The Quantitative Analysis of the Dynamics and Structure of Terminologies
Terminology and Lexicography Research and Practice (TLRP)

Terminology and Lexicography Research and Practice aims to provide in-depth studies and background information pertaining to Lexicography and Terminology. General works include philosophical, historical, theoretical, computational and cognitive approaches. Other works focus on structures for purpose- and domain-specific compilation (LSP), dictionary design, and training. The series includes monographs, state-of-the-art volumes and course books in the English language.

Editors

Marie-Claude L’ Homme
University of Montreal

Kyo Kageura
University of Tokyo

Consulting Editor

Juan C. Sager

Volume 15

The Quantitative Analysis of the Dynamics and Structure of Terminologies
by Kyo Kageura
The Quantitative Analysis of the Dynamics and Structure of Terminologies

Kyo Kageura
University of Tokyo

John Benjamins Publishing Company
Amsterdam / Philadelphia
# Table of contents

Acknowledgements xi
Preface xiii

Part I. Background

CHAPTER 1
The sphere of terminology 3
1.1 Lexicology: Its definition and status 3
   1.1.1 The definition of lexicology 3
   1.1.2 Vocabulary as a concrete object of language and lexicology 4
   1.1.3 Vocabulary and utterances 6
1.2 Terminology: Its definition and status 9
   1.2.1 What are terms? 9
   1.2.2 Terminology and terms 10
1.3 Vocabulary and terminology 11
   1.3.1 Vocabulary and terminology as a set 11
   1.3.2 Characteristics of words and terms 12
1.4 Quantitative approaches to lexicology and terminology 13
   1.4.1 Existing quantitative work on lexical items 13
   1.4.2 The position, framework and structure of the book 14

CHAPTER 2
The nature of terminological data 17
2.1 Terminological data 17
   2.1.1 The original dataset 17
   2.1.2 Identification of morphemes 18
   2.1.3 Data used in the analysis 20
2.2 Distributional characteristics of morphemes in terminologies 23
   2.2.1 Frequency distributions of morphemes 24
   2.2.2 Zipf’s law and the distribution of morphemes 27
2.3 Term-length distributions 35
   2.3.1 Basic nature of term-length distributions 35
   2.3.2 Term lengths and types of origin of morphemes 37
Part II. Distributional dynamics

CHAPTER 3
The mathematical framework and the status of data
3.1 The expectation of growth: A naive observation 44
3.2 The basic mathematical framework 45
  3.2.1 The urn model 45
  3.2.2 Binomial distribution 47
  3.2.3 Frequency spectrum and expected number of types 50
3.3 The standard way of estimation and the status of terminological data 52
3.4 Amount of unseen items and discounting 56
  3.4.1 An intuitive explanation 57
  3.4.2 Good-Turing estimation 59
3.5 Data, samples and discounting for terminologies 64

CHAPTER 4
The dynamics of morphemes in terminologies 67
4.1 Developmental profiles 67
  4.1.1 The sample size factor and the developmental profile 67
  4.1.2 Binomial interpolation and extrapolation 71
4.2 Some preparations 74
  4.2.1 Examining the randomness assumption 74
  4.2.2 Re-introducing the level of terms 80
4.3 The dynamics of morphemes in the terminologies of the six domains 81
  4.3.1 Points for observation 81
  4.3.2 Cross-domain observations 82
    4.3.2.1 Relationships between morpheme tokens and morpheme types 82
    4.3.2.2 The dynamics of morphemes vis-à-vis the growth of terminology 86
  4.3.3 Observations relating to types of origin 91
    4.3.3.1 Relations between morpheme tokens and morpheme types 92
    4.3.3.2 The dynamics of morphemes vis-à-vis the growth of terminology 95
4.4 LNRE models 102
4.5 Re-examining the qualitative difference in types of origin 104
CHAPTER 5
Interpretative and epistemological examination 109
5.1 The status of the data 110
  5.1.1 The data as the object of the study 110
  5.1.2 The data as a sample of the object of the study 112
5.2 The dynamics of terminology: Structure and event 114
  5.2.1 A core theory and surrounding factors 115
  5.2.2 The developmental model and the evolution of terminology 116
5.3 Epistemological implications of interpolation and extrapolation 117
  5.3.1 The “potential” data 117
  5.3.2 Morphemes with lost identity revisited 118
  5.3.3 What we have as and how we perceive the data 120

