

Cambridge Textbooks in Linguistics

The Lexicon

James Pustejovsky
Olga Batiukova

The Lexicon

What is the lexicon, what does it contain, and how is it structured? What principles determine the functioning of the lexicon as a component of natural language grammar? What role does lexical information play in linguistic theory?

This comprehensive yet accessible introductory textbook aims to answer these three questions, and explores the relation of the lexicon to grammar as a whole. It offers a critical overview of major theoretical frameworks, and puts forward a unified treatment of lexical structure and design. The text can be used for introductory and advanced courses, and for courses that touch upon different aspects of the lexicon, such as lexical semantics, lexicography, syntax, general linguistics, computational lexicology, and ontology design.

One of the authors' key objectives is to provide students with a set of tools which will enable them to work with lexical data for all kinds of purposes, and they have included an abundance of exercises and in-class activities for this purpose, which also ensure that students are actively engaged with the content and effectively acquire the necessary knowledge and skills they need.

JAMES PUSTEJOVSKY is the TJX Feldberg Chair in Computer Science at Brandeis University. He has published numerous books on Generative Lexicon theory, lexical semantics, temporal and spatial semantics, and linguistic annotation.

OLGA BATIUKOVA is Associate Professor in the Department of Spanish Philology at the Autonomous University of Madrid. She has published extensively on lexical theory and its impact on lexicography, and the syntax–lexicon interface and verbal aspect.

CAMBRIDGE TEXTBOOKS IN LINGUISTICS

General Editors: P. AUSTIN, J. BRESNAN, B. COMRIE, S. CRAIN,
W. DRESSLER, C. EWEN, R. LASS, D. LIGHTFOOT, K. RICE, I. ROBERTS,
S. ROMAINÉ, N. V. SMITH.

The Lexicon

In This Series:

- R. CANN *Formal Semantics*
 J. LAVER *Principles of Phonetics*
 F. R. PALMER *Grammatical Roles and Relations*
 M. A. JONES *Foundations of French Syntax*
 A. RADFORD *Syntactic Theory and the Structure of English: A Minimalist Approach*
 R. D. VAN VALIN JR AND R. J. LAPOLLA *Syntax: Structure, Meaning and Function*
 A. DURANTI *Linguistic Anthropology*
 A. CRUTTENDEN *Intonation Second edition*
 J. K. CHAMBERS AND P. TRUDGILL *Dialectology Second edition*
 C. LYONS *Definiteness*
 R. KAGER *Optimality Theory*
 J. A. HOLM *An Introduction to Pidgins and Creoles*
 G. G. CORBETT *Number*
 C. J. EWEN AND H. VAN DER HULST *The Phonological Structure of Words*
 F. R. PALMER *Mood and Modality Second edition*
 B. J. BLAKE *Case Second edition*
 E. GUSSMAN *Phonology: Analysis and Theory*
 M. YIP *Tone*
 W. CROFT *Typology and Universals Second edition*
 F. COULMAS *Writing Systems: An Introduction to Their Linguistic Analysis*
 P. J. HOPPER AND E. C. TRAUGOTT *Grammaticalization Second edition*
 L. WHITE *Second Language Acquisition and Universal Grammar*
 I. PLAG *Word-Formation in English*
 W. CROFT AND A. CRUSE *Cognitive Linguistics*
 A. SIEWIERSKA *Person*
 A. RADFORD *Minimalist Syntax: Exploring the Structure of English*
 D. BÜRING *Binding Theory*
 M. BUTT *Theories of Case*
 N. HORNSTEIN, J. NUÑES AND K. GROHMANN *Understanding Minimalism*
 B. C. LUST *Child Language: Acquisition and Growth*
 G. G. CORBETT *Agreement*
 J. C. L. INGRAM *Neurolinguistics: An Introduction to Spoken Language Processing and Its Disorders*
 J. CLACKSON *Indo-European Linguistics: An Introduction*
 M. ARIEL *Pragmatics and Grammar*
 R. CANN, R. KEMPSON AND E. GREGOROMICHELAKI *Semantics: An Introduction to Meaning in Language*
 Y. MATRAS *Language Contact*
 D. BIBER AND S. CONRAD *Register, Genre and Style*
 L. JEFFRIES AND D. MCINTYRE *Stylistics*
 R. HUDSON *An Introduction to Word Grammar*
 M. L. MURPHY *Lexical Meaning*
 J. M. MEISEL *First and Second Language Acquisition*
 T. MCENERY AND A. HARDIE *Corpus Linguistics: Method, Language and Practice*
 J. SAKEL AND D. L. EVERETT *Linguistic Fieldwork: A Student Guide*
 A. SPENCER AND A. LUÍS *Clitics: An Introduction*
 G. CORBETT *Features*
 A. MCMAHON AND R. MCMAHON *Evolutionary Linguistics*
 B. CLARK *Relevance Theory*
 B. LONG PENG *Analyzing Sound Patterns*
 B. DANCYGIER AND E. SWEETSER *Figurative Language*
 J. BYBEE *Language Change*
 S. G. THOMASON *Endangered Languages: An Introduction*
 A. RADFORD *Analysing English Sentences Second edition*
 R. CLIFT *Conversation Analysis*
 R. LEVINE *Syntactic Analysis*
 I. PLAG *Word-Formation in English Second edition*
 Z. G. SZABÓ AND R. H. THOMASON *Philosophy of Language*
 J. PUSTEJOVSKY AND O. BATIUKOVA *The Lexicon*

Earlier issues not listed are also available.

The Lexicon

JAMES PUSTEJOVSKY

Brandeis University, Massachusetts

OLGA BATIUKOVA

Universidad Autónoma de Madrid



CAMBRIDGE
UNIVERSITY PRESS

University Printing House, Cambridge CB2 8BS, United Kingdom

One Liberty Plaza, 20th Floor, New York, NY 10006, USA

477 Williamstown Road, Port Melbourne, VIC 3207, Australia

314–321, 3rd Floor, Plot 3, Splendor Forum, Jasola District Centre, New Delhi – 110025, India

79 Anson Road, #06–04/06, Singapore 079906

Cambridge University Press is part of the University of Cambridge.

It furthers the University's mission by disseminating knowledge in the pursuit of education, learning, and research at the highest international levels of excellence.

www.cambridge.org

Information on this title: www.cambridge.org/9780521839327

DOI: [10.1017/9780521839327](https://doi.org/10.1017/9780521839327)

© James Pustejovsky and Olga Batiukova 2019

This publication is in copyright. Subject to statutory exception and to the provisions of relevant collective licensing agreements, no reproduction of any part may take place without the written permission of Cambridge University Press.

First published 2019

Printed in the United Kingdom by TJ International Ltd. Padstow Cornwall

A catalog record for this publication is available from the British Library.

Library of Congress Cataloging-in-Publication Data

Names: Pustejovsky, J. (James) author. | Batiukova, Olga, author.

Title: The lexicon / James Pustejovsky, Brandeis University, Massachusetts; Olga Batiukova, Universidad Autónoma de Madrid.

Description: Cambridge, United Kingdom; New York, NY: Cambridge University Press, 2019. | Series: Cambridge Textbooks in Linguistics | Includes bibliographical references and index.

Identifiers: LCCN 2018038011 | ISBN 9780521839327 (Hardback) | ISBN 9780521547956 (Paperback)

Subjects: LCSH: Lexicology.

Classification: LCC P326 .P87 2019 | DDC 413.028–dc23

LC record available at <https://lccn.loc.gov/2018038011>

ISBN 978-0-521-83932-7 Hardback

ISBN 978-0-521-54795-6 Paperback

Cambridge University Press has no responsibility for the persistence or accuracy of URLs for external or third-party internet websites referred to in this publication and does not guarantee that any content on such websites is, or will remain, accurate or appropriate.

