Text Analytics with Python

A Practical Real-World Approach to Gaining Actionable Insights from Your Data

Dipanjan Sarkar

Apress®

Text Analytics with Python

A Practical Real-World Approach to Gaining Actionable Insights from your Data

Dipanjan Sarkar

Apress®

Text Analytics with Python: A Practical Real-World Approach to Gaining Actionable Insights from Your Data

Dipanjan Sarkar
Bangalore, Karnataka
India

ISBN-13 (pbk): 978-1-4842-2387-1 ISBN-13 (electronic): 978-1-4842-2388-8
DOI 10.1007/978-1-4842-2388-8

Library of Congress Control Number: 2016960760

Trademarked names, logos, and images may appear in this book. Rather than use a trademark symbol with every occurrence of a trademarked name, logo, or image we use the names, logos, and images only in an editorial fashion and to the benefit of the trademark owner, with no intention of infringement of the trademark.

The use in this publication of trade names, trademarks, service marks, and similar terms, even if they are not identified as such, is not to be taken as an expression of opinion as to whether or not they are subject to proprietary rights.

While the advice and information in this book are believed to be true and accurate at the date of publication, neither the authors nor the editors nor the publisher can accept any legal responsibility for any errors or omissions that may be made. The publisher makes no warranty, express or implied, with respect to the material contained herein.

Managing Director: Welmoed Spahr
Lead Editor: Mr. Sarkar
Technical Reviewer: Shanky Sharma

Coordinating Editor: Sanchita Mandal
Copy Editor: Corbin Collins
Compositor: SPi Global
Indexer: SPi Global
Artist: SPi Global

Distributed to the book trade worldwide by Springer Science+Business Media New York, 233 Spring Street, 6th Floor, New York, NY 10013. Phone 1-800-SPRINGER, fax (201) 348-4505, e-mail orders-ny@springer-sbm.com, or visit www.springeronline.com. Apress Media, LLC is a California LLC and the sole member (owner) is Springer Science + Business Media Finance Inc (SSBM Finance Inc). SSBM Finance Inc is a **Delaware** corporation.

For information on translations, please e-mail rights@apress.com, or visit www.apress.com.

Any source code or other supplementary materials referenced by the author in this text are available to readers at www.apress.com. For detailed information about how to locate your book's source code, go to www.apress.com/source-code/. Readers can also access source code at SpringerLink in the Supplementary Material section for each chapter.

Printed on acid-free paper

This book is dedicated to my parents, partner, well-wishers, and especially to all the developers, practitioners, and organizations who have created a wonderful and thriving ecosystem around analytics and data science.

Contents at a Glance

Contents

About the Author

Dipanjan Sarkar is a data scientist at Intel, the world's largest silicon company, which is on a mission to make the world more connected and productive. He primarily works on analytics, business intelligence, application development, and building large-scale intelligent systems. He received his master's degree in information technology from the International Institute of Information Technology, Bangalore, with a focus on data science and software engineering. He is also an avid supporter of self-learning, especially through massive open online courses, and holds a data science specialization from Johns Hopkins University on Coursera.

Sarkar has been an analytics practitioner for over four years, specializing in statistical, predictive, and text analytics. He has also authored a couple of books on R and machine learning, reviews technical books, and acts as a course beta tester for Coursera. Dipanjan's interests include learning about new technology, financial markets, disruptive startups, data science, and more recently, artificial intelligence and deep learning. In his spare time he loves reading, gaming, and watching popular sitcoms and football.

About the Technical Reviewer

Shanky Sharma Currently leading the AI team at Nextremer India, Shanky Sharma's work entails implementing various AI and machine learning–related projects and working on deep learning for speech recognition in Indic languages. He hopes to grow and scale new horizons in AI and machine learning technologies. Statistics intrigue him and he loves playing with numbers, designing algorithms, and giving solutions to people. He sees himself as a solution provider rather than a scripter or another IT nerd who codes. He loves heavy metal and trekking and giving back to society, which, he believes, is the task of every engineer. He also loves teaching and helping people. He is a firm believer that we learn more by helping others learn.

Acknowledgments

This book would definitely not be a reality without the help and support from some excellent people in my life. I would like to thank my parents, Digbijoy and Sampa, my partner Durba, and my family and well-wishers for their constant support and encouragement, which really motivates me and helps me strive to achieve more.

This book is based on various experiences and lessons learned over time. For that I would like to thank my managers, Nagendra Venkatesh and Sanjeev Reddy, for believing in me and giving me an excellent opportunity to tackle challenging problems and also grow personally. For the wealth of knowledge I gained in text analytics in my early days, I would like to acknowledge Dr. Mandar Mutalikdesai and Dr. Sanket Patil for not only being good managers but excellent mentors.

A special mention goes out to my colleagues Roopak Prajapat and Sailaja Parthasarathy for collaborating with me on various problems in text analytics. Thanks to Tamoghna Ghosh for being a great mentor and friend who keeps teaching me something new every day, and to my team, Raghav Bali, Tushar Sharma, Nitin Panwar, Ishan Khurana, Ganesh Ghongane, and Karishma Chug, for making tough problems look easier and more fun.