Part III. Tropistic structure

CHAPTER 6
Terminological structure and network representations 125
6.1 Compounding, term formation and terminological structure 126
  6.1.1 From term formation to the construction of terminologies 126
  6.1.2 Terminology as a network of terms 128
  6.1.3 The status of morphemes in the terminological structure 128
6.2 The tropistic nature of terminology 130
  6.2.1 Arbitrariness and the degree of tropism 130
  6.2.2 The tropistic nature of terminology 131
6.3 Terminological structure and “tropistic networks” 133
  6.3.1 Networks and tropism 133
  6.3.2 Systematicity and tropism 135
  6.3.3 Tropistic networks and types of origin 138
  6.3.4 Networks, conceptual structure and tropism revisited 139
6.4 Formal aspects of tropistic networks 140
  6.4.1 Basic notions concerning a graph/network 141
  6.4.2 The basic nature of tropistic networks 142

CHAPTER 7
The tropistic nature of terminologies 149
7.1 The framework of observation 151
  7.1.1 Aspects of tropistic networks 151
  7.1.2 Descriptive indices of networks 153
7.2 The scope of tropism 157
7.2.1 Cross-domain observations 159
   7.2.1.1 The scope of tropism from the point of view of the ratio of connected terms 160
   7.2.1.2 The scope of tropism from the point of view of the largest component 162
   7.2.1.3 The scope of tropism from the point of view of the distribution of components 163
7.2.2 Observations from the point of view of types of origin 166
   7.2.2.1 The scope of tropism from the point of view of the ratio of connected terms 166
   7.2.2.2 The scope of tropism from the point of view of the largest component 168
   7.2.2.3 The scope of tropism from the point of view of the distribution of components 168
7.3 The tropistic nature of the largest components 169
7.3.1 The strength of tropism 171
   7.3.1.1 Cross-domain observations 172
   7.3.1.2 Observations from the point of view of types of origin 176
7.3.2 Structural characteristics of tropistic networks 179
   7.3.2.1 Cross-domain observations 179
   7.3.2.2 Observations from the point of view of types of origin 183
7.4 Summary observations of the tropistic nature of terminologies 184

CHAPTER 8

The status of morphemes in terminological structures 187
8.1 Methodological framework 187
   8.1.1 The status of morphemes in putative terminologies 188
   8.1.2 The morphological network 189
   8.1.3 Morphological networks of the actual terminologies 192
8.2 Isolates and small components 196
8.3 The status of morphemes in the largest component 199
   8.3.1 Points of observation and indices for the largest components 199
      8.3.1.1 Status of morphemes as seen from their locations in the network 199
      8.3.1.2 The status of morphemes in terms of their relative importance 200
8.3.1.3 Relationships between borrowed and native morphemes 202
8.3.2 The status of borrowed and native morphemes 203
8.3.2.1 Conceptual versatility 204
8.3.2.2 Conceptual prestige 209
8.3.2.3 Mixing degree 213
8.4 Summary observations 214

Part IV. Conclusions

CHAPTER 9
Quantitative approaches to terminology in perspective 219
9.1 An indicative summary of the present study 219
9.1.1 The theoretical framework 219
9.1.2 The choice of methodology 220
9.1.3 Descriptive results 221
9.1.4 The epistemological framework 221
9.2 Remaining issues and directions for further research 222
9.2.1 Some technico-theoretical issues 222
9.2.2 Structure and history 223
9.2.3 Towards integrated descriptions of structural growth 224
9.3 Implications for related studies 224
9.3.1 Descriptive studies of terminology and vocabulary 224
9.3.2 Possible contributions to applications 225

Bibliography 227
Author index 237
Subject index 239
Acknowledgements

This work could not have been completed without the help of many people. First and foremost, I am deeply indebted to Professor Juan C. Sager for his constant encouragement with this project and for his detailed comments on earlier drafts of the text. His help has been invaluable in honing and strengthening every aspect of this book.

My thanks also go to my friends and colleagues in the field of terminology, especially to Professor Beatrice Daille (LINA-CNRS, University of Nantes), Professor Emmanuel Morin (LINA-CNRS, University of Nantes), and Professor Marie-Claude L’Homme (University of Montreal), and in the field of computational linguistics, especially to Professor Satoshi Sato (Nagoya University). I owe much to the input and guidance I received more than 10 years ago from Dr. Harald Baayen, then at the Max-Planck Institute for Psycholinguistics and currently Professor at the University of Alberta, regarding the quantitative methodologies adopted in Chapters 3 and 4.