Contents

<i>List of figures</i>	<i>page</i> xii
<i>List of tables</i>	xiii
<i>Preface</i>	xv
<i>Acknowledgments</i>	xix
Part I The Lexicon in Linguistic Theory	1
1 Introduction	3
1.1 Overview	3
1.2 What Is the Lexicon?	3
1.3 What Is a Word?	6
1.4 Lexeme, Word Form, and Grammatical Word	10
1.5 What Is in a Lexical Entry?	10
1.6 The Role of Empirical Data in Lexicon Research	11
1.6.1 Naturally Occurring Data	12
1.6.2 Naturally Elicited Data	13
1.6.3 Experimentally Elicited Data	14
1.6.4 Data in the Linguistic Research Cycle	16
1.7 Goals of Linguistic Theory and the Notion of Grammar	18
1.8 Summary	19
1.9 Further Readings	20
1.10 Exercises	21
2 Lexicon and Syntax	24
2.1 Overview	24
2.2 What Is Syntax?	24
2.3 Syntactically Relevant Lexical Features and Lexically Dependent Syntactic Phenomena	30
2.3.1 Syntactic Category	31
2.3.2 Semantic Features and Semantic Types	37
2.3.3 Countability	38
2.3.4 The Interaction of Lexical Features	45
2.4 Summary	46
2.5 Further Readings	47
2.6 Exercises	47

3	Lexicon in Syntactic Frameworks	49
3.1	Overview	49
3.2	The Lexicon from Aristotle to Structuralism	49
3.3	Generative Grammar	56
3.4	Head-Driven Phrase Structure Grammar (HPSG)	64
3.5	Lexical–Functional Grammar (LFG)	70
3.6	Construction Grammar	74
3.7	Lexicon in Its Own Right: Generative Lexicon	78
3.8	Summary	88
3.9	Further Readings	89
3.10	Exercises	90
4	Lexicon and Semantics	93
4.1	Overview	93
4.2	What Is Semantics?	93
4.2.1	Sentence Meaning and Inference	94
4.2.2	Word Meaning and Lexical Relations	98
4.3	Conceptual Meaning	99
4.3.1	Referential Models of Semantics	100
4.3.2	Representational Theories of Meaning	102
4.3.3	Rich Denotational Models of Semantics	104
4.4	Associative or Connotative Meaning	108
4.5	Literal and Non-Literal Meaning	110
4.6	Summary	112
4.7	Further Readings	113
4.8	Exercises	114
5	Lexicon in Semantic Frameworks	115
5.1	Overview	115
5.2	Formal Semantics	115
5.3	Conceptual Semantics	121
5.4	Cognitive Linguistics	126
5.5	Prototype Theory	131
5.6	Natural Semantic Metalanguage	137
5.7	Summary	140
5.8	Further Readings	140
5.9	Exercises	142
Part II	Lexical Structures	145
6	The Structure of a Lexical Entry	147
6.1	Overview	147
6.2	Linguistic versus Extralinguistic Knowledge in the Lexical Entry	147

6.2.1	Conventional Dictionary Definitions and Their Types	149
6.3	Linguistic Structures	151
6.3.1	Typed Feature Structure	151
6.3.2	Argument Structure and Predication	153
6.3.3	Event Structure	158
6.3.4	Qualia Structure	161
6.4	Conceptual Structures	164
6.4.1	Image Schemas	164
6.4.2	Frames and Scenes	166
6.4.3	Conventionalized Attributes	168
6.5	Properties of Lexical Entries	169
6.5.1	Ambiguity, Polysemy, and Homonymy	169
6.5.1.1	Tests of Sense Distinctness	171
6.5.1.2	Polysemy and Homonymy	173
6.5.1.3	Regular Polysemy	175
6.5.2	Vagueness and Underspecification	177
6.6	Summary	178
6.7	Further Readings	179
6.8	Exercises	180
7	Semantic Typing and Decomposition	182
7.1	Overview	182
7.2	Strategies for Lexical Decomposition	182
7.3	Semantic Types	188
7.4	Qualia Structure As Semantic Types	194
7.4.1	Natural Types and Artifactual Types	194
7.4.2	Complex Types	198
7.5	Summary	200
7.6	Further Readings	201
7.7	Exercises	201
8	Argument Structure	204
8.1	Overview	204
8.2	The Basics of Function–Argument Behavior	204
8.2.1	Formalizing Argument Structure	204
8.2.2	Argument Structure and Lexical Meaning in the Predicate	206
8.2.3	Arguments and Adjuncts	214
8.2.4	Unexpressed Arguments	218
8.3	Argument Structure in Verbs	219
8.3.1	Semantic Roles and Related Notions	220
8.3.2	Verbal Valence	224
8.4	Argument Structure in Nouns	227
8.5	Argument Structure in Adjectives	234

8.6	Argument Structure in Prepositions	236
8.7	Summary	240
8.8	Further Readings	242
8.9	Exercises	243
9	Lexical Aspect, Tense, and Modality	245
9.1	Overview	245
9.2	Events As Arguments of Natural Language Predicates	245
9.3	Internal Temporal Makeup of Events: Event Structure	247
9.3.1	Perspectives on Event Decomposition and Event Structure	248
9.3.2	Event Type as a Property of Phrases and Sentences	256
9.3.3	Event Type Diagnostics	258
9.3.4	Eventive Readings of Nouns and Adjectives	263
9.4	Tense–Aspect Details	268
9.5	Modality	273
9.6	Summary	279
9.7	Further Readings	281
9.8	Exercises	282
Part III	Lexicon As a System	285
10	General Architecture of the Lexicon	287
10.1	Overview	287
10.2	Lexical Architecture	287
10.3	Syntactic Type Hierarchy	289
10.4	Semantic Type Hierarchy	292
10.5	Lexical-Semantic Relations	300
10.6	Morphology and Word Structure	311
10.7	Summary	316
10.8	Further Readings	318
10.9	Exercises	319
11	Compositionality in the Mapping from the Lexicon to Syntax	321
11.1	Overview	321
11.2	The Basics of Compositionality	321
11.3	Apparent Violations of Compositionality	328
11.4	Main Approaches to Compositionality	336
11.5	Non-Compositional Constructions	346
11.6	The Lexicon and Grammar: Outstanding Problems	351
11.7	Summary	355
11.8	Further Readings	356
11.9	Exercises	357

<i>Answers to Selected Exercises</i>	359
<i>Online Resources</i>	378
<i>Glossary</i>	380
<i>References</i>	397
<i>Subject Index</i>	413
<i>Name Index</i>	419

Figures

3.1	Agent nominalization lexical rule	<i>page</i> 68
3.2	Multiple inheritance hierarchy	69
3.3	Ditransitive construction	76
4.1	Concept FLOWER	102
5.1	The Parallel Architecture of the language faculty	122
5.2	Trajector–landmark alignment	128
5.3	Vantage point	128
5.4	Types of entities in Cognitive Linguistics	129
5.5	Naming preferences and basic level categories	135
6.1	UNCONTAINMENT image schema	164
6.2	INTO image schema	166
7.1	Aristotelian ontology	189
7.2	Classic upper ontology	189
7.3	ENTITY subtypes	193
7.4	Multiple inheritance in artifactual types	197
8.1	The Actor–Undergoer hierarchy	224
9.1	Geometric analysis of event types	253
10.1	Syntactic type hierarchy	290
10.2	Individuation tree	294
10.3	Ontological types with individuation hierarchy	295
10.4	Semantic type hierarchy (fragment)	296
10.5	Ontological types with qualia specification and individuation	296
10.6	Semantic subtyping of ENTITY	299
10.7	-er nominalization rule	314
10.8	-al adjectivization rule	315
11.1	COUNT–MASS alternation rule	337
11.2	Causative–inchoative alternation rule	338

Tables

2.1	Differences between syntax and lexicon	page 30
2.2	Grammatical features	36
3.1	Distinctive features in phonology	55
3.2	Lexical field of cooking words, componential analysis	55
3.3	Types of constructions in CG	75
3.4	Lexical entry of <i>school</i>	88
3.5	Lexicon–syntax relationship in syntactic theories	89
4.1	Entailment	95
4.2	Logical conjunction	96
4.3	Logical relations through entailment	97
4.4	P presupposes Q	97
4.5	Extension of different types of linguistic expression	100
4.6	Concepts of different types of linguistic expression	104
4.7	Denotational semantic meanings of different types of linguistic expression	106
5.1	Full-blown and defective lexical items in the Parallel Architecture	123
5.2	Semantic definition of syntactic categories in Radical Construction Grammar	130
5.3	Coding of propositional act constructions in Radical Construction Grammar	130
5.4	Goodness-of-example ratings of concrete categories	133
5.5	Family resemblances in the category FRUIT	134
5.6	NSM semantic primitives (English exponents)	138
6.1	Core event types	161
6.2	THEFT frame, FrameNet	167
7.1	Semantic types	193
8.1	Illustration of the θ -criterion	221
10.1	Types of part-of relations	308
10.2	Conjugation of the German verb <i>öffnen</i> ('open')	311
11.1	Syntactic categories and their semantic types (part A)	326
11.2	Syntactic categories and their semantic types (part B)	327
11.3	Types of MWEs	348
11.4	Collocation examples in English, Spanish, and Russian	349

Preface

This book is the first textbook on the lexicon in the Cambridge Textbooks in Linguistics series. It is intended as an in-depth introduction to the theory and structure of the lexicon in linguistics and linguistic theory. The text offers a comprehensive treatment of lexical structure and design, the relation of the lexicon to grammar as a whole, and to methods of interpretation driven by the lexicon.

