



TERMINOLOGY and LEXICOGRAPHY
RESEARCH and PRACTICE 21

The Corporate Terminologist

Kara Warburton

John Benjamins Publishing Company

The Corporate Terminologist

Terminology and Lexicography Research and Practice (TLRP)

ISSN 1388-8455

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.

For an overview of all books published in this series, please see benjamins.com/catalog/tlrp

Editors

Marie-Claude L'Homme
University of Montreal

Kyo Kageura
University of Tokyo

Volume 21

The Corporate Terminologist
by Kara Warburton

The Corporate Terminologist

Kara Warburton

University of Illinois at Urbana-Champaign

Copy Editor

Emma Warburton

John Benjamins Publishing Company

Amsterdam / Philadelphia



The paper used in this publication meets the minimum requirements of the American National Standard for Information Sciences – Permanence of Paper for Printed Library Materials, ANSI Z39.48-1984.

DOI 10.1075/tlrp.21

Cataloging-in-Publication Data available from Library of Congress:
LCCN 2020057218 (PRINT) / 2020057219 (E-BOOK)

ISBN 978 90 272 0849 1 (HB)
ISBN 978 90 272 6009 3 (E-BOOK)

© 2021 – John Benjamins B.V.

No part of this book may be reproduced in any form, by print, photoprint, microfilm, or any other means, without written permission from the publisher.

John Benjamins Publishing Company · <https://benjamins.com>

Table of contents

List of figures	XI
List of tables	XIII
Glossary	XV
Typographical conventions	XIX
Preface	XXI
PART 1	
Foundations of terminology	1
CHAPTER 1	
What is terminology management?	3
Terminology and terminography	3
Terminology management	8
CHAPTER 2	
Theories and methods	11
Theories	11
Onomasiology and semasiology	14
Thematic versus ad-hoc terminography	17
Prescriptive and descriptive terminography	18
Reflections on theory and practice	20
CHAPTER 3	
Principles	21
Univocity	21
Concept orientation	23
Term autonomy	26
Data granularity, elementarity and integrity	27
Repurposability	28
Interchange	29
Data categories	30

PART 2

Commercial terminography 33

CHAPTER 4

Definition, motivation, challenges 35

The commercial environment 35

Does commercial content contain terminology? 36

Motivation for managing terminology 42

The historical confines to translation 46

The terminologist as a working professional 47

The advent of XML 49

Lack of suitable models 50

The value of corpora 52

Terminology problems and challenges 53

Relevant literature 57

CHAPTER 5

Terms in commercial content 63

Terms considered by word class 63

Terms considered by length 65

Proper nouns 66

Variants 67

CHAPTER 6

Applications 73

Where can terminology be used? 73

Content management 74

Translation 77

Authoring 78

Search 81

Extended applications 89

CHAPTER 7

Towards a theoretical framework 93

Statement of the problem 93

Termhood and unithood 94

Microcontent 101

Elements of a new theory and methodology 105

PART 3

Planning a corporate terminology initiative	109
--	------------

CHAPTER 8

The proposal	111
---------------------	------------

Organizational position	111
-------------------------	-----

Standards and best practices	113
------------------------------	-----

Users and their roles	117
-----------------------	-----

Stakeholder engagement	120
------------------------	-----

The authoring community	121
-------------------------	-----

Business case	123
---------------	-----

Implementation plan	128
---------------------	-----

Approval	129
----------	-----

CHAPTER 9

The process	131
--------------------	------------

Access mechanisms and user interfaces	131
---------------------------------------	-----

Stages and workflows	134
----------------------	-----

The terminology audit	135
-----------------------	-----

Inclusion criteria	137
--------------------	-----

CHAPTER 10

Data category selection	141
--------------------------------	------------

Computer-assisted translation	142
-------------------------------	-----

Controlled authoring	145
----------------------	-----

Concept relations	151
-------------------	-----

Search	154
--------	-----

Subsetting	156
------------	-----

Data category proposal	158
------------------------	-----

CHAPTER 11

The terminology management system	163
--	------------

Standalone or integrated	163
--------------------------	-----

Core features	164
---------------	-----

Languages and scripts	168
-----------------------	-----

Term entry	169
------------	-----

Import and export	170
-------------------	-----

Views	172
-------	-----

Search	173
--------	-----

Access controls	174
Relations	175
Workflows, community input	176
Administrative functions	177
 PART 4	
Implementing and operating the termbase	179
 CHAPTER 12	
Create the termbase	181
The data model	181
Controlling access	183
Views and filters	184
Workflows	185
 CHAPTER 13	
Launch the termbase	191
Initial population	191
Beta test	195
Launch	196
Documentation and training	197
Community outreach	198
 CHAPTER 14	
Expand the termbase	201
Term extraction	201
Concordancing	206
Target language terms	210
New concepts	211
 CHAPTER 15	
Maintain quality	215
The termbase-corpus gap	215
Field content	222
Backups	225
Leveraging opportunities	226

Conclusion and future prospects	229
Further reading and resources	231
Bibliography	235
Index	245

List of figures

Figure 1.	Terminological entry and lexicological entry	6
Figure 2.	Lexicological entry	6
Figure 3.	Terminological entry	7
Figure 4.	Sign, signifier and signified	22
Figure 5.	Two terms in one term field	27
Figure 6.	Percentage of enterprise respondents who realize benefits from managing terminology	59
Figure 7.	Unfettered variants	70
Figure 8.	Activities that are part of content management	75
Figure 9.	Concepts within a sentence	76
Figure 10.	Enterprise search	82
Figure 11.	Query correction in Google	84
Figure 12.	Autocomplete	85
Figure 13.	Autocomplete without company terminology	86
Figure 14.	Faceted search without company terminology	87
Figure 15.	Termhood and unithood	95
Figure 16.	Word cloud	102
Figure 17.	Marketing slogans	102
Figure 18.	Some Microsoft® applications	103
Figure 19.	Concept: hybrid cloud	104
Figure 20.	Game server – key concepts	105
Figure 21.	Organizational placement of the terminology initiative	113
Figure 22.	A standardized term from ISO	114
Figure 23.	Example of ITS markup	122
Figure 24.	Percentage of respondents collecting specific types of information	142
Figure 25.	A synset containing two terms	146
Figure 26.	English synset showing usage status	147
Figure 27.	Active CA showing recommended term	147
Figure 28.	Concept map	152
Figure 29.	SEO metadata in a term entry	155
Figure 30.	SEO metadata in a term entry	155
Figure 31.	A multi-level hierarchy of subject fields	157
Figure 32.	Structure of a terminological entry	182

Figure 33.	Basic process-oriented workflow	188
Figure 34.	CSV file	192
Figure 35.	Specifying the separator	193
Figure 36.	CSV file after import	193
Figure 37.	Entry in TBX format	194
Figure 38.	An unstructured glossary entry	194
Figure 39.	Cell containing two terms	195
Figure 40.	Partial list of the word patterns extracted by TernoStat from a sample corpus	204
Figure 41.	A concordance from TernoStat	206
Figure 42.	High ranking unigrams with and without comparison to a reference corpus	208

List of tables

Table 1.	Lexicology and terminology	5
Table 2.	Onomasiology and semasiology	14
Table 3.	Prescriptive and descriptive terminography	19
Table 4.	Usage status values used in CA	146
Table 5.	Semantic relations	152
Table 6.	Concept-level data categories	159
Table 7.	Language-level data categories	159
Table 8.	Term-level data categories	160
Table 9.	Core design / Data model	165
Table 10.	Data points	168
Table 11.	Integration	168
Table 12.	Term entry	169
Table 13.	Import	170
Table 14.	Export	171
Table 15.	Views	172
Table 16.	Search	173
Table 17.	Access controls	174
Table 18.	Relations	175
Table 19.	Workflows, community input	176
Table 20.	Administrative functions	177
Table 21.	Frequency of multiword terms before and after boundary adjustments	219
Table 22.	Frequency of terms before and after adjustments	219
Table 23.	Common errors in termbases	222

Glossary

The following terms are used in this book with the meaning indicated herein. Readers are also directed to the ISO Online Browsing Platform (www.iso.org/obp) for additional terms and definitions.

ATE

See *automated term extraction*.

automated term extraction (ATE)

The identification of terms in a corpus through automated means.

CA

See *controlled authoring*.

CAT

See *computer-assisted translation*.

computer-assisted translation (CAT)

Translation carried out on a computer using a dedicated software. Most CAT software tools include a translation memory (TM), a termbase, file filters, and quality assurance checks, among other features. Also called *computer-aided translation*.

concept entry

An entry in a termbase. It describes one and only one concept, and includes all terms that denote that concept, as well as information about the concept (such as a definition), and information about the terms (such as their part of speech). Also called *terminological entry*.

concept orientation

Principle whereby an entry in a termbase describes one and only one concept.

content model

The type of content that is permitted in a termbase field, such as free text or pick-list.

controlled authoring (CA)

Authoring carried out according to rules of grammar, style, and vocabulary. It is usually done with dedicated software.

corpus

A body of texts. Plural: *corpora*.

Darwin Information Typing Architecture (DITA)

An XML-based, end-to-end architecture for authoring, producing, and delivering readable information as discrete, typed topics. (Oasis DITA FAQ).

data category

Specification of a type of terminological data that is used for structuring terminological entries or terminology resources. (ISO 1087:2019). Examples of data categories include definitions, parts of speech, usage notes, subject fields, and so forth.

data model

The structure of a termbase entry and the specification of permissible data categories, their content model, and their placement in the entry structure.

DITA

See *Darwin Information Typing Architecture*.

entry

See *concept entry*.

General Theory of Terminology (GTT)

The classical theory of terminology considered by some to be the original theory of terminology, which is focused on normalization, conceptual description, concept systems, and onomasiological approaches. For more information, see *Theories*.

GTT

See *General Theory of Terminology*.

Internationalization Tag Set (ITS)

A set of attributes and elements that provide internationalization and localization support in XML documents.

ITS

See *Internationalization Tag Set*.

language for general purposes (LGP)

Natural language used in everyday communication about non-specialized topics.

language for special purposes (LSP)

Natural language used in communication between experts in a domain and characterized by the use of specific linguistic means of expression. (ISO 1087:2019).

LGP

See *language for general purposes*.

LSP

See *language for special purposes*.

microcontent

Instance of very small content. Terms are a form of microcontent.

MWT

See *multiword term*.

multiword term (MWT)

A term comprised of more than one word, for example: *nuclear power plant*.

Natural Language Processing (NLP)

Field covering knowledge and techniques involved in the processing of linguistic data by a computer. (ISO 24613-1:2019).

NLP

See *Natural Language Processing*.

search engine optimization (SEO)

The use of techniques to raise the ranking of a website when it is searched using particular search keywords.

search keyword

A word or words people use to search for content on the world wide web.

SEO

See *search engine optimization*.

SL

See *source language*.

source language (SL)

Language serving as a starting point in translation work or in a search for term equivalents in another language.

target language (TL)

Language of a translation (document) or of an equivalent for a term existing in a source language.

TBX

See *TermBase eXchange*.

TEI

See *Text Encoding Initiative*.

term autonomy

The principle whereby all terms in a concept entry are considered independent sub-units and can be described using the same set of data categories. (ISO 26162-1:2019).

term candidate

A linguistic unit extracted by a term extraction tool, prior to its validation.

term extraction

See *automated term extraction*.

termbase

See *terminology database*.

TermBase eXchange (TBX)

The International Standard (ISO 30043) for representing and exchanging terminological information in an XML-compliant format.

termhood

The degree to which it is determined that a linguistic unit justifies being included in a termbase due to its domain-specificity or some other reason.

terminological entry

See *concept entry*.

terminological resource

See *terminology resource*.

terminology database (termbase)

A database that contains concept entries.

terminology management system (TMS)

Specially-designed software for managing terminology databases.

terminology resource

Any information medium, document, file, etc., that contains terminology and/or information about terminology. Often this term refers to a terminology database, but it can also refer to the various output products of a terminology database, such as a glossary, dictionary for a CAT or CA tool, list of terms, and so forth. Also called *terminological resource*.

Text Encoding Initiative (TEI)

A consortium which collectively develops and maintains a standard for the representation of texts in digital form.

TL

See *target language*.

TMS

See *terminology management system*.

TM

See *translation memory*.

TMX

See *Translation Memory eXchange*.

translation memory (TM)

Electronic collection of source language and target language segment pairs.

Translation Memory eXchange (TMX)

An XML specification designed for the exchange of translation memory data between CAT tools.

Typographical conventions

The following typographical conventions are used in this book:

- Examples of terms are in italics, e.g. *smart phone*
- Italics are also used for emphasis, for instance, when a concept is first introduced
- Field labels in a terminology management system or other software are in bold, e.g. **Part of speech**
- Permissible values of fields are shown in italics, e.g. *noun*
- Quotations from other sources are in quotation marks

Note: To avoid the awkward plural form of acronyms, such as “TMSs” for “terminology management systems,” we have chosen to keep the singular form even when the acronym is used in a plural sense.

Preface

This book is intended for anyone who is interested in learning about how to manage terminology to support global communication in the modern era. The focus is on best practices for managing terminology in the private rather than public sector, although many of the principles and guidelines described are applicable to both. It addresses employees of companies or institutions who are taking on the role, formally or informally, of a *terminologist* – that is to say, they have been asked to manage terminology for the organization. They may have other titles, such as writer, information developer, translator, software localizer, translation project manager, and content manager. They may have no background in or knowledge of terminology management.

This book is also addressed to seasoned terminologists who do have experience and/or a formal education in terminology management. Although they may already be familiar with the traditional theories and principles of terminology, we challenge the validity and applicability of some of those theories and principles for enterprise-scale terminology management. They will gain some fresh ideas about how to further optimize the terminological resources that they are responsible for.

The principles and methods described in this book may not only help corporate terminologists do their jobs, but also help them to exceed expectations and demonstrate their true value to their employers.

Key decision makers in an enterprise's global content management strategy are particularly encouraged to read this book, since terminology management is a cornerstone of such a strategy. Any executive who wants to explore how to leverage language resources to benefit translation, authoring, and even enterprise search, will find this book useful.

If you think that managing terminology does not concern you, before you put this book down, we encourage you to read *Does commercial content contain terminology?*, *Motivation for managing terminology*, and *Microcontent*. In these sections, we demonstrate how terminology in commercial settings is quite different from terminology as is conventionally viewed in academic circles, and even by many terminologists themselves. Terminology in commercial settings is *microcontent* which manifests in different ways and can benefit various business applications. Managing this microcontent is critical for commercial enterprises and other organizations, not only to support their success in multilingual global markets,

but also to optimize communications, content retrieval, customer satisfaction, and other core requirements even within a single-language market.

Fundamental principles and methods are introduced at the elementary level, so no prior knowledge of terminology management is necessary. However, some topics are not covered in depth. Furthermore, since the focus is on terminology management in commercial settings, the broader field of terminology management is not addressed comprehensively. Thus having some knowledge of terminology management would be an advantage to readers. To address any gaps, we provide suggestions for continued learning in *Further reading and resources*.

In more than 25 years of experience as a terminologist working in various commercial settings,¹ I have struggled with how to apply the established theories, methodologies, and approaches – which as a classically trained terminologist I wholeheartedly support – for creating and managing terminology resources in companies. I quickly realized that many of those theories, methodologies, and approaches are unsuitable for commercial settings. After searching at length, I also discovered that there are no documented practices specifically for developing terminological resources, i.e. terminology databases, for commercial applications.

With guidelines for managing terminology in commercial settings generally lacking in the literature, I noticed that terminologists working in companies often turn to the ISO committee that sets standards in this area – Technical Committee 37 – assuming that its prescribed methods should be comprehensive and apply globally. And so I became involved in ISO TC37, hoping that those methods would help me too, and that perhaps I could contribute in a small way to their development. However, the terminology standards and specifications produced by ISO TC37 are focused on terminology harmonization and standardization in a multilingual (translation-oriented) context, and for the most part they further the principles and methods of General Theory of Terminology (GTT) (to be explained later). This should not be a surprise, given the historical background of the committee. But through trial and error, I discovered that many of those traditional principles and methods do not work well in commercial environments. In a company, translation is only one of many potential applications of terminology resources, and some of the principles behind the GTT are often unjustified, difficult to apply and sometimes even contrary to commercial goals.

Terminology resources developed to support business goals and requirements need to be *multipurpose*, but the mainstream body of knowledge in terminology tends to lack this vision. This is a recurring theme in this book.

To manage terminology in commercial settings, one needs to be prepared to diverge from convention. When the opportunity presented itself, I challenged

1. 35 years in the terminology field if education is counted.

the applicability of the ISO TC37 tenets to practical terminology management in industrial and commercial settings and was met with opposition. Documented best practices for managing terminology in commercial environments simply do not exist. And so after nearly 20 years of being actively involved in ISO TC37, and more than 25 years working as a terminologist in commercial settings, I decided to write this book.