A lot of the content in this book would not have been possible without Christine Doig Cardet, Brandon Rose, and all the awesome people behind Python, Continuum Analytics, NLTK, `gensim`, `pattern`, spaCy, `scikit-learn`, and many more excellent open source frameworks and libraries out there that make our lives easier. Also to my friend Jyotiska, thank you for introducing me to Python and for learning and collaborating with me on various occasions that have helped me become what I am today.

Last, but never least, a big thank you to the entire team at Apress, especially to Celestin Suresh John, Sanchita Mandal, and Laura Berendson for giving me this wonderful opportunity to share my experience and what I've learned with the community and for guiding me and working tirelessly behind the scenes to make great things happen!

Introduction

I have been into mathematics and statistics since high school, when numbers began to really interest me. Analytics, data science, and more recently text analytics came much later, perhaps around four or five years ago when the hype about Big Data and Analytics was getting bigger and crazier. Personally I think a lot of it is over-hyped, but a lot of it is also exciting and presents huge possibilities with regard to new jobs, new discoveries, and solving problems that were previously deemed impossible to solve.

Natural Language Processing (NLP) has always caught my eye because the human brain and our cognitive abilities are really fascinating. The ability to communicate information, complex thoughts, and emotions with such little effort is staggering once you think about trying to replicate that ability in machines. Of course, we are advancing by leaps and bounds with regard to cognitive computing and artificial intelligence (AI), but we are not there yet. Passing the Turing Test is perhaps not enough; can a machine truly replicate a human in all aspects?

The ability to extract useful information and actionable insights from heaps of unstructured and raw textual data is in great demand today with regard to applications in NLP and text analytics. In my journey so far, I have struggled with various problems, faced many challenges, and learned various lessons over time. This book contains a major chunk of the knowledge I’ve gained in the world of text analytics, where building a fancy word cloud from a bunch of text documents is not enough anymore.

Perhaps the biggest problem with regard to learning text analytics is not a lack of information but *too much* information, often called *information overload*. There are so many resources, documentation, papers, books, and journals containing so much theoretical material, concepts, techniques, and algorithms that they often overwhelm someone new to the field. What is the right technique to solve a problem? How does text summarization really work? Which are the best frameworks to solve multi-class text categorization? By combining mathematical and theoretical concepts with practical implementations of real-world use-cases using Python, this book tries to address this problem and help readers avoid the pressing issues I’ve faced in my journey so far.

This book follows a comprehensive and structured approach. First it tackles the basics of natural language understanding and Python constructs in the initial chapters. Once you’re familiar with the basics, it addresses interesting problems in text analytics in each of the remaining chapters, including text classification, clustering, similarity analysis, text summarization, and topic models. In this book we will also analyze text structure, semantics, sentiment, and opinions. For each topic, I cover the basic concepts and use some real-world scenarios and data to implement techniques covering each concept. The idea of this book is to give you a flavor of the vast landscape of text analytics and NLP and arm you with the necessary tools, techniques, and knowledge to tackle your own problems and start solving them. I hope you find this book helpful and wish you the very best in your journey through the world of text analytics!

CHAPTER 1

Natural Language Basics

We have ushered in the age of Big Data where organizations and businesses are having difficulty managing all the data generated by various systems, processes, and transactions. However, the term *Big Data* is misused a lot due to the nature of its popular but vague definition of "the 3 V's"—volume, variety, and velocity of data. This is because sometimes it is very difficult to exactly quantify what data is "Big." Some might think a billion records in a database would be Big Data, but that number seems really minute compared to the *petabytes* of data being generated by various sensors or even social media. There is a large volume of unstructured textual data present across all organizations, irrespective of their domain. Just to take some examples, we have vast amounts of data in the form of tweets, status updates, comments, hashtags, articles, blogs, wikis, and much more on social media. Even retail and e-commerce stores generate a lot of textual data from new product information and metadata with customer reviews and feedback.

The main challenges associated with textual data are twofold. The first challenge deals with effective storage and management of this data. Usually textual data is unstructured and does not adhere to any specific predefined data model or schema, which is usually followed by relational databases. However, based on the data semantics, you can store it in either SQL-based database management systems (DBMS) like SQL Server or even NoSQL-based systems like MongoDB. Organizations having enormous amounts of textual datasets often resort to file-based systems like Hadoop where they dump all the data in the Hadoop Distributed File System (HDFS) and access it as needed, which is one of the main principles of a *data lake*.

The second challenge is with regard to analyzing this data and trying to extract meaningful patterns and useful insights that would be beneficial to the organization. Even though we have a large number of machine learning and data analysis techniques at our disposal, most of them are tuned to work with numerical data, hence we have to resort to areas like *natural language processing* (NLP) and specialized techniques, transformations, and algorithms to analyze text data, or more specifically *natural language*, which is quite different from programming languages that are easily understood by machines. Remember that textual data, being highly unstructured, does not follow or adhere to structured or regular syntax and patterns—hence we cannot directly use mathematical or statistical models to analyze it.

Electronic supplementary material The online version of this chapter (doi:10.1007/978-1-4842-2388-8_1) contains supplementary material, which is available to authorized users.