The book examines the structure of lexical information within the context of both traditional and newly emerging views on the role of the lexicon in linguistic theory. No single syntactic framework in linguistics is assumed. Rather, the linguistic phenomena are observational, while the proposed mechanisms and explanations to account for these data can be seen as generally theory-neutral. Nevertheless, a strong theoretical claim is made for specific, generatively oriented, structural principles, characterizing the behavior of lexical items in relation to grammar.

Audience

This book is written for anyone interested in knowing more about the lexicon and its general design, and in using lexical data for specific tasks, both theoretical and applied. In other words, it aims at answering the following questions: (1) what is the lexicon, what does it contain, and how is it structured? (2) what principles determine the functioning of the lexicon as a component of natural language grammar? and (3) what role does lexical information play in linguistic theory (e.g., how is lexical information exploited by other components of grammar)? The approach taken is strongly data-oriented and has as one of its essential goals to provide the reader with a set of tools enabling work with lexical data for all kinds of purposes.

Having a linguistics background is not required, although a basic (introductory course level) knowledge of grammar will make the formal details and representations easier to follow. For this purpose, we suggest *A Student's Introduction to English Grammar* (Huddleston and Pullum 2005) and other references listed in Section 1.9 of Chapter 1. Two more textbooks we highly recommend for acquiring additional background on different aspects of lexical theory are *The Lexicon: An Introduction* (Ježek 2016), which covers a wide range of topics related to the

internal structure of the lexicon and its usage, and *Lexical Meaning* (Murphy 2010), which deals with word meaning and different approaches to its study.

This textbook can be used for both introductory and advanced courses in the study of the lexicon and for courses that touch upon different aspects of the lexicon, such as: lexical semantics, lexicography, syntax, general linguistics, computational lexicology, ontology design, etc. The book is structured in such a way that its contents are easily customizable to a specific course level, set of objectives and requirements, and instructional strategies.

Features

The overall structure of the book is uniform, where each chapter has an *Overview* and a *Summary* section, which are meant to help the student by adding context and connecting the various strands of the topic. The former is intended to introduce the content covered in the chapter and to contextualize it with respect to previously presented material. The latter sums up succinctly the main points addressed in the chapter, thus allowing the reader to reflect on what has been learnt. The *Glossary* may be used as review material, too, as it contains definitions of the key concepts introduced in the book.

Extensive problem sets offered in the *Exercises* section at the end of each chapter, as well as questions labeled as DISCUSS, which are integrated into the exposition of different topics, ensure that the students are actively engaged with the essential content and effectively acquire the necessary skills and knowledge of the topic. The discussion points are meant to ensure that the students approach the topics presented critically and open-mindedly, and are better suited for in-class activities and group discussion. The exercises may be preferably used for individual assignments and their ultimate goal is to provide the necessary feedback on the learning process results. Answers are provided in *Answers to Selected Exercises*.

Two more features are framed as boxes: CLUE/COMMENT and WARNING. The CLUE/COMMENT boxes (indicated by an image of a magnifying glass) contain additional comments (definitions, historical remarks, data from languages other than English, etc.). The WARNING boxes (indicated by an image of an exclamation mark) are meant to prevent common terminological misunderstandings and introduce important exceptions to the generalizations presented in the text.

The *Further Readings* sections provide lists of references on the topics covered in each chapter. *Primary references* are more appropriate for advanced students and as additional sources for the instructor, while *secondary references* are helpful for students who need to strengthen their background or look for a more detailed explanation of the contents included in the corresponding chapters. Additional resources available online are provided in *Online Resources*, which is structured in four sections: *General Links*, *Lexical Databases and Lexicons*, *Modern Theoretical Frameworks*, and *Corpora and Corpus Query Systems*.

The following brief cross-reference guide may be useful if the instructor plans on covering specific aspects of the lexicon in class:

- *word* and other key concepts in lexical study: Chapter 1
- lexical semantics: Chapters 4–7
- lexicon–syntax relations: Chapters 2, 3, 8, 11
- lexicography: Chapters 1, 6
- word meaning: Chapters 4–7
- lexical-semantic relations: Chapters 4, 10
- lexical ontology and type system: Chapter 10
- lexical morphology: Section 10.6
- syntactically and semantically oriented approaches to the lexicon: Chapters 3, 5
- lexical structures: Chapters 6–9
- verbal semantics and syntax: Sections 2.2, 2.3, 6.3.2, 6.3.3; Chapters 8, 9
- nominal semantics and syntax: Sections 2.3, 6.3.4; Chapter 7
- lexical data and empirical methods in lexicon research: Section 1.6

Organization

A general “bottom-up” pedagogic approach has been followed in organizing the book, which we believe allows one to integrate the new ideas and topics with the previous ones as the book proceeds.

Part I (*The Lexicon in Linguistic Theory*) contains three introductory chapters on the key concepts in lexical study (Chapter 1), the relation between lexicon and syntax (Chapter 2), and the relation between lexicon and semantics (Chapter 4). Chapters 3 and 5 show how these concepts have been treated in different theoretical frameworks.

The component pieces of the lexical entry are dealt with in Part II (*Lexical Structures*): they are presented at a general level in Chapter 6, and are further detailed and applied to specific syntactic categories in Chapters 7, 8, and 9.

The various threads introduced in the above chapters come together in Part III (*Lexicon As a System*), in two chapters dealing with higher levels of lexical and syntactic organization: the general architecture of the lexicon (Chapter 10) and the mapping from the lexicon to syntax and its effects on compositional interpretation of complex linguistic expressions (Chapter 11).

Conventions

The following typographical conventions are used in this book:

Bold face

Indicates new terms, followed by definitions when necessary.

Italics

Indicate previously introduced technical terms and signal a metalinguistic use of an expression (e.g., *book* used to refer to the word *book* in English rather than to the physical object this word represents).

SMALL CAPS

Small capital letters signal major conceptual and linguistics categories, as well as semantic types (e.g., EVENT, LOCATION, HUMAN, etc.).

*, #, ?

When preceding a linguistic expression, these symbols signal different acceptability judgments (i.e., whether a native speaker would judge such an expression as “good” or acceptable, or not):

* marks *grammatically* unacceptable (ungrammatical) sequences

marks *semantically* and *pragmatically* odd expressions

?, ?? marks expressions which are not as bad as those following a star (*), but which still are perceived as more or less odd or inadequate syntactically, semantically, and sometimes pragmatically.



This icon signifies a general note, an additional comment, or a clue.



This icon indicates a warning or caution.

Acknowledgments

We are very grateful to everyone at Cambridge University Press who helped us see through this project: in the first place, Andrew Winnard, and also Lisa Pinto, Rosemary Crawley, Melissa Shivers, and David Bonamy. Many friends and colleagues from both sides of the Atlantic gave us their feedback on different parts of the manuscript or at some point discussed with us the issues it deals with: Mark Verhagen, Elena de Miguel, Annie Zaenen, Christiane Fellbaum, Lauri Karttunen, Nicholas Asher, Pier Marco Bertinetto, Elisabetta Ježek, Nikhil Krishnaswamy, Jan Hajic, Martha Palmer, and Kiyong Lee. We would also like to thank the anonymous Cambridge University Press reviewers for their valuable input. We are also in debt to our students, past and present, who we always kept in mind while writing this textbook.

James adds: I would very much like to thank my family, Sophie and Zachary, for putting up with this project for so many years. I would also like to thank Karen, for her patience and support over the past two years while we finished this book. And of course infinite thanks to Olga, for ensuring that the book would be finished at all!

Olga adds: I would like to thank my husband, Iván, for his support, encouragement, and patience in spite of all the time and energy that my work takes away from him, and also Noa for having let me finish the book and for making me the happiest person on earth with her smile. Lots of love and gratitude to my family and especially my parents Viacheslav and Valentina, who always believed in me and helped me believe in myself. And many thanks to James for having offered me the chance to undertake this exciting project with him and for everything I learnt from him.

