TERMINOLOGY and LEXICOGRAPHY
RESEARCH and PRACTICE 15

The Quantitative Analysis of the Dynamics and Structure of Terminologies

Kyo Kageura

John Benjamins Publishing Company

The Quantitative Analysis of the Dynamics and Structure of Terminologies

Kyo Kageura
University of Tokyo

John Benjamins Publishing Company
Amsterdam / Philadelphia

Library of Congress Cataloging-in-Publication Data

Kageura, Kyo, 1964-
The Quantitative Analysis of the Dynamics and Structure of Terminologies /
Kyo Kageura, University of Tokyo.
p. cm. (Terminology and Lexicography Research and Practice, issn 1388-8455 ; v. 15)
Includes bibliographical references and indexes
1. Lexicography--Data processing. 2. Terms and phrases--Data processing.
3. Linguistics--Statistical methods. 4. Computational linguistics. I. Title.
P327.5.D37K34 2012
401'.4--dc23 2012039713
isbn 978 90 272 2339 5 (Hb ; alk. paper)
isbn 978 90 272 7246 1 (Eb)

John Benjamins Publishing Co. · P.O. Box 36224 · 1020 me Amsterdam · The Netherlands
John Benjamins North America · P.O. Box 27519 · Philadelphia pa 19118-0519 · usa

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.

Volume 15

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

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, Computerm) 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,[1] 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.[2]

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.

Table of contents

Part II. Distributional dynamics