The Library and Information Science Laboratory of the University of Tokyo, with which I am affiliated, provides me with an excellent research environment. I am grateful for the intellectually stimulating conversations I have had with faculty staff and students in the lab, and would especially like to express my appreciation to Professor Akira Nemoto and Professor Anthony Hartley (visiting professor at our lab from the University of Leeds).

The research in Chapters 2 to 4 and Chapter 8 was carried out with the generous support of the National Institute of Informatics (NII) research cooperation project “Development of practically usable terminological resources using a variety of textual resources” (2009–2010). Discussions I had with Professor Akiko Aizawa (NII) within the NII research cooperation project framework were a constant source of inspiration and encouragement.

The research in Chapters 6 and 7 was carried out with the support of the Japan Society for the Promotion of Sciences (JSPS) grant-in-aid (A) 21240021 “Developing an integrated translation-aid site that provides comprehensive reference resources for translators.”

Two reviewers kindly took the time to evaluate the draft of this book and give me detailed comments. Although I could not take into account all of their suggestions, I believe the modifications and augmentations I made in response to their
comments improved the content of the book. I would like to express my appreciation to them. Needless to say, the responsibility for any remaining shortcomings is mine alone.

The writing of this book was delayed partly due to the earthquake and tsunami that hit Japan on March 11, 2011 and the subsequent accident at the Fukushima Daiichi nuclear plant, which was caused by negligence on the part of the Tokyo Electric Company (TEPCO). I felt compelled to do my part to counter the ongoing misinformation campaign that seeks to portray the nuclear plant accident as trivial, by publishing a small book in Japanese devoted to a critical examination of how the media and so-called “experts” have responded to the accident. During this period, I received many messages of encouragement from colleagues all over the world in the field of terminology and computational linguistics, which helped me a great deal in maintaining the motivation to complete this book. I also received moral support from researchers sharing a similar point of view on the nuclear accident, and especially would like to thank the following people, all of whom I regard as persons of integrity: Professor Susumu Shimazono, Professor Masaki Oshikawa, Professor Shuichi Kito, Professor Ayumu Yasutomi, and Professor Shigeo Kodama of the University of Tokyo, and Associate Professor Hazuki Ishida, Associate Professor Akiko Endo, Associate Professor Koji Nagahata, Associate Professor Ryota Koyama, and Associate Professor Shinobu Gogo of Fukushima University. I would also like to thank Associate Professor Keita Tsuji of Tsukuba University for the timely support he provided on various occasions.

Finally, I would like to thank Stephanie Coop, my life partner, for her consistent help and support in every aspect of my life.

Kyo Kageura
August 20, 2012
Preface

Terminology and lexicology

We have recently been witnessing a growth in interest in the study of lexicology and lexicography in linguistics, as well as in practical and theoretical studies of terminology, in accordance with the rapid growth in universal communication and specialised knowledge.

In the research sphere, this can be observed simultaneously in several related fields, such as linguistics, natural language processing, translation studies and terminology. There is a proliferation of journals (e.g. *International Journal of Lexicography*, *Lexicology: An International Journal on the Structure of Vocabulary*, *Terminology: An International Journal of Theoretical and Practical Issues in Specialised Communication*, *Lexicon Forum*), book series (e.g. Oxford Studies in Lexicography and Lexicology; John Benjamin’s Terminology and Lexicology Research and Practice), reference books and textbooks (e.g. Atkins and Rundell 2008; Cruse et al. 2002/2005; Fontenelle 2008; Hartmann 2003; Sterkenburg 2003; Svensén 2009; Wright and Budin 1997/2001), and academic conferences (e.g. Euralex, Asialex, Terminology and Knowledge Engineering, Terminology and Artificial Intelligence, Compterm) devoted to these topics.

A look at these journals, conference proceedings, etc. reveals the existence of several research trends. Firstly, there are traditional qualitative studies – both theoretical and descriptive – in lexicology and terminology, which, in general, deal either with a limited set of lexical items or with lexical forms. As such, they do not directly address vocabulary or terminology as a whole. Secondly, there are studies that address vocabulary as a set. Two approaches can be identified in this latter type of research. On the one hand, there is the applied approach, which is motivated by practical concerns such as compiling dictionaries or terminologies. Methodologically, this type of work incorporates whatever is necessary for the practical aim. On the other hand, there is also a large amount of work devoted to automatic computational processing of lexical items or terms, such as automatic term extraction, automatic thesaurus construction, etc. Most studies in this latter category implicitly regard vocabulary or terminology as an element dependent on texts; they try to extract certain types of units such as terms and/or related information from textual corpora, without explicitly determining the desiderata for the final product,
which could be a lexicon or a terminology. As a result, this computational work currently tends to fall short of practical usability in real-world situations.

If, following Maeda (1989) and Mizutani (1983), we see lexicology and terminology as essentially the study of (a) not forms but substance or actual existence and (b) not individual lexical items or an arbitrarily chosen small number of lexical items but vocabulary or terminology as a coherent set, then it is practical lexicological or terminological work that addresses the sphere of lexicology or terminology more directly, because for such work to be practically useful it must deal with a substantial number of lexical items coherently and consistently. Being essentially applied, however, such work does not explicitly constitute a theoretical study of vocabulary or terminology.

We can recognise a lacuna here: there is a paucity of theoretical work on vocabulary or terminology as a set. Directly targeting vocabulary or terminology as a set is all the more important because “language qualifies ... as a complex system” (Ninio 2006: 147), and vocabularies themselves can qualify as such, as they “are emergent phenomena in the sense that they are the spontaneous outcome of the interactions among the many constituent units” and “are not engineered systems put in place according to a definite blueprint” (Barrat et al. 2008: 47). While terminologies in general tend towards systematicity compared to general vocabularies, deliberate planning only acts at the microscopic level, and even if social control is applied in the form of recommendations or regulations by academic societies, it is carried out in hindsight rather than in accordance with some kind of pre-existing blueprint, and affects only a small portion of terminological phenomena. Terminologies, therefore, can also be regarded as complex systems.

While we have so far talked about both general vocabulary and specialised terminology – because the issues discussed up to this point are common to both – the present study focuses on terminology, not general vocabulary, and, within this hitherto underaddressed area of study, seeks clarification of the nature of terminologies as a set, although it does not explicitly deal with terminologies as complex systems. While some of the methods and assumptions adopted in the study may be applicable only to terminologies, it is still hoped that the work as a whole will provide some useful methodological insights into the study of general vocabulary as well.

1. There is considerable ambiguity regarding what is meant by “lexicology.” For instance, unlike Maeda (1989) and Mizutani (1983), Geeraerts (1994) does not require that lexicology should deal with a coherent vocabulary as a set. We will examine this point in Chapter 1.

2. As most of us know that the utility of dictionaries depends, among other factors, on the very choice of entries, we can reasonably expect that professional lexicographers possess some important theoretical understanding of the nature of vocabulary as a whole, but little published work exists in this regard.
Quantitative approach

One possible approach we can naturally resort to in order to explore this area of study is a quantitative one, as quantitative approaches have successfully been used for describing, characterising or modelling a range of complex collective phenomena. In addition, since the pioneering work by Zipf (1935; 1949) and Yule (1944), the quantitative approach to language analysis, characterisation and modelling has established its own footing in linguistics, especially when dealing with actual language data or corpora. In Russia, the Czech Republic, Germany and Japan, there are strong research communities with long traditions of work devoted to quantitative linguistics.

Nevertheless, quantitative linguistics per se seems to be in a rather ambiguous situation at present. For one thing, with the rapid growth of statistical approaches to computational linguistics and natural language processing, especially since the 1990s, which aim, to some extent at least, to model languages for the sake of language processing (cf. Charniak 1993; Manning and Schütze 1999), the number of quantitative studies directed at the theoretical understanding and modelling of language phenomena seems to be in decline, in both relative and absolute terms.

The situation is aggravated by the fact that quantitative methods, as opposed to modellings, have become widely and easily accessible, due in great part to the ready-to-use statistical packages that have become available at no or low cost. Paradoxically, this seems to have created a tendency for quantitative analysis to be used in a much wider range of studies in linguistics while at the same time reducing the relative number and range of in-depth quantitative studies of language that aim at promoting understanding of language itself, rather than of individual language phenomena.

Although there are important and sound contributions based on the quantitative approach to languages (Baayen 2001; Lebart et al. 1997; Mizutani 1983; Tuldava 1995), it is nevertheless the case that the potential of this approach is neither fully understood nor exploited, the situation with regard to terminology being no exception. While many qualitative and computational studies exist, only a few (e.g. Kageura 2002; Sanada 2004) seriously pursue the quantitative modelling or description of terminology with a due theoretical perspective. This book, which explores the potential of the quantitative approach to terminology, is an attempt to fill this gap.

While by saying this we share Dr. Samuel Johnson’s view that quantitative material “brings everything to a certainty which before floated in the mind indefinitely,” we do not wish to claim that quantitative approaches provide a magic solution for everything. Nevertheless, we believe that the quantitative approach, even if it cannot by itself capture all the important theoretical features of vocabulary and terminology (or even if it may just be a ladder that should be discarded
after one scales the wall), it is not only useful but also essential for anybody who seriously wishes to deal with such complex phenomena as terminology.

The context and the framework of the present study

While the present work is completely independent and self-contained, it is still useful to give the direct context from which it arose. The antecedent of the present work is Kageura (2002). Since its publication, we have received a number of comments and questions at a variety of levels, mainly from researchers in terminology and computational linguistics.

Among the major comments and questions, two are concerned with theoretical and methodological issues:

1. A request to clarify further the status of “dynamic” quantitative analyses of terminologies on the basis of the distribution of morphemes, both in terms of the methodological framework and in terms of epistemological implications;
2. Questions regarding the connection between conceptual analyses and quantitative analyses, which point out that while quantitative analyses can be regarded as describing the overall characteristics of terminologies, conceptual analyses remain essentially at the level of individual terms, and that the connection between the individual descriptions and the interpretation of the results in terms of the terminologies as a whole is supported only by the fact that the entire terminological data, not a sample, is dealt with and quantitative information is provided.

Readers will find some direct and indirect responses to these points, mainly in Part II for the first question and in Part III for the second question. In a sense, the present work takes up from the topic dealt with in the second part of Kageura (2002), i.e. the quantitative observation of terminological growth, and works back from there to the topic dealt with in the first part, i.e. the conceptual structure represented by terminologies, but at a rather different level.

Another question raised is concerned with the phenomena of terminology:

3. What is the status and role of borrowed morphemes in Japanese terminologies, which were mentioned in Kageura (2002) but not fully explored?

Borrowing or the use of loanwords is a common occurrence in many languages (Haspelmath and Tadmor 2009) and is sometimes held to be “one of the primary forces behind changes in the lexicon of many languages” (Malmkjaer 1991: 208). In terminology, borrowing constitutes an important mechanism for creating new terms, which is reflected in the fact that the standard textbooks on terminology contain discussions on borrowing (Rey 1995; Sager 1990), and borrowing has been
studied in a variety of domains in many languages (Benson 1958; Karabacak 2009; Milić and Sokić 1998; Zhiwei 2004). In some languages, borrowed items often not only constitute new terms but also are incorporated into the repository of lexical items that contribute to creating new complex terms by compounding. Japanese is one of these languages in which borrowed items or morphemes play an important role in terminologies (Kageura 2003; Nomura and Ishii 1989b; Shioda 2002), and several studies have dealt with the status of borrowed terms or morphemes in Japanese terminologies (Ishii 2007; Kageura 2002; Kageura 2006; Nomura and Ishii 1989b; Otani 2007; Otani 2008). Against this backdrop, the status and role of borrowed and native morphemes within the system of terminologies constitutes the focal point of concern in this study.

Let us assert here the theoretical standpoint of the present study. It is first and foremost descriptive. Although the methodologies adopted can be interpreted as elucidating a model of terminological growth in relation to the constituent morphemes, especially in Part II, the present work is concerned with the description of existing terminologies, not providing models of terminology construction (this should become clear after reading Part III). As was argued in Kageura (2002) and will be confirmed in due course in the present study, the concept of terminology precedes the concept of individual terms, and, as already discussed, terminology as a whole, rather than individual terms, should be explicitly addressed in the study of terminology. Unfortunately, however, given the sheer size of terminologies, it is not possible to “see” what they are like directly, and the descriptions of terminologies to date have been mainly concerned with counting such basic features as the length of terms, distribution of term length, distribution of morphemes, etc. Against this backdrop, the present work aims at proceeding one step further in the quantitative description of terminologies as a whole, given that we cannot “see” the characteristics of terminologies directly. As such, it is concerned with using methodological aids to observe what we cannot see straightforwardly, rather than revealing the underlying mechanisms of terminology construction or developing models to capture these mechanisms. What was kept in mind in carrying out the present work was the framework given in Foucault (1968), in which he stated:

La question que pose l’analyse de la langue, à propos d’un fait de discours quelconque, est toujours: selon quelles règles tel énoncé a-t-il été construit, et par conséquent selon quelles règles d’autres énoncés semblables pourraient-ils être construits? La description du discours pose une toute autre question: comment se fait-il que tel énoncé soit apparu et nul autre à sa place? 3

3. “The question that the analysis of langue raises, in the face of a certain fact of discourse, is always: from what kind of rules was this énoncé constructed, and, consequently, from what kind of rules can other énoncés that resemble this one be constructed? The description of discourse
While fully acknowledging the simplification of his statement in the present context, it is still useful to state here that the study of terminologies is somewhat inclined to the “description du discours,” because, unlike sentences or language expressions in general, terminologies and vocabulary cannot be reduced to a set of abstract rules from which an infinite range of well-formed terms can be constructed; vocabularies and terminologies are essentially what are always and already there in the world as concrete entities. While it is not illegitimate to talk about what terminologies could be like or what kinds of terms are well-formed, the essence of terminologies nevertheless always consists of what we actually have at a given time in a given society for a given language which are, though saying this sounds like an outright oxymoron, terminologies. This is the underlying theoretical concern of the present study.

Incidentally, this explains why this work also occupies a place in library and information science (the author is affiliated with the Library and Information Science Laboratory of the University of Tokyo): library and information science also asks, in the face of a certain piece of recorded data, information and/or knowledge, how it is that this particular piece of data, information and/or knowledge, and nothing else in its place, came to exist. To the extent that the study of terminology deals with existing terms and the realistic possibility of new terms coming into existence, it has much in common with the perception of language and information in library and information science.

A note on typographical conventions

In the literature on linguistics and terminology, especially in work referring to meanings or concepts and linguistic symbols, there are typographical conventions in which meanings or concepts are indicated using double quotes and symbols are written in italics. This work does not follow these conventions, adopting instead an easy-going approach in which individual linguistic items and important terms are indicated using double quotes or as they are (in the case of Japanese). There are two main reasons for this decision:

1. Although the concept/symbol dichotomy is assumed in the background, the main arguments and discussions in this book relate not to the relationships between concepts and symbols per se, but to the structures of terminologies, which are defined over the surface form while at the same time the underlying

---

raises a completely different question: how is it that this énoncé, and nothing else in its place, appeared?” (my translation)
conceptual structure is assumed. It is thus not always easy to rigidly distinguish between concepts and symbols.

2. The individual linguistic examples referred to are mostly Japanese, and italicising them or double-quoting them would make the typography unnecessarily complex.
PART I

Background
CHAPTER 1

The sphere of terminology

This chapter clarifies the basic status and nature of lexicology and terminology, in order to give a general epistemological framework within which the ensuing technical discussions can be situated. In the following we will first define the concept of lexicology, and then characterise the status of terminology. Although the present study is concerned with terminology and not vocabulary in general, we begin by clarifying the scope of lexicology, for this (a) should help clarify the scope of terminology and (b) may assist readers whose main interest lies in lexicology to adopt the framework and methodologies outlined in this book to study general vocabulary.

1.1 Lexicology: Its definition and status

1.1.1 The definition of lexicology

Following Crystal (1992), we define “lexicology” as “the study of a language’s lexicon”; “lexicon” is “the vocabulary of a language (technically, its lexical items or lexemes), especially when these are listed in a dictionary as a set of lexical entries”; and “vocabulary” is “the set of lexical items (‘words’) in a language.” In these definitions, the word “set,” which is used twice, i.e. once in the definition of “lexicon” and once in the definition of “vocabulary,” plays an essential role.

This is not necessarily the case in other definitions of “lexicology.” For instance, in the definition of “lexis and lexicology,” McCarthy (1991) says:

> The study of lexis is the study of the vocabulary of languages in all aspects: words and their meanings, how words relate to one another, how they may combine with one another, and the relationships between vocabulary and other areas of the description of languages, the phonology, morphology, and syntax.

This characterisation of “the study of the vocabulary of languages” does not explicitly regard vocabulary as a set. The same sort of ambiguity in relation to the status of vocabulary can be identified in Geeraerts (1994) and the introductory explanation of lexicology given by Halliday (2004).

Without intending to claim that the definition given by Crystal (1992) should be the definition of lexicology, and fully appreciating that there can be several
definitions of and a variety of interpretations of the research field, in this study we use “lexicology” to mean “the study of the set of lexical items in a language,” regarding the basic target of lexicology to be vocabulary as a “set,” not individual lexical items or a small group of lexical items. Accordingly, the present study is concerned with the characteristics of vocabulary or terminology as a set, or the characteristics of items as sets constituting a vocabulary or terminology. What we call lexicology here, therefore, is essentially different from the definition adopted in studies that deal with words or a limited group of words, such as studies of word formation or of lexical semantics.

A note on the usage of the term “vocabulary” in the present work may also be useful here. While “vocabulary” may mean a coherent subset of a lexicon of a language defined according to certain extralinguistic features (thus it is possible to say that “terminology” is a “vocabulary” of a specialised domain), in the discussions below, the term is basically used to refer to the entire set of lexical items in a language, i.e. as a synonym for the “lexicon” of a language.

1.1.2 Vocabulary as a concrete object of language and lexicology

Defining lexicology and claiming that this constitutes an area of study worth pursuing presupposes that vocabulary as a set has an independent and autonomous existence in language. This leads us back to the fundamental discussion about the object of linguistic studies, especially given the current trends in linguistics in which prominence is given to texts or utterances and the lexicon or vocabulary is often regarded as a secondary derivation from texts or utterances, as represented by dictionaries.

In order to grasp the status of the vocabulary in language, let us return to the “starting phase” of modern linguistics. In his third lecture on general linguistics given at the University of Geneva from 1910 to 1911, Ferdinand de Saussure, in an attempt to identify *langue* as an autonomous object of linguistics, stated (Saussure 1910/11: 268–269):

L’exécution restera individuelle, c’est là que nous reconnaîtrons le domaine de la parole. C’est la partie réceptive et coordinate (qui est sociale), voilà ce qui forme un dépôt chez les différents individus, lequel arrive à être appréciablement conforme chez tous les individus.

C’est cette sphère-là qui nous représente la sphère de la langue. Ce sont ces milliers d’images verbales associées chez les individus à autant de concepts placées en regard. On peut dire qu’en prenant un individu nous aurons dans le seul exemplaire l’image de ce qu’est la langue dans la masse sociale. Si nous pouvions examiner
le dépôt des images verbales dans un individu, conservées, placés dans un certain ordre et classement, nous verrions là le lien social qui constitue la langue.¹

Note that *langue* is delineated here in contrast with *parole*, which is an individual execution of the language faculty. Without entering into a discussion of the distinction between spoken and written language, in our immediate context we can project this contrast upon the distinction between texts or utterances on the one hand and vocabulary or dictionaries on the other, a point on which we will elaborate shortly.

After isolating the sphere of *langue* within each person and characterising it as a social deposit, Saussure elaborates on the basic nature of *langue* as follows (Saussure 1910/11: 271–273):

1. *(La langue est) un objet définissable et séparable de l’ensemble des actes de langage. ...... (2) La langue est étudiable séparément; il n’est pas indispensable de considérer les autres éléments du langage pour étudier la langue. ...... (3) La langue ainsi délimitée est un objet de nature homogène. ...... (4) Dans la langue nous avons un objet de nature concrète.”²

and concludes that (Saussure 1910/11: 273):

Donc cet objet est non seulement de nature concrète, mais d’une espèce qui permet l’étude directe, à peu près comme celle de papillons classés dans une boîte de collectionneurs. Nous pouvons fixer ce qui est relatif à la langue. Grâce à ce caractère on peut dire en somme qu’un dictionnaire et une grammaire sont une image admissible, convenable de ce qui est contenu dans la langue.” (my translation)³

---

¹. “Execution (of language) remains personal, and it is there that we recognise the domain of *parole*. What is social is the receptive and coordinative part; this is where a deposit is formed in different persons, which ends up being identical to a considerable extent among all persons. / It is this sphere which represents the sphere of *langue*. It is these many verbal images associated with as many corresponding concepts. We can say that in any one person we will find an image of what *langue* consists of in the social mass. If we could examine the deposit of verbal images – conserved, placed in a certain order and classification – in a single person, we would see there the social bond that constitutes *langue*.” (my translation)

². “(1) *Langue* is an object which can be defined and separated from the totality of the acts of language. ...... (2) *Langue* can be studied separately; in order to study *langue*, it is not indispensable to take other elements of language into consideration. ...... (3) *Langue*, delimited in this way, has a homogeneous nature. ...... (4) In *langue* we are dealing with a concrete object.” (my translation)

³. “Thus this object is not only of a concrete nature but also of a kind that allows direct study, like butterflies classified in collector’s case. We can identify what is part of *langue*. Due to this characteristic, it can be said all in all that a dictionary and a grammar are an acceptable and relevant image of what is contained in *langue*.” (my translation)
Of the two elements of *langue* identified by Saussure, grammar can be regarded as a computational mechanism that constitutes the abstract part of *langue* (Saussure 1910/11: 295–299), while the dictionary or vocabulary can be regarded as constituting the concrete part. Wilks et al. (1996) contains a related discussion on the social status of dictionaries.

Being a field of study that deals with vocabulary, which is a social deposit consisting of actual and concrete lexical items, lexicology naturally attains a concrete nature in itself; lexicology should start from a concrete set of vocabulary, and, even when it talks about the possible states of vocabulary, it should always be bound to its actual state. Although which “lexical” items belong to the vocabulary and which do not cannot be decided fully, the concreteness of the vocabulary is the basic starting point of lexicology (Maeda 1989).

This contrasts with the study of word formation or word forms. The question of whether, for instance, “a very large reddish-orange prize-winning pumpkin with manifold admirers” is a well-formed expression or not may be a relevant topic in the study of word formation but not in lexicology, as it is obvious that this expression does not in itself belong to the vocabulary and lexicology is concerned with the set of lexical items. We can postulate the point as follows: while the study of word formation and word forms can be concerned with the acceptable form of words, just as grammar or syntax is concerned with the acceptable form of sentences, the study of lexicology and terminology is necessarily concerned with the existing set of lexical items and, even when it talks about the possible expansion of vocabulary or terminology, it is concerned with the realistic possibility of existence, rather than the abstract well-formedness of certain putative individual words or lexical constructions thought up by a linguist.

### 1.1.3 Vocabulary and utterances

In the field of natural language processing, there are many studies devoted to the extraction of lexical items or information related to lexical items from texts, such as automatic monolingual term recognition and extraction (e.g. Daille 2001; Frantz and Ananiadou 1999; Heid 1999; Jacquemin 2001; Nakagawa 2001; Yoshida and Nakagawa 2005), bilingual or multilingual term extraction (e.g. Bernhard 2006; Daille and Morin 2005; Fung 1995; Gaussier 1998; Morin et al. 2007) or thesaurus construction (e.g. Bourigault and Jacquemin 1999; Fox et al. 1988; Grabar and Zweigenbaum 2004; Grefenstette 1994; Lin 1998; Tsuji and Kageura 2006). Although this study deals with terminology as a set and does not deal with the usage of lexical items in texts, it is convenient here to summarise briefly the relationship between vocabulary and texts, in order to clarify the position of the present work.

Figure 1.1 shows the basic relationship between vocabulary and utterances. Although it portrays a cyclical interaction between the vocabulary of a language
and the set of utterances, further elaboration is needed in relation to the theoretical and practical status of vocabulary and texts. For simplicity, let us exclude the personal or individual dimension.

Firstly, when we talk about “the vocabulary of a language,” we are referring to a social deposit from which, together with grammar and other linguistic and extra-linguistic factors, not only existing utterances have been produced but also potential utterances can be produced within the idiosyncratic state of that particular language. For instance, if we observe utterances, it would most probably be noted that the words “man” and “woman” are not used symmetrically, with the only difference being that of sex. Their use is likely to reflect the social context that affects the actual utterances. But the lexical meanings of these words, at their core, are identical to the sex difference, as can be seen in most dictionaries. Within the sphere of the set of utterances in Figure 1.1, therefore, potential utterances and existing utterances should be distinguished.

It is on the basis of this existing set of utterances that the vocabulary of a language is updated. In this process, lexical items newly incorporated into the vocabulary are bound to the actual, not the potential, utterances (the lexical items that become obsolete are also bound to the actual utterances, this time negatively). The moment a lexical item is incorporated into the vocabulary and is being used to produce utterances, however, it becomes distanced from its immediate context as established by the original set of utterances. The vocabulary thus maintains its independence from actual utterances and its correspondence to potential utterances. The change in vocabulary, compared to the change in the set of actual utterances, is much slower; vocabulary and utterances have a quite different time scale. Note also, as we argued above, that the vocabulary cannot be divided into actual and