D. Sarkar, *Text Analytics with Python*, DOI 10.1007/978-1-4842-2388-8_1

Before we dive into specific techniques and algorithms to analyze textual data, we will be going over some of the main concepts and theoretical principles associated with the nature of text data in this chapter. The primary intent here is to get you familiarized with concepts and domains associated with *natural language understanding*, *processing*, and *text analytics*. We will be using the Python programming language in this book primarily for accessing and analyzing text data. The examples in this chapter will be pretty straightforward and fairly easy to follow. However, you can quickly skim over Chapter 2 in case you want to brush up on Python before going through this chapter. All the examples are available with this book and also in my GithHub repository at `https://github.com/dipanjanS/text-analytics-with-python` which includes programs, code snippets and datasets. This chapter covers concepts relevant to natural language, linguistics, text data formats, syntax, semantics, and grammars before moving on to more advanced topics like *text corpora*, NLP, and text analytics.

Natural Language

Textual data is unstructured data but it usually belongs to a specific language following specific syntax and semantics. Any piece of text data—a simple word, sentence, or document—relates back to some natural language most of the time. In this section, we will be looking at the definition of natural language, the philosophy of language, language acquisition, and the usage of language.

What Is Natural Language?

To understand text analytics and natural language processing, we need to understand what makes a language "natural." In simple terms, a *natural* language is one developed and evolved by humans through natural use and communication, rather than constructed and created artificially, like a computer programming language.

Human languages like English, Japanese, and Sanskrit are natural languages. Natural languages can be communicated in different forms, including speech, writing, or even signs. There has been a lot of scholarship and effort applied toward understanding the origins, nature, and philosophy of language. We will discuss that briefly in the following section.

The Philosophy of Language

We now know what a natural language means. But think about the following questions. What are the origins of a language? What makes the English language "English"? How did the meaning of the word *fruit* come into existence? How do humans communicate among themselves with language? These are definitely some heavy philosophical questions.

The *philosophy of language* mainly deals with the following four problems and seeks answers to solve them:

- The nature of meaning in a language
- The use of language
- Language cognition
- The relationship between language and reality

- *The nature of meaning in a language* is concerned with the semantics of a language and the nature of meaning itself. Here, philosophers of language or linguistics try to find out what it means to actually "mean" anything—that is, how the meaning of any word or sentence originated and came into being and how different words in a language can be synonyms of each other and form relations. Another thing of importance here is how structure and syntax in the language pave the way for semantics, or to be more specific, how words, which have their own meanings, are structured together to form meaningful sentences. *Linguistics* is the scientific study of language, a special field that deals with some of these problems we will be looking at in more detail later on. Syntax, semantics, grammars, and parse trees are some ways to solve these problems. The nature of meaning can be expressed in linguistics between two human beings, notably a sender and a receiver, as what the sender tries to express or communicate when they send a message to a receiver, and what the receiver ends up understanding or deducing from the context of the received message. Also from a non-linguistic standpoint, things like body language, prior experiences, and psychological effects are contributors to meaning of language, where each human being perceives or infers meaning in their own way, taking into account some of these factors.

- *The use of language* is more concerned with how language is used as an entity in various scenarios and communication between human beings. This includes analyzing speech and the usage of language when speaking, including the speaker's intent, tone, content and actions involved in expressing a message. This is often termed as a *speech act* in linguistics. More advanced concepts such as the origins of language creation and human cognitive activities such as language acquisition which is responsible for learning and usage of languages are also of prime interest.

- *Language cognition* specifically focuses on how the cognitive functions of the human brain are responsible for understanding and interpreting language. Considering the example of a typical sender and receiver, there are many actions involved from message communication to interpretation. Cognition tries to find out how the mind works in combining and relating specific words into sentences and then into a meaningful message and what is the relation of language to the thought process of the sender and receiver when they use the language to communicate messages.

- *The relationship between language and reality* explores the extent of truth of expressions originating from language. Usually, philosophers of language try to measure how factual these expressions are and how they relate to certain affairs in our world which are true. This relationship can be expressed in several ways, and we will explore some of them.

One of the most popular models is the *triangle of reference*, which is used to explain how words convey meaning and ideas in the minds of the receiver and how that meaning relates back to a real world entity or fact. The triangle of reference was proposed by Charles Ogden and Ivor Richards in their book, *The Meaning of Meaning*, first published in 1923, and is denoted in Figure 1-1.