Part I

The Lexicon in Linguistic Theory

1 Introduction

1.1 Overview

In this introductory chapter we outline the basic concepts covered and terms used in this book. We first introduce the notion of *mental lexicon* and explain how it is different from the word listings familiar from most dictionaries. The next issue we address is the definition of *word* as a basic unit of language: we present the different criteria of wordhood, and lay out a preliminary approach to the notion of *lexical entry* and its structure (dealt with in more detail in Chapter 6). In order to explain how different kinds of lexical data should be treated and how they can be used to validate theoretical claims, we offer a general typology of lexical data coupled with the research methodology usually adopted in each case. Finally, we present a general vision of the concept of *natural language grammar* and establish the role of the lexicon as one of its core modules.

1.2 What Is the Lexicon?

Loosely speaking, the **lexicon** is a collection of words capturing the knowledge that speakers and hearers have about basic lexical expressions in a language. There is no question that our repertoire of words (what we will call our *mental lexicon*) is large, containing up to 250,000 lexical entries. This is based on fairly conservative estimates of speaker competence with *active* vocabulary (do I know how to use it?) and *passive* vocabulary (do I understand it?). For example, an average speaker's lexicon might contain at least 5,000 different verbs, 20,000 distinct nouns, and over 5,000 adjectives. Combine this with an additional 20,000 compound forms and at least 200,000 distinct proper names, and the lexicon grows fairly large indeed. This certainly indicates that the organization of our mental lexicon has the capacity for storing large amounts of information.

Furthermore, words are retrieved from the mental lexicon surprisingly fast: when reading, it takes us less than half a second to decide whether the sequence of letters we see is a real word (e.g., *parrot*) or a non-word (e.g., *varrot*), a problem known as a *lexical decision task* in psycholinguistics. This would be impossible if words were just randomly heaped up in our mind. Rather, this suggests that the lexicon is a highly organized and very complex system.

The notion of a lexicon as a reference list of words is familiar to anyone who has used a dictionary or studied a foreign language. However, there are a number of important differences between the mental lexicon and conventional dictionaries, designed by people for specific purposes. In the words of Hanks (2000), “checklist [numbered list] theories in their current form are at best superficial and at worst misleading. If word meanings do exist, they do not exist as a checklist.” Hanks believes that the very structure of dictionary definitions has given people a false impression of how language is used, and certainly how it is organized. Most book dictionaries emulate a *sense-enumerative lexicon* strategy, which involves essentially lists of words, wherein each lexical entry represents a different word, and each lexical entry is structured as a list of senses. *Homonyms* (words having the same sound and spelling but different, unrelated meanings, cf. Section 6.5.1) are treated as different words (i.e., different lexical entries), and logically related meanings of the same acoustic and written form are treated as different senses of the same *polysemic* word within the same lexical entry. For example, the following entries from the Merriam-Webster Online Dictionary illustrate how both cases (polysemy and homonymy) are dealt with in conventional dictionaries:

(1) ¹**race** *noun*

1. the act of running
2. a strong or rapid current of water flowing through a narrow channel
- ...

²**race** *verb*

1. to compete in a race
2. to go, move, or function at top speed or out of control
- ...

³**race** *noun*

1. a breeding stock of animals
2. a family, tribe, people, or nation belonging to the same stock
- ...

As we can see, there are three lexical entries corresponding to *race*, two of which are nominal and one verbal. The ‘act of running’ sense and the ‘current of water’ sense are treated as related, they are included in the same lexical entry, and so are the different submeanings included in the other two entries. The ‘act of running / water current’ meaning cluster is treated as unrelated to the other nominal sense (referring to a group of common descent), or to the verbal meanings. Is this division justified? Most speakers would probably agree that the first and third lexical entries are not logically related. But what about the first and second entries? Can we state in good faith that the act of running and the running competition are totally unrelated? Hardly: both are obviously related, and the noun *race* is even used in the definition of the verb *race*.

The lexical entries and word senses in such listed representations are *isolated* and *compartmentalized*: there is no straightforward way to establish a connection even between related word senses in the same lexical entry, let alone different

lexical entries. To a large extent, this is due to the fact that the lexical meanings are presented as *atomic*: even though in each case there is a specification of the more general class to which the defined entity, property, or event belong ('act', 'current', 'stock' in (1)), there is no systematic indication of how these classes and, consequently, the defined terms are arranged and related among themselves.

A further indication of why this is not how the mental lexicon is organized is the order of lexical entries. Alphabetical order is convenient for human search and use. However, it does not seem to reflect the characteristics or properties of everyday language use, such as frequency and semantic and morphological relatedness. For example, the proximity of one word to another in the alphabetical order seems to have little to do with how they are related to each other: *abdicate* and *accolade*, or *table* and *taboo*, are relatively close in a dictionary, but this is only because they share the first letter or two, and not because they are related semantically or syntactically (cf. Section 1.6).

Another important property of the mental lexicon is that we naturally assign words to categories and larger conceptual classes. While such groupings (referred to as *hypernyms*, see Section 10.5) are typically captured in print dictionaries, humans can easily distinguish instances of such categories – that is, typical representatives of a class – from those which are not as typical (cf. Section 5.5). Thus, we know that an *apartment* or a *house* are more typical kinds of *housing* than, say, a *tent* or a *barrack*. This has proven to be an important distinction, giving rise to *typicality effects* in experimental psycholinguistic studies. However, the closeness of the defined term to the prototype is not accounted for in conventional dictionaries: both *apartment* and *barrack* are defined as *housing*.

Typicality Effects in Dictionaries

Of course, there are exceptions to this characterization even among conventional dictionaries, for example the *Diccionario del español usual en México* (Lara Ramos 1996). Anna Wierzbicka also advocates enriching the lexicographical entries with the information related to stereotypes and other socially and culturally relevant information in her natural semantic metalanguage proposal (Wierzbicka 1985a, see Section 5.6).

The list of entries and senses registered in the dictionaries (even very large ones) is finite and thus incomplete for obvious reasons: only the *actual* uses of words in a language will ever appear in a dictionary, usually after they have been circulating for some time. The *potential* uses of a word (sense extensions, contextual modulations, and so on) are not reflected in a dictionary, or even predictable, although we continuously witness the spontaneous creation of novel senses of existing words or even new words, which is an important source of linguistic creativity.

Finally, the information on how words function in context is very limited in book dictionaries, and it does not allow relating the lexical features of the

described item to their syntactic projection. The only syntactically relevant feature provided in the example (1) is the syntactic category (verb or noun). It does not explain, for example, why the noun *race* can be the direct object of the verb *witness* or can be modified by the adjective *fast* (unlike many other nouns, such as *table*: we cannot *witness a table* and there is no such thing as *fast table*).

From our discussion above, it should be clear that conventional dictionaries are not an accurate reflection of the mental lexicon: print dictionaries provide an incomplete, static view of our knowledge of words and how they are used in language. By contrast, it appears that the mental lexicon is flexible, extensible, and much more richly structured.

1.3 What Is a Word?

The notion of word is deeply rooted in the Greco-Latin grammatical tradition. It is assumed to be a basic unit of language, which demarcates the border between morphology and syntax: morphology deals with the internal composition of words and syntax deals with the combination of words. Also, it seems easy to grasp intuitively: we know a word when we see one, so we think. When presented with the following list, a native speaker of English would: surely identify (2a) as a word; reject (2b) because she has never heard it before and because it doesn't look like an English word; claim that (2c) is an expression composed of four words, as is (2d) (probably with a little hesitation about the status of *a*, to which we will return later).

- (2) a. orange
b. aeegf
c. Give it to me!
d. become a compulsive liar

DISCUSS

Do you agree with this? Whether or not your opinion is the same, try to justify your answer. Use the following as guiding questions:

- What criteria did you use to decide whether or not a given sequence of letters is a word? Did you consistently use the same criteria for all the cases?
- Should we define *orange* as one or two words, depending on whether it refers to the fruit or the color? Are both senses related?
- Would equivalent linguistic expressions in other languages qualify as words, too?
- Would your answer be the same if you perceived these expressions acoustically and not visually?

As you probably realized at this point, a considerable amount of complexity is hiding behind this intuitively simple notion, especially if we try to find a more technical definition of “word”. To a large extent, this complexity is due to the inherently intricate nature of the word, which is a highly structured object: it has internal constituents (studied in morphology) and it is used to build bigger linguistic expressions (studied in syntax); it conveys meaning (studied in semantics); and it has graphic and acoustic identity (as studied in orthography, and phonetics and phonology, respectively). Depending on which of these perspectives on wordhood we adopt, we will end up with the following definitions:

(3) **Orthographic word**

A sequence of letters written consecutively, with no blank spaces.

