1 Introduction

1.1 The syntactic complexity of noun phrases

Linguists generally agree that noun phrases (NPs) can be more or less complex. From a syntactic point of view, there is thus little doubt that *the man I saw* in (2) is more complex than the non-postmodified NP *the man* in (1).

(1) the man
(2) the man I saw

What exactly makes NPs more or less complex, however, is much more of an unresolved issue. Different options suggest themselves. Is it their length (measured in terms of words, graphemes, morphemes, or syllables)? Is it the number of phrases these NPs contain? Or is it the fact that some NPs are sentential (like *the man I saw*) while others are not? Alternatively, syntactic complexity may be defined as a product of all these three factors combined rather than as one of them in isolation.

It is the aim of this book to compare different measures of the syntactic complexity of NPs and to explore how strong each of them is when isolated from the rest. This comparison will be conducted on the basis of a linear and a hierarchical parameter of the syntactic complexity of NPs. The linear one is their length and the hierarchical one is the type of postmodifier following the head of the NP (like *I saw* following the head noun *man*). While the term 'postmodifier' traditionally implies the functional distinction between modifiers and complements, it is here used as a purely structural category, denoting what Huddleston, Pullum et al. (2002: 329) call 'post-head dependents', namely all structural elements following the head of the NP. Postmodifiers may thus be phrasal, as in *the man from next door/the friends of John*, or clausal (sentential), as in *the man I saw* in Example (2). Alternatively, there may be no postmodifier at all, as in *the man* in (1).

Comparisons of the syntactic parameters NP-length and NP-structure are very rare in the literature, and I will review the few studies that have made the effort to tease apart the individual factors and test each of them for their independence (see Section 1.3). This is not surprising given the fact

that different quantitative measures of syntactic complexity (like word counts and phrasal nodes) are very highly correlated (see Section 1.2) and that the outcome of such a comparison strongly depends on the syntactic framework applied (see Section 3.2.1). In this study, I will classify all NPs according to their type of postmodifier, thereby avoiding any theoretical commitment to specific theories of syntax. While I will suggest nine hypotheses stating which type of NPs are more and less complex (see Section 3.3), the quantitative analyses (see Chapters 5–9) will either have to confirm or reject the complexity scale suggested.

The most important theoretical contribution of my study is that it brings in a qualitative dimension to the discussion of the syntactic complexity of NPs which has, so far, been rather neglected. Since the presence or absence of a verb phrase (VP) is not sufficiently captured by the number of (phrasal) nodes in the NP (see Section 3.2.1), I will ask what the relevance of a VP is for the definition of an NP's syntactic complexity. On the basis of detailed analyses of the different types of structural categories I apply (see Section 3.2.2), I will set up the following claim, which will subsequently be tested against empirical data: the syntactic complexity of NPs cannot sufficiently be described via quantitative measures (length and phrasal node counts); rather, it needs an additional qualitative dimension which is the presence or absence of a VP.

Overall, my study will now compare three different parameters of the syntactic complexity of NPs which are (a) the length of the NPs, (b) their structural complexity measured in terms of phrasal nodes, and (c) their quality of being either sentential or non-sentential. In order to test the strength of each of my three parameters, I will conduct corpus-based analyses on four different syntactic variables, all of which occur with more and less complex NPs. This methodological approach has only few parallels in the literature. The studies that come closest to my approach are Grafmiller and Shih (2011), Wasow and Arnold (2005; 2003), and Wasow (2002). Yet, no one has so far devoted a book-length treatment to the question of how best to define the syntactic complexity of NPs from a usage-based perspective.[1]

The variables under investigation in this study are the topic-restricting *as far as* construction of the type illustrated in (3), two cases of word-order variation exemplified in (4) and (5), and the optional occurrence of the infinitive marker *to* in the context of *help*, which is shown in (6). A more detailed description of these variables will be provided in Section 2.1.

[1] Keizer (2007) has explored a variety of NP-structures containing two nominal elements in terms of their form, meaning, and use. Her study has not, however, measured the complexity of these structures. Wasow (2002) has studied how the grammatical weight of postverbal constituents determines their word order in a sentence. His comparison of length and structure is, however, limited to a single chapter. My book will point out what he has to say on factor isolation (of length and structure).