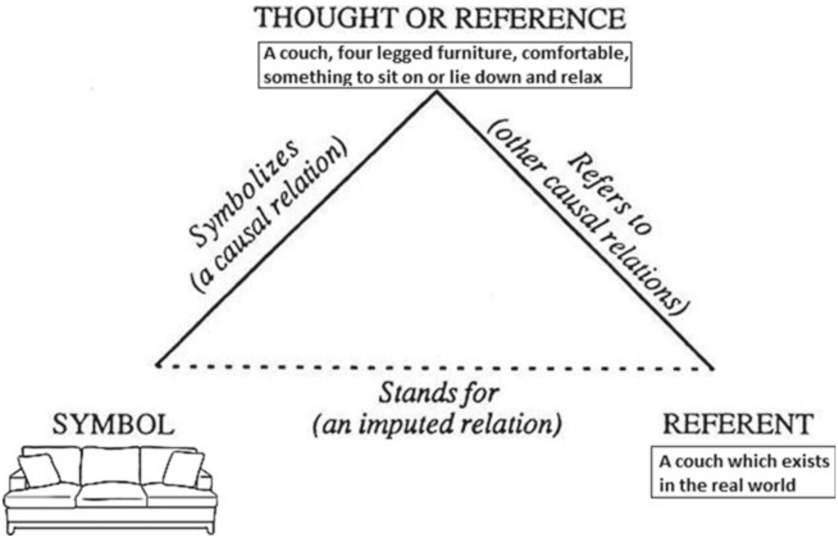

THOUGHT OR REFERENCE
A couch, four legged furniture, comfortable, something to sit on or lie down and relax
Symbolizes (a causal relation)
Refers to (other causal relations)
Stands for (an imputed relation)
SYMBOL
REFERENT
A couch which exists in the real world

Figure 1-1. *The triangle of reference model*

The triangle of reference model is also known as the *meaning of meaning* model, and I have depicted the same in Figure1-1 with a real example of a *couch* being perceived by a person which is present in front of him. A *symbol* is denoted as a linguistic symbol, like a word or an object that evokes thought in a person's mind. In this case, the *symbol* is the couch, and this evokes thoughts like *what is a couch, a piece of furniture that can be used for sitting on or lying down and relaxing, something that gives us comfort*. These thoughts are known as a *reference* and through this reference the person is able to relate it to something that exists in the real world, termed a *referent.* In this case the referent is the couch which the person perceives to be present in front of him.

The second way to find out relationships between language and reality is known as the *direction of fit*, and we will talk about two main directions here. The *word-to-world* direction of fit talks about instances where the usage of language can reflect reality. This indicates using words to match or relate to something that is happening or has already happened in the real world. An example would be the sentence *The Eiffel Tower is really big,* which accentuates a fact in reality. The other direction of fit, known as *world-to-word*, talks about instances where the usage of language can change reality. An example here would be the sentence *I am going to take a swim*, where the person *I* is changing reality by going to take a swim by representing the same in the sentence being communicated. Figure 1-2 shows the relationship between both the directions of fits.

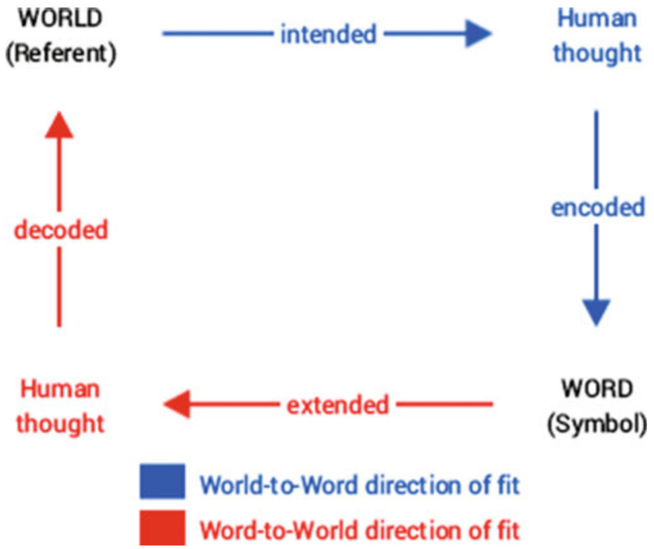


Figure 1-2. *The direction of fit representation*

It is quite clear from the preceding depiction that based on the referent that is perceived from the real world, a person can form a representation in the form of a symbol or word and consequently can communicate the same to another person, which forms a representation of the real world based on the received symbol, thus forming a cycle.

Language Acquisition and Usage

By now, we have seen what natural languages mean and the concepts behind language, its nature, meaning, and use. In this section, we will talk in further detail about how language is perceived, understood, and learned using cognitive abilities by humans, and finally we will end our discussion with the main forms of language usage, discussed in brief as *speech acts*. It is important to not only understand what natural language denotes but also how humans interpret, learn, and use the same language so that we are able to emulate some of these concepts programmatically in our algorithms and techniques when we try to extract insights from textual data.

Language Acquisition and Cognitive Learning

Language acquisition is defined as the process by which human beings utilize their cognitive abilities, knowledge, and experience to understand language based on hearing and perception and start using it in terms of words, phrases, and sentences to communicate with other human beings. In simple terms, the ability of acquiring and producing languages is language acquisition.