In the above examples, (2a) would certainly count as a word according to this definition, which seems to confirm our initial intuitions. But so would (2b), which certainly seems counterintuitive because we cannot assign any meaning to this sequence of letters (we take up the meaning component of wordhood shortly). (2c) and (2d), in turn, have four orthographic words, but we would probably have to answer differently when dealing with other languages: the Spanish equivalent of *give it to me* is written with no spaces (*dámelo*, lit. ‘give-me-it’), and *become a compulsive liar* is expressible with a single verb in Russian: *izolgat’sja*. Does it make sense? As it turns out, there is much cross-linguistic variation in the definition of word.

(4) **Phonological word**

“A string of sounds that behaves as a unit for certain kinds of phonological processes, especially stress or accent” (Aronoff and Fudeman 2011: 40).

The stress criterion is relevant in English and in many other languages (although there are many other phonological criteria, depending on the language) where a phonological word must have a stress. But, again, we find a counterexample in (2c): under normal circumstances, *give it to me* is pronounced with just one main stress, on *give*, which makes the whole expression one phonological word and seems to indicate that *it*, *to*, and *me* are not treated like words by the phonology. These kinds of linguistic elements, which are syntactically (and sometimes orthographically, as in English) independent, but which cannot stand alone phonologically, are called *clitics*: they usually need to be incorporated into the prosodic structure of adjacent words.

🔍 *Prosodic or supersegmental features* are stress, rhythm, pitch, volume, and intonation. They characterize strings of sounds or phonemes rather than individual sounds.

Consider the ways that stress can convey different kinds of meaning: in (a) below the pronoun *it* is used *metalinguistically* (we talk about the word *it* instead of using it to refer to a real-world entity); the preposition *to* in (b) carries

contrastive stress, to highlight the opposition between *to* and *from*; finally, the personal pronoun *me* in (c) is used after a pause in an elliptical construction.

- a. What does *it* stand for in this sentence?
- b. I said I was returning *to* Madrid, not *from* Madrid.
- c. What do you think about this book? – Who, *me*?

(5) **Content words** (lexical categories)

Content words encode a rich conceptual meaning, related to the real world.

(6) **Function words** (functional categories)

Function words denote much more abstract, language-internal meaning.

The lexical versus functional category distinction is a fundamental one in syntactic studies, and we are following the syntactic perspective on it in this section (although it is not the only one, as will be demonstrated).

Lexical categories are also called *content words* or *semantic words* because they carry meanings that denote or describe things in the world: they refer to real entities (*table, sun, book, crowd*), imaginary entities (*unicorn, horcrux, holodeck*), events (*blow, coronation; teach, explode*), and properties (*whiteness, wisdom; tired, pretty*). In most approaches, lexical categories include nouns, adjectives and verbs. The lexical categories largely overlap with the so-called “open-class lexical items”, the latter term referring to the fact that new words can be introduced into the lexicon by speakers of the language, and that they can also fall into disuse. The acquisition of open-class words is a life-long process for every speaker. It is also usually a conscious one: even young children can explain what *table* or *milk* means, and can point to the objects in the real world denoted by these words.

Functional categories, on the other hand, represent a very restricted and synchronically stable (closed) repertoire: the speaker cannot coin a new determiner or auxiliary verb voluntarily. The following syntactic classes are usually considered functional categories: prepositions (e.g., *in, to, of*), conjunctions (e.g., *and, or, until*), complementizers (*that, if, whether*), pronouns (*she, him, themselves*), determiners and quantifiers (e.g., *a, the, her; most, more, several*), and auxiliary verbs (*do, be*).

The function words are acquired for life in early childhood and their content is much less accessible consciously, which is one of the reasons why adult second language learners usually have a hard time acquiring these categories even if their vocabulary is extensive in general: how would you explain to your Czech friend what *the* or *a* means (Czech has no indefinite/definite distinction)? The meaning of these words seems to be language-internal and very abstract: e.g., *a* introduces new entities belonging to a certain class into the discourse (8a), and *the* establishes definite reference: it appears in nominal groups referring to uniquely identifiable entities, which are known from previous contexts (8b):

- (7) a. I saw {a/ the} cat.
 b. Viděl jsem kočku.
- (8) a. I just saw *a* cat. (in a context where there was no cat before)
 b. I just saw *the* cat. (in a context where a cat had been mentioned or alluded to previously)

Some function words, however, do have a much more transparent meaning. The semantic content of spatial prepositions (*under, on, in, between*, etc.), for example, is much easier to grasp. While the distinction between lexical and functional words is not a clear one in many cases, it has far-reaching consequences for syntactic theory and the articulation of the relationship between syntax and the lexicon: although both kinds of elements are sound–meaning pairings, roughly, function words belong to the realm of syntax, since they are instrumental in combining different words in phrases and sentences, and lexical words belong to the realm of lexicon, understood as the repository of different kinds of idiosyncratic, not uniquely linguistic, information. If we compare syntactic structures with buildings (again, following the syntactic perspective on this opposition), the lexical words would be the bricks or building blocks, and the function words would be the cement or the glue that holds the bricks together. Without the function words, the syntactic structure would fall apart. For example, the sentence (2c) above would turn into the generally uninterpretable *Give!* if we remove the function words.

Considering how complex and typologically heterogeneous an entity a *word* is, we might ask whether it should still be considered a linguistic technical term at all. Researchers working in different subfields of linguistics seem to opt for a positive answer to this question. Independently of which criterion of wordhood (orthographic, phonological, or grammatical) prevails in any given language, there seems to be reliable psychological evidence that ‘word’ is an important *cognitive* unit. One of the proofs is the *word superiority effect* detected in psycholinguistic studies: letters are processed faster and more accurately when they are embedded in a word rather than presented alone or within a random sequence of letters, which indicates that words have some kind of access advantage as compared to non-words or single letters.

At this point, we are ready to introduce the definition of ‘word’ as adopted throughout the book. Unless otherwise noted, we assume that ‘word’ used in the context of linguistic theory has the following definition:

(9) **Word**

A “meaning–form” pairing (i.e., association of an acoustic or graphical form and meaning) used in forming a sentence in a language and intuitively recognized by native speakers as the basic unit of meaningful speech.

 This definition comes from the classical Saussurean definition of *linguistic sign*, which has been used in linguistics since the beginning of the twentieth century: “the linguistic sign unites a concept [signified] and a sound–image [signifier]” (Saussure 1959).

1.4 Lexeme, Word Form, and Grammatical Word

After this detailed review of all the things that *word* stands for, you may still wonder whether there is a way of referring to a word in all its richness, with all these nuances? In fact, there is one. *Lexeme* is the term used to refer to the word as an abstract linguistic unit, which represents all the information we associate with a lexical item out of context, i.e., its inherent features. It is abstract because words are always used in context (except the *metalinguistic uses*), and because none of the words we actually hear or read are lexemes: they are (physical or acoustic) *word forms* or *grammatical words*.

We can also talk about lexemes as *types* and word forms as *tokens*. The type–token distinction is familiar to anyone who has dealt with multiple instances of the same kind of object: coins, apples, books, etc. If you buy two copies of the same book, you have two *tokens*, but just one *type*. In the actual use of linguistic utterances we make reference to linguistic types but each use of the type is a specific token. For example, in (10), the number of **tokens** of the **type** *the* is 3: i.e., # of types = 1, # of tokens = 3.

(10) **The** boy bought **the** pizza at **the** store.

In order to refer to a lexeme (e.g., in a dictionary), one of its forms is used conventionally, which is called the *lemma* or the *citation form*. For example, in most European languages the infinitive is used as the citation form for verbs (*go*, *break*), the singular masculine form is used for adjectives and participles (*green*, *broken*), and the singular is used for the nouns (*woman*, *house*).

Once the lexeme is inserted into the syntactic context, it becomes a *grammatical word* or *word form*. In the sentence, the lexeme acquires additional syntactic features. In the following example, there are two word forms corresponding to the lexeme *to be* (*were* and *was*) and two word forms corresponding to the lexeme *week* (*week* and *weeks*):

(11) **Were** you in Cape Cod last week? – Yes, I **was** there for the last two weeks.

In both cases, the verb acquires the past tense features and, in addition, it must agree in person and number with the subjects *you* and *I*, respectively. The noun, on the other hand, is realized in the singular and plural grammatical number forms (*week*, *weeks*, respectively).

