupy a given rank. This is why the absolute number of ranks varies across different structural measures. Figure 1, then, visualizes the results of the three structural measures by indicating relative ranks to enhance

	Christine corpus			Susanne corpus		
sentence #	SL	NODE	ISC	SL	NODE	ISC
1	7 (8)	17 (7)	3 (4)	21 (10)	39 (14)	12 (7)
2	6 (9)	11 (11)	1 (6)	38 (3)	68 (3)	21 (1)
3	6 (9)	11 (11)	1 (6)	33 (4)	56 (5)	14 (5)
4	17 (2)	34 (2)	5 (2)	29 (6)	49 (7)	13 (6)
5	2 (12)	2 (17)	0 (7)	21 (10)	41 (12)	9 (10)
6	3 (11)	7 (14)	1 (6)	21 (10)	44 (9)	11 (8)
7	2 (12)	2 (17)	0 (7)	39 (2)	76 (2)	16 (3)
8	5 (10)	11 (11)	1 (6)	22 (9)	40 (13)	11 (8)
9	7 (8)	12 (10)	2 (5)	20 (12)	36 (16)	7 (12)
10	16 (3)	33 (3)	9 (1)	11 (15)	22 (19)	8 (11)
11	14 (4)	27 (5)	4 (3)	13 814)	26 (18)	4 (13)
12	2 (12)	3 (16)	0 (7)	25 (7)	42 (11)	14 (5)
13	19 (1)	35 (1)	9 (1)	23 (8)	43 (10)	8 (11)
14	9 (7)	15 (8)	4 (3)	52 (1)	90 (1)	18 (2)
15	6 (9)	12 (10)	2 (5)	20 (11)	36 (16)	9 (10)
16	6 (9)	9 (13)	1 (6)	20 (11)	37 (15)	11 (8)
17	6 (9)	12 (10)	2 (5)	14 (13)	28 (17)	7 (12)
18	7 (8)	14 (9)	3 (4)	32 (5)	61 (4)	15 (4)
19	1 (13)	1 (18)	0 (7)	29 (6)	48 (8)	10 (9)
20	1 (13)	1 (18)	0 (7)	29 (6)	54 (6)	11 (8)
21	6 (9)	10 (12)	3 (4)			
22	1 (13)	1 (18)	0 (7)			
23	3 (11)	7 (14)	2 (5)			
24	2 (12)	5 (15)	1 (6)			
25	11 (5)	21 (6)	3 (4)			
26	10 (6)	17 (7)	2 (5)			
27	14 (4)	29 (4)	3 (4)			
28	2 (12)	5 (15)	1 (6)			
29	6 (9)	11 (11)	2 (5)			
30	5 (10)	9 (13)	1 (6)			

Table 1. Syntactic Complexity scores (ranks in brackets) according to different structural measures in the Susanne Corpus (written English) and the Christine Corpus (spoken English)

comparability. This means that ranks returned by the NODE measure were taken as baseline and the ranks returned by the other two measures were then plotted with regard to their relative relation to the NODE measure ranking. Sentences on the X axis were ordered according to the baseline ranking suggested by the NODE measure.

A cursory glance at these tables and especially the graphs indicates that the three structural measures rank the data in an indeed surprisingly conform fashion. Assuming — as it was done here — that NODE is conceptually and psychologically the 'most real' structural measure, it can be seen that neither SL nor ISC produce substantially significant outliers in either dataset. This visual impression is confirmed when Spearman's correlation coefficient (*Spearman's ρ*) is computed. This statistical procedure provides a means to assess the linear relationship between two variables. The resulting coefficient can range between -1 (which would indicate a perfect negative linear relationship) and +1 (which would indicate a perfect positive linear relationship). Actual correlation coefficients can be seen from table 2.

Overall, SL and ISC are better proxies for NODE in spoken data than in written data. Also, SL slightly outperforms ISC in both datasets. It should be pointed out though that with Spearman's
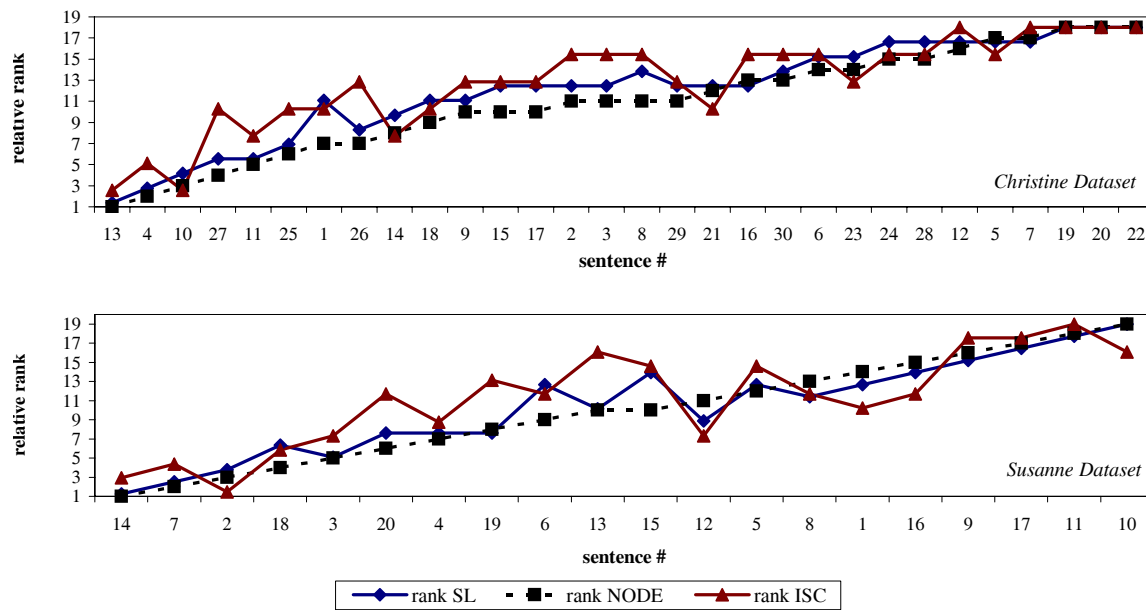
Figure 1. Relative Syntactic Complexity ranks according to different structural measures in the Susanne corpus (written English) and Christine Corpus (spoken English)

	SL vs. NODE	ISC vs. NODE
Susanne corpus (written English)	.976**	.836**
Christine corpus (spoken English)	.989**	.916**

** significant at the .01 level

Table 2. Correlation coefficients for rankings suggested by different structural measures

ρ in the .8/.9 range, both measures are highly correlated to NODE. Both SL and ISC, therefore, are near-perfect proxies of NODE.

5. Summary and conclusion

In conclusion, I have suggested that measuring sentence length (SL) and, albeit to a slightly lesser degree, computing an index of syntactic complexity (ISC) do an excellent job in approximating a node count (NODE), the structural measure of syntactic complexity which is probably the most 'real' one cognitively. In essence, thus, the three structural measures really gauge the same thing. Interestingly, Wasow (1997) has found to be much the same true for measuring weight effects ("It is very hard to distinguish among various structural weight measures as predictors of weight effects. Counting words, nodes, or phrasal nodes all work well", Wasow, 1997: 102). Importantly, though, the three structural measures presented here differ markedly in what data they require and in the expenditure of manual coding they necessitate. SL just requires counting words, which even any standard word processor is able to perform without human help. ISC requires manual coding, unless the data source is POS tagged. NODE also requires manual coding, unless the data source is syntax tagged. More often than not, the researcher is confronted with data that are not tagged at all. What I hope to have provided evidence for in this paper is that the most economic method — determining length in words — to assess

syntactic complexity is by all means one that is nearly as accurate as the more sophisticated and cognitively, conceptually, or even psychologically 'more real' methods.

References

Arnold J., Wasow T., Losongco A. and Ginstrom R. (2000). Heaviness vs. Newness: The Effects of Structural Complexity and Discourse Status on Constituent Ordering. *Language*, vol. (17/1): 28-55.

Beaman K. (1984). Coordination and Subordination Revisited: Syntactic Complexity in Spoken and Written Narrative Discourse. In Tannen D. and Freedle R. (Eds), *Coherence in Spoken and Written Discourse*: 45-80.

Behagel O. (1909). Beziehungen zwischen Umfang und Reihenfolge von Satzgliedern. *Indogermanische Forschungen*, vol. (25/110).

Bock K. (1982). Towards a cognitive psychology of syntax: Information processing contributions to sentence formulation. *Psychological Review*, vol. (89): 1-47.

Bock K. and Kroch A. (1989). The isolability of syntactic processing. In Carlson G. and Tanenhaus M. (Eds), *Linguistic structure in language processing*.

Chomsky N. (1975). *The logical structure of linguistic theory*. Chicago University Press.

Ferreira F. (1991). Effects of Length and Syntactic Complexity on Initiation Times for Prepared Utterances. *Journal of Memory and Language*, vol. (30/2): 2110-2233.

Givón T. (1991). Markedness in grammar: distributional, communicative and cognitive correlates of syntactic structure. *Studies in Language*, vol. (15/2): 335-370.

Hawkins J. (1990). A Parsing Theory of Word Order Universals. *Linguistic Inquiry*, vol. (21/2): 223-261.

Hawkins J. (1994). *A Performance Theory of Order and Constituency*. Cambridge University Press.

Johnson N. (1966). On the relationship between sentence structure and the latency in generating the sentence. *Journal of Verbal Learning and Verbal Behavior*, vol. (5): 375-380.

Kubon V. (2001). A Method for Analyzing Clause Complexity. *Prague Bulletin of Mathematical Linguistics*, vol. (75): 5-28

Mondorf B. (2003). Support for More-Support. In Rohdenburg G. and Mondorf B. (Eds), *Determinants of Grammatical Variation in English*: 251-304.

Rickford J., Denton M., Wasow T. and Espinoza J. (1995). Syntactic Variation and Change in Progress: Loss of the Verbal Coda in Topic-Restricting *As Far As* Constructions. *Language*, vol. (71/1): 102-131.

Rohdenburg G. (1996). Cognitive complexity and increased grammatical explicitness in English. *Cognitive Linguistics*, vol. (7): 149-182.

Schleppegrell M. (1992). Subordination and Linguistic Complexity. *Discourse Processes: A Multidisciplinary Journal*, vol. (15/1): 117-131.

Wasow T. (1997). Remarks on grammatical weight. *Language Variation and Change*, vol. (9): 81-105.

Wasow T. and Arnold J. (2003). Post-verbal constituent ordering in English. In Rohdenburg G. and Mondorf B. (Eds), *Determinants of Grammatical Variation in English*: 119-154.