The history of language acquisition dates back centuries. Philosophers and scholars have tried to reason and understand the origins of language acquisition and came up with several theories, such as language being a god-gifted ability that is passed down from generation to generation. Plato indicated that a form of word-meaning mapping would have been responsible in language acquisition. Modern theories have been proposed by various scholars and philosophers, and some of the popular ones, most notably B.S. Skinner, indicated that knowledge, learning, and use of language were more of a behavioral consequent. Human beings, or to be more specific, children, when using specific words or symbols of any language, experience language based on certain stimuli which get reinforced in their memory thanks to consequent reactions to their usage repeatedly. This theory is based on *operant* or *instrumentation conditioning*, which is a type of conditional learning where the strength of a particular behavior or action is modified based on its consequences such as reward or punishment, and these consequent stimuli help in reinforcing or controlling behavior and learning. An example would be that children would learn that a specific combination of sounds made up a word from repeated usage of it by their parents or by being rewarded by appreciation when they speak it correctly or by being corrected when they make a mistake while speaking the same. This repeated conditioning would end up reinforcing the actual meaning and understanding of the word in a child's memory for the future. To sum it up, children try to learn and use language mostly behaviorally by imitating and hearing from adults.

However, this behavioral theory was challenged by renowned linguist Noam Chomsky, who proclaimed that it would be impossible for children to learn language just by imitating everything from adults. This hypothesis does stand valid in the following examples. Although words like *go* and *give* are valid, children often end up using an invalid form of the word, like *goed* or *gived* instead of *went* or *gave* in the past tense. It is assured that their parents didn't utter these words in front of them, so it would be impossible to pick these up based on the previous theory of Skinner. Consequently, Chomsky proposed that children must not only be imitating words they hear but also extracting patterns, syntax, and rules from the same language constructs, which is separate from just utilizing generic cognitive abilities based on behavior.

Considering Chomsky's view, cognitive abilities along with language-specific knowledge and abilities like syntax, semantics, concepts of parts of speech, and grammar together form what he termed a *language acquisition device* that enabled humans to have the ability of *language acquisition*. Besides cognitive abilities, what is unique and important in language learning is the syntax of the language itself, which can be emphasized in his famous sentence *Colorless green ideas sleep furiously*. If you observe the sentence and repeat it many times, it does not make sense. *Colorless* cannot be associated with green, and neither can ideas be associated with green, nor can they sleep furiously. However, the sentence has a grammatically correct syntax. This is precisely what Chomsky tried to explain—that syntax and grammar depict information that is independent from the meaning and semantics of words. Hence, he proposed that the learning and identifying of language syntax is a separate human capability compared to other cognitive abilities. This proposed hypothesis is also known as the *autonomy of syntax*. These theories are still widely debated among scholars and linguists, but it is useful to explore how the human mind tends to acquire and learn language. We will now look at the typical patterns in which language is generally used.

Language Usage

The previous section talked about speech acts and how the direction of fit model is used for relating words and symbols to reality. In this section we will cover some concepts related to speech acts that highlight different ways in which language is used in communication.

There are three main categories of speech acts: *locutionary*, *illocutionary*, and *perlocutionary* acts. *Locutionary* acts are mainly concerned with the actual delivery of the sentence when communicated from one human being to another by speaking it. *Illocutionary* acts focus further on the actual semantics and significance of the sentence which was communicated. *Perlocutionary* acts refer to the actual effect the communication had on its receiver, which is more psychological or behavioral.

A simple example would be the phrase *Get me the book from the table* spoken by a father to his child. The phrase when spoken by the father forms the locutionary act. This significance of this sentence is a directive, which directs the child to get the book from the table and forms an illocutionary act. The action the child takes after hearing this, that is, if he brings the book from the table to his father, forms the perlocutionary act.

The illocutionary act was a directive in this case. According to the philosopher John Searle, there are a total of five different classes of illocutionary speech acts, as follows:

- *Assertives* are speech acts that communicate how things are already existent in the world. They are spoken by the sender when he tries to assert a proposition that could be true or false in the real world. These assertions could be statements or declarations. A simple example would be *The Earth revolves round the Sun*. These messages represent the word-to-world direction of fit discussed earlier.
- *Directives* are speech acts that the sender communicates to the receiver asking or directing them to do something. This represents a voluntary act which the receiver might do in the future after receiving a directive from the sender. Directives can either be complied with or not complied with, since they are voluntary. These directives could be simple requests or even orders or commands. An example directive would be *Get me the book from the table*, discussed earlier when we talked about types of speech acts.
- *Commisives* are speech acts that commit the sender or speaker who utters them to some future voluntary act or action. Acts like promises, oaths, pledges, and vows represent commisives, and the direction of fit could be either way. An example commisive would be *I promise to be there tomorrow for the ceremony*.
- *Expressives* reveal a speaker or sender's disposition and outlook toward a particular proposition communicated through the message. These can be various forms of expression or emotion, such as congratulatory, sarcastic, and so on. An example expressive would be *Congratulations on graduating top of the class*.

- *Declarations* are powerful speech acts that have the capability to change the reality based on the declared proposition in the message communicated by the speaker\sender. The usual direction of fit is world-to-word, but it can go the other way also. An example declaration would be *I hereby declare him to be guilty of all charges*.

These speech acts are the primary ways in which language is used and communicated among human beings, and without even realizing it, you end up using hundreds of them on any given day. We will now look at linguistics and some of the main areas of research associated with it.