1.5 What Is in a Lexical Entry?

In order to understand what information needs to be represented in a lexical entry, imagine for a second that you are given a list of new terms in English, which you have to memorize and learn how to use in context. What kind of data would you need to include for every lexical item, in order to achieve this goal? As we will see in more detail in Chapter 6, there are at least four

types of information in a lexical entry: *orthographic* (how do I write the word?); *phonological* (how do I pronounce the word?); *semantic* (what does the word mean?); and *syntactic* (what other words is this one compatible with in a phrase or a sentence?).

Let's take as an example the word *table*. We first need to specify its spelling (*table*) and pronunciation (/ˈteɪbəl/). As we have seen, semantic information is often provided through paraphrases called *definitions* in the dictionaries. The following is the definition of *table* taken from the Merriam-Webster Online Dictionary:

- (12) **table**: a piece of furniture with a flat surface that is designed to be used for a particular purpose.

Like most dictionary definitions, this one is composed of two parts: the *definiendum*, which is the word being defined; and the *definiens*, which is how it is defined. The *definiens* makes reference to two features: (a) the larger class in which the word meaning is included (i.e., *furniture*); and (b) the set of distinctive features which differentiates this concept from the other members of the same class (i.e., *with a flat surface* and *designed to be used for a particular purpose*).

As shown at the beginning of this chapter, the syntactic information provided in the dictionaries is usually limited to the syntactic category: e.g., *table* is a noun. This is an important piece of data, from which we can deduce (with the help of syntactic rules) that *table* is projected in the syntax as the head of the noun phrase. As such, it is compatible with certain kinds of complements (adjectival and prepositional phrases: [_{NP} [_{AP} red] table], [_{NP} table [_{PP} by the window]]), and it can be used as a complement of higher-level functional projections, for example within a determiner phrase (DP): [_{DP} the/three [_{NP} red tables]]. However, it says nothing about more fine-grained requirements this noun imposes on its complements. Recall the ungrammatical **fast table* example: if any kind of adjective was compatible with *table*, *fast* should be able to combine with it, just as well as the adjectives *red* or *nice*, but this is not the case. Also, it does not account for the interpretation of *useful table*, for example: the speakers know that it means 'useful for a particular purpose' (writing, cooking, etc.), but this meaning cannot be inferred from the dictionary definition.

Print dictionaries are a useful source of linguistic data, but as we will see in the following chapters the structure of the lexical entry in the mental lexicon requires a much richer and much more complex representation. Chapter 6 deals with this topic in detail and shows formally how the different kinds of information are connected in a lexical entry (in particular, how the word meaning components affect the syntactic behavior of the word).

1.6 The Role of Empirical Data in Lexicon Research

Determining the appropriate kind of data is, of course, critical for the development of a theory in any empirically oriented field of scientific research.

The theory being constructed must be evaluated against the data we have chosen as relevant, in order to be validated. As we will see in this section, the very nature of the linguistic data determines the linguistic methodology to a great extent.

1.6.1 Naturally Occurring Data

Linguists deal with two kinds of data: finite expressions, which we can list exhaustively (at any given point in time, the set of **phonemes** and **morphemes** in a given language is stable) and potentially infinite expressions, which cannot be exhaustively enumerated (combinations of words as phrases and sentences).

- (13) **Phoneme:** an abstract unit of the phonetic system which is perceptually distinct from other sounds in the same language. Phonemes are not endowed with meaning but they can distinguish one word from another. For example, /b/ and /p/ are two English phonemes which are similar (both are plosive bilabial consonants) but are perceived as different by speakers (/b/ is voiced and /p/ voiceless) and can distinguish words, as in *bet* and *pet*.
- (14) **Morpheme:** the smallest linguistic unit with a lexical or grammatical meaning. Morphologically simple words can have just one morpheme (*sun*, *get*, *black*) and morphemes can be smaller than a word (e.g., *un-*, *-believ-*, and *-able* in *unbelievable*).

The first kind of data are directly observable and can be characterized more or less completely for a given language. The second kind of data, however, due to the creative nature of language, are only observable in part, by sampling from available resources: (a) our own language intuitions (*naturally elicited* data); or (b) attested data that come from linguistic *corpora* of some sort (*naturally occurring* data). Lexicographers and corpus linguists take *naturally occurring* data to be the primary source for making generalizations and theoretical statements about a language.

- (15) **Corpus:** A collection of linguistic elements (usually texts) selected and ordered according to explicit linguistic criteria in order to be used as a representative sample of the language.

🔍 The plural of *corpus* is *corpora*.

In other words, a corpus is a finite snapshot of a natural language. In order to be a source for creating a reliable model of linguistic reality, the corpus must be *balanced*, i.e., it must contain quantitatively proportional samples of different kinds of texts, according to the medium of publication (journals, newspapers, books, brochures, etc.), genre (poetry, theater, fiction, essay, etc.), topic (biology, literature, arts, etc.), and geographical origin (e.g., American versus British English, or European versus Latin American Spanish). The corpora allow us to establish the existence and relative frequency of different linguistic phenomena, and they

are especially useful for the study of contextual distributions of different lexical items, which is one of the central issues in lexicological and lexicographic research.

Current linguistic corpora can be very large, consisting of millions, hundreds of millions, and even several billion words; they can also be processed relatively easily for simple search and retrieval operations. They are usually *annotated* with various metadata (author, document production time, source, genre, language) and linguistic information (part of speech, word sense, and other syntactic and semantic information).

1.6.2 Naturally Elicited Data

As far as theoretical linguistics is concerned, the source and scope of the data examined will have a direct impact on the generality and breadth of the explanations that the theory makes about language. If you merely look into what is directly observable, you risk eventually coming across a new piece of data which can refute your theory. If you want to go further, you have to find a way to *elicit* (build or obtain) those data which do not occur naturally (for example, because they are infrequent or very specific).

The needed data can be elicited *naturally* or *experimentally*. Naturally elicited data are usually referred to as *introspective* in theoretical linguistics: linguists often exploit their native competence to generate linguistic expressions and evaluate their acceptability. Introspection plays a very important role in generative linguistic studies, inasmuch as the linguistic competence of *any* speaker is representative of the linguistic competence in a given language. In the words of Chomsky (1965: 3), “linguistic theory is concerned primarily with an ideal speaker-listener, in a completely homogeneous speech community, who knows its language perfectly and is unaffected by such grammatically irrelevant conditions as memory limitations, distractions, shifts of attention and interest, and errors (random or characteristic) in applying his knowledge of the language in actual performance.”

Let us now examine, operationally, what the procedure is for generating linguistic examples in the absence of naturally occurring data. Assume that we are interested in describing the syntactic behavior of adjectives in an English corpus, particularly those that describe events, such as *likely*, *unlikely*, and *probable*. The corpus contains many sentences containing the target adjectives, including those shown in (16).

- (16) a. It is *likely* that Mary is coming to the party.
 b. It is *probable* that spring will arrive early.
 c. Mary is *likely* to come to the party.

We notice a syntactic correlation between (16a) and (16c), where the subject of the embedded sentence (*Mary*) can appear as the subject of the entire sentence. However, we observe no sentences where *probable* behaves in a manner similar to *likely* in (16c), as illustrated in (17) (where NA is “not attested”).

- (17) (NA) Spring is *probable* to arrive early.

By virtue of our own linguistic competence, we can introspect regarding the acceptability of sentences that are not attested in the corpus. Specifically, (17) doesn't appear because it is ungrammatical, and this only confirms our own linguistic intuitions regarding the adjective's grammatical behavior, giving support to a classification that separates adjectives like *probable* from those behaving like *likely*.

🔍 Another valuable if minor source of empirical data are *slips of the tongue*, which can either be registered by the linguist in natural speech or else elicited. They provide valuable clues on how the lexical items are generated and retrieved, and on how they are related among themselves. These speech production errors are not produced randomly: they have been shown to be highly constrained and to affect certain kinds of segments (e.g., only phonemes and phoneme clusters present in normal speech can appear in the slips of the tongue).

There are many types of speech errors, but the ones a lexicologist is most interested in are those where the morphemes within a morphologically complex word are somehow misplaced or blended, as in (a) and (b); where a wrong word is retrieved from the mental lexicon (synonym, antonym, associated word, as in (c) and (d); or a new (possible but nonexistent) word is created, as in (e):

- a. infinity clause → infinitive clause
- b. momentaneous (blend of two synonyms, *momentary* and *instantaneous*)
- c. This room is too damn *hot – cold* (antonyms)
- d. the two *contemporary*, er – sorry, *adjacent* buildings (similar notions in the time and space dimensions)
- e. groupment → grouping

(Fromkin 1971, Fromkin 1973)

1.6.3 Experimentally Elicited Data

Experimentally elicited data are usually obtained as a result of a *questionnaire* or task proposed by the experimenter to a group of native speakers. The former is used in all subfields of linguistics, where the researcher might: ask the subjects explicitly about a specific issue of interest (e.g., *Which of the following words have a negative connotation for you?*); or focus on the data needed in order to investigate this issue (e.g., by asking the subjects to fill in the blanks in a *cloze test*, a fragment of text with certain words removed, or to rate the linguistic expressions according to their acceptability, etc).

In *psycholinguistics*, data are elicited from a wide variety of language experiments. The ultimate goal of the experiment is to establish a relationship between dependent and independent variables. To this end, the independent variables are manipulated and the dependent variables are measured simultaneously, while holding constant other factors (the control variables) which can have an impact

on the dependent variables. The tested data in psycholinguistic experiments are usually controlled for frequency and length (among other variables): other things being equal, frequent and shorter words are processed faster than low-frequency and long words, and we do not want this difference to affect the results of the experiment. The following list includes some of the most common experimental setups:

- (18) a. *letter identification*: what letter is it?
- b. *picture naming, word naming*: name the picture or read aloud the word
- c. *reading*: read the following sequence
- d. *lexical decision*: is this sequence of letters a real word?
- e. *semantic decision*: does the word refer to a living or a nonliving entity? (animacy), are the two words synonymous? (synonymy), etc.

These tasks can be combined with specific experimental techniques (eye-tracking, ERP, fMRI). *Priming* is one of the most popular psycholinguistic techniques. It is based on the *priming effect*, defined as “improvement in speed or accuracy to respond to a stimulus [*target*], when it is preceded by a semantically [morphologically, phonologically, orthographically, etc.] related stimulus [called *prime*] (e.g., *cat-dog*) relative to when it is preceded by a semantically unrelated stimulus (e.g., *table-dog*)” (adapted from McNamara 2004). Thus, the presence of priming is indicative of relatedness between the prime and the target in the mental lexicon of the speakers.

These effects are recorded and quantified using either *offline* (focusing on the result of processing) or *online* (focusing on the course of processing) measurement techniques. The latter includes the following strategies:

- (19) a. *Eye-tracking*. This records eye movements during a task, based on the assumption that eyes do not fixate on visual stimuli randomly, but stop at what is being processed by the brain and only for as long as necessary.
- b. *Neuroimaging techniques* (fMRI: Functional Magnetic Resonance Imaging, and PET: Positron Emission Tomography). These are based on changes in brain metabolism: the brain areas activated as a response to a processing task can be detected and visualized because they demand more oxygen and therefore attract more blood flow.
- c. *Neurophysiological measures* (ERP: Event-Related Potentials). These measure the electric responses of the brain to certain external events, measured using electroencephalography. Certain types of ERP components are known to be related (although not exclusively) to language processing.

It is important to recognize that both naturally occurring and elicited data have significant drawbacks. Naturally occurring data are partial, in that they do not always contain the phenomena we want to investigate: as a result, it is very difficult to control all the variables (dialect, style, register, frequency, among others) and make them compatible with a unified theoretical framework. Importantly, naturally occurring data do not contain *negative data*, i.e., unacceptable linguistic

expressions which would allow us to show what is *not* well-formed in a natural language (e.g., if we do not find the expression *fast table* in a corpus, we would not know if it is not possible at all or if it is just not attested in this particular corpus). The set of elicited data, on the other hand, usually reflects the linguistic competence of a very limited number of speakers, and hence is incomplete or unrepresentative of the larger linguistic community. This is why it is commonly assumed in contemporary linguistics research that both kinds of data should be combined in different stages of the research cycle.

1.6.4 Data in the Linguistic Research Cycle

When formulating a linguistic hypothesis with language data, a fairly regular cycle of development is employed, summarized below.

1. Identification of the phenomenon to be analyzed and/or formulation of the problem to be solved;
2. Formulation of an initial hypothesis or outline theory, usually backed up by a limited and manageable set of introspective (naturally elicited) data;
3. Testing on a much larger set of naturally occurring data (corpora) or experimentally elicited data;
4. Quantitative and qualitative data analysis;
5. Refinement/modification of the initial hypothesis or formulation of a new one.

Stage (1) concerns the questions we ask ourselves in lexicological research. This is an important point, since the researcher's choice of method will be substantially influenced by the nature of this question. The major goals of lexicological research can be separated into the following areas:

- (20) a. *lexical semantics*: what is word meaning, how can it be represented, how can lexical ambiguity and polysemy be accounted for?
- b. *the interaction of words in context*: what syntactic and semantic constraints does a lexical item impose on its neighbors in a sentence, what lexical features are syntactically relevant, how is sentence meaning computed?
- c. *the internal structure of the lexicon*: how are words organized in the mental lexicon, what is the nature of lexical relations?

Of course, the research question raised in a concrete linguistic study must be much more specific than these general goals: we can investigate the selectional requirements of the verbs of cutting, the syntactic behavior of mass and count nominals, the hypernymic relations among types of artifacts, etc.

Regarding stage (2), the hypothesis we put forward must fulfill certain basic requirements. It should be:

- (21) a. *falsifiable*: it must be refutable by linguistic data;
- b. *empirically adequate*: it must account for the actual data;

- c. *able to generate predictions*: it must account for the kinds of linguistic expressions that are *possible*, beyond the actual data used to test the initial hypothesis.

Stage (3) states that, whenever possible, we need to test our hypothesis on large data sets to assess the robustness of the observed phenomenon and minimize the impact of random and irrelevant factors which can affect our conclusions (e.g., individual differences between speakers: age, educational level, dialectal and idiolectal specificities, etc.).

In stage (4), the quantitative data analysis refers to statistical interpretations of the obtained data. We need to check if the manipulation of independent variables affects the dependent variable and if the registered effect is statistically significant, i.e., if it is not pure chance. Qualitative data analysis relates the obtained results to the initially formulated hypothesis: do they confirm or disconfirm it? Furthermore, how do these results compare to those obtained in similar or related studies; and what is their bearing on the research of this particular linguistic phenomenon? At stage (5), we loop back to the beginning of the cycle to refine or abandon our theoretical assumptions.

II Interpreting Data in the Context of Theory

Empirical data must be treated with caution, both when collecting them (choosing balanced corpora and building balanced data sets for experiments) and when interpreting them relative to a linguistic theory.

- Make sure that the data are as *ecological* as possible, i.e., that the conditions in which they are obtained are close to the natural conditions (e.g., when the speaker is not being taped or is not aware that he is being taped). For the same reason, in psycholinguistic experiments preference should be given to simple and non-invasive tasks, which minimize the chance that the tested speaker's response is due to a conscious decision (e.g., eye-movement experiments are more ecological than semantic decision tasks).
- Elicited data are the product of linguistic competence (which, by definition, cannot be directly observed), and hence can be compatible with several underlying linguistic mechanisms or models of language representation. For instance, the slip of the tongue (*groupment* instead of *grouping*) could have happened because the speaker could not access the existing lexeme in his or her mental lexicon (because of time constraints, for example) or because this speaker was looking for a specific meaning nuance encoded by the suffix *-ment* but not by *-ing*.
- Since we have no direct access to what the mental lexicon really is, the models of language representation we formulate are usually metaphorical in nature: we use syntactic *trees* to represent the phrase

structure, or lexical *networks* and *fields* to visualize how words are linked to one another. There is no question that some of these metaphors are very useful and can even provide us with the clues to the nature of the investigated phenomenon, but we should not forget that this is only a useful approximation to how language works: this is not necessarily how language is actually represented in the brain.

1.7 Goals of Linguistic Theory and the Notion of Grammar

In the next four chapters, we position the lexicon within the larger context of the grammar of a language. We begin by characterizing the different approaches to the lexicon through the way the theory accesses lexical information, in order to discuss the “use” of the lexicon by the syntax and semantics. In particular, we see how information linked to individual words can have an impact on two of the main components of the grammar: syntax (contributing to the structure of a sentence) and semantics (contributing to the interpretation of words and word combinations).

Since most linguistic theories approach the study of the lexicon from a particular point of view and tend to focus on the interaction or interface of the lexicon with specific linguistic modules or sub-disciplines, we will provide an overview of the models of lexicon assumed in some of the most influential syntactically and semantically based frameworks. Our main goal will be to present the changing view of the lexicon within linguistic theory: from a passive storage of atomic items to a complex, genuinely creative, and dynamic component of the language.

The aim of linguistic theory is to characterize and explain the behavior of language as a complex human activity. We can identify three goals that will figure prominently in our investigation of language and what role the lexicon plays in it:

- *Provide an adequate description of linguistic phenomena:* this constitutes the first phase of linguistic research. We *describe* the observable properties of linguistic items and structures before developing a theory which would *account* for these properties.
- *Give an explanation of these phenomena:* we design *linguistic models* which would account for the previously identified properties of language phenomena: why do these properties occur? We use the term *linguistic model* because the inner workings of the language faculty are hidden from us: although the native speakers *can* speak and understand their language, the knowledge underlying this ability is *subconscious* and they cannot explain *why* and *how* they do it. Theoretical linguistics does not concern itself with whether and how the

linguistic rules and mechanisms are represented in our brain (brain physiology is the realm of neurology). Instead, it aims at providing abstract representations (i.e., models) compatible with the external manifestation of the language faculty. Recall from Section 1.6.4 that any theory or hypothesis must be *falsifiable*, *empirically adequate*, and must be able to *generate predictions*.

- *Contribute to our understanding of the cognitive underpinnings of language*: natural language is a cognitive ability, and linguistic research should help unveil how it is acquired, what (language specific and/or general cognitive) mechanisms are involved in its processing, and how linguistic knowledge can be represented.

In contemporary linguistics, *grammar* is a term referring to a human's cognitive system that enables one to speak and understand language. It encompasses everything we need to know in order to create well-formed expressions in language and interpret them, and also to judge whether the expressions encountered conform to the grammatical constraints of our language. Most linguists view grammar as itself consisting of distinct modules or systems, either by cognitive design or for descriptive convenience. These areas usually include: syntax, semantics, morphology, phonology, and the lexicon. Knowledge of pragmatics and language use are not uniformly considered linguistic in nature, but we will briefly refer to the interaction between lexical knowledge and pragmatic information in the last chapter of the text (Section 11.6). Throughout the book, we will refer to *lexical competence* as the ability to fluently recognize and use words.



Traditional Meaning of Grammar

In traditional grammatical studies, *grammar* tends to refer only to syntax and morphology, since its goals are purely descriptive: listing and classifying the different kinds of word-internal (morphological) and word-external (syntactic) structures in a given language.

1.8 Summary

In this chapter, we have introduced the general concepts that we will examine in this book. We stressed the following points:

- The *mental lexicon* is both complex and highly structured. Print dictionaries do not reflect many of its core properties: the relatedness of lexical senses and lexical items; the impact of lexical-semantic properties on syntax; semantic flexibility; lexical creativity; and so on.

- The concept of *word* can be defined from different perspectives (orthographic, phonological, grammatical, etc.), each of which focuses on a different facet of its complex nature and singles out partially different linguistic units. In spite of these differences, the word is a cognitively substantiated unit.
- The *lexical entry* in the mental lexicon comprises all the information we need in order to be able to use words in context, which includes their form (pronunciation and spelling) as well as their semantic and syntactic features.
- *Empirical data* are crucial in lexicological research: any model of the lexicon must be evaluated relative to the data; further, the nature of the linguistic data determines the linguistic methodology. The data can be *naturally occurring* (as in linguistic corpora), *naturally elicited* (introspective), or *experimentally elicited* (through experimentation). The different kinds of data complement each other in linguistic research.
- The lexicon is one of the core modules of *natural language grammar* understood as a human cognitive system, that enables native speakers to speak and understand their language. *Lexical competence* is the part of linguistic competence that enables the speakers to recognize and use words.

1.9 Further Readings

If you had trouble understanding the linguistic terminology and concepts used in this chapter, we strongly recommend strengthening your background. The following references can be helpful for this purpose:

TOPIC	SECONDARY REFERENCES
Basic linguistic terminology and concepts	Denham and Lobeck (2012) , Fromkin et al. (2018) , Akmajian et al. (2017) , Huddleston and Pullum (2005)
Introduction to the organization of the mental lexicon	Aitchison (2012) (Part I)
Different approaches to wordhood	Dixon and Aikhenvald (2003) , Aronoff and Fudeman (2011) (Chapter 2)
Introduction to morphological theory	Aronoff and Fudeman (2011)
Overview of the state of the art in psycholinguistic research and introduction to experimental design	Spivey et al. (2012) , Warren (2012) , Sandra (2009a) , Sandra (2009b)

1.10 Exercises

1. The following concepts are expressed through single words in languages other than English (they have been taken from the book *They have a word for it* by Howard Rheingold, 2000). Put to test your lexical creativity and create English words with these meanings. Say what linguistic resources you used: combination of existing morphemes (from English or other languages), phonological elements which seem to somehow evoke the meaning you need, existing words with a new meaning, borrowings from other languages, etc. Think about why a language might have a word for some concepts but not for others.
 - a. The prevailing mood or spirit in a society of a certain period (German: *Zeitgeist*, noun)
 - b. Mother who pushes her children into academic achievement a little too hard (Japanese: *kyoikumama*, noun)
 - c. The unspoken truth of certain social situations that everyone knows but nobody talks about directly (Kiriwina, New Guinea: *mokita*, noun)
 - d. Peace offerings for wives from guilty husbands (German: *Drachenfutter*, noun)
 - e. An extraordinarily energetic, talented, competent woman (Yiddish: *berrieh*, noun)
 - f. Excessively narrow-minded technical expert (German: *Fachidiot*, noun)
 - g. Elegantly beautiful, endowed with aesthetic value which grows over the years (Japanese: *shibui*, adjective)
 - h. Something that is all fouled up, especially as the result of an attempt to fix it (Yiddish: *farpotshket*, adjective)
 - i. To slake an emotional or spiritual thirst, to be revitalized (Thai: *sabsung*, verb)
 - j. To understand things and thus take them lightly, philosophically (Chinese: *ta*, verb)
 - k. To consciously do nothing, refrain from doing something (e.g., in politics) (Chinese: *wei-wu-wei*, verb)
2. In each sequence, identify the orthographic, phonological, and grammatical words:

a. I won't come.	c. washing machine
b. Unfrigginbelievable!	d. my ex
3. Provide examples of linguistic expressions (in English and in other languages) wherein

- a. the phonological and the grammatical words coincide
 - b. a phonological word consists of more than one grammatical word
 - c. a grammatical word consists of more than one phonological word
4. We have already defined what a word is, in terms of linguistic theory. The linguist Yuen Ren Chao argues that there is another definition of ‘word’ that we should consider, when studying language from a socio-cultural perspective, namely the *sociological word*. He writes:

By the “sociological word” I mean that type of unit, intermediate in size between a phoneme and a sentence, which the general, nonlinguistic public is conscious of, talks about, has an everyday term for, and is practically concerned with in various ways. It is the kind of thing which a child learns to say, which a teacher teaches children to read and write in school, which a writer is paid for so much per thousand, which a clerk in a telegraph office counts and charges so much per, the kind of thing one makes slips of the tongue on, and for the right or wrong use of which one is praised or blamed. Thus it has all the social features of the common small change of everyday speech which one would call a “word” in English. (Chao 1968).

 - Do you agree with the author?
 - In what other aspects of our life do we realize that a word is not just a linguistic unit and that its cultural significance goes beyond the scope of linguistics?
 - How does this importance manifest itself in other languages and cultures?
5. The following poem, credited to Richard Thompson, is composed of actual quotes from the former US President George W. Bush. Detect, classify, and analyze the different slips of the tongue and provide a tentative account of their source.

Make the Pie Higher

I think we all agree, the past is over.
 This is still a dangerous world.
 It’s a world of madmen
 And uncertainty
 And potential mental losses.
 Rarely is the question asked
 Is our children learning?
 Will the highways of the internet
 Become more few?
 How many hands have I shaken?
 They underestimate me.
 I am a pitbull on the pantleg of opportunity.
 I know that the human being and the fish

Can coexist.
Families is where our nation finds hope
Where our wings take dream.
Put food on your family!
Knock down the tollbooth!
Vulcanize society!
Make the pie higher!
Make the pie higher!

6. What source of linguistic data and what methodology would you use to investigate the following issues?
- a. Do speakers perceive the following verbs as more or less related to each other: *snore*, *strive*, *doze*, *shiver*, *tremble*?
 - b. What semantic groups of nominals do the verbs *fix*, *mend*, and *repair* usually take as their direct objects?
 - c. Do we process words that share their roots with many other words (e.g., *word*: *wordy*, *wording*, *wordless*, *word-book*, etc.) faster than words with only a few morphologically related lexical items (e.g., *lion*: *lioness*)?