Our aim is not, however, to resolve the disconnect between established theories and methods and the practical needs of companies. Rather, the aim is to present those theories and methods with a critical eye, and to relate them somehow to the day-to-day challenges faced by corporate terminologists. In doing so we hope to improve understanding about the dynamic environment in which corporate terminologists work and the manifold opportunities that it offers.

It is difficult to find the right term to characterize the type of terminologist for whom the principles and practices described herein would resonate. Later in this book we will see a clear distinction between the principles and practices that are necessary and effective for developing terminology resources that serve public interests, such as those developed by governments to serve the citizenry, and those that are necessary and effective for supporting more institutional, production-oriented goals. This book addresses the latter. Those goals are common in any institution, organization, company, enterprise, industry, and so forth, that is concerned with producing content at large scales and in multiple languages in an efficient and cost-effective manner. But in this digital era, their interests go even beyond that to include concerns such as content retrieval (search engine optimization) and content management.

For lack of a better term, we have chosen *corporate* for this sector. It should be understood, however, that other types of large organizations, such as non-governmental organizations and professional associations, can also apply the principles and methods described in this book.

To place our propositions in the wider historical and professional context of terminology as a discipline and practice, Part 1 begins by introducing the key notions of terminology and terminography, and reviews some of the classical theories and methodologies for managing terminology. In Part II we move on to the commercial environment, its unique challenges, and how terminology work can benefit various commercial processes and applications. We critically redefine what a term really is and introduce the more suitable notion of *microcontent*. To help readers appreciate why managing terminology is necessary, we present some examples of what can happen when terminology is not managed. We conclude this look into the commercial world with some proposals of how classical theory and methods could be adapted to meet the demands of commercial stakeholders.

We then cover some key principles, many adopted to some degree from the classical approach, that should be respected by the corporate terminologist.

Part III covers the preparation and planning phases of a corporate terminology management initiative. We start with the proposal itself, which requires building relationships with management and stakeholders, developing and presenting a business case, and identifying users and their roles. A terminology audit is conducted in an attempt to identify which types of information the terminology database should contain. During this phase, the terminology management system needs to be selected, so we present a detailed description of the various functions to consider.

Part IV is the implementation phase, followed by continuous operation. Topics covered include designing and developing the termbase and its various objects (views, filters, etc.) and of course adding terms. Regular communications and feedback ensure that stakeholders remain engaged and management is satisfied. The termbase needs to grow continuously while its quality and integrity are maintained. Complementary tools such as for term extraction and concordancing are essential for both of these objectives. Finally, we emphasize that the terminologist needs to continuously seek opportunities to leverage the terminology resources in new ways.

The final few pages synthesize what we feel are the most important take-aways from this book. We also share some thoughts on what the future might hold for corporate terminologists.

PART 1

Foundations of terminology

Before we can begin to elaborate on terminographical practices that are most efficient for commercial settings, we need to ensure that we have a good understanding of the basic concepts, theories, methods, and principles that have shaped terminology as a practice and a discipline.

What is terminology management?

Terminology and terminography

This section presents a few conventionally-accepted definitions of what *term* and *terminology* mean, relevant to the field and practice of terminology management.

There are various definitions of *terminology* reflecting different theoretical views and perspectives.² In general, *terminology* refers to either the scholarly discipline (related to linguistics, lexicology, translation, and other language-related fields), or a set of terms, such as the terminology of the internet, which includes the terms *blog*, *cookie*, *cache*, *pop-up*, *internet service provider*, its acronym *ISP*, and so forth. The International Organization for Standardization (ISO) distinguishes between the two, where *terminology* refers to the set of terms, and *terminology science* is defined as a “science studying terminologies, aspects of terminology work, the resulting terminology resources, and terminological data” (ISO 1087:2019).

What about the actual practice of “managing” these sets of terms? What do we call the work that terminologists actually do? As will be shown later, their main (but not only) task is creating and maintaining *terminology resources*. Terminology resources are, according to ISO 1087:2019, “collections of terminological entries” which can be in paper format, such as printed glossaries and dictionaries. Nowadays, terminology resources more often exist in the form of terminology databases, also known as *termbases*. ISO refers to the work centering on termbases as *terminography* which it defines as “terminology work aimed at creating and maintaining terminology resources” (ISO 1087:2019).

Corporate terminologists do more than just create and manage termbases. They are engaged in a broader range of responsibilities including planning, budgeting, designing, staffing, ensuring quality control, and integrating the terminology resources into extended corporate applications and processes. ISO 1087 uses the broader terms *terminology work* and *terminology management* for “work concerned with the systematic collection, description, processing and presentation of concepts and their designations.” Some of these activities, such as term collection, do not involve termbases directly. Furthermore, it elaborates in a note that “Terminology work includes term extraction, concept harmonization, term harmonization and

2. See Cabré, 1996, for a detailed description of the meanings of *terminology*.

terminography” (our emphasis). Wright and Budin (1997) define *terminology management* broadly as “any deliberate manipulation of terminological information”, and they define *terminography* as “the recording, processing, and presentation of terminological data acquired by terminological research.”

Thus, *terminology work* and *terminology management*,³ which are synonymous according to ISO, are broader in scope than *terminography*. They include *terminography* (the work with termbases) among other tasks. But those tasks as listed in the ISO note seem to be confined to activities that are directly related to the terminological data itself, such as collecting, describing, and presenting concepts and their designations. Other tasks and responsibilities mentioned above that relate more to the overall coordination of a corporate-wide terminology initiative (planning, staffing, integration, etc.) are overlooked, at least in these formal ISO definitions. By not acknowledging the full scope of responsibilities that corporate terminologists often shoulder, the ISO definition of *terminology management* is inaccurate. Interestingly, ISO does not provide any definition for *terminologist*, the name of the actual profession.

Returning to terminology as a science, according to convention, it is concerned with *terms*, which are distinguished from other members of the lexis (i.e. *words*) because they are confined to distinct *subject fields*.⁴ ISO 704:2009 defines *term* as “a designation consisting of one or more words representing a general concept in a special language in a specific subject field”.⁵

The distinction between *terms* and *words* is fundamental to understanding what terminology as a practice is all about. Terminology’s closest related discipline, *lexicology*, is concerned with the whole lexicon (i.e. all the words) of a language. Lexicology is defined in the Oxford dictionary as “that branch of knowledge which treats of words, their form, history, and meaning.” Thus the difference between terminology and lexicology is that the former is confined to the lexicon of subject fields (fields of special knowledge), and the latter is concerned with the general lexicon of a language. Later we will see how this distinction, while universally recognized by scholars, has proven difficult to apply in commercial environments.

Lexicologists are researchers and scholars in the field of lexicology. The practical work associated with lexicology, referred to as *lexicography*, is the prepara-

3. We prefer the term *terminology management* as it sounds more professional.

4. See, for example, Sager 1990, Wright et al 1997, Rondeau 1981, Cabré 1999-b, Rey 1995, Kageura 2015.

5. Here, *general concept* refers to a class or category of objects, such as *island*, as opposed to an *individual concept*, such as *Prince Edward Island*.

tion of dictionaries. Lexicographers prepare dictionaries that represent the whole lexicon of a language.

One can make a similar distinction between theory and practice in the field of terminology management. Terminology is increasingly recognized as a distinct scholarly discipline (Rey 1995: 50; Dubuc 1992: 3; Budin 2001: 16; Cabré 1996: 16; Santos and Costa 2015), although this recognition is not universal (see Condamines 1995: 225; Cabré 2000; Sager 1990; Kageura 1995 and 2015; Packeiser 2009: 36). It has its own set of theories and research interests, distinct from those of lexicology, and indeed dedicated courses in terminology management have been offered at a handful of universities for some time now. Terminologists then would be the experts in this scholarly discipline. As lexicography is the practice of creating dictionaries, *terminography*, as we noted above, can be considered as the practice of recording terminologies (often in electronic form in databases) and this work is carried out by *terminographers*.

Table 1. Lexicology and terminology

	Lexicology	Terminology
Scholar	lexicologist	terminologist
Practice	lexicography	terminography
Practitioner	lexicographer	terminographer
Product	dictionary	terminology database
Focus	words	concepts

In a dictionary, the information provided about a word, such as the part of speech, etymology, pronunciation, and definition, is contained in a structure called an *entry*. Similarly, the information about a term that is contained in a terminology database is also called an entry. But there is one fundamental difference. In a dictionary, which is the product of lexicology, entries are *word oriented*. In terminology, the entries are *concept oriented*. Concept orientation will be further described later.

Figure 1 shows these two divergent perspectives.

The lexicological entry describes all the meanings of a word; often they are numbered (1), (2), etc. Since many words are polysemic, many of the entries in dictionaries describe multiple concepts. In contrast the terminological entry describes only one concept, but it also includes all the words used to denote that concept, i.e. synonyms. Most dictionaries, on the other hand, do not include synonyms in the entry (those that do are referred to as *thesauri*).

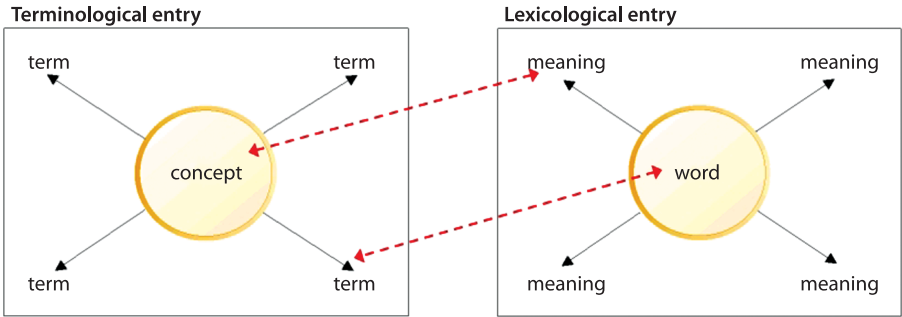


Figure 1. Terminological entry and lexicological entry

Figure 2 shows a typical dictionary-like entry.⁶

balance [bal-uhns] [SHOW IPA](#) [🔊](#)

[SEE SYNONYMS FOR balance ON THESAURUS.COM](#)

noun

- 1 a state of equilibrium or equipoise; equal distribution of weight, amount, etc.
- 2 something used to produce equilibrium; counterpoise.
- 3 mental steadiness or emotional stability; habit of calm behavior, judgment, etc.
- 4 a state of bodily equilibrium:
He lost his balance and fell down the stairs.
- 5 an instrument for determining weight, typically by the equilibrium of a bar with a fulcrum at the center, from each end of which is suspended a scale or pan, one holding an object of known weight, and the other holding the object to be weighed.
- 6 the remainder or rest:
He carried what he could and left the balance for his brother to bring.
- 7 the power or ability to decide an outcome by throwing one's strength, influence, support, or the like, to one side or the other.
- 8 (in winemaking) the degree to which all the attributes of a wine are in harmony, with none either too prominent or deficient.

Figure 2. Lexicological entry

Here, it is clear that the focus is on the word and its various meanings. In this example, eight different meanings of the term *balance* (noun) are listed.⁷ Note also that for synonyms the site provides a link to a thesaurus.

6. Source: dictionary.com.

7. In the actual entry that this example is sourced from, there are 17 different meanings described, all of which could not all fit into a screenshot.

Figure 3 shows a sample terminological entry of the same term *balance*.⁸

Domain	• Finance and Private Sector Development ◦ Finance
Definition	The amount of money present in a financial repository, such as a savings or checking account, at any given moment.

 English	
account balance	
balance on account	
balance	

 Russian	 
остаток на счете	
баланс счета	
сальдо	
результат	
остаток	

Figure 3. Terminological entry

Here, it is clear that the meaning is restricted to a financial one. Note that there are three terms (synonyms) in English and three in Russian.⁹ The entry describes only one meaning and includes all the terms that denote that meaning.

There are significant differences in methodology which have helped to further distinguish lexicology and terminology.¹⁰ The methods used by terminographers to investigate and manage terminologies are, according to established theory, clearly different from those used by lexicographers to describe the lexicon (Cabr  1999-b:37–38 and 1996:25; Sager 1990:3). In practice however, terminographers adopt some methods characteristic of lexicography, of which many are particularly relevant for managing terminology in commercial environments. These methodological differences will be further explored later.

8. Source: Interword Technology AB.

9. The actual entry that this example is sourced from contains five Russian terms, but they could not fit into this screenshot. It also contains more information about the terms, but that information was hidden to enable a reasonably-sized screenshot.

10. For a deeper discussion of the differences between terminology and lexicology, see: Cabr  1996 and 1999, Rey 1995, Riggs 1989, Dubuc 1997, Kageura 2015, and L’Homme 2006.

Today, the term *terminography*, while still unfamiliar to many, is starting to be recognized to refer to the work involved in managing terminological resources (i.e. termbases). However, the term *terminographer* has not gained wide acceptance as compared to its counterpart *lexicographer*; it occurs in scholarly literature but is rarely used in practice. The person who manages terminology is almost always referred to as a terminologist, we believe because the theoretical development and methodological practice have a short history and are therefore less mature than their counterparts lexicology and lexicography. There are many more lexicographers in the world than there are terminographers, many more dictionaries than termbases.

In summary, we use the term *terminology management* in a broad sense, to include terminography (working with termbases), the use of various other tools (such as concordancers and term extraction tools, which will be discussed later), and overseeing the entire corporate terminology initiative including its integration into extended corporate systems. These and other responsibilities merge to create the fascinating role of the corporate terminologist.

Terminology management

There are differing opinions about what is involved in managing terminology. *ISO 704:2009 – Terminology work – Principles and methods*, one of the most frequently cited ISO standards governing terminology management, “establishes the basic principles and methods for preparing and compiling terminologies.” Its major topics include how to identify and characterize concepts, create concept systems, write definitions, classify designations, and create new terms. This reference provides a solid foundation in principles and methods that reflect the mainstream *General Theory of Terminology*, principles that are followed by many terminologists: “The aim of terminological work is to define concepts and to make explicit the semantic relations among them in concept systems with a view to standardizing and thereby optimizing specialized communication” (Schubert 2011: 27). By confining the goal to *standardization*, Schubert also confines its definition to the perspective advocated by the General Theory of Terminology. As such, it fails to recognize the unique nature of terminological work that is carried out in companies and other production-oriented settings, where productivity gains and cost savings in the development of multilingual content are the main goals. In these environments, the quality of commercial content and information is measured in terms of ease of use (comprehension), optimal retrievability (through search), and ease of translation, characteristics that are often achieved through the use of a range of tools and processes, including properly

managed termbases. But formal *standardization* is not a goal in commercial environments, most commercial termbases do not include many semantic relations (although that is regrettable, as will be shown later), and concept systems are almost never elaborated. To say that terminology management involves defining concepts, making semantic relations, and creating concept systems, all with a standardization focus, does not reflect the reality faced by corporate terminologists.

For L'Homme, terminography comprises the various types of terminology work that would be carried out by a terminologist (2004: 45). She defines it as “the collection of activities the objective of which is to describe terms in specialized dictionaries or term banks” (2004: 21, our translation). She classifies these activities into the following areas:

1. Preparing a corpus.
2. Identifying terms.
3. Collecting information about the terms (from the corpus, and from reference materials).
4. Analyzing and organizing the information.
5. Encoding and further organizing the information in a database.
6. Managing the terminological data: adding, deleting, or correcting data in a termbase.

This view recognizes the importance of the tasks of corpus building, term identification (which often involves term extraction tools), and terminology research, much of which occurs outside of a termbase. The LISA study of industry-based terminology management includes a similar range of activities, all oriented towards the goal of improving the use of terminology in the organization (Warburton 2001a: 4). In this book, *terminography* has been adopted to refer to the applied side of terminology, i.e. terminology work, including aspects such as manipulation of corpora, term identification, and distribution activities in addition to working directly in a termbase.

Theories and methods

To appreciate the somewhat uncertain framework that corporate terminologists work in today, it is necessary to have a reasonable understanding of the theoretical foundations of the field and the methodologies that have evolved to support those theories. In this chapter we provide an overview of the various theories that are shaping terminology as a discipline and terminography as a practice.

Theories

Although there were some previous efforts to elaborate and structure the nomenclatures of specialized domains such as in the pure sciences, what is widely acknowledged as the original *theory* of terminology was developed in Vienna by Eugen Wüster in the middle of the twentieth century, culminating in 1979 in his treatise the *Introduction to the General Theory of Terminology and Terminological Lexicography*.¹¹ The collection of Wüster's ideas became known as the *General Theory of Terminology (GTT)*.¹² Several alternative theories have emerged in recent decades, and their legacies have broadened our understanding of terminology as a scholarly discipline and practice (see Temmerman 2000, Cabré 2003, L'Homme 2004 and 2019, Faber 2011, Warburton 2014). However, to this day the GTT has dominated mainstream knowledge (Bourigault and Jacquemin 2000: section 9.2.2). In this and subsequent sections, we will therefore focus on the GTT, but we will also provide some insight into how the other theories might contribute to a better understanding about terminology management in commercial environments.