Linguistics

We have touched on what natural language means, how language is learned and used, and the origins of language acquisition. These kinds of things are formally researched and studied in linguistics by researchers and scholars called *linguists*. Formally, *linguistics* is defined as the scientific study of language, including form and syntax of language, meaning, and semantics depicted by the usage of language and context of use. The origins of linguistics can be dated back to the 4th century BCE, when Indian scholar and linguist Panini formalized the Sanskrit language description. The term *linguistics* was first defined to indicate the scientific study of languages in 1847, approximately before which the term *philology* was used to indicate the same. Although a detailed exploration of linguistics is not needed for text analytics, it is useful to know the different areas of linguistics because some of them are used extensively in natural language processing and text analytics algorithms. The main distinctive areas of study under linguistics are as follows:

- *Phonetics*: This is the study of the acoustic properties of sounds produced by the human vocal tract during speech. It includes studying the properties of sounds as well as how they are created and by human beings. The smallest individual unit of human speech in a specific language is called a *phoneme.* A more generic term across languages for this unit of speech is *phone*.

- *Phonology*: This is the study of sound patterns as interpreted in the human mind and used for distinguishing between different phonemes to find out which ones are significant. The structure, combination, and interpretations of phonemes are studied in detail, usually by taking into account a specific language at a time. The English language consists of around 45 phonemes. Phonology usually extends beyond just studying phonemes and includes things like accents, tone, and syllable structures.

- *Syntax*: This is usually the study of sentences, phrases, words, and their structures. It includes researching how words are combined together grammatically to form phrases and sentences. Syntactic order of words used in a phrase or a sentence matter because the order can change the meaning entirely.

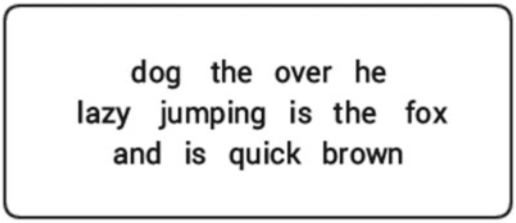
- *Semantics*: This involves the study of meaning in language and can be further subdivided into lexical and compositional semantics.
    - *Lexical semantics*: The study of the meanings of words and symbols using morphology and syntax.
    - *Compositional semantics*: Studying relationships among words and combination of words and understanding the meanings of phrases and sentences and how they are related.
- *Morphology*: A *morpheme* is the smallest unit of language that has distinctive meaning. This includes things like words, prefixes, suffixes, and so on which have their own distinct meanings. Morphology is the study of the structure and meaning of these distinctive units or morphemes in a language. Specific rules and syntaxes usually govern the way morphemes can combine together.
- *Lexicon*: This is the study of properties of words and phrases used in a language and how they build the vocabulary of the language. These include what kinds of sounds are associated with meanings for words, the parts of speech words belong to, and their morphological forms.
- *Pragmatics*: This is the study of how both linguistic and non-linguistic factors like context and scenario might affect the meaning of an expression of a message or an utterance. This includes trying to infer whether there are any hidden or indirect meanings in the communication.
- *Discourse analysis*: This analyzes language and exchange of information in the form of sentences across conversations among human beings. These conversations could be spoken, written, or even signed.
- *Stylistics*: This is the study of language with a focus on the style of writing, including the tone, accent, dialogue, grammar, and type of voice.
- *Semiotics*: This is the study of signs, symbols, and sign processes and how they communicate meaning. Things like analogy, metaphors, and symbolism are covered in this area.

Although these are the main areas of study and research, linguistics is an enormous field with a much bigger scope than what is mentioned here. However, things like language syntax and semantics are some of the most important concepts that often form the foundations to natural language processing. The following section looks at them more closely.

Language Syntax and Structure

We already know what language, syntax, and structure indicate. Syntax and structure usually go hand in hand, where a set of specific rules, conventions, and principles usually govern the way words are combined into phrases, phrases get combines into clauses, and clauses get combined into sentences. We will be talking specifically about the English language syntax and structure in this section because in this book we will be dealing with textual data that belongs to the English language. But a lot of these concepts can be extended to other languages too. Knowledge about the structure and syntax of language is helpful in many areas like text processing, annotation, and parsing for further operations such as text classification or summarization.

In English, words usually combine together to form other constituent units. These constituents include words, phrases, clauses, and sentences. All these constituents exist together in any message and are related to each other in a hierarchical structure. Moreover, a sentence is a structured format of representing a collection of words provided they follow certain syntactic rules like grammar. Look at the bunch of words represented in Figure 1-3.



dog the over he
lazy jumping is the fox
and is quick brown

Figure 1-3. *A collection of words without any relation or structure*

From the collection of words in Figure 1-3, it is very difficult to ascertain what it might be trying to convey or mean. Indeed, languages are not just comprised of groups of unstructured words. Sentences with proper syntax not only help us give proper structure and relate words together but also help them convey meaning based on the order or position of the words. Considering our previous hierarchy of sentence → clause → phrase → word, we can construct the hierarchical sentence tree in Figure 1-4 using *shallow parsing*, a technique using for finding out the constituents in a sentence.

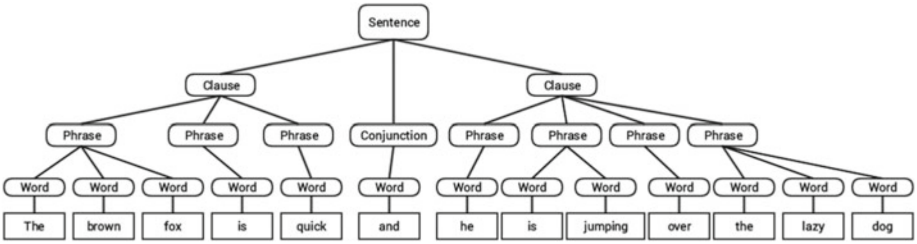


Figure 1-4. *Structured sentence following the hierarchical syntax*

From the hierarchical tree in Figure 1-4, we get the sentence *The brown fox is quick and he is jumping over the lazy dog*. We can see that the leaf nodes of the tree consist of words, which are the smallest unit here, and combinations of words form phrases, which in turn form clauses. Clauses are connected together through various filler terms or words such as conjunctions and form the final sentence. In the next section, we will look at each of these constituents in further detail and understand how to analyze them and find out what the major syntactic categories are.

Words

Words are the smallest units in a language that are independent and have a meaning of their own. Although morphemes are the smallest distinctive units, morphemes are not independent like words, and a word can be comprised of several morphemes. It is useful to annotate and tag words and analyze them into their parts of speech (POS) to see the major syntactic categories. Here, we will cover the main categories and significance of the various POS tags. Later in Chapter 3 we will examining them in further detail and looking at methods of generating POS tags programmatically.

Usually, words can fall into one of the following major categories.

- *N(oun)*: This usually denotes words that depict some object or entity which may be living or nonliving. Some examples would be *fox*, *dog*, *book*, and so on. The POS tag symbol for nouns is *N*.
- *V(erb)*: Verbs are words that are used to describe certain actions, states, or occurrences. There are a wide variety of further subcategories, such as auxiliary, reflexive, and transitive verbs (and many more). Some typical examples of verbs would be *running*, *jumping*, *read*, and *write*. The POS tag symbol for verbs is *V*.
- *Adj(ective)*: Adjectives are words used to describe or qualify other words, typically nouns and noun phrases. The phrase *beautiful flower* has the noun (N) *flower* which is described or qualified using the adjective (ADJ) *beautiful*. The POS tag symbol for adjectives is *ADJ*.
- *Adv(erb)*: Adverbs usually act as modifiers for other words including nouns, adjectives, verbs, or other adverbs. The phrase *very beautiful flower* has the adverb (ADV) *very*, which modifies the adjective (ADJ) *beautiful*, indicating the degree to which the flower is beautiful. The POS tag symbol for adverbs is *ADV*.

Besides these four major categories of parts of speech, there are other categories that occur frequently in the English language. These include pronouns, prepositions, interjections, conjunctions, determiners, and many others. Furthermore, each POS tag like the noun (N) can be further subdivided into categories like singular nouns (NN), singular proper nouns (NNP), and plural nouns (NNS). We will be looking at POS tags in further detail in Chapter 3 when we process and parse textual data and implement POS taggers to annotate text.

Considering our previous example sentence (*The brown fox is quick and he is jumping over the lazy dog*) where we built the hierarchical syntax tree, if we were to annotate it using basic POS tags, it would look like Figure 1-5.

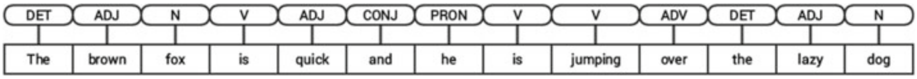


Figure 1-5. *Annotated words with their POS tags*

In Figure 1-5 you may notice a few unfamiliar tags. The tag DET stands for *determiner*, which is used to depict articles like *a*, *an*, *the*, and so on. The tag CONJ indicates *conjunction*, which is usually used to bind together clauses to form sentences. The PRON tag stands for *pronoun*, which represents words that are used to represent or take the place of a noun.

The tags N, V, ADJ and ADV are typical open classes and represent words belonging to an open vocabulary. *Open classes* are word classes that consist of an infinite set of words and commonly accept the addition of new words to the vocabulary which are invented by people. Words are usually added to open classes through processes like *morphological derivation*, invention based on usage, and creating *compound lexemes*. Some popular nouns added fairly recently include *Internet* and *multimedia*. *Closed classes* consist of a closed and finite set of words and do not accept new additions. Pronouns are a closed class.

The following section looks at the next level of the hierarchy: phrases.

Phrases

Words have their own lexical properties like parts of speech, which we saw earlier. Using these words, we can order them in ways that give meaning to the words such that each word belongs to a corresponding phrasal category and one of the words is the main or head word. In the hierarchy tree, groups of words make up *phrases*, which form the third level in the syntax tree. By principle, phrases are assumed to have at least two or more words, considering the pecking order of words ← phrases ← clauses ← sentences. However, a phrase *can* be a single word or a combination of words based on the syntax and position of the phrase in a clause or sentence. For example, the sentence *Dessert was good* has only three words, and each of them rolls up to three phrases. The word *dessert* is a noun as well as a *noun phrase*, *is* depicts a verb as well as a *verb phrase*, and *good* represents an adjective as well as an *adjective phrase* describing the aforementioned dessert.

There are five major categories of phrases:

- *Noun phrase (NP)*: These are phrases where a noun acts as the head word. Noun phrases act as a subject or object to a verb. Usually a noun phrases can be a set of words that can be replaced by a pronoun without rendering the sentence or clause syntactically incorrect. Some examples would be *dessert*, *the lazy dog*, and *the brown fox*.

- *Verb phrase (VP)*: These phrases are lexical units that have a verb acting as the head word. Usually there are two forms of verb phrases. One form has the verb components as well as other entities such as nouns, adjectives, or adverbs as parts of the object. The verb here is known as a *finite verb*. It acts as a single unit in the hierarchy tree and can function as the root in a clause. This form is prominent in *constituency grammars*. The other form is where the finite verb acts as the root of the entire clause and is prominent in *dependency grammars*. Another derivation of this includes verb phrases strictly consisting of verb components including main, auxiliary, infinitive, and participles. The sentence *He has started the engine* can be used to illustrate the two types of verb phrases that can be formed. They would be *has started the engine* and *has started*, based on the two forms just discussed.

- *Adjective phrase (ADJP)*: These are phrases with an adjective as the head word. Their main role is to describe or qualify nouns and pronouns in a sentence, and they will be either placed before or after the noun or pronoun. The sentence *The cat is too quick* has an adjective phrase, *too quick*, qualifying *cat*, which is a noun phrase.

- *Adverb phrase (ADVP)*: These phrases act like adverbs since the adverb acts as the head word in the phrase. Adverb phrases are used as modifiers for nouns, verbs, or adverbs themselves by providing further details that describe or qualify them. In the sentence *The train should be at the station pretty soon*, the adjective phrase *pretty soon* describes when the train would be arriving.

- *Prepositional phrase (PP)*: These phrases usually contain a preposition as the head word and other lexical components like nouns, pronouns, and so on. It acts like an adjective or adverb describing other words or phrases. The phrase *going up the stairs* contains a prepositional phrase *up*, describing the direction of the stairs.

These five major syntactic categories of phrases can be generated from words using several rules, some of which have been discussed, like utilizing syntax and grammars of different types. We will be exploring some of the popular grammars in a later section. *Shallow parsing* is a popular natural language processing technique to extract these constituents, including POS tags as well as phrases from a sentence. For our sentence *The brown fox is quick and he is jumping over the lazy dog*, we have obtained seven phrases from shallow parsing, as shown in Figure 1-6.

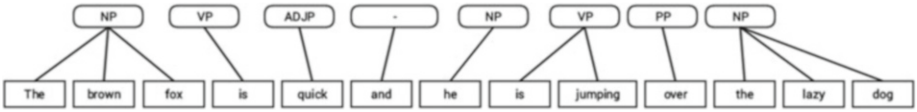


Figure 1-6. *Annotated phrases with their tags*

The phrase tags fall into the categories discussed earlier, although the word *and* is a conjunction and is usually used to combine clauses together. In the next section, we will be looking at clauses, their main categories, and some conventions and syntactic rules for extracting clauses from sentences.

Clauses

By nature, clauses can act as independent sentences, or several clauses can be combined together to form a sentence. A *clause* is a group of words with some relation between them that usually contains a subject and a predicate. Sometimes the subject is not present, and the predicate usually has a verb phrase or a verb with an object. By default you can classify clauses into two distinct categories: the *main clause* and the *subordinate clause*. The main clause is also known as an *independent* clause because it can form a sentence by itself and act as both sentence and clause. The subordinate or *dependent* clause cannot exist just by itself and depends on the main clause for its meaning. They are usually joined with other clauses using dependent words such as subordinating conjunctions.

With regard to syntactic properties of language, clauses can be subdivided into several categories based on syntax:

- *Declarative*: These clauses usually occur quite frequently and denote statements that do not have any specific tone associated with them. These are just standard statements, which are declared with a neutral tone and which could be factual or non-factual. An example would be *Grass is green*.
- *Imperative*: These clauses are usually in the form of a request, command, rule, or advice. The tone in this case would be a person issuing an order to one or more people to carry out an order, request, or instruction. An example would be *Please do not talk in class*.
- *Relative*: The simplest interpretation of *relative* clauses is that they are subordinate clauses and hence dependent on another part of the sentence that usually contains a word, phrase, or even a clause. This element usually acts as the antecedent to one of the words from the relative clause and relates to it. A simple example would be *John just mentioned that he wanted a soda*, having the antecedent proper noun *John*, which was referred to in the relative clause *he wanted a soda*.
- *Interrogative*: These clauses usually are in the form of questions. The type of these questions can be either affirmative or negative. Some examples would be *Did you get my mail?* and *Didn't you go to school?*
- *Exclamative*: These clauses are used to express shock, surprise, or even compliments. These expressions fall under *exclamations*, and these clauses often end with an exclamation mark. An example would be *What an amazing race!*