(3) **As far as** the weather **is concerned/goes/Ø**, they say it's going to rain.
(4) They **have taken** these men **prisoner**./They **have taken prisoner** these men.
(5) **Notwithstanding** the bad weather, we're going for a walk./ The bad weather **notwithstanding**, we're going for a walk.
(6) He **helps** those people **(to)** get a job.[2]

What all these four variables have in common is that the constructions occur with a dependent NP and that these NPs (e.g. *the weather*, *these men*, *the bad weather*, *those people*) may vary in terms of their syntactic complexity. While the NPs selected in Examples (3)–(6) are relatively short and structurally simple, they may well be extended to long and structurally complex NPs which include clausal elements (e.g. *those people who live next door*). As my book will show, all of the four variables are sensitive to the effects of NP-complexity in the sense that complex NPs show a preference for one type of variant and simple NPs for the other. What is different for each variable, however, is the point at which each variable shifts from variant A to variant B.

While it is perfectly clear that variation is never governed by one factor alone (here, NP-complexity), my methodological approach allows me to concentrate on the effects of NP-complexity on variation and, on the basis of qualitative and quantitative corpus findings, to say whether one parameter (e.g. length) simply is the epiphenomenon of another (e.g. structure in terms of phrasal node counts and the quality of being sentential).[3] The assumption is that if one parameter has a strong effect on the distribution of the variants in a given case of variation, this type of syntactic complexity will have to figure prominently in a definition of the NP's syntactic complexity. I will also explore what the strength of each parameter depends on. As we will see, an answer to this question can only come from a comparison of all four variables (see Chapter 10).

In the remainder of the introduction, I would like to introduce the three parameters that I will employ to measure the syntactic complexity of the NPs investigated. By means of illustration, I will now return to Examples (1) and (2), which are here repeated for the sake of convenience.

(1) the man
(2) the man I saw

One way in which the two phrases in (1) and (2) differ is in terms of their length. While *the man* consists of only two words, *the man I saw* has a total of

[2] Bold print in these and all the ensuing examples of this book is my own.
[3] A similar research question has been phrased by Rosenbach (2005) with respect to the rivalry of animacy and weight effects: is one of these effects simply an artefact of the other?

Figure 1.1 Complexity scale in terms of length

Figure 1.2 Complexity scale in terms of phrasal nodes

four words. The length of NPs can alternatively be measured in terms of graphemes, morphemes, or syllables, and I will compare different measures of NP-length in Section 3.1 of this book. What we can conclude from the comparison of Examples (1) and (2) at this point is: counting the number of words in an NP is one way to account for its syntactic complexity. Figure 1.1 illustrates a complexity scale based on the number of words in an NP.

The length of the NPs is, however, not the only way to distinguish between more and less complex NPs. In addition to having more words, the NP in (2) also has a more complex internal structure than the NP in (1) in that it contains a subordinate relative clause (*I saw*). In the literature, we find various ways to account for the structural complexity of an NP. The approach that I wish to introduce is based on counting the number of phrasal nodes in an NP. This means I will count the number of phrases that constitute a superordinate phrase (e.g. an NP like *the man from next door* contains a subordinate PP *from next door*, which, again, contains the NP *next door*). For now, it is sufficient to realise that phrases have a hierarchical structure because some words in a phrase belong more closely together than others (so-called 'constituents'). In parallel to the complexity scale in terms of word counts, I can now set up a complexity scale in terms of phrasal node counts. This is illustrated in Figure 1.2. As in Figure 1.1, the higher the number of items (i.e. phrasal nodes), the more complex the NP.

The comparison of the Examples in (1) and (2) has revealed that the length and the structural complexity of an NP can be closely correlated, meaning that as one increases, the other also increases. In other words: long NPs (such as *the man I saw*) tend to be structurally complex and vice versa. There are, however, also NPs where the relation between the structure and the length of the NPs is less clear, and where it is difficult to decide how we can best account for the syntactic complexity of the phrases. Examples (7) and (8) illustrate this situation.

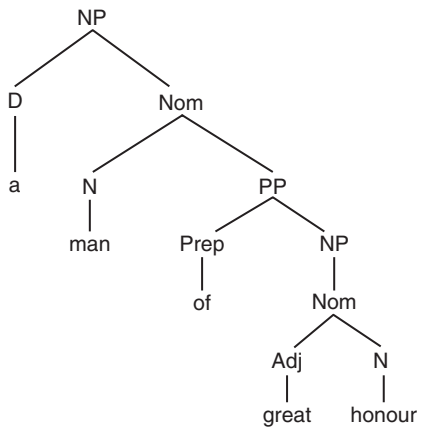

Figure 1.3 NP+PP

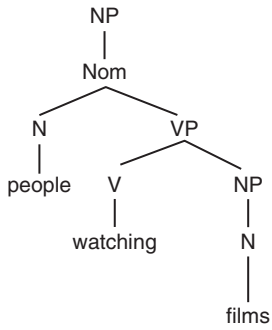

Figure 1.4 NP+non-finite clause

(7) a man of great honour
(8) people watching films

Which of these two NPs is more complex? Different answers suggest themselves: on the basis of word counts, the construction in (7), which contains a postmodifying prepositional phrase (henceforth: NP +PP), is more complex than the construction in (8), which involves a non-finite clause (henceforth: NP+non-finite clause). While we have five words in (7), (8) contains no more than three words. Looking at the number of phrasal nodes, we see that both NPs have exactly the same number, namely three. The number of phrasal nodes and their corresponding hierarchical structures are illustrated in Figures 1.3 and 1.4.[4] What the

[4] The syntactic trees follow the conventions of *The Cambridge Grammar of the English Language* (CGEL).

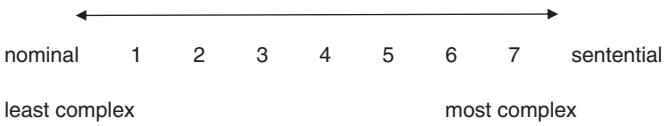

Figure 1.5 Complexity scale in terms of the quality of being sentential

comparison of word counts and phrasal node counts shows us is: different models of syntactic complexity provide different results.

Let us now look into the third dimension of the syntactic complexity of NPs applied in this book: the quality of being sentential. Being sentential means that the NP involves a clause (a VP or clause in a tree diagram) as part of its structure. On the basis of this criterion, we can distinguish between NPs which contain clauses, such as *the man I saw* in (2) or *people watching films* in (8) and those without, such as *the man* in (1) and *a man of great honour* in (7). If we further assume that sentential NPs are more complex than non-sentential ones because sentences are more complex than non-sentences (which are words or combinations of words), the structure NP+non-finite clause in (8) should be more complex than the non-sentential NP in (7). This conclusion supports none of the two lines of argumentation outlined above because it claims that phrasal nodes do not have the same weight: VPs weigh more than non-VPs.

A scale of complexity based on the degree to which an NP is being sentential is illustrated in Figure 1.5. While we may so far have assumed that an NP either is sentential or not (involving a clause or not), there are in fact degrees to which an NP can be sentential. This concerns, for instance, the difference between NPs with finite and non-finite clauses (e.g. in *the man I saw* in (2) and *people watching films* in (8)). Ross (2004 [1973]) has developed a series of test frames which help us to decide on the degree to which an NP is sentential. I will elaborate on these test frames in Section 1.4. Here, we only need to know that the more tests attesting to the quality of being sentential are passed, the further towards the right of the scale the NP should occur.

To summarise, we have seen that the syntactic complexity of NPs cannot be determined by one parameter alone. Rather, there are (at least) three different parameters involved in the distinction between more and less complex NPs:

1. the length of the NPs;
2. their structural composition measured in terms of (phrasal) node counts;
3. the degree to which NPs are sentential.

Apart from focussing on the syntactic complexity of NPs, this study will explore what effects selected other language-internal and external factors have on variation in the case of our four variables. Among them is the regional

contrast between British and American English (henceforth, BrE and AmE), the contrast between speech and writing, and the influence that the discourse status of NPs has on the distribution of the variants. The investigation of these factors is motivated by the assumption that variation is never governed by one factor alone. In order to adequately describe the effects of syntactic parameters on variation, we therefore need to isolate them from rivalling constraints. A diachronic perspective on variation eventually answers the question of what the role of NP-complexity is in language change.

It is the aim of this chapter to look more closely into the properties associated with each of the three parameters (1) the length of NPs, (2) their number of phrasal nodes, and (3) their degree of being sentential. In Section 1.2, I will start with an overview of the strong correlation between length and structure as two prominent measures of syntactic complexity. I will then ask what evidence there is (a) for the fact that structure (hierarchy) is independent of length and for (b) that length is independent of structure (Section 1.3). Section 1.4 will subsequently deal with Ross' scale of nouniness (2004 [1973]) and the question of what it means for NPs to be more and less sentential. Finally, Section 1.5 provides an outline of the book's structure.

1.2 Factor correlation: length and structure

In the literature, we find ample attestations of the strong correlation between the length and the structural weight of constituents in the sense that longer phrases are structurally more complex than shorter ones and vice versa. In this section, we will look at two types of evidence coming (a) from correlation coefficients between length and structure calculated for various studies on word-order variation (e.g. Wasow 1997; 2002) and (b) from Hawkins' theory of processing efficiency (e.g. 1994; 2004). I will start with Wasow's account.

Wasow (1997: 93; 2002: 32) calculates correlation coefficients for measures of length and structural complexity. Length, in his studies, is measured in terms of word counts and structure in terms of either nodes or phrasal nodes. The coefficients compare both words to node counts and words to phrasal node counts. They are based on three different types of variation, which are heavy NP-shift (HNPS), dative alternation (DA), and particle movement (PM). The alternation for HNPS is illustrated in (9a) and (9b), for the DA in (10a) and (10b), and for PM in (11a) and (11b).

(9a) They communicated **the next step** to us. (Wasow 2002: 58)
(9b) John took into account **only the people he knew**. (Wasow 2002: 33)
(10a) Gorbachev's second-in-command, Vice President Anatoly Lukyanov, gave **fatherly counsel** to the party. (Wasow 1997: 83)
(10b) I gave to Mary **the valuable book that was extremely difficult to find**. (Wasow 2002: 42)

Table 1.1 *Correlation coefficients for weight measures in three data sets (Wasow 1997: 93, 2002: 32)*

	HNPS	DA	PM
Words and nodes	0.94	0.96	0.99
Words and phrasal nodes	0.96	0.97	0.95

(11a) French President François Mitterrand sent an envoy to pick **the communiqué** up. (Wasow 1997: 83)

(11b) Pat looked up **where to go**. (Wasow 2002: 58)

The examples in (9)–(11) illustrate that long and structurally complex NPs tend to occur towards the end of the sentence. In each of the cases, the longer structure in the b-examples is also structurally more complex than the shorter one in the a-examples. This qualitative finding is supported by quantitative evidence: Table 1.1 shows us that the correlation between length and structure is extremely high for all three constructions and for the comparison between words and both types of structural weight measures (nodes and phrasal nodes).

Against these findings it is not surprising that many researchers apply the most economic operationalisation of syntactic complexity: measuring the length of phrases in terms of word counts (cf., e.g., Arnold et al. 2000; Szmrecsanyi 2004; Rosenbach 2005;[5] Jäger and Rosenbach 2006; Kreyer 2006; Bresnan et al. 2007; Bresnan and Ford 2010). Word counts are not only easier to count manually than the number of nodes or phrasal nodes, but word counts can also often be automatised. Hawkins, too, in his theory of processing efficiency, takes word counts over specific structural domains as a suitable proxy of the structural complexity of a phrase (see, e.g., Hawkins 1994; 2001; 2004).[6]

Hawkins' approach to processing efficiency shows that the structure and length of syntactic constituents are closely correlated and that speakers of head-initial languages such as English prefer to have short elements precede long ones in constituent ordering. This word-order preference was already captured in Behaghel's Law of Growing Elements (1909) and Quirk et al.'s Principle of End Weight (1972: 943), but has received considerable elaboration in Hawkins' theory of processing efficiency since.[7] I will illustrate

[5] We should note here that Rosenbach (2005: 617) is one of the few researchers who explicitly points out that it is still a matter of debate whether length and structure are independent factors.

[6] Hawkins' efficiency theory defines processing preferences for language comprehension. Whether it also holds for production models yet needs to be tested (see Hawkins 1994: 425–7; 2001: 5).

[7] The Principle of End Weight will be discussed more extensively in Section 2.2.1.

Hawkins' idea of processing efficiency by comparing the alternation in (12) and (13), which involves two prepositional phrases (PPs) following an intransitive verb (*looked*).

(12) The gamekeeper $_{VP}$[looked $_{PP1}$ [through his binoculars] $_{PP2}$[into the blue but slightly overcast sky]].
1 2 3 4 5

(13) The gamekeeper $_{VP}$[looked $_{PP2}$[into the blue but slightly overcast sky] $_{PP1}$[through his binoculars]]
1 2 3 4 5 6 7
8 9 (Hawkins 2001: 4)

In Hawkins' model, the parsing of words proceeds over what he calls the Constituent Recognition Domain (CRD) of a phrase.[8] This is the structural domain that needs to be parsed in order to recognise the immediate constituents (ICs) of a phrase. In our case, the hearer needs to recognise the ICs of the VP, which are V, PP_1 and PP_2. As Examples (12) and (13) illustrate, the CRD is much smaller in (12), where it consists of five words, than in (13), where it has nine words. We will now look in more detail at how the parsing of constituents works.

Hawkins' model assumes that, as soon as we encounter the head of a phrase (e.g. a verb of a verb phrase), the respective structural node is activated (the VP). In our example, encountering *looked* in (12) activates the VP and encountering *through* activates the first PP. If users' prime motivation in processing is to be as efficient as possible, they should, no doubt, prefer the structure in (12), where they need to parse five words (*looked through his binoculars into*) in order to recognise the three ICs of the VP, while nine words (*looked into the blue but slightly overcast sky through*) have to be parsed in order to recognise these constituents in (13).

The metric that is applied to measure the processing efficiency with which hearers recognise the ICs of a CRD is the IC-to-nonIC or IC-to-word ratio[9] that is summarised in Hawkins' Early Immediate Constituents (EIC) Principle: 'The human parser prefers linear orders that minimize CRDs (by maximizing their IC-to-nonIC [or IC-to-word] ratios), and in proportion to the minimization difference between competing orders' (Hawkins 2001: 5). The IC-to-word ratios are maximised if the number of words that have to be parsed in order to recognise the immediate constituents are kept to a minimum. This leaves us with an IC-to-word

[8] The term 'parsing' exclusively refers to processing strategies of the hearer (comprehension). It does not apply to the production processes of the speaker.

[9] Hawkins explains that IC-to-word ratios are simplified procedures for calculating IC-to-nonIC ratios, which take into account all terminal and non-terminal nodes in a CRD (see, e.g., Hawkins 2001: 4).

ratio of 3/5 = 60% for Example (12) and with a ratio of 3/9 = 33% for Example (13). Since higher ratios indicate greater parsing efficiency, (12) should be clearly preferred over (13). Hawkins argues that, on a psycholinguistic level, higher and lower IC-to-nonIC ratios correspond to the amount of processing imposed on our working memory. He claims that the reason for why the parser prefers more minimal processing domains is that they put fewer demands on our working memory.

As concerns the relation between word counts and structural complexity, Hawkins' model attests to a very strong correlation between the two, supporting Wasow's correlation coefficients provided above (Wasow 1997; 2002). According to Hawkins (1994: 74), 'More words means more structure: each new word in a domain adds one terminal node, plus a pre-terminal category node, and possibly more non-terminal nodes besides.' In this framework, the number of words which need to be parsed in order to recognise the ICs of a phrase can thus be considered a suitable proxy of the structural complexity of a constituent.

To summarise, this section has provided empirical evidence for the fact that the length and the structural complexity of syntactic phrases are highly correlated. Two pieces of evidence have been adduced in favour of this correlation: (a) the correlation coefficients between length and structural measures that apply to various cases of variation and (b) Hawkins' theory of the processing preferences of listeners. There is, however, at least some empirical evidence running counter to this claim. I will review such evidence in Section 1.3, where I will ask whether length and structure can both have an independent status in terms of the syntactic complexity of NPs.

1.3 Factor isolation: length vs structure

Since measures of length and structural complexity are very highly correlated (see the evidence provided in Section 1.2), it is extremely difficult to tease apart the effects that each of them has on given cases of variation. There are, however, a few studies conducted in a variationist and a psycholinguistic framework which have isolated the effects of length and structure. Among them are Ferreira (1991), Rickford et al. (1995), Wasow (2002), Wasow and Arnold (2003; 2005), and Grafmiller and Shih (2011). In sum, they show that the relative importance of length and structural complexity depends on the construction under investigation.

The question of whether structure has a status independent of length has already been raised by Chomsky (1975). My brief overview of studies isolating the effects of length and structure will begin with his grammaticality judgement, followed by large-scale empirical evidence collected in a psycholinguistic setting and, subsequently, by evidence coming from a corpus-linguistic/sociolinguistic framework. For the latter part, I will focus