Wüster was an electrical engineer with no formal background in a language-related field. However, as an engineer operating in the multilingual European context, he became aware of the challenges of multilingual communication in specialized fields. Cross-lingual and even intra-lingual communication is impeded, particularly in specialized fields, when multiple terms within one language are

11. For a fairly comprehensive account of the history of terminology as a discipline, see Rondeau 1981.

12. The GTT is also known as the *Traditional Theory* and the *Vienna School of Terminology*.

used to express the same concept. He rightly felt that communication among specialists would be facilitated, and translation made easier, if each concept was designated by only one term in a given language, that is, if terms could be standardized. Synonymy was a problem and should be minimized. He realized that the lexicographical methods used to produce dictionaries were inadequate for documenting and managing potentially many designations of concepts. His seminal dictionary *The Machine tool*, published in 1968, included an alphabetical index for handy lookup purposes, but the main content comprised a systematically-arranged, concept-oriented layout that effectively showed all synonyms in multiple languages as well as illustrations. And with a single view of all terms used in all languages for a given concept, choosing and mandating a *standard* term, Wüster's interest, became possible.

Through his substantial life's work, Wüster contributed significantly to the foundations of terminology science and terminology standardization as they are known today.

According to the GTT, totally clear communication, a goal of specialized language, can be achieved by establishing a one-to-one relationship between terms and concepts. This theory therefore favors *biunivocity*, whereby a linguistic form corresponds to one concept only, and a concept is expressed by only one linguistic form (L'Homme 2004: 27). The focus of study is the concept, to which terms are secondarily assigned as designators (Cabr  2003: 166–167; L'Homme 2005: 1114). Concepts occupy fixed positions in a universal, language-independent concept system, where they are hierarchically related to other concepts. The approach is *onomasiological*: identify and define the concept first, then afterwards find terms to denote it. The goal is normalization (standardization) (L'Homme 2005: 1114–1115; Cabr  1996: 25). Terms are designations of objects which are conceptualized systematically. A terminology (set of terms) must correspond to a conceptual system (Rey 1995: 140).

Since the mid 1990s, these and other tenets of the GTT have been subject to some criticism (L'Homme 2005: 1115; Packeiser 2009: 29). The main critique is that the GTT does not take into account authentic language in dynamic use (see Pearson 1998, Temmerman 1997 and 2000, L'Homme 2004, Cabr  1999-b, Collet 2004, Packeiser 2009). In failing to do so, it lacks empirical foundation: "Since the tenets put forward by traditional terminology do not equip the term with features that result in the required behavior in discourse, it can be concluded that these tenets are not borne out by empirical data" (Collet 2004: 102). Packeiser (2009) extends the lack of empirical foundation to some of the basic tenets of the GTT, such as the unique relationship between terms and concepts as distinct from the relationship between words and their meaning: "The principles of concept and term are not explained scientifically." (17) and "The general theory of terminol-

ogy does not provide a single theoretical explanation of the formal relationship between concept and terms.” (35). Concepts are studied outside of their use in communicative settings. Another critique is that the GTT is focused on terminology standardization, which restricts its applicability to other aims. “The principles and methods of Traditional Terminology coincide with the principles and methods for the standardisation of terminology. Traditional standardisation-oriented Terminology should widen its scope” (Temmerman 2000: 37).

In questioning some of the basic tenets of the GTT, a number of alternative views on terminology theory and methodology emerged. These include the Socio-Cognitive Theory (Temmerman), the Lexico-Semantic Theory (L’Homme), the Textual Terminology Theory (Bourigault and Slodzian), Frame-based Terminology (Faber et al, 2005) and the Communicative Theory (Cabré). These theories present various challenges to the GTT, and they particularly dispel the universality of conceptual systems (here, universality refers to language-independence, in other words, concepts are extra-linguistic). They stress, respectively, the role of human experience in formulating concepts, the lexical properties and structures of terms as members of the lexical system, the importance of corpora for discovering and validating terminological units, the notion from frame-semantics that entire conceptual networks form a frame of understanding, and the role of pragmatic conditions and communicative situations in creating or establishing terms. These theories, if one can call them so, maintain that the communicative role of terms, the purpose of the communication, the textual environment, and the wider conceptual framework can contextually-condition meaning. These new perspectives have practical and positive repercussions for commercial terminography.

Granted, during Wüster’s time, there were no computers, no documents in digital form, no corpora, and no Internet – elements of modern society that we now take for granted. Taking into account authentic language in use is much easier with these tools and resources. Although it failed to represent terminology as a fully independent discipline (Packer 2009: 15, 82), to be fair, the GTT laid a solid foundation towards this goal by establishing principles and methods that allowed terminology to be distinguished from lexicology. The subsequent theories, whether they validate or invalidate the GTT, certainly further that distinction.

In the following sections, we will demonstrate that terminography today, particularly in commercial environments, is increasingly text-driven, semasiological, descriptive, and ad-hoc. These properties run counter to some of the founding principles of the GTT, which suggests that commercial terminography has, in response to technical demands and expanded uses, evolved into a different kind of terminology work. We will also present additional tenets of traditional theory and discuss to what degree they apply to commercial terminography.

Onomasiology and semasiology

The GTT significantly influenced methodology. As previously mentioned, according to the GTT, the terminologist identifies, delimits, and defines concepts first while building and analyzing the broader concept system into which the concepts can be conveniently classified. Only then can the terminologist identify the designators (terms) that denote those concepts. Term delimitation is therefore based on concept delimitation. This approach, referred to as *onomasiological*, is one of the basic tenets of the GTT. It contrasts with the *semasiological* approach, which is used in lexicography. Lexicographers describe words and their various meanings. Terminologists, according to the GTT, evaluate and describe the meaning (concept) first and then name it with a term secondarily.

Table 2. Onomasiology and semasiology

	Onomasiology	Semasiology
Question to answer	How do you describe object X? What are the properties of object X? How do you define the concept that corresponds to the object X? What are the names for the concept X?	What does the word Y mean?
Starting point	object, concept	word
Objective	Find names for the concept	Find meanings of the word
Processes	Analyze concept features. Create hierarchical concept diagrams. Distinguish related concepts. Write definitions according to strict rules. Determine (and possibly eliminate) synonyms.	Discover instances in use. Document usage. Identify and analyze usage. Observe textual environment.

The onomasiological approach produces a *concept system* in the form of a diagram that shows how concepts are related. This is why the onomasiological approach is favored for creating “knowledge representations” (Santos and Costa 2015). The diagram frequently includes an indication of each concept’s *delimiting characteristics*. A delimiting characteristic is one that distinguishes a concept from related concepts. Producing such a diagram requires careful analysis of concept properties. The ability to organize concepts into such a hierarchical system is based on the premise that concepts are universal, i.e. language independent. In other words, it should be possible to adorn each box in the diagram with a *term* in any and all languages. Accordingly, research about the concepts should be carried

out in a way that avoids a language bias. The use of trusted resources in any one particular language to the exclusion of others is thus discouraged.

And yet the universality of concepts has been disputed and even empirically disproven with myriad examples from socially- or culturally-conditioned fields such as law, cuisine, and education. Temmerman's Socio-Cognitive theory of terminology, which is based on cognitive semantics and prototype theory, holds that onomasiological concept description out-of-context is insufficient for describing many conceptual categories and the terms that denote them (1998, 2000). She provides examples from the domain of life sciences of concepts that do not have a universal "prototype". "The traditional Wüsterian analysis which postulates that concepts should be treated as if they existed objectively and independent of human understanding and language is misleading and in need of revision." (1998: 87).

For example, there is no standard or official equivalency between diplomas and degrees in France and the United States, and this is an impediment to bilateral cooperation in education. Consequently, students wishing to study abroad often face bureaucratic obstacles to obtain formal recognition of their academic credentials and to gain acceptance into foreign education programs. Concepts in corporate content frequently denote unique products and services, some of which are difficult to transfer to other cultures. The term *transcreation* is more accurate than *translation* for describing translations that require cultural adaptations, as is frequently the case for marketing content. Translations of a web page promoting the Watch, for example, show additions, deletions, and replacements for concepts relating to the use of the watch for tracking exercise and making purchases.

In commercial communications, most if not all of the content is created first in one source language (SL), and then translated into multiple target languages (TL). It is difficult to avoid the language bias in such conditions. Terminology work does not start *before* the content is written, when ideas are formulating, as that would be an unattainable ideal. Terms thus already exist in the SL across the company's materials, with all their imperfections as the case may be (errors, inconsistencies, proliferation of synonyms, etc.). All the terminologist can hope to do is to help minimize problems in the SL, such as excessive synonymy (too many different terms used for the same thing) or unmotivated variants (these and other problems are further discussed in *Terminology problems and challenges*), and support translators in determining TL equivalents. Terminologists in such environments must therefore adopt a semasiological approach: collect terms that are already in use, check how they are used (via textual evidence), identify problems, and assist authors and translators to minimize those problems.

Terminographers in general, but particularly those who work in commercial environments (which are not strictly normative), usually adopt a semasiological

approach to their work in practice: first they identify a term in a text, next they figure out what it means and how to fully document it in a termbase, and eventually they determine how to translate it. Acknowledging this fact is a hard pill to swallow, since doing so is a direct challenge to the GTT and an affront to its elevated status. But the exclusivity of the onomasiological approach in terminography has already been challenged in the literature. There are cases where the onomasiological approach cannot be applied in terminology work (Packeiser 2009:17). Some scholars even acknowledge that the semasiological approach is *more common* in terminology work (Cabr  1999, Sager 2001, Temmerman 1998 and 2000, L'Homme 2005, Rey 1995, Bowker 2003). To quote Temmerman (p. 230),

Even though – in practice – terminographers have always started from understanding as they had to rely on textual material for their terminological analysis, one of the principles of traditional Terminology¹³ required them to (artificially) pretend that they were starting from concepts.

The rapid development of computing technologies has no doubt contributed to a general shift towards a textual approach to terminography, through the availability of corpora in digital form and natural language processing (NLP) tools enabling their manipulation. Indeed, the increased use of computers for managing terminology has contributed to narrowing the gap between terminographic and lexicographic methods because of easier access to large-scale corpora, which has given a more important role to the linguistic context of terms (Cabr  1999-b:163; Sager 1990:136).

Nonetheless, the onomasiological approach is not completely irrelevant in commercial terminography. Occasionally it is necessary to organize a family of related terms into a concept system, and carefully analyze and contrast the concepts in order to establish clear terminology in both the SL and the TL. Due to the time and effort involved, this work is only justified in cases where there is a business impact, such as when, due to confusing terminology, customers are calling the support line, placing orders for the wrong product, or unable to complete an order placement. In one software company, for example, the names of the various *fixes* that a customer could download from the company's support website were inconsistent and confusing. They varied depending on the product they were related to, the features they contained, and the support program they were associated with. Web content developers and technical support needed guidance from the terminologist about these names. In cases like this, an onomasiological evaluation of the concept system is necessary. Another example is the names of various types of computer virus: trojans, worms, spyware, ransomware, and more.

13. Temmerman uses the term *traditional Terminology* to refer to the GTT.

Consumers cannot be expected to understand the difference between these various threats, and how to best protect themselves, if the terms are confusing, used inconsistently, and not adequately explained. Creating a concept system in the onomasiological sense can therefore be helpful to clarify and distinguish a set of important, related terms.

Thematic versus ad-hoc terminography

To study concepts onomasiologically and build a concept system is an approach that is referred to as *thematic* (L'Homme 2004: 45; Meyer and Mackintosh 1996: 280) or *systematic* (Cabr  1999-b: 129, 151; Dobrina 2015: 180). This approach takes the view that terms and their underlying concepts are members of a system, and therefore their existence and meaning can only be confirmed in comparison to other members (terms/concepts) in the same system. According to the GTT, and reinforced by the onomasiological approach, concepts can only be studied systematically, that is, as members of the logical and coherent concept system of the subject field (Rondeau 1981: 21; Rey 1995: 145). In practical situations, such as when a translator or a writer needs to decide what term to use, what a term means, and so forth, this approach is rarely followed. "Every practitioner knows that such a method is highly artificial" (Rey 1995: 45). Faber (2011: 11) even observes that creating concept systems is not a standard procedure in developing terminological resources in general: "The explicit representation of conceptual organization does not appear to have an important role in the elaboration of terminological resources." More often, a task- and text-driven approach is adopted, whereby an individual term is studied in the context of its use. This approach is referred to as *ad-hoc* (Wright et al 1997: 147; Cabr  1999-b: 129, 152) or *punctual* (Cabr  2003: 175; Picht and Draskau 1985: 162; Alcina 2009: 7).

Thematic or systematic terminography aims to document the terms used in a subject field or particular topic area comprehensively. Some have even elevated this approach to terminography in general: "the cornerstone of terminography is a detailed analysis of a domain's conceptual structures" (Meyer and Mackintosh 2000: 134). It is often conducted in the form of a project (Dobrina 2015: 180). The terminologist studies the field thoroughly, using reliable resources, and often consults subject-matter experts. Determining term equivalents in different languages requires a comparable field study, using resources that were written originally in those languages (i.e. not translated documents), and sometimes consultation of subject-matter experts who work in those languages and geographical markets. Due to the effort involved, this type of work is undertaken only when justified. It

is more typical of research environments such as university studies or publicly-funded language planning efforts undertaken by governments.

Ad-hoc terminography aims to solve an immediate need or problem, usually involving only one concept at a time. Solutions are usually found quite quickly by doing an internet search or checking a few documents. Subject-matter experts are only consulted when the available resources do not provide a definitive answer.

In commercial terminography, there is yet another scenario that is neither ad-hoc nor thematic: large scale term extraction. Term extraction, to be covered later in this book, is the process whereby a large number of terms are extracted from a corpus with the aid of a computer program. This is not ad-hoc, since many terms are involved. Nor is it strictly thematic, since the corpus used typically is not confined to a particular topic area or subject field, but rather corresponds to a collection of company documents that were hastily assembled without following the formal corpus compilation methods advocated by corpus linguistics. Due to the limitations of the tools used, the extracted terms never represent the system of concepts comprehensively. Lastly, the work actually done with the extracted terms is minimal, more closely reflecting ad-hoc methods. In the field of terminology, no name has yet been proposed for this type of work as differentiated from ad-hoc and systematic.

Prescriptive and descriptive terminography

It should be understood that the onomasiological approach served the primary goal of the GTT: standardizing terms. Evaluating the features of concepts and establishing concept systems enabled the terminologist to also establish and prescribe normative labels (terms) for those concepts, in multiple languages. A working approach that aims to standardize or prescribe terms (and deprecate others) is referred to as *prescriptive*. The goal is to accept and promote some terms, and reject or discourage others, in the interests of clear and unambiguous communication. The GTT is essentially prescriptive in nature (Faber and Rodríguez 2012: 12). Yet today, except in the strictest normative content types such as legal documents, standards produced by ISO and other such organizations, and official policies enacted by governments, terminology standardization is not the primary motivation for developing and managing terminological resources. This fact will be further discussed in *Motivation for managing terminology*.

In contrast to prescriptive terminography, *descriptive* terminography aims to describe the terms that are used in a particular environment, and avoids making recommendations about what constitutes proper, correct, or even permitted term usage. The goal is to collect and document as many terms as possible. The

approach aims to develop an inventory of the terms and other expressions used in a particular socio-linguistic environment or community of practice. It is therefore typically adopted for researching and documenting languages, a sub-field known as *socioterminology*.

Table 3. Prescriptive and descriptive terminography

	Prescriptive	Descriptive
Perspective	Judgmental	Observational
Objective	Control usage of terms. Eliminate or reduce the use of synonyms.	Describe the terms that people use. Document synonyms.
Processes	Rank synonyms according to preference. Provide usage notes.	Demonstrate usage with contextual (corpus) evidence.

Commercial terminography adopts a hybrid mixture of descriptive and “loosely” prescriptive approaches. On the one hand, as will be shown later, the termbase serves multiple purposes in the organization, each of which requires different types of terms. When it comes to deciding what lexical units should go into the termbase, a comprehensive, inclusive approach is therefore required. Terms are collected in large quantities, given a cursory check, and imported into the termbase, with, at best, often only a context sentence. This is characteristic of the descriptive approach. On the other hand, most companies are interested in developing content that adheres to certain corporate guidelines; they want employees to use “approved” terms. Regulating language in the discourse of corporate communications is aimed at increasing clarity, avoiding errors, and improving general efficiency. But it is not standardization in the formal Wüsterian sense, as it allows deviations from those regulations in the form of linguistic variety, where justified by the communicative situation and context. “Language must have *regulating instruments*, but not necessarily standardization ones” (Santos and Costa, 2015, our emphasis).

Identifying problems and helping writers and translators minimize those problems requires some prescriptive techniques, such as ranking synonyms according to a company’s preference. Imposing certain terminology choices is often done with the assistance of software for controlled authoring (CA) and computer-assisted translation (CAT). That is characteristic of the prescriptive approach. One can generally observe that most information in corporate termbases is descriptive and a minor proportion is prescriptive.

Reflections on theory and practice

Terminologists working in commercial settings need to realize that the classical theories and methodologies for terminography are frequently at odds with practical needs. While not abandoning tradition altogether, they should use common sense when determining the best approach for a given situation. Some principles and methods inherited from classical theory work well for modern corporate terminography, while others less so.

For example, the principle of concept orientation, which will be described later, is one of the greatest contributions that the GTT has made to the field of terminology management and it should be strictly adhered to. Corporate termbases that fail to uphold this principle are seriously flawed.

On the other hand, according to the GTT's thematic approach and onomasiological focus, terminologists should research concepts, identify delimiting characteristics, structure them in hierarchical concept systems, and write definitions according to strict rules. Corporate terminologists rarely perform these tasks because often they are impractical and unnecessary in production environments.

The contrasting ad-hoc, semasiological approach is the norm in corporate terminography. This approach produces terminological resources that are largely descriptive. Nevertheless, prescriptive methods are used for applications such as CA and to some degree CAT.

The availability of sophisticated computer applications and large-scale corpora in digital format have brought changes to terminographical methods that are beneficial in commercial environments, and commercial terminologists need to embrace these developments.

Principles

In this chapter, we describe some of the key principles for managing terminology and for developing termbases. Termbases developed according to these principles are more repurposable, which is one of the objectives of corporate terminography. The principles described in this chapter are, therefore, applicable in commercial environments.

Univocity

Many if not most words in a language can have multiple meanings, a property known as *polysemy*. But one of the tenets of the GTT is that a term shall be univocal, meaning that it can have only one meaning. This is referred to as the *univocity* principle. How do we explain this, given polysemy in language?

In general language, words can have different meanings because they occur in different application areas: a *bridge* in dentistry is different than *bridge* the card game or the *bridge* on a violin, for example. But in terminology, the view of a terminological unit is application- or domain-specific: each of the three cases of *bridge* is considered a distinct term. In the following paragraphs we attempt to explain the relationship between a term's surface form (its written form), its meaning, and its domain-specificity. The domain-specificity of terms helps to establish their univocity; it confines their meaning to a semantically-constrained framework.

A term's meaning is a fundamental part of its identity as a term. The notions of *signifier*, *signified*, and *sign*, taken from Saussurian linguistics (De Saussure 1916), can help to demonstrate this point. The signifier is the auditory (sound) or written (text) representation of the concept. The signified is the concept that the signifier evokes in our mind. Together, the signifier and the signified form the sign. The sign, the signifier and the signified are in a bound relationship.

A term is a sign in the Saussurian sense; it is the bound relationship between a word form and *one* meaning of that word form. The relationship between the word form and the concept for a particular sign is monoreferential; this means that a term has one and only one meaning, it is, indeed, univocal. Two terms might *look* the same, because they have identical signifiers, but if they have different meanings (signifieds) then they are actually different terms. In Figure 4, *dog*

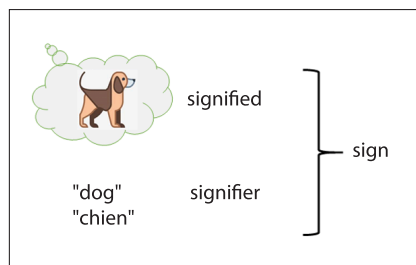


Figure 4. Sign, signifier and signified

the domestic mammal is a term. But *dog* in the field of engineering, i.e. “a tool or part of a tool that prevents movement or imparts movement by offering physical obstruction or engagement of some kind,”¹⁴ is yet another term. This interpretation is in line with that of Rondeau (1981: 22–23), Dubuc (1992: 26) and other classic terminology theorists.

Another example will help to further demonstrate this key principle. The word *port* has multiple meanings in different subject fields: (1) wine-making (the strong wine), (2) computer hardware (the connection port), (3) computer software (a verb meaning to adapt a software program so that it can run on a different operating system), (4) marine industry (a maritime facility with wharves), and (5) sailing (the left side of the boat), to mention just a few. Each one of these binary instances, the spelling p-o-r-t associated with one of those meanings, is a term. What we have here is five terms not one.

Terminology and terminography focus on the sign as their object of study. Each sign is a term – a binary association between a word form and a meaning. This accounts for the term *bridge* in dentistry and the term *bridge* in music. The word *bridge* in itself is not a term, it is only a signifier. It becomes a term when it is attached to a meaning in a *language for special purposes (LSP)* such as music or dentistry. There are probably dozens of different terms in the English language that have the signifier *bridge*, as there are terms that have the signifier *dog* and that refer to something other than man’s best friend.

Depecker (2015) describes the relationships between the conceptual and the linguistic sides of terms. Based on the 2000 version of ISO 1087, which has since been updated, he uses the term *designation* to refer to the linguistic part of the signifier-signified relationship. “A term is made of a designation associated with a concept.” But the binary relationship necessary in a term’s identity remains.

By viewing the univocity principle from within the confines of specific LSPs or subject fields we can affirm that terms are univocal. Indeed, most proponents

14. Wikipedia.

of the GTT agree that the univocity principle applies *within* an LSP. The bridge in dentistry is one term, the bridge on a violin is another term, and so forth.

A problem occurs when terminologists and other people contributing to a termbase perceive the signifier as the term. This leads them to include more than one meaning in the termbase entry, which violates concept orientation, an essential property of termbases (this principle is described in the next section). Unfortunately, this occurs quite often, usually because the individuals concerned have not had any training or background in terminology management, and thus they rely on their general knowledge of how dictionaries are structured to guide their work. But dictionaries focus on the signifier, not the sign. What these individuals produce is lexicography, not terminography, and the termbase degrades in quality and in its ability to fulfill its purposes.

It is important for terminologists to understand the univocity principle and how it sets semantic boundaries for what a term actually is. Henceforth, in this book, when we use the term *term*, we are usually referring to the *sign*... a mapping between a lexical unit in written form and one of its meanings.

Concept orientation

Concept orientation refers to the principle whereby a terminological entry describes one concept only and includes all the terms that convey that concept. This is why terminological entries are frequently called *concept entries*. If a concept can be expressed by more than one term, as is the case with synonyms, spelling variants, abbreviations, and equivalents in different languages, all those terms must be included in the entry, in all languages. Concept orientation differs from the word-based structure used in dictionaries (sometimes called lexical orientation or word orientation), in which an entry describes all the meanings of a word (see Figure 1). Concept orientation provides a single view of synonyms, variants, and equivalents in different languages for a given concept, and this is why it is fundamental for managing multilingual terminology. Concept orientation is a principle of the GTT and an ISO standard for terminology databases (see ISO 16642, ISO 30042 and ISO 26162).

The principle of concept orientation means that in the interface of the terminology management software, a separate field for synonyms, labeled as such, is not needed, and the presence of one is in fact considered poor practice. A term is recognized as being a synonym of another term by virtue of the fact that the two terms are found in the same termbase entry. A synonym is merely a term recorded in the same entry as another term... that it is a synonym need not be

explicitly stated. Likewise, a term is recognized as a translation¹⁵ of another term by virtue of the fact that the two terms are in the same entry, but in two different languages. These practices are reflected in Figure 3, where the entry shows terms but no labels such as “Synonym” or “Translation.” Because of this principle, terminological entries are language neutral and multidirectional. This means that from the perspective of translation activities, in a termbase any language can play the role of the source language and any other language the target.

It should be emphasized once more that each terminological entry corresponds to only one concept. Since words (not terms, refer to *Univocity*) can be polysemic, we need to ensure that we do not describe those different meanings in the same entry, as this violates concept orientation. Instead, we need to create separate entries for each meaning. The fact of the matter is that each meaning is likely to require different terms in other languages. For example, depending on the meaning, which is constrained by subject field, the French equivalent of *dog* can be *chien* (domestic mammal), *toc* (metallurgy), *agrafe* (fastener), *croc* (forestry), *crabot* (mechanical engineering), *clameau* (building construction), *taquet* (mining) and more. This is the reason why concept orientation is essential for managing multilingual terminology and supporting the needs of translators.

We will later demonstrate that concept orientation is also critical for developing terminological resources that are multi-purpose. Polysemy and homonymy occur in commercial texts and are rarely a problem (provided that they are not excessively widespread). However, in conjunction with terminology’s focus on concepts, the univocity principle motivated the terminographic practice of concept orientation for terminological resources. This concept-based view of terms helps the company “gain a handle” on polysemy and homonymy, so that they do not become a problem. Concept orientation therefore serves three key purposes of a termbase: translation, repurposability, and communication quality.

It is important to understand the difference between concept orientation and the onomasiological approach, as these two concepts can sometimes be confounded. Concept orientation refers to the structure of a terminological entry (centered on the concept) and the fact that it contains information describing only one concept. It serves to distinguish terminography from lexicography. The onomasiological approach, as described in *Onomasiology and semasiology*, refers to a working approach whereby terminologists study concepts first (by identifying their properties, differentiating features, relations to other concepts, and so forth) *before* identifying the terms that denote those concepts. Both concept orientation

15. We have used the word “translation” here for simplicity purposes, but to be precise, terms are not *translated*. Instead, *target language equivalents* are found. The various terms in a termbase entry that are expressed in different languages are semantic *equivalents*.

and the onomasiological approach are a legacy of the GTT. As we will explain later, in corporate terminography, concept orientation is a critical principle, and for that we are indebted to the theoretical foundations of terminology for providing a solid structural framework. But the onomasiological approach is impracticable in commercial settings, does not serve the many diverse aims of corporate terminography, and would deny the important potential of large-scale commercial corpora in digital format for the discovery of terms.

An example of repurposability is *controlled authoring* (CA). Controlled authoring refers to methods and procedures applied at the authoring stage to improve quality and consistency (this topic will be covered in more detail in *Authoring*). Concept orientation is an absolute necessity for termbases if they are to support CA processes and software. The software needs to tell users when they have used a “bad” term and what “good” term they should use in its place. Only by putting both the bad terms and good terms in the same termbase entry, which means adhering to concept orientation, can the software make this link.

Many company termbases erroneously contain redundant fields such as **Synonym** and **Translation**. Many also record synonyms in different entries, or multiple meanings of the same word in one entry. Finally, it is not uncommon to find terms documented in text fields, such as in a **Comment** field. This is very unfortunate, since terms documented in such text fields will never be retrieved by using the search function, which includes the autolookup function of a CAT tool or CA software. These and other errors violate the principle of concept orientation. We will discuss these and other common pitfalls later.

Figure 3 shows a basic concept-oriented entry.

Concept orientation is essential for developing multi-purpose terminological resources for companies. It enables the resources to be used for CA, controlled translation, SEO, and other purposes *at the same time*. This fundamental principle is the basis for termbase design and drives several practices such as the documentation and ranking of synonyms.

Concept orientation is a very important, fundamental principle. However, there are cases where it simply cannot be respected. One case in point occasionally occurs when the termbase is used to support translators. Sometimes there are multiple terms in the SL and multiple terms in the TL in the same entry – a common situation due to synonymy – and there is a need to ensure that for a specific SL term one particular TL term must always be used as its translation. For example, one entry contains terms A and B in English and terms C and D in Spanish. For the purpose of translating the company’s materials, if term C is the preferred translation of term A, and term D is the preferred translation of term B, then somehow this information needs to be conveyed to translators. By virtue of the concept orientation principle, all terms in the entry are “equal” and there is

no standard implementation for this type of pairing of specific terms within the entry.

It is difficult to control this mapping of terms within an entry by using existing terminology standards such as TBX (ISO 30042) and the DatCatInfo data category repository¹⁶ (these resources will be described later). The entry could include a note describing this term-mapping in a field called **Transfer comment** or **Note**, but this is awkward and not particularly effective in CAT tools. A solution that works very well in CAT tools is to put the term pairs AC and BD in two separate entries rather than all combined in a single entry. Indeed, in a comprehensive survey carried out among translators by Allard (2012), nearly 50 percent of respondents, in particular freelancers, adopted the approach to “organize their term records by equivalent pair rather than by concept.” This is a violation of concept orientation and should only be adopted when absolutely necessary. Interestingly, the figure was lower for translators working in an institutional setting, where 35 percent create separate records for each synonymous word pair. This suggests that translators working in institutional settings are more inclined to respect concept orientation due to its strategic importance.

Allard (2012:206) notes that this practice could be avoided if CAT tools included the functionality for establishing equivalent pair links within each record. We add that this functionality should also be available in all TMS.

Term autonomy

Term autonomy refers to the principle whereby all terms in an entry can be described with the same level of detail. This means that the same number and type of fields available to describe one term in an entry must also be available for any other term in the entry. It also means abiding by certain best practices when adding terms to the entry.

A common violation of term autonomy is when two terms are entered in one **Term** field. It is very important to enter each term separately. Figure 5 shows an entry that contains this type of error.¹⁷ Here, two terms have been entered in the **Term** field, i.e. *electrostatic discharge* and *ESD*:

16. datcatinfo.net

17. Source: Interverbum Technology AB

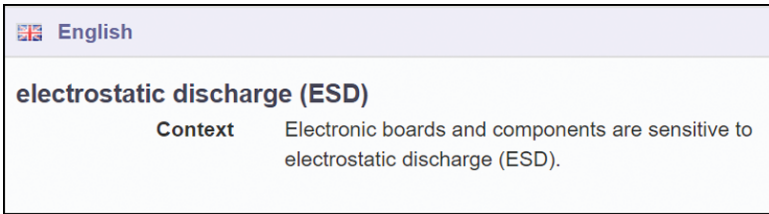


Figure 5. Two terms in one term field

Another violation of term autonomy occurs when terms that are considered incorrect or are not recommended for use in the organization's communications are excluded from the termbase. By “not recommended,” we are referring to cases of synonyms, where the organization has decided that one is recommended and the other is not, in an effort to drive consistency. Such not-recommended terms must be included in termbases so that users are informed that they are not recommended. A dedicated field, **Usage status**, can indicate that the term is not recommended. More information on this practice will be provided later.

Data granularity, elementarity and integrity

Data granularity is a data modeling principle according to which data is recorded in as granular (fine-grained) a way as possible, without of course adversely affecting usability by imposing over-granularity. A terminology management software that respects data granularity provides a separate storage container in the database (corresponding to a field in the interface) for each distinct type of information. For example, data granularity is violated when there is a field called **Grammar** for recording various kinds of grammatical information: part of speech, gender, transitivity, and so forth. Another example is when the software does not allow for the inclusion of distinct fields for recording the sources of information, forcing the user to indicate, for example, the source of a definition within the **Definition** field itself.

Data elementarity refers to the principle whereby a given database field contains only one type of information. This principle is violated, for instance, if the database contains only one text field for descriptive information. Users therefore include various types of information in this field, such as definitions, explanations, sample sentences (contexts), examples, and usage notes. Data elementarity and data granularity are closely related, almost synonymous concepts.

Data integrity is assured when the system automatically controls information that can be controlled. The use of limited-value fields (also called picklists)

wherever possible is one way to ensure data integrity. For instance, a domain (subject field) taxonomy can be set up as a limited-value field. This ensures that the user selects a subject field from a pre-defined list rather than typing it in by hand, which prevents different users from making typos, or entering the same value in different ways. An example of poor data integrity is when a text field is provided for recording the part of speech value. One user might enter *noun* while another may enter *n* or *n..* Such incompatible values cannot be used as filters for search or for export. Another form of data integrity is when the system automatically records administrative information such as dates, names (of users who have added/deleted/changed information), change logs (so that changes can be reversed), and so forth.

When designing a termbase, you must establish the appropriate type of data that can be entered in each field. Limited-value (picklist) fields, as opposed to free-text fields, are preferred because they increase data integrity and significantly improve the usability of the system in various ways:

- Choosing a value takes less time than typing one; productivity is improved.
- A default value can be set for the most common values, such as *noun* for **Part of speech**, saving data entry time.
- Limited-value fields can be used as search filters and export filters. This is a valuable feature as it enables the termbase to be a single source for a number of smaller focused terminology resources, such as for CAT and CA software. Using a limited-value field for domains (subject fields) enables the generation of domain-specific terminology resources, which are more effective for end-users in certain situations, such as interpreters.

Repurposability

Although not a key principle of terminology management in the conventional sense, *repurposability* is becoming increasingly important particularly in the digital age. Repurposability refers to the ability of the termbase to serve multiple purposes. As we will show later, terminology data stored in termbases comprises very small pieces of linguistic information that can, if properly set up, be used for a multitude of purposes in an enterprise. The conventional and still almost exclusive use of terminology resources has been to support translation work. We maintain that that vision is too narrow. Repurposability must be first and foremost in the terminologist's mind when designing a corporate termbase. It is *the* most important goal. Repurposability will be a recurring topic throughout this book.

Interchange

Interchange refers to the ability to move terminological data from a TMS into other systems or applications, and vice-versa. Interchange is an important consideration when deploying a terminology process in a company since it is the key to repurposability and therefore, to the return on investment.

The most common interchange scenario is when terminological data is exported from the TMS into a file, which is then imported into another application or system. Another is when terminological data is imported from a file into the TMS. The file acts as an intermediary medium for the data. This form of interchange is often used for exchanging data with both people, such as translators, and systems, such as a CA software.

There can also be a seamless, direct connection between the TMS and the target application or system, which eliminates the need for a physical file to act as an intermediary medium. This is typically achieved by means of an application programming interface (API). With the increasing adoption of the web as the hosting environment for many database applications, including termbases, and the maturity of standards for web protocols, web service architectures such as REST and SOAP are making these forms of exchange more commonplace.

Most TMS provide the ability to export the termbase to Excel format, or a compatible tab-delimited or character-delimited format which can be opened in Excel or another spreadsheet program such as Open Office Calc. Most people who work with terminology, including translators, are familiar with spreadsheets so this format has become very common. However, spreadsheets have limitations, which will be described in *Initial population*. Unfortunately, some of the simpler TMS do not support any other viable export format.

For file-based interchange to work, the format of the intermediary file needs to be sufficiently robust to handle all the types of information in the termbase, and sometimes spreadsheets fail in this regard. A more robust interchange format is *TermBase eXchange*, which is the XML markup language specifically designed for terminological data (ISO 30042, 2008, 2019). Enterprise-scale demands for interchange require the use of TBX. TBX is also recommended, above spreadsheets, for a termbase that contains entries that have more than a few languages and/or multiple terms within languages (synonyms). Spreadsheets can be very cumbersome in this case. Thus, corporate terminologists are well-advised to ensure that the TMS that they select for managing the termbase supports TBX as an import and export format.

More information about file-based importing and exporting will be provided in *Initial population*.

Data categories

Most termbases contain a wide range of information types to describe terms, such as definitions, usage notes, grammatical descriptors, and links between related terms. These types of information are called *data categories*. There are over a hundred different data categories possible for a termbase: the 2008 version of the TBX standard included 112. In 1999, ISO TC37 published an inventory of hundreds of data categories (ISO 12620). In 2009 this standard was revised and the inventory was moved into an electronic database, named DatCatInfo, which is now available on the Internet.¹⁸ This Data Category Repository (DCR), as it is also generically known, has expanded beyond terminology to include a wide range of linguistic domains (morphology, syntax, ontology, etc.) and thus has grown to include thousands of data categories.

For use in termbases, data categories are organized into three groups: conceptual, terminological, and administrative. Conceptual data categories describe concepts; they provide semantic information. Examples include definitions, and subject fields (e.g. *law* to designate a legal term). Terminological data categories describe terms, for example, usage notes, part of speech and context sentences. Administrative data categories include the name of the person who added some information or the date that it was added.

Certain data categories are commonly found in termbases, such as part of speech and definition. However, some applications require data categories that others do not. For example, CA requires usage status values (such as: *preferred*, *admitted*, *restricted*, *prohibited*) in order to guide writers in using terms that are recommended over others in accordance with company style. Some unique data categories may be required for machine translation (MT) systems, such as those pertaining to transitivity and inflections of verbs. CAT tools may require data categories for tracking stages in the translation review process, i.e. when the translation of a term is initially proposed or finally approved. Data categories for subsetting the termbase into practical sections, such as company departments or clients, are very useful as they allow filtering the termbase for search and export purposes. A repurposable termbase should be comprehensive, i.e. it needs to include the superset of all the data categories required for all potential uses.

Terminologists may find it difficult to navigate this quagmire of data categories. For that reason, a consortium of terminologists working in large organizations, TerminOrgs,¹⁹ developed a recommendation, *TBX-Basic*, which includes about 20 data categories commonly required in termbases, along with guidelines

18. datcatinfo.net

19. terminorgs.net

for their use. TerminOrgs also publishes other useful documents for terminologists, such as the *Terminology Starter Guide*.

TBX-Basic provides a good set of data categories to start with. Allard's study (2012:178) confirms that this set of data categories is also appropriate for developing terminological resources for translators, who are the primary user group for a corporate termbase. Nevertheless, caution should be exercised since this set of data categories is unlikely to suffice for a corporate termbase that is truly repurposable. The requirements for any particular corporate termbase should be based on an analysis of existing resources, processes, and needs in the company: authoring processes, translation processes, tools used, stakeholders, and more. These aspects are covered in Planning a corporate terminology initiative.

Data categories also have different kinds of *content models*, that is, the type of content allowed. There are three content models:

- open – allows any text and is used for fields that contain free text such as definitions and contexts
- constrained – restricted to a certain pattern or format, for instance, date fields
- limited value – (also called picklists) restricted to a set of pre-defined values. Used, for instance, for part of speech (noun, verb, adjective, and so forth), and for usage status (preferred, admitted, restricted, prohibited).

Another important field in a termbase is one that contains links, or *relations*, to other terms or other entries. Data categories for hierarchical relations such as broader and narrower terms exist in the DatCatInfo DCR and in TBX. However, how these data categories are implemented in a termbase, or in other words, what content model they adhere to, has never been formally described in any standard or guideline.²⁰ They could be viewed as the limited value type rather than open, since you can only link to a term or entry from among the existing terms or entries in the termbase. It could be that relations require a hybrid content model: both free text, such as to specify a URI, and limited value, to “pick up” a target term from among the available ones in the termbase. It is also possible that none of the three content models adequately implement relations.

The design of relational fields, which is somewhat challenging, has been left to the ingenuity of software designers and engineers. Only the more advanced TMS include them. One market-leading system is infamous for its failure to implement relational fields properly. In this system, the link to another entry is established based on the surface form (the signifier) of a term in the target entry. To create the link, in the linking dialog, the user types the term that is in that entry. Problems occur, however, when the termbase contains multiple entries that

20. As of the writing of this book, ISO TC37/SC3 is developing a guideline for relations.

feature this signifier, which is a common scenario due to the presence of homographs in most languages. In the system we describe, when there are two or more entries that contain the signifier of the term that the user wishes to link to, the system automatically establishes that link to the entry that has the lowest internal term ID, which is often not the correct one. This is rarely the correct one, meaning that the user has no real control over where the link is pointing. The issue is difficult to describe in simple terms... suffice it to say that this problem renders the relational fields of this system entirely dysfunctional. The issue is described here to demonstrate to corporate terminologists that relational fields should be carefully tested before a TMS is purchased.

More information about data categories is provided in Data category selection.

Commercial terminography

In the previous part we described the traditional theories, methods, and principles of terminology as a discipline and practice. We also raised some questions about their applicability to commercial terminography. In this part, we provide further justification for raising those questions by demonstrating the unique challenges of the commercial environment. In particular, we look closely the features, requirements, motivations, and challenges of managing terminology in commercial environments and describe the various applications where terminology is or can be used to support commercial activities. We conclude this part with some proposals for adapting conventional theories and methodologies to meet the needs of commercial terminography.

Definition, motivation, challenges

In this chapter we describe aspects of the commercial environment that influence how terminology work is conducted. We also prove that commercial content contains terminology, since that fact has been the subject of challenges and debate. The reasons why a company would choose to undertake a terminology management program are laid out.

The commercial environment

The terms *commercial*, *corporate*, *company*, *enterprise* or *corporation* are used more or less interchangeably to include any organization that produces content in a way that can be described as highly production-oriented: large in volume, time-constrained, driven by strategic organizational objectives, and usually multi-lingual. These organizations are usually but not necessarily profit centers, as they also include non-governmental organizations (NGOs) such as the World Bank and the World Health Organization.²¹

Public entities and governments, which are often driven by socio-linguistic concerns such as language planning, the preservation of minority languages, and citizen rights, do not have the same needs with respect to terminology resources and terminology management as production- and revenue-driven enterprises. Governments are less constrained by market factors which are so fundamental to commercial activities. Governments are national in scope whereas the companies that can benefit from terminology management are typically active multinationally if not globally. Governments were actually among the first organizations to actively manage terminology,²² and have therefore developed mature processes that fit public service well. Some of the practices recommended for corporate terminology are not appropriate in governmental settings.

The commercial environment presents the following characteristics:

- limited budgets which are tightly controlled
- time and labor constraints

21. Thus the word *commercial* is not perfectly suited, but a better alternative was not found, and most organizations that fit the profile are indeed commercial.

22. For example, the Government of Canada.

- large volumes of content
- extensive use of multi-media
- frequent staff changes
- global marketing vision
- product or service portfolio that is constantly changing and growing
- dramatic changes presented by mergers and acquisitions.

In the content production pipeline, various tools and technologies are already used to support tasks such as CA, CAT, content management, desktop publishing, translation management, terminology management, translation memory, concordancing, term extraction, SEO, accessibility testing, voice recognition and generation, machine translation, and quality assurance. Natural Language Processing (NLP) research is continually developing new technologies that could be used to further optimize content areas such as in content forensics and sentiment and opinion analysis. A company's knowledge assets are "curated" in *ontologies*.²³ Content professionals need to be prepared to work in these new areas.

In this context the use of technology is paramount to surviving market pressures. In fact, success requires innovation. The commercial workplace is highly technical and becoming more so. Companies see technical innovation as a key differentiator and competitive advantage. The commercial environment makes heavy use of technology to automate, or semi-automate, production processes, including authoring, translation, and publishing, in order to increase productivity, reduce costs, and shorten time-to-market. Market pressures are the reason why the private sector often adopts new technologies long before the public sector does.

Does commercial content contain terminology?

In *Theories* and *Terminology and terminography*, we presented the classic notion of what a term *is*. A key feature of that classic interpretation is that a term represents a concept that is confined to a subject field, the content of which is expressed in a language for special purposes (LSP). This semantic restriction is what distinguishes terms from words of the general lexicon, according to conventional thought. And as shown in *Terminology and terminography* it also serves to distinguish terminology and terminography from lexicology and lexicography.

23. See for example the SBVR specification produced by the Object Management Group: omg.org/spec/SBVR/About-SBVR/

After reviewing the literature, we can summarize the GTT notion of *term* as follows: terms are designations in languages for special purposes of concepts, which correspond to objects in the real world, and which can be classified systematically. This view certainly describes hierarchically organized designations of concrete concepts such as one finds in nomenclatures like the Linnaean biological taxonomy. It also suggests that terms are primarily nouns, which of course they are. Indeed, the GTT originated from the need to denote highly delineated concepts in science and technology, which are primarily nouns. However, we will demonstrate that if commercial termbases included only those linguistic units that meet the GTT definition of *term*, they would be too small to serve their intended purposes.

In the scholarly literature, there is a lack of consensus on the notion of *term* (L’Homme 2019: 55). Terms are, however, generally described as existing within the confines of LSP (ISO TC37; Cabré 1999-b: 32, 36, 79, 114; Picht and Draskau 1985: 97; Rondeau 1981: 21; Sager 1990: 19; Dubuc 1992: 3, 25, 26; Wright 1997: 13; Rey 1995: 95; Teubert 2005: 96; Meyer and Mackintosh, 2000: 111; L’Homme 2019: 57). This leads to some important questions. Does the language used in commercial settings constitute an LSP? Does it correspond to a subject field? Because if not, then perhaps it does not contain any terminology, and therefore, there is no terminology to manage, no need for corporate terminologists, and no need for this book. And if it does not contain terminology, then perhaps the methods of terminography do not apply, and we should be looking at lexicography for more suitable approaches.

The benefits of managing terminology in commercial settings are gaining recognition. “It is a common practice in companies and industries to maintain terms to promote efficient communication and to improve quality control” (Kageura 2015: 49). “Managing terminology” is now frequently talked about in commercial contexts. Yet these fundamental questions remain unanswered, which leaves commercial terminography without a solid theoretical foundation.

In the following paragraphs, we will demonstrate that commercial language does constitute an LSP and therefore contains terminology. Then later, in *Termhood and unithood*, we will challenge the classical notion of what a term is, at least, insofar as how this notion can be applied to commercial terminography. Our first approach is to examine the difference between LSP and LGP.

LSP are usually studied in contrast to the so-called general language, or *LGP* (*language for general purposes*) (for example, Picht and Draskau 1985: 1, Kocourek 1982: 31; Bowker and Pearson 2002: 25; Warburton 2015). Adherence to a subject field (also called a *domain*) is a key criterion that differentiates LSP from LGP (Rondeau 1981: 30; Sager 1990: 18; Kocourek 1982: 26; Kittredge and Lehrberger

1982:2; Rogers 2000:8; Bowker and Pearson 2002:25; Meyer and Mackintosh, 2000:134).

Thus, terms express concepts that can be classified into subject fields. The collection of terms and other linguistic expressions that form the language used in a subject field corresponds to the LSP of that subject field. The relationship between terms and subject fields is widely acknowledged in the literature (for example, Cabré 1999-b:9, 114 and 1996:17; Nagao 1994:399; Dubuc 1997:38, Kageura 2002:2, 12; Rogers 2000:4; Pearson 1998:36; Kageura 2015). Cabré (1999-b:81) claims that “the most salient distinguishing feature of terminology in comparison with the general language lexicon lies in the fact that it is used to designate concepts pertaining to special disciplines and activities,” and this view is echoed by other scholars (e.g. L’Homme 2004:64; Dubuc 1997:38; Wright 1997:13; Sager 1990:19). Thus, there is a consensus a defining feature of terms is their association with a subject field, and by extension an LSP.

General language, for its part, is the collection of words and expressions that do not refer to a specialized activity in the context in which they are used (Rondeau 1981, 26, translated). Bowker and Pearson describe LGP as “the language we use every day to talk about ordinary things in a variety of common situations” (2002:25). Bellert and Weingartner (1982) define everyday language simply as the language that does not satisfy the necessary conditions of scientific texts; this leaves the notion of general language open to quite a range of different interpretations. They also add that the background of the interlocutor requires only logical and “commonplace knowledge” (1982:229).

These fuzzy definitions of general language do not indicate that it does not contain terminology. Today, “ordinary things” can be quite technical and “common situations” highly specialized. How many among us non-specialists have had to navigate the confusing terminology of networks and routers as we call technical support when our wireless internet goes down? Would the myriad of technical features and functions of smart phones today *not* be “commonplace knowledge”? Most five-year-old children are conversant in this jargon. Moreover, Meyer and Mackintosh (2000) describe the phenomenon of *de-terminologization*, whereby “a lexical item that was once confined to a fixed meaning within a specialized domain is taken up in general language,” (p. 112) often with a shift in meaning. They even assert that “de-terminologization may have terminologizing effects” (p. 133). The language that “we use every day” surely includes our interaction with the Internet, which for many of us amounts to several hours daily. Does it not contain terminology?

Language as a whole can be seen as a system of sub-languages having different functional purposes (Kocourek 1982:14). Galisson and Coste (1976:583) speak of three sub-languages: everyday language, special language, and aesthetic language.

That was 45 years ago, before *all* that we know today. Is it possible there are still only three?

Rondeau adopts the term *technical or scientific communication* to refer to the corpora in which terms are found (1981: 16). He adds that these corpora – where terms are found – include the full range of both pure and applied sciences, and all techniques, technologies and specialized activities carried out by humans (crafts, professions, trades, occupations, hobbies, leisure activities, etc.). For Rondeau, an LSP is the language used in any communication that can be characterized as specialized in a broad sense. Cabré (1996: 22) shares this view, by including professional domains and industry among specialized fields where terms are found.

Following Rondeau, some scholars extend the notion of subject field to professional activities carried out in business, industry, companies, and professional settings (Cabré 1999-b: 35; Rey 1995: 139, 144; L'Homme 2019: 5). Rey acknowledges the existence of terminology for “practical situations” and mentions the “terminology of a firm” as a typical case (p. 144). Thus the notion of subject field, and consequently of LSP, is imprecise (Condamines 1995: 227).

LSP are also considered to have restricted linguistic properties. Cabré (1999-b: 61) defines LSP as “linguistic codes that differ from the general language and consist of specific rules and units.” Hoffman (1979: 16) refers to these codes as “linguistic phenomena,” while Picht and Draskau (1985: 3) speak of “a formalized and codified variety of language.” Both Sager (1990) and Rondeau (1981) also use linguistic and stylistic properties to characterize LSP. These include textual characteristics (concision, precision, depersonalization, economy, appropriateness, and referentiality), lexical patterns (especially preponderance of nominal structures), dominance of written form over verbal, frequency of figures, and so forth. Coincidentally these are also features of commercial language.

Other experts emphasize the importance of the communicative context for LSP. LSP occur “within a definite sphere of communication” (Hoffman 1979: 16). Cabré (1999-b: 63) refers to “interlocutors in a communicative situation.” She also points out that terms (and by extension LSP) need to be studied in the framework of specialized communication, and indeed terms do not even exist “prior to their usage in a specific communicative context” (2003: 188, 190). Pavel (1993: 21) observes that LSP are becoming less confined to communicative settings that are restricted to specialists and they now, thanks to mass media, extend to the public at large across the private and public sectors, including academia and industry.

LSP also differ from general language in communicative function. They are strictly informative (Cabré 1999-b: 68) to allow objective, precise, concise, and unambiguous exchange of information (Sager 1990: Section 4.2). This is, again, also characteristic of commercial content. In contrast, general language is evocative, persuasive, imaginative, and even deceptive (Cabré 1999-b: 74).

Cabré (1999-b: 63, 68) suggests that any type of text that varies from general language text can be considered a special language text, i.e. an instantiation of LSP. She identifies three common characteristics of special languages: limited number of users (which are defined by profession or expertise), a formal or professional communicative situation, and an informative function.

Another differentiating factor is how language competence is acquired. Some scholars believe that competence in an LSP requires effort over and above the innate knowledge we have of general language, as specialists in the domain all demonstrate. The use of an LSP “presupposes special education and is restricted to communication among specialists in the same or closely related fields” (Sager et al 1980: 69). That was 40 years ago. Today, anyone can become a pseudo-specialist in all kinds of techniques by watching a few YouTube videos. Picht and Draskau (1985: 11) support this view but they simply state that users acquire the LSP voluntarily. This is to account for communication acts of a more didactic nature between specialists and initiates, which are also rich in terminology. Autodidacticism, which is experiencing a surge thanks to technology, means that more and more people are doing precisely that today.

To summarize the previous paragraphs, the following properties of LSP have been cited in the literature:

- it is domain-specific
- it exhibits a closed set of linguistic properties (vocabulary, syntax, style, etc.)
- it is used in a specific communicative context for a specific communicative function
- it is consciously acquired.

According to these broadly-recognized properties, we maintain that the language used in most companies does indeed constitute an LSP, although few scholars state this explicitly. For example, while there is a consensus that an LSP is restricted to a subject field, the definition of subject field has broadened from a highly-structured objectivist hierarchy of science and technology to an experiential delimitation that is both context- and application-dependent. Even nearly 40 years ago, Rondeau (1981: Section 3.2.2), who espoused the GTT, acknowledged that subject fields span all areas of human activity and are not restricted to scientific and technical disciplines.

Companies operate in economic sectors that reflect different degrees of specialization; some are obviously more specialized than others. The texts produced in most companies describe tangible products, services, and activities, often within a limited number of industrial or economic sectors, which could be viewed as subject fields according to the latter’s broader interpretations. Commercial communication often adheres to specific linguistic rules and styles; many

companies have a style guide, and some are automatically implementing the style rules through CA software. The written form predominates. With a few notable exceptions such as marketing material, the information that companies produce is strictly informative, and often didactic. While the target audience of this information – often consumers – may not have undertaken special education, they are actively engaging with the goal of acquiring knowledge, services, or tangibles. Depending on its level of specialization and technical register, a company's informational content could therefore fit into one or another of the three communicative contexts that Pearson (1998: Section 1.9) describes for sub-languages: expert-to-initiates, relative expert-to-uninitiated, or teacher-pupil. A specifications sheet about a computer-assisted design (CAD) software is most likely to be read by someone already familiar with this technology, such as an architect (expert-to-initiate). A web page introducing a new smart phone is likely to be read by someone who knows a little bit about smart phones but nothing about this particular one (relative expert-to-uninitiated). An online help topic is showing the reader how to complete a task (teacher-pupil). Respectively, these different genres call for decreasing degrees of technical content (thus, terminology) and increasing simplicity of style.

Even though Rey recognizes the importance of clearly delineated scientific and technical subject fields as the defining feature of terms in the classical sense, he also accepts more pragmatic criteria. He proposes the term *terminological field* to designate the scope of the object of study in a terminographic project, “regardless whether the subject field is theoretical, thematic, a set of activities or a set of needs instigated by a professional group or even a particular firm, or the terminological content of a corpus of texts” (1995: 144–145). We suggest that terminological field resonates well for terminography in commercial settings.

The boundaries of subject fields and LSPs are not clearly defined, and even what constitutes an LSP is subject to ongoing debate. While the GTT views LSP as a vertical sub-language based purely on semantic differentiation (such as medicine or law), a number of scholars have characterized LSP in terms of stylistic conventions (precision, concision, etc.) and communicative function (professional, informative, etc.). Several even mention industrial and commercial texts specifically as examples of LSP. There is sufficient evidence to claim with confidence that the language used in a company exhibits properties of an LSP, and therefore it contains terminology.

The significant contribution of *de-terminologization*, mentioned earlier, to the vocabularies of commercial language suggests that there exists a degree of terminologization in the latter. Examples of de-terminologization abound in commercial language. For example, in the term *virtual employee*, *virtual* is adopted

from computer science. Other examples include *virus*,²⁴ *depression* (in economics), *domino effect*, and *static* and its antonym *dynamic* (from physics). At what point in this journey from special language to general language does a term cease to be a term? At what point is the term so general in meaning that it is no longer useful to proactively manage it as in a termbase? This point is undetermined.

As we move further into the age of the networked knowledge society, where consumer products for example become more and more technical, education even of children is delivered by computing means, the media is omnipresent, and tourists can book travel to outer space, the boundaries between LGP and LSP becomes more and more blurred. It would be short-sighted to adhere to the historical notion that terminology today is restricted to specialized technical and scientific fields.

We can now agree that commercial content contains terms, or at least lexical units, that if “managed” in a certain way can support a company’s interests in areas such as authoring, translation, and content retrieval. Later, we will attempt to resolve the apparent conflict between the classic definition of what a term is (i.e. based solely on specialized meaning), and the more pragmatic view that must apply to commercial terminography (i.e. based on purpose).

Motivation for managing terminology

In this section we present the main reasons why a company would decide to invest in a terminology management initiative. Ideas for creating a business case will be presented in *Business case*.

There are four main reasons why companies are interested in managing their terminology, and all are very pragmatic and business-focused:²⁵

1. increase employee productivity
2. reduce production costs
3. reduce time-to-market
4. increase customer satisfaction.

Increasing employee productivity reduces production costs and time-to-market at the same time. These three objectives can be more or less equated or are at least interdependent. Several studies about the benefits of managing terminology have

24. When de-terminologization passes through different disciplines, such as *virus* from medicine to computing, it is more accurately referred to as *trans-disciplinary borrowing*.

25. For a comprehensive evaluation of these and other factors that demonstrate the business case for managing terminology, refer to Schmitz and Straub, 2010.

emphasized time savings on the part of employees, particularly writers and translators (see for example Champagne 2004, Warburton 2001, Schmitz and Straub 2010). Employees who have access to a trusted source of company terminology in an easy-to-use lookup tool (nowadays usually a website) spend much less time researching when they have questions or doubts about a term or its usage than employees who lack such a resource and are left to search randomly across disparate sources of information. Time savings equate increased productivity, and the time saved can be applied to other high-priority tasks. Champagne estimates that providing translators with a trusted source of information about terms (via suitable resources and tools) can increase their productivity by as much as 20 percent. In a study conducted by SDL, a translation firm and CAT tool developer, a whopping 87 percent of translator respondents claimed that their productivity would improve with access to terminology resources (Hurst, 2009).

Consider the case of acronyms, which are widespread in most organizations. The same acronym used in any two organizations is likely to refer to different things; across an entire industry or between different industries the number of different meanings multiplies. In IBM, the term *APAR* is an Authorized Program Analysis Report, which usually results in the issuing of a *PTF*, or program temporary fix. In Thales Nederland, a Dutch manufacturer of naval defense systems, an *APAR* is an Active Phased Array Radar, while *PTF* refers to Precise Time Facilities. But even within the same organization, a given acronym can have multiple expanded forms. In the World Bank, the acronym *ADR* has four expansions: after debt rescheduling, American depositary receipt, asset depreciation range, and adverse drug reaction. The *UNEP* corresponds to United Nations Environment Program, but in one organization's corpus alone this term was found written ten different ways, alternating between British and American spelling, hyphenations, plurals, and other variations. A central, trusted reference for organizational terminology can shorten the time it takes people to demystify acronyms and all other types of terms and use them correctly and consistently.

For translators, the time savings and productivity gains are most evident. Today, most translators use computer-assisted translation (CAT) tools. These translation aids typically comprise a translation memory (TM) and a terminology database (termbase),²⁶ among other features. The TM stores sentences that were translated in the past, and the termbase contains the organization's terminology. During a new translation project, if a sentence in the source text matches or closely matches a sentence in the TM, the existing translation is shown to translators so that they do not have to retranslate it from scratch. If a term in the sentence matches a term in the termbase, the TL equivalent term from the termbase is

26. In some CAT tools the termbase is known by other names: glossary, dictionary, etc.

shown to translators as well. The termbase brings great value to the whole process by enabling reuse of translations at the level of lexical units rather than just full sentences, which is the limitation of TM. CAT tools are designed to increase translator productivity, and they do this most efficiently with a termbase containing sufficient terminology.

In the industry-oriented literature, there is evidence that using consistent terms improves content quality and product usability and reduces costs (Schmitz and Straub 2010; Kelly and DePalma 2009; Fidura 2013, Dunne 2007, Childress 2007). When terminology in the company's SL is consistent, the leverage rate of the TM increases, which reduces translation costs; this argument – increasing TM leverage through more consistent SL terminology – is one of the metrics used in the localization industry to justify investment in terminological resources (Schmitz and Straub 2010: 18, 20, 25, 26, 29, 57–58, 293).

Without a doubt, the use of TM reduces translation costs; translating a sentence that has a full match in the TM costs less than a sentence that has a partial match, and much less than a sentence that has no match at all. The ratio of full matches versus partial and non-matches increases when the source text is better controlled in the areas of style, grammar, word usage, and terminology. And yet, the results of at least two studies indicate that inconsistencies in terminology and word usage in source texts remain widespread (Schmitz and Straub 2010, Hurst 2009). If these studies are correct, inconsistent terminology is having a considerable impact on the cost of translation, and therefore, reducing inconsistencies has cost-saving potential in commercial environments. More information about terminology inconsistencies and other problems are described in *Terminology problems and challenges*.

Indeed, saving translation costs by increasing the leverage of TM is one of the main reasons why organizations decide to implement CA. But it should be noted that CA also has direct benefits in the authoring process itself. It can increase the productivity of company writers by prompting them with suggestions and helping them fix mistakes as they write, as well as reduce editorial demands, increase the quality of the final output, and more. But, as will be shown later, in order to work properly and attain these benefits, CA tools require access to a specially-designed termbase.

Marketing professionals frequently claim that the longer it takes to get a product into a market, the more likely competitors will get there first and seize market share. In today's fast-paced production process, each additional day that product release is delayed can translate into lost revenue. The goal in globalized marketing is simultaneous shipment (simship) of all language versions of a product. The agile software development model emerged in reaction to concerns about product release times. But a key factor in global release schedules is obviously translation

and localization. There are significant pressures to reduce translation times in order to retain or grow market share. A corporate termbase saves employees time in authoring and translation, and therefore supports the goal of shortening time to market.

Managing translation costs is a significant motivation for managing terminology in the source and target languages, but it is not the only motivation. Lombard (2006) and Dunne (2007), among others, have made convincing arguments justifying the management of SL terminology from an authoring standpoint. Language that is clear, easy-to-understand, concise, consistent, and (in the case of marketing material) powerful, resonates with customers. Language that is the opposite, namely clumsy, wordy, incoherent and complicated, confuses and frustrates customers. Language problems can lead to service calls, legal challenges resulting from copyright or trademark infringement, and in the most severe situations, injury or death due to product misuse.

Reduction of *defective terminology* (its total elimination seems to be an unattainable ideal), which is an umbrella term for a range of terminology problems that Dunne describes (2007), supports the realization of all four objectives listed at the beginning of this section. JP Morgan reported at various LISA conferences that an error in the source, if left undetected, multiplies exponentially in multiple language versions as well as multiple delivery media. Whether or not those claims have any empirical basis, the underlying argument that preventing terminological problems is cheaper and easier than fixing them after they occur has never been challenged or disputed.

Customer satisfaction translates into customer loyalty, which in turn drives revenue. When used in marketing, documentation or other product materials, language that is sloppy, inconsistent, or confusing gives the impression that the product itself is of poor quality, which is damaging to the company's reputation. Any problems in the SL are magnified in all target languages, which is why high standards in authoring are vitally important. This leads to the observation that terminology management in commercial settings should give equal if not additional emphasis to the SL compared to the TL. We wholeheartedly support this premise. However, as will be proven in the following section, terminology management as a practice has been historically dominated by the translation industry and therefore focused on the TL, with the SL too often neglected

Last but certainly not least, many companies demonstrate commitment to quality and customer service by implementing a quality management system and having it certified as compliant with ISO's flagship quality management standard, *ISO 9001: Quality management systems – Requirements*.²⁷

27. ISO website: iso.org/news/2015/09/Ref2002.html

With over 1.1 million certificates issued worldwide, ISO 9001 helps organizations demonstrate to customers that they can offer products and services of consistently good quality. It also acts as a tool to streamline their processes and make them more efficient at what they do. (ISO press release, 2015²⁸)

Compliance with ISO 9001 is a significant achievement that companies proudly announce. Since many of the benefits relate to quality, managing a company's terminology should be viewed as an integral part of quality management, and therefore, part of its ISO 9001 compliance strategy.

The historical confines to translation

Historically, terminology management has been viewed as an activity confined to the translation industry (Condamines 1995:220; Bowker 2002:290). Termbases were primarily created to support translators, and continue to exist for this reason today. The few terminology management courses offered in colleges and universities are virtually locked into translation programs, meaning that only aspiring translators are likely to enroll. Consequently, people entrusted with the role of terminologist are almost always translators working in the company's translation department. Moving the terminology effort to a higher level in the company's organizational structure requires the backing of a VP of Globalization, or someone of similar ranking, who has commitment and vision. This ensures that the company can have the greatest corporate-wide impact through authoring, translation, and other global initiatives such as SEO.

Another consequence of terminology's close ties to translation is the assumption that being a translator also makes one a de-facto terminologist, when in fact, due to the lack of terminology courses in many university-level translation programs (Teubert 2005: 97), translators may have no background in terminology management whatsoever. Terminology work is not translation. In fact, theoretically at least, unlike documents, paragraphs, and sentences, we do not "translate" terms. Each language can have a unique way to express a concept, one that is not *modelled* on the way that any other language expresses it. For example, a fire fighter in French is a *pompier* (literally, a *pumper*), not a *lutteur de feux*. Research is required to find the *equivalent* term that is used in the target language community. Let us repeat, as this is important, terms are not "translated."²⁹

28. *ibid*

29. In spite of our defense of this principle, i.e. that terms are not "translated," occasionally in this book we still use expressions such as "translated term" to avoid continually repeating the more awkward "target language equivalent."

An example may help to demonstrate why terms are not *translated*. The French equivalents of the English terms *garage* and *sale* are *garage* and *vente*. It would seem logical to translate *garage sale* as *vente de garage* but this is incorrect and even nonsensical. In Canada, the correct term is *vente-débaras* and in France, it is *vide-grenier*.

Having a translation background is obviously beneficial for terminologists, since the primary reason for managing terminology remains to help translators, and termbases are generally multilingual. However, the virtually exclusive association between terminology and translation has largely prevented terminography from penetrating the authoring stage of content development, where it is often needed most (see Lombard 2006). The value that termbases could contribute to other application areas, such as content management and information retrieval, is also often overlooked.

The terminologist as a working professional

It seems that new job titles and careers are springing up constantly in response to the growing demands of global communication. Writers are now called information developers, translators (especially those working on software) are often called localizers, and there are new roles in information “architecture,” search engine optimization (SEO), user-centered design (UCD), content management, agile product development, and cloud computing, to mention a few. Systems analysts design, create, and manage information systems. The umbrella vocation *information professional* refers to a wide range of job titles assigned to people who preserve, organize, and disseminate information. Certainly terminologists fit into that category. To some degree or another, these professionals, who work variously with information and language resources, are engaged in the field of *information science*. Many of these practices, and the contributions that terminology resources can make to them, will be discussed in *Extended applications*.

Nowadays, in global companies, one can expect to see a VP of Globalization in the Boardroom. In commercial contexts, the term *globalization* does not have the negative connotation that it often acquires in political discourse. It is the upper-most member of the triad: localization – internationalization – globalization. The term refers to the various processes and strategies put in place to ensure that the company can operate and compete in global markets. A cornerstone of such strategies is, of course, translation. A company has to translate its products into the languages of its target markets, and terminology management has always been a key enabler of efficient translation.

In recent years, though, companies, especially large multinationals that have accumulated extensive volumes of translation memories, have begun to seek ways to further streamline their globalization efforts. People began to realize that areas such as SEO, CA, and global content management have an impact on the effectiveness of the globalization strategy. And they began to also realize that these applications benefit from structured terminology resources in the form of termbases.

All of these vocations and focus areas work with language in one form or another. SEO uses keywords, which are essentially a specific type of terms. CA applies terminology consistency, as well as localization-aware word usage and style in the SL. Global content management requires triggers and a complex set of metadata to manage content throughout the globalization pipeline. These different pieces of information correspond to an extended view of terminology that we will describe later in this book. It is what we call *microcontent*.

So where does the terminologist stand?

The job title Terminologist is rarely encountered. In a survey conducted by the Localization Industry Standards Association of users and providers of localization services, a sector where managing terminology would seem to be important, only 12 percent of respondents claimed to employ a terminologist. The remaining 88 percent delegated any terminology management work to staff with other responsibilities (Lommel and Ray 2007: 30). Information professionals generally have little knowledge about terminology as a field of study, and lack awareness that terminological resources can benefit information production as well as their own work. Professional organizations dedicated to technical writing are large, such as Tekom with 8,250 members and the Society for Technical Communication (STC) with over 6,000 members. Yet terminology and terminologists are hardly mentioned within these organizations. Sager states boldly that “only in Canada can one speak of a body of independently trained professionals who work as terminologists” (1990: 220). He is no doubt alluding to the Translation Bureau of the Government of Canada, which at the moment this book is being written, employs dozens of terminologists to support its more than 1,000 translators and interpreters (these numbers are intentionally vague, as they change with every notable shift in government).

Are the professional opportunities for terminologists in commercial settings really that bad today? Granted, the sources cited above are dated (2007, 1990). Due to the aforementioned interests, motivations, and concerns, the situation is improving. With the realization that developing a terminology database requires special knowledge and skills, numerous companies now employ terminologists, and their ranks are growing. Take a look, for example, at the membership roster

of the industry stakeholder group, TerminOrgs.³⁰ It includes the likes of Canon, Microsoft, and Pitney-Bowes, to mention just a few. Of its 22 members (as of this writing), only one represents a translation company, and only one other a government. The remaining 20 are large organizations that fit the profile addressed by this book, and they employ terminologists. If these 20 are exemplars, there are many more like them.³¹ Terminology is gradually gaining recognition as a key enabler of global communications. And as organizations explore NLP technologies such as automatic content classification and machine translation, the importance of structured terminology resources becomes ever more evident. As will be demonstrated further in *Extended applications*, corporate terminologists in some cases carry out the tasks of information scientists and would therefore benefit from acquiring the skills of information professionals in the wider sense.

The advent of XML

Replacing unstructured and antiquated word processing formats, XML has become the new norm as the source format for virtually everything, from documents and help systems in DITA to translation packaging in XLIFF, terminology in TBX, translation memory in TMX and even text segmentation rules in SRX. Even this book is written in XML.

XML enables informational content to be structured in meaningful ways and at highly granular levels. This is why we are witnessing a proliferation of XML markup languages that address various types of documents and use cases, from invoices and legal contracts to recipes.

Covering XML in detail is beyond the scope of this book; there are plenty of useful resources available to learn about it. Suffice it to say that XML is a *markup language*, which is not a language in the linguistic sense but rather a computer language for representing text. Actually, to be more precise XML is a **meta** markup language, since it is a framework for creating other markup languages that are specifically designed for certain types of documents. As a markup language, it includes *elements*, *attributes*, and other computer-readable codes that are used to annotate various types of content in documents and other carriers of information, such as titles, lists, paragraphs, graphics, and even terms. Corporate terminologists need to be proficient in XML.

30. terminorgs.net

31. Not every large organization that employs terminologists is a member of TerminOrgs, of course.

One can expect the increased adoption of XML in the information industry to have a major impact on the opportunities to leverage terminology, and in fact this has been occurring for more than a decade. The Text Encoding Initiative (TEI) introduced a methodology for identifying terms in source content, and this practice was further developed in the Internationalization Tag Set (ITS), which added markup for instructions to translators. A form of ITS was adopted by DITA (Darwin Information Typing Architecture), which is a popular XML markup language for developing commercial content. Terms encapsulated with such markup can be automatically extracted and used in other processes, such as localization readiness testing, building a project glossary, submitting candidates for SEO keywords, and populating a termbase. Corporate terminologists must realize that the scope of their work extends beyond managing a termbase. In fact, they are responsible for identifying terms at the source and monitoring their passage through the content production process.

Lack of suitable models

Historically, termbases were outside the realm of commerce. Most of the early termbases (once known as term banks) were developed by and for the public sector. This is due to a number of conditions that were previously described, including terminology's traditional normative focus, its close ties to translation, and its social dimension supporting national and cultural identity. Those termbases tended to be funded by and designed for public interests such as for language planning (particularly for minority languages) (Temmerman 1997: 53; Rey 1995: 51), translating government documents, and delivering public services. The primary user group of these termbases is translators (Nkwenti-Azeh 2001: 604). The largest and most mature termbases in the world fit this description, for example, Termium Plus (Government of Canada), IATE (European Union), UNTERM (United Nations), and the EuroTermBank for Eastern European languages. These termbases have certain particularities, such as a subject field coverage that reflects the range of public services and a structural model intended to serve the needs of translators.³² Termbases designed to serve purposes that are not revenue-driven may be ill-suited to private sector applications, where virtually every task undertaken and every item of data stored is subject to a return-on-investment scrutiny. They may contain types of information that would be considered impractical or unnecessary in commercial settings. Conversely, they

32. ISO has characterized this type of terminology work as *socioterminology* (ISO/TR 22134:2007).

may lack information types that are important for producing commercial content, such as certain sub-setting values for tracking project-specific terms, and meta-data required by CA and CAT software.

In 1990 Sager already remarked that the large institutional termbases of that era have a “problem of orientation,” by serving only a very specific user group, and “they encounter difficulties in adapting to the requirements of the many new user groups who have emerged” (p. 166). Three decades later, with the emergence of NLP applications requiring terminological resources, XML as a powerful markup language, and large-scale corpora, these new user groups have undoubtedly multiplied in number and type. Nkwenti-Azeh (2001) describes the various types of data required for different users of a terminological resource. He observes that there is little information in existing termbanks that can be used to support the needs of NLP (p. 609), noting for example that little attention has been paid to recording textually-conditioned variants and subject fields, both important for NLP applications (the former establishes a semantic link between two terms, and the latter facilitates meaning disambiguation). He also notes that few termbanks record *collocations* (words that frequently co-occur). Condamines (2007: 136) observes that NLP applications are consumers of semantically-related terminological resources (which she calls *termino-ontological* resources). Even though commercial enterprises are increasingly looking to NLP applications to manage their vast volumes of content, commercial termbases that contain the required semantic relations are few and far between. We maintain that this is a legacy of the translation focus, which does not emphasize semantic relations.

Rey (1995:164) notes that large termbases were developed by international organizations like the UN or UNESCO and regional organizations like the European Union and NATO. As with official multilingual states such as Canada, these organizations require substantial translation services and develop large termbases to meet their needs. (We might add that they also share the mandate of serving the linguistic needs of society at large, which is not the case for companies.) Rey acknowledges that this translation view is reflected in the structures of the termbases, and feels the situation needs to change:

It would be necessary for the directors of these services to become fully aware of the problems and the power of good terminological support not only for translation but also for the improvement of the quality of discourse in these institutions and consequently the quality of their information.

In summary, the major institutional termbases of the world were designed without considering the commercial applications of terminological resources. Therefore, they may not be suitable models for termbases designed to support commercial needs. This lack of good models exacerbates other gaps in the development of

commercial terminography as a viable practice, gaps which this book can help to bridge. As Sager states (1990: 220),

In the absence of a systematically trained profession and clearly documented methodologies the production of terminological information proceeds along different paths as required by each organisation that collects and processes terminology.

The value of corpora

We have stated that the semasiological approach is common in commercial terminography. This means that corporate terminologists are finding terms in company materials, which form the company *corpus* (plural: *corpora*). However, while there is growing recognition that corpora are useful for selecting terms and obtaining information about terms, the actual use of corpora – except in random, ad-hoc searches – remains low among terminologists in the private sector. Most corporate terminologists come from varied backgrounds and are often thrown into the position with little formal training. They have not studied corpus linguistics nor used large-scale corpora and corpus-analysis tools. More troubling, however, is if influenced by conventional theory, corporate terminologists are unlikely to realize that they even *should* work with corpora.

Another challenge is actually obtaining a representative corpus. In many companies, especially large ones, it may not be possible to gain access to the company corpus in full; by virtue of its size and organizational structure, the corpus is typically distributed across different content repositories. When terminologists do carry out research to validate a term, due to these factors typically they do so on a subset of the full company corpus, such as the documentation for a given product. Since statistical evidence is more reliable the larger the data set analyzed, validating terms based on a subset of the full corpus will result in some terms being selected for the termbase that are not optimized for large-scale repurposability.

Another barrier to obtaining the corpus is file formats. Corpus analysis tools need to “parse” the corpus, and for this the files need to be in plain text format. File types such as txt, html, and xml work fine, but PDF, PowerPoint, Excel, MSWord files and similar file types do not. When asked to provide files for term extraction, which is a process that also requires plain text format, company representatives often supply PDF files, not realizing that PDF is not the original format. And it is surprising how often they are unable to provide the original format, simply because they do not know where those files might be.

The company's corpus is the most important resource for the corporate terminologist. Simply put, the terms in the termbase should reflect the corpus and vice-versa. The value of including a term in the termbase, which is the basis for determining the ROI of the whole initiative, can only be established empirically by determining the term's relevance to the corpus, and in relation to the relevance of other terms. Corporate terminologists, and the terminology initiative itself, would therefore benefit greatly from adopting a systematic corpus-based approach to term identification, as opposed to relying only on random, ad-hoc searches in uncontrolled conditions. This means collecting a comprehensive and representative sampling of the company's materials, preprocessing them (converting file formats, checking encoding, etc.), and analyzing the corpus with proper corpus analysis tools.³³ It can be challenging for the corporate terminologist to overcome organizational and technical barriers in order to implement a truly corpus-based approach to term identification, but the rewards are worth the effort. Corporate terminologists will gain the respect of peers and management by demonstrating the value of corpora in providing an empirical foundation for their work and the need to establish processes that enable their use.

Relationships to the corpus as an indicator of the quality and value of a termbase will be further addressed in *The termbase-corpus gap*, where we show how a company's corpus can be used to optimize the content of its termbase.

Terminology problems and challenges

The importance of managing terminology becomes apparent when terminology is mismanaged and errors occur.

Four main types of terminology problems or errors that frequently occur in commercial communications are described in this section.

1. Unintentional synonyms
2. Intentional synonyms that have only one equivalent in the TL
3. Distinct terms that have only one equivalent in the TL
4. Issues with proper nouns.

We have previously stated that terminology inconsistencies have an impact on costs and customer satisfaction. The word "inconsistency" has a negative connotation and is therefore typically used to refer to synonyms that occur unintentionally as opposed to synonyms that are used for a valid reason. A term inconsistency occurs when people use different terms to refer to the same thing and this different

33. Such as Wordsmith Tools, Antconc, etc.

usage was unintended and unmotivated. This can occur at the authoring stage (SL) or in translation (TL). Unfortunately, when an instance of inconsistent terms occurs in a SL document, the problem is almost always repeated in the translated version, and often, results in even more inconsistencies. For instance, a concept represented by two terms in the SL may become represented by three or more terms in the TL. As a consequence, translations are often perceived as lower in quality compared to the original.

It is important to distinguish between two types of synonyms, as they are treated differently in commercial terminography, and impact it differently as well. For this purpose, we have coined the terms *variant synonym* (variant), and *lexical synonym*. A variant shares properties with its so-called main counterpart term at the surface level; it is in some manner lexically derived from the latter. Variants therefore include abbreviations, acronyms, short forms, rearranged versions of multiword terms (e.g. *transmission mode* versus *mode of transmission*), spelling variants (*color/colour*, *organization/organisation*), and terms with minor adjustments such as the presence or absence of spaces (e.g. *check box* versus *checkbox*), hyphenations (e.g. *e-mail* versus *email*), spelling (e.g. *login* versus *logon*) and morphological features (e.g. *application program interface* versus *application programming interface*).

A lexical synonym is a synonym that has no similarity with the surface form of the term of which it is a synonym, such as *soccer* (American English) and *football* (British English), or *aspirateur* (Metropolitan French) and *balayeuse* (Canadian French) for vacuum cleaner. As these examples show, lexical synonyms often reflect regional variations. However, this is not always the case. The US Census Bureau uses the terms *enumerator* and *census taker* interchangeably. And in the US, the Supplementary Nutrition Assistance Program is colloquially known as *food stamps*.

Synonyms, whether intended or not, can cause confusion. Consider the terms *Unemployment Insurance* and *Employment Insurance*, used in Canada, where the only difference between the terms is the prefix “un.” This Latin prefix of negation suggests that these two terms have distinctly different and probably even opposite meanings. And yet, they refer to the exact same concept – a social security payment made by the Canadian government for eligible citizens during periods of unemployment, to compensate for lost income. For a long time, the program was formerly known as *Unemployment Insurance*, and this term became entrenched in colloquial usage. But to avoid the negative connotation associated with being unemployed, at some point a bureaucrat had the bright idea to change the term to *Employment Insurance*, failing to consider that 30 million people were already

familiar with *Unemployment Insurance*.³⁴ The new term never fully took hold, certainly not among the older generation, and many people today still refer to the program as Unemployment Insurance. If one government writer inadvertently uses the old term and another the new one, translators will see the two different English terms, and they may assume that they refer to two different insurance programs. In French, employment is *emploi* and unemployment is *chômage*, resulting in *assurance emploi* and *assurance chômage*. Clearly this would be a serious translation error. But by recording the two English terms in one termbase entry, the terminologist can make it clear that they are synonyms and should therefore have only one French equivalent. Thankfully, Canadian government terminologists do precisely that, through Termium Plus, and therefore Canadian translators are unlikely to make this mistake.

Unintended synonyms also occur in the TL when there are none in the SL. For example, in one software product, the word *graphic* was rendered in Portuguese in some places as *gráfico* and in others as *imagem*. What appears at first as an innocuous inconsistency may have unanticipated consequences. For example, the software user may wonder whether these two different terms for graphics correspond to different file formats as well. And consider how this affects the search function and the index of a help system, where different files discussing the same topic are now using different terms. This problem is quite common, particularly when there are multiple translators involved and they are not provided with access to a termbase. It is a reason why users of translated information tend to have a lower satisfaction rate than users of the original version.

But using synonyms in corporate communications is not always a bad idea. Some are justified and need to be preserved in both the SL and TL. Terminological variation is a stylistic technique used to realize economy (through the use of acronyms and abbreviations), to avoid repetition, to convey emphasis, and to improve textual coherence and cohesion (these strategies will be discussed further in *Variants*). In commercial texts, terminology consistency is desired between versions of a product, between related products, and between various communication media (for example, the user guide and the marketing material should be consistent when referring to product features). But terminology diversification is sometimes necessary for market differentiation (Corbolante and Irmeler 2001: 534–535). Furthermore, some SEO experts claim that the use of variants and synonyms in a document can improve its retrieval rate in search engines by matching a text with more user queries (Thurow 2006, Seomoz 2012, Strehlow

34. The concept is (or was) also colloquially referred to as *pogey*, which may be derived from the Scottish *pogie* meaning *workhouse*.

2001b: 434). There are obviously cases where lexical synonyms and variants are necessary.

So how can a translator know, when encountering lexical synonyms and variants in the SL, whether they are unintended, and therefore there is no need to maintain such variation in the TL, or intended, in which case they should be preserved (if possible)? A termbase can provide translators with the information they need to make the right decision.

Distinct terms that could be perceived as synonyms are another challenge faced by translators. One of our favorite examples involves the terms *manager* and *administrator*. These two terms were used in a software product that was sold to companies for creating their online *stores* in the early days of internet shopping. The two terms had distinct meanings. The store *Manager* was a product module featuring functions for managing the store, and the store *Administrator* was the name of the title or role of the person who used those functions. They were key terms that occurred thousands of times in hundreds of different files, which were distributed to two or three translators for each of the nine languages that the product was being translated into. Each translator translated only a portion of the product files, leaving gaps in knowledge of the product, and they may not be aware that the two terms have different meanings. After all, generally speaking, there is little difference between an *administrator* and a *manager*, the two terms are virtually interchangeable in most contexts, and the use of both in the product may be a simple inconsistency between different writers, an issue that translators encounter frequently due to lack of terminology control in the SL. Moreover, in some languages there is only one term to designate both *administrator* and *manager*. Consequently, in the translated versions of the product the two terms became one, for example, the French term *gestionnaire* was used for both. This made the product virtually unusable. The release to global markets was delayed while thousands of errors were fixed. This terminology error occurred during a particularly sensitive period in the history of software marketing, and thus the financial impact due to lost market share was likely significant.

Proper nouns – the names of companies, products, programs, and other entities – are particularly challenging. There is no hard and fast rule as to if, or when, they should be translated. The same proper noun may even be translated into one language while in another the SL term is retained. Some proper nouns have acronym forms, such as *Canada Revenue Agency (CRA)*. Different translation models can apply: translate the full form but not the acronym, translate the full form and the acronym, or do not translate either. For instance, the French version of the aforementioned term is fully translated, both the full form and the acronym: *Agence du revenu du Canada (ARC)*. In contrast, in the case of the

International Organization for Standardization (ISO),³⁵ the acronym is not translated in French: *Organisation internationale de normalisation (ISO)*.

Dunne defines *defective terminology* as terminology that is “incorrect, inconsistent and/or ambiguous,” where the latter refers to “the use of the same term for different concepts, or terminology whose meaning is unclear, indefinite or equivocal” (2007: 33). He provides examples of each category and describes their inherent risks and costs. Through an analysis of the cognitive translation process he demonstrates that, without adequate support, translators are frequently unable to detect defective terminology which is then propagated in the target languages.

The only way to effectively handle these challenges is to provide both writers and translators with guidance on how to use terminology. This takes the form of a properly designed terminology database and a proactive terminology management program.

Relevant literature

In previous chapters we have frequently quoted from the scholarly literature especially when describing traditional theories, methodologies, and principles, which are well documented. This section provides an overview of publications that focus on terminology management in commercial settings.

The body of literature about terminology management in commercial settings is very limited. This is because companies – and just a few of them at best – only began managing terminology in the late 1990s or early 2000s. Other reasons for the lack of literature include the dominance of the traditional theories and practices in the field, the near exclusive focus on translation, and the fact that most terminologists working in companies have struggled to determine best practices.

In the early 2000s, the Localization Industry Standards Association (LISA) was the first organization to consider terminology management as a core activity in an enterprise communications strategy. It should be noted that, in spite of its name, LISA was not strictly an association of only “localization” companies.³⁶

LISA is the premier not-for-profit organization in the world for individuals, businesses, associations, and standards organizations involved in language and

35. Note how, in this case, the letters *ISO* do not match the word order of the organization’s official name. Since there is a tendency to formulate the full form of acronyms according to the order of their letters, most people refer to the organization, erroneously, as the International Standards Organization. The acronym that would match the organization’s official name, *IOS*, is more difficult to pronounce, which is possibly why *ISO* was preferred.

36. Here, localization is more widely interpreted to include translation.

language technology worldwide. LISA brings together IT manufacturers, translation and localization solutions providers, and internationalization professionals, as well as increasing numbers of vertical market corporations with an international business focus in finance, banking, manufacturing, health care, energy and communications. Together, these entities help LISA establish best practice guidelines and language technology standards for enterprise globalization. (LISA, 2005)

LISA's membership therefore included "some of the largest and best-known companies in the world" (LISA 2005). For over two decades,³⁷ LISA held conferences and workshops, published industry reports, and developed standards and best practices. Its most important works include TMX (Translation Memory eXchange), TBX (TermBase eXchange),³⁸ SRX (Segmentation Rules eXchange), and the LISA QA Model (for translation quality).

LISA established a special interest group (SIG) that was focused on terminology management. It comprised members mostly from large enterprises who sought to establish guidelines for managing terminology as part of a global content management strategy. The Terminology SIG also held workshops about terminology management, which were attended by nearly 100 companies, many from the Fortune 500 list.

The Terminology SIG conducted surveys which resulted in three published reports³⁹:

- Terminology Management in the Localization Industry. Results of the LISA Terminology Survey (2001)
- Terminology Management: A study of costs, data categories, tools, and organizational structure (2003)
- Terminology Management Survey: Terminology Management Practices and Trends (2005)

The first describes the results of a survey in which 75 respondents – both localization service providers and global enterprises – answered nearly 80 questions about if and how they manage terminology and about how important doing so is for their business. The focus of this survey was establishing the role and value of terminology management for both the translation industry and the wider content

37. After its dissolution in 2011, LISA's role as industry representative was filled by other organizations such as the Globalization and Localization Association (GALA), Localization World (LocWorld), and TAUS (Translation Automation User Society).

38. This later became an ISO standard: 30042.

39. These reports are now available from TerminOrgs: terminorgs.net/External-Resources.html.

management fields. This survey was the first of its kind, seeking to determine the state, conditions, value, and requirements of terminology management in various commercial settings.

One of the key findings of this survey is how respondents in the enterprise category that operate a termbase answered questions about the value of the initiative. Eighty-six percent indicated that managing enterprise terminology improves the quality of their offerings, 70 percent that it increases employee productivity, 58 percent that it saves costs, and 53 percent that it improves the company's competitive edge. (LISA, 2001)

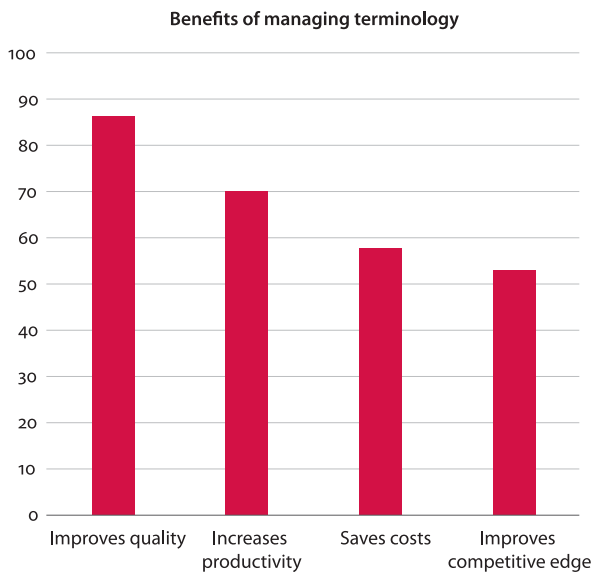


Figure 6. Percentage of enterprise respondents who realize benefits from managing terminology

The following is a quote from the conclusion of this report:

Whether or not they have a termbase, respondents representing organizations directly involved in localization recognize the benefits of performing proactive terminology management. Most perceive terminology management as a broader activity than simply translating terms. Terminology begins at the source through such initiatives as controlled English and source language terminology monitoring, and continues through product localization and distribution in target markets, passing through a wide range of tools and formats. It involves various players such as product designers and developers, writers, and translators. Organizations that maintain a termbase do so because it enables them to manage this complex process most effectively.

In an effort to learn more about terminology management in global companies, two years later the SIG conducted a survey with seven large enterprises, which resulted in the second report. Granted, with only seven respondents, this survey was quite limited. But one should also consider that at that time, few global companies were managing their terminology or had a terminology database. The seven respondents were considered leaders in this area. All respondents had an inhouse termbase that was concept-oriented and supported semantic relations. The universal support of these two critical features underscores the importance of synset structures and concept relations in large termbases designed to support a range of enterprise applications. More about these features will be described in Data category selection.

As LISA's member organizations began to realize that they too might benefit by managing their terminology, there was interest in obtaining more detailed information about terminology management in the localization industry, and the Terminology SIG conducted yet another survey in 2005. This survey had over 80 respondents and asked more detailed questions about practices and data structures. One interesting finding is that 75 percent of respondents indicated that terminology is systematically managed in their company.

In 2004, the Government of Canada commissioned a study which resulted in the report *The Economic Value of Terminology. An Exploratory Study* by Guy Champagne. The study's aims were to "determine the economic value of the terminology function in businesses and organizations, in terms of revenue, percentage of return on investment and cost reduction." Twelve large Canadian firms participated in the study. One of the interesting findings is that "terminology research is required for four to six percent of all words in a text." While this may at first seem a small figure, considering that large organizations can produce hundreds of millions of words a year, and the average time it takes a writer or translator to research a term is 20 minutes, the cost of this research in terms of lost productivity could be astronomical. In 2010, for example, it was estimated that one large company we studied produced 430 million words a year. This means that employees could potentially be spending over five million hours of time researching terminology. Fortunately, this company provides employees with access to a comprehensive termbase that cuts this research time to a fraction. These particular estimates could be challenged and need further validation, but they certainly support the basic premise that providing termbases to employees increases productivity.

In 2010, the German publisher TC and More, in partnership with Tekom (the German association for technical communication), released a publication entitled *Successful terminology management in companies : Practical tips and guidelines : Basic-principles, implementation, cost-benefit analysis, and system overview*,

co-authored by Daniela Straub and Klaus-Dirk Schmitz. This book provides useful information especially for newcomers to the field. A focus is on demonstrating the return on investment (ROI) of terminology work. The inconsistent use of terminology is repeatedly cited as having a significant impact on communication efficiency, the leverage of translation memory, customer loyalty, and other areas. The publication also includes a comparative analysis of the terminology management tools available at that time.

After the dissolution of LISA, the Terminology SIG continued operations as an independent thinktank, TerminOrgs,⁴⁰ which is short for “Terminology for large organizations.” It is the only group in the world representing terminologists who work in large institutional settings (mostly private industry). TerminOrgs is a “consortium of terminologists who promote terminology management as an essential part of corporate identity, content development, content management, and global communications in large organizations,” and its mission is “to raise awareness about the role of terminology for effective internal and external communications, knowledge transfer, education, risk mitigation, content management, translation and global market presence, particularly in large organizations” (TerminOrgs website). The membership includes an impressive list of major world companies. TerminOrgs has produced a number of publications and best practices, including the *TBX-Basic Specification*, the *Terminology Starter Guide*, and others about skills and training opportunities. Terminologists who work in similar environments should monitor this organization for additional resources as they become available.

A number of reports about managing terminology in commercial settings have been published by companies that have a vested interest in this topic. These companies may sell a TMS, a CAT tool, or a CA tool, or they may provide related services. Demonstrating the importance of managing terminology is a marketing strategy to sell their products or services. Market research companies, for their part, have also published papers about terminology management, but keep in mind that they are in the business of selling their reports, and they may not have direct expertise in this field. Some of these reports are based on solid research and are very reliable, others less so. Caution should be exercised when using these reports to substantiate proposals, such as a business case. It is always best to find a second source for any statistic or argument that was presented by a vested company. Nevertheless, given the limited sources in the scholarly and fully independent literature, corporate terminologists cannot afford to dismiss this type of information and other forms of so-called grey literature.

40. terminorgs.net

Terms in commercial content

In this chapter, we look at the types of terms that occur in commercial content, from various perspectives.

Terms considered by word class

It is generally agreed in the literature that, with respect to word class (part of speech), terms are predominantly nouns (Cabr  1999-b: 36, 70, 112; Rey 1995: 29, 136; Kocourek 1982: 71; Condamines 2005: 44; Daille et al 1996: 2077; Faber et al 2005; Kageura 2015: 48, L'Homme 2019: 61). These include single-word nouns and multiword nouns, which are sometimes called noun phrases or nominal phrases. (Multiword terms are described in more detail in the next section.)

According to ISO 704:2009, concepts “depict or correspond to objects or sets of objects.” Given that terms “designate or represent a concept,” the conclusion to be drawn according to this standard is that terms are primarily nouns. In reference to other word classes possibly being terms, ISO 704 uses the terms *verbal designation* and *adjectival designation*, but only once, and these terms seem somewhat contradictory since *designation* is defined in ISO 1087 in terms of concepts, which are in turn defined in terms of characteristics, which are in turn defined in terms of properties, and a *property* is defined as a “feature of an object,” with all examples of *objects* given in ISO 1087 being nouns. The GTT, upon which ISO 704 is largely based, gives prominence to nouns in its study of terminology.

But emphasizing denomination of extra-linguistic concepts and objects as an essential motivation of terms results in the near exclusion of non-nominal forms from terminography, which is rejected by defenders of the Lexico-semantic Theory and other text-based approaches. Bourigault and Slodzian, for example, stated that the exclusive selection of nouns (as terms) is incompatible with what can be observed in specialized texts (1999:31). They encourage the extension of terminological description to other word classes, particularly verbs. Faber et al (2005) also note that verbs play a crucial role. “This is due to the fact that a considerable part of our knowledge is composed of EVENTS and STATES, many of which are linguistically represented by verbs.”

In her article on verbs and adjectives (2002), L'Homme uses examples from the IT domain to demonstrate that certain verbs are associated with domain-

specific nouns, such as *launch a browser*. Jacquemin (2001: 273) notes that “studies in terminology have generally focused on noun phrases, but other categories convey important concepts in documents.”

Furthermore, many verbs and adjectives can themselves be nominalized. For example, *to treat* (v) versus *treatment* (n), and *compatible* (adj) versus *compatibility* (n). Hence, it could be argued that the corresponding verbal and adjectival forms themselves are terminologically interesting.

As other text types where terms are prevalent, commercial texts on the whole are dominated by noun structures. Nevertheless, sufficient challenges have been raised in the literature about the primacy of the noun word class among terms to warrant an examination of the prevalence of domain-specific verbs and other word classes in commercial content. We suggest that domain-specific verbs, and possibly adjectives and adverbs to a lesser degree, exist and need to be proactively managed as well as nouns.

The computing domain, for example, is rich in domain-specific verbs (Thomas 1993:59). At IBM, a leader in corporate terminography, verbs were found to be important because of their prevalence in software user interfaces. IBM’s term extraction tool, originally designed to extract only nouns, was enhanced to extract verbs.

One only needs to consult a few company websites to observe that in commercial texts, domain-specific concepts are also expressed by non-nouns, and many of the nouns themselves express intangible concepts, such as those referring to processes and states. Consider the following paragraph from a document about Microsoft SQL Server 2008, in which we have underlined some terms⁴¹:

Take advantage of SQL Server 2008 R2 features such as partitioning, snapshot isolation, and 64-bit support, which enable you to build and deploy the most demanding applications. Leverage enhanced parallel query processing to improve query performance in large databases. Increase query performance further with star schema query optimizations.

Of the 17 term candidates, three are verbs (build, deploy, leverage), and the remaining 14 are nouns or noun phrases. But of these 14 nominal structures, five express verbal concepts (partitioning, isolation, processing, performance and optimization). Only seven express tangible objects (SQL Server, snapshot, application, parallel query, query, database, star schema). What is further interesting is that the document in question is a data sheet, which is a type of document that is normally dominated by tangible concepts because it describes features of a product. And yet the incidence of verbal concepts is relatively high.

41. It should be noted here that no two terminologists would underline the same set of words. Identifying what are *terms* is a fairly subjective exercise.

One might question the decision to consider a verb like *build* to be terminologically interesting, as it could be considered a word from the general lexicon, as used in “build a house.” However, all terminological work is done for a given purpose and target audience, and in commercial settings, translation, as the traditional motivation for terminology work, is the main reason why terms are extracted and recorded in a termbase – i.e. to construct terminological resources for use by translators when translating the company materials. The verb “build” when used with its collocate “application” in the context of software is a specific usage of build that may require a different translation than the one that conveys the general meaning; for that reason, it is of interest to translators and therefore of interest at the moment of term identification or extraction. Indeed, in the French version of this same document, the verbs *créer* (create) and *développer* (develop) are used with application, instead of the general lexicon equivalent of build, which is *construire*. Thus, a verb that might appear on the surface to belong to the general lexicon may in fact have a domain-specific meaning or usage which needs to be taken into consideration when determining target-language equivalents.

In computing fields, verbs are significant because of their prevalence in software user interfaces. Verbs such as open, save, view, export, and print are very common. And yet, considered as belonging to the general lexicon, these verbs would have no terminological interest according to traditional theory. Dirk-Schmitz (2015) maintains that user interface labels are specialized concepts in their own right and are candidates for inclusion in termbases. “There is no doubt that software user interface terms used in menus and dialog boxes like *File* or *Options* represent concepts in the traditional terminological view.” In software products, which are often translated into dozens of languages, and where consistency between the user interface and the help system is critical, these are terms of interest for pragmatic purposes.

Finally, it should be noted that the inclusion of non-nouns in a termbase is also necessary to meet the needs of CA, SEO and other applications.

To summarize, verbs, adjectives, and sometimes adverbs join nouns as candidates for terminology management in commercial environments.

Terms considered by length

Terms comprising more than one word are abundant in termbases. They are generally referred to as *multiword terms* (MWT), but also as *complex terms*, *terminological phrases*, *phrasal terms*, and *compounds*.

When examining a corpus for terms, or scanning the list of terms in a termbase, the predominance of MWT compared to single-word terms is apparent. Kageura (2015: 48), citing research, estimates that MWT account for about 70 to

80 percent of terminology. L’Homme, for her part, notes that 77 percent of the terms in one specialized glossary are MWT (2019: 63). Indeed, their predominance has been observed empirically through statistical studies. Research conducted by Maynard and Ananiadou finds that the average length of terms is 1.91 words (2001: 25).⁴² Justeson and Katz (1995) noted that terms in the form of bigrams (two words, e.g. *sewing machine*) are more frequent than unigrams (one word, e.g. *machine*).⁴³ Daille et al (1996: 204) take this further and demonstrate that bigrams are actually more frequent than terms of any other length. Nakagawa and Mori (1998, 2002) share this view, claiming that 85 percent of term candidates are technical noun phrases consisting of two or more words.

Evidence shows that multiword terms are a major focus in corporate termbases. An investