1 | *Introduction*

In this chapter, we provide an overview of what this book is about, how it connects with what you likely already know about English grammar, and how each chapter is structured to lead you from learning new concepts and skills to applying those concepts and ultimately to designing your own research projects.

1.1 What This Book Is About

"Discourse syntax" is the cover term for the study of all aspects of syntactic form that we can only explain with reference to the surrounding discourse. The objects of study are syntactic forms, constructions, or phenomena that can only be fully accounted for by taking the surrounding discourse (also referred to as "text," regardless of written or spoken mode) into consideration. This is where this book is fundamentally different from a general introduction into English syntax, which first and foremost deals with the structure of English sentences without considering any surrounding discourse. For example, in an introductory English syntax class, one would discuss that in English the subject position cannot be left empty (hence we have "dummy subjects," like *there* or *it*), that the object comes after the verb, and that questions require subject–auxiliary **inversion**. One might also discuss that under certain circumstances the object may come first in the sentence and one might introduce the name for such a construction (**topicalization**). Example (1a), taken from the freely available Corpus of Contemporary American English (COCA; Davies 2008), illustrates this construction; Example (1b) would be its canonical syntax counterpart.

(1) a. The second batch I only microwaved for 4 minutes covered and that seemed to work. (COCA, Blog, 2012)
     b. I only waved the second batch for 4 minutes covered and that seemed to work.

3

What one normally does not study in such a class is under which
discourse conditions topicalization occurs, which can leave the impres-
sion that word-order variations like topicalization are simply a stylistic
choice. But that is not the case! We cannot give a proper account of
topicalization without specifying the discourse conditions that warrant
it. The same holds for other constructions that you are probably
already familiar with, like passivization, inversion, or left dislocation,
which are all, in some way, non-canonical clauses, that is, clauses that
are derived from more basic, or canonical, patterns through a change of
word order or the addition of extra lexical material. Canonical and
non-canonical patterns of a clause typically share the same propos-
itional meaning (or **proposition**), which is to say that they underlie the
same truth conditions. There are no circumstances under which only
sentence (1a) would be true, but not (1b), and vice versa. The second
kind of phenomena considered in discourse syntax are constructions
and elements that cannot be described as variations of another pattern
but whose very function is to hold the discourse together and weave
connections beyond the sentence, such as **conjunctions** and **coordin-
ators**. We refer to the first set of phenomena as *grammar in discourse*
and the second as *grammar of discourse*. Discourse syntax is therefore a
field of linguistic study that describes and investigates patterns of
syntax in language use, with the focus on illuminating the role of the
surrounding text, rather than on the role of the speaker's individual
preferences or choices.

1.2 How This Book Connects with What You Already Know

Perhaps you are working with this book because it has been
assigned as the textbook or supplementary reading for one of
your classes. If that is the case, the chances are that you already
have some knowledge of English syntax. You have probably stud-
ied parts of speech (nouns, verbs, adjectives, conjunctions, etc.),
tests for the constituency of phrases (noun phrases, verb phrases,
and so on) and functions they take in the clause (subject, object,
adverbial/**adjunct**). You know that *before* in (2b) is a preposition,
which makes the italicized sequence in (2b) a preposition phrase,
while *before* in (2a) is a subordinator (conjunction), which makes
the italicized sequence in (2a) no phrase at all.

(2) a. We've got about 10 weeks *before the trial* comes up in January. (COCA, Movies, 1996)
     b. By its very definition, any expertise a juror gained *before the trial* cannot be extrinsic. (COCA, Academic, 2017)

You are likely familiar with the distinction between simple and complex sentences and the functions that subordinate clauses can take. The *that*-clause in (3a) is a content clause (traditionally also known as an object clause), while the one in (3b) is a relative clause (modifying the noun *plan*).

(3) a. He said that the progress of the action plan will be reviewed by him every month. (COCA, Web, 2012)
     b. Choose the plan that's right for you. (COCA, News, 2019)

You also probably know that English is a language that has the requirement that the subject position be filled (in finite, or tensed, clauses) and that this is the reason we have sentences like (4b), with a dummy subject. What you may not have studied are the circumstances under which speakers may choose to produce a sentence like (4a), with the *what*-clauses serving as a subject, rather than the construction known as **extraposition** in (4b).

(4) a. What had happened was clear. (COCA, Magazine, 1993)
     b. It was perfectly clear what had happened. (COCA, Web, 2012)

You may also have heard that English is considered a strict word order language. The function of a phrase is indicated by its position in the clause rather than by a case-marking affix, as in other Germanic languages. In German, unlike in English, the function of a phrase can be expressed through morphology. For example, in (5), it doesn't matter if we put the phrase *den Jungen* (the boy) in a preverbal or postverbal position. The accusative case marking *-n* indicates that the phrase functions as the direct object.

(5) a. Den Jungen kenne ich nicht.
         The$_{ACC}$ boy$_{ACC}$ know$_{1ST-PERSON-PRESENT}$ I not.
         "I don't know the boy"
     b. Ich kenne den Jungen nicht.
         I know$_{1ST-PERSON-PRESENT}$ the$_{ACC}$ boy$_{ACC}$ not.
         "I don't know the boy"

English used to be like that. Old English, the earliest recorded form of English (roughly from the fifth to the eleventh century), was a language in

which grammatical relations could be expressed through inflection. However, by around the twelfth century, much of the case inflection in Old English had become optional. How and exactly why this happened is a complicated question, due to the limited availability of data. What we do know is that this "deflexion" (Allen 2016) happened most rapidly in the northern parts of England, where there was close contact with Scandinavians. It seems, therefore, that language contact is an accelerating factor, if not the initiating one. The further loss of inflectional morphology during the Middle English period led to a decrease in flexibility in word order for **canonical sentences**, which ultimately paved the way for non-canonical patterns to be associated with specific communicative purposes. One example would be the increase of constructions that allow for non-agents to fill the subject position and become the entity that the sentence is "about" (also known as the **topic**), such as passives, cleft sentences, and middle constructions (*This jam spreads easily*). In other words, in a language with a strong sense of canonical syntax, non-canonical patterns have to be motivated, and, more often than not, that motivation lies in the surrounding text. As unlikely as it sounds: A language with strict word order, like Modern English, is the perfect ground to study syntax beyond the sentence boundary.

1.3 How This Book Is Structured

This book is divided into three parts: Part I: *Foundations* (Chapters 1 and 2) introduces the concept of canonical and non-canonical syntax in English and situates the book in the context of variationist linguistics – the idea that linguistic variation does not occur at random, but is highly structured and influenced by linguistic and extralinguistic factors. Of all those factors, we will focus on the role of the surrounding discourse of different registers and on **genre** conventions. As a reference model for linguistic terminology, we will use the *Cambridge Grammar of the English Language* (Huddleston & Pullum 2002), which also comes in a less comprehensive student version (Huddleston & Pullum 2005). Another grammar we will often refer to is the *Grammar of Spoken and Written English* (Biber et al. 2021), which provides rich usage data on grammatical features in four register varieties (conversation, fiction, news, and academic prose). There is also a student version of the previous edition (Biber et al. 1999) of this grammar (Biber, Conrad & Leech 2002).

Reference Grammars of English

There is a myriad of grammars of the English language. While they all, in theory, describe the same facts, they do not all have the same goal. Some grammars intend to provide guidance to learners, some resemble style guides, and some use rather idiosyncratic terminology. We are choosing the *Cambridge Grammar of the English Language* as our point of reference because it is a comprehensive, synchronic account of English grammar, written with the goal of incorporating "as many as possible of the insights achieved in modern linguistics" (Huddleston & Pullum 2002: xv), without assuming great familiarity with theoretical linguistics. Using the *Cambridge Grammar* as our **reference grammar** means that we will adopt its terminology along with the premise that the role of the linguist is to "describe and not prescribe" (Huddleston & Pullum 2002: 2). By contrast, we will use the *Grammar of Spoken and Written English* (henceforth *GSWE*, Biber et al. 2021) mostly as a resource on usage data. The *GSWE* is based on data from a proprietary **corpus** (40 million words of text representing four main register categories) and, along with a description of the linguistic system, also provides information, including quantitative data, on how a grammatical feature is used in a particular situational variety of English. The *GSWE* itself closely follows Quirk et al.'s *A Comprehensive Grammar of the English Language* (1985) in its terminology. An earlier version of the *GSWE* was published as the *Longman Grammar of Spoken and Written English* (Biber et al. 1999). As you can see, a lot of thought goes into the writing of grammars, and the analysis of grammar writing has become a linguistic field in its own right, called "grammaticography."

Part II: *Grammar in Discourse* (Chapters 3–5) focuses on grammatical phenomena *in* discourse and looks at how syntactic patterns that you are most likely already familiar with (like **topicalization** or **particle shift**) are realized in different discourse situations. The underlying assumption here is that the reason for choosing one syntactic option

over another lies in the surrounding discourse. The three chapters in this part move from phenomena that affect the beginning of the sentence (such as any kind of fronting operation, Chapter 3), over phenomena in the core sentence (anything from subject to object position, Chapter 4), to complex sentence endings (any kind of construction that shifts material toward the end of the sentence, like extraposition, Chapter 5). If the names of these constructions don't mean anything to you at this point, don't worry, every construction that we highlight will be defined and illustrated with data from linguistic corpora (and we will also show you how to work with these corpora).

Part III: *Grammar of Discourse* (Chapters 6–9) looks at the grammar *of* discourse, in particular the way in which sentences are connected, and also discusses syntactic phenomena conditioned by or brought about by genre conventions. Chapter 6 deals with the most obvious way of connecting sentences, namely through extra linguistic material known as "**connectives**" (such as coordinators). Chapter 7 looks at how we can underspecify information in a given sentence (through the use of **pronouns** and elliptical constructions) because it refers to something already mentioned before. Chapter 8 introduces the concept of "**discourse markers**," a series of elements (not necessarily unified in their appearance or syntactic behavior) that have the function of structuring the discourse. Lastly, Chapter 9 goes beyond the local discourse and discusses the role that register or genres as varieties of discourse have for shaping the syntactic characteristics of a text. You will see that certain constructions and phenomena may be used not so much because of any relationship with the surrounding discourse, but because the text is part of a whole discourse situation, and the relationship between discourse participants or the purpose of the text may determine the use of syntactic constructions. For example, in scientific abstracts we have an overrepresentation of passive-**voice** constructions because in many scientific fields there is still a convention to make the object of inquiry (the thing that is studied) the subject of a sentence, rather than the person who led the inquiry (the **agent**).

We like clarity in the classes that we teach and have aimed to achieve the same in this book. Each chapter will start with a definition of learning outcomes for the chapter. Each chapter will then introduce core concepts and questions necessary for the subsequent discussion, such as non-canonical word order or discourse grammar and **cohesion**. This introductory part is followed by a discussion of two or three selected syntactic phenomena or constructions that illustrate the

concept under discussion, plus one subsection presenting exemplary empirical work (mostly based on corpus data, but occasionally also on experimental data). That way, you will gain insight into what is being studied in a field without being overburdened with all the complexities around the issue right away. References come both from recent work and from classics in the field. The chapters conclude with a summary, recommendations for further reading, and two kinds of exercises: level-one exercises will help you practice your analytic skills by applying the concepts introduced in the chapter, and level-two exercises will ask you to create or interpret data and guide you toward thinking about designing your own research. Throughout the text we will let you know when you should be ready to take on which exercise with this study icon:

For these exercises, we do not presuppose any prior experience with corpus-linguistic methodology or statistics and we include practical information on carrying out corpus-based research (how to do it, how to present and interpret data, but also which problems to watch out for) throughout the book. These tips will be marked with the toolbox icon.

The chapters are rounded out with squibs on interesting, but adjacent questions of language usage, clearly set apart from the main text in boxes labeled "Good to Know," marked with the owl icon:

The index and glossary at the end will help you with studying specific concepts (terms that are printed in bold in the text when occurring for the first time have their own entry in the glossary). The index focuses on those pages where a concept is introduced or elaborated on. Overall,

the chapter structure of the book makes it possible to cover the whole book in the course of one semester or to cover selected parts and use the book as a secondary textbook or resource book.

As the linguistic world we live in continues to develop, presenting us with new words, new modes of expression, and new communicative needs, we hope you will share our excitement about studying syntax beyond the sentence. It is our goal to empower you to pursue projects of your own that show that syntax is a lot more than a set of rules holding a sentence together.

Further Reading

- The developing field of grammaticography is discussed in Ameka, Dench & Evans (2006). Articles in this edited volume deal with aspects like the role of linguistic theory in grammar writing and the distinction between native and non-native speakers.
- If you are interested in learning more about the language-internal and external factors that made English the strict word order language it is today, you will find the articles in *The Cambridge Handbook of English Historical Linguistics* (Kytö & Pahta 2016) useful, especially Cynthia Allen's article "Typological change: investigating loss of inflection in early English" (2016). The book also offers overviews on theoretical frameworks, such as historical pragmatics or generative grammar, and on methodologies, such as corpus linguistics and philological methods.

2 | *Concepts, Data, and Methods*

2.1 Introduction: Why Discourse Syntax?

In introductions to linguistics, we all learn that syntax is the study of rules that form sentences. If you are used to tree diagrams to visualize syntactic structures, you know that they never go beyond the boundary of a sentence. But we don't speak in isolated sentences (actually, we often don't speak in sentences at all) and syntactic phenomena do not stop at the sentence boundary. To demonstrate how strongly the structure of sentences is influenced by the surrounding discourse, let's look at the sentences in Figure 2.1, taken from the beginning of an op-ed in *The New York Times* (August 2020) and numbered arbitrarily. Let's put them in the order that seems most natural to you. Note that this text is an opinion piece – it does not recount a series of events. Therefore, in arranging the sentences, you cannot rely on your extra-linguistic knowledge of how certain types of events unfold (for example, we know that in a car accident the crash comes before anyone is taken to the hospital). Rather, you will apply your intuitive knowledge about discourse syntax. Keep track of the markers that you rely on for making your choice and write down the sequence of numbers that seems right to you.

Perhaps your thought process looks something like this: Sentence #6 does not make a good first sentence – a connective like *and* would typically not be the first word in a text. You also probably did not put sentence #4 first – a demonstrative pronoun like *that* needs to have an antecedent and it wouldn't be clear what *that* refers to if sentence #4 came first. Similarly, you probably ruled out sentences #1, #3, and #5 as potential first sentences because they have personal pronouns (*it, her, she*) whose reference would be equally unclear. Sentence #8 is an elliptical structure (there is no verb in the main clause), which points to this sentence not being the first sentence in this text either. This leaves us with sentences #2 and #7. Sentence #2 has an indefinite noun phrase as its subject (*Hundreds of thousands of*
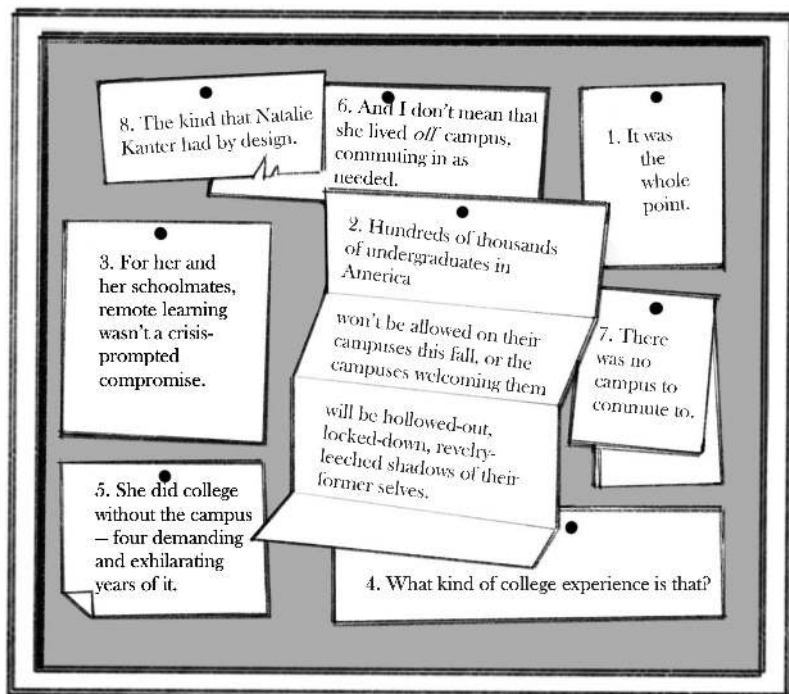
11

Figure 2.1 Deconstructed paragraph
(Bruni, 2020)

undergraduates in America) and tells us something about them. This is achieved through the use of a **passive construction** (*won't be allowed*). Sentence #7 uses a dummy pronoun (*there*) as a subject and thus does not really establish an entity the text could be about. That makes #2 the most natural first sentence. Once the topic of students and their return-to-college experience is established, sentence #4 is the one that is the best continuation of that topic. Sentence #4 is in question format, and sentence #8 is the best answer to that question because it picks up on the same topic, choosing the same noun, *kind*, but this time with a definite article, *the*. And so on.

Go to Exercise 1 to see if your sequence matches the one of the original op-ed. There will also be the opportunity to reconstruct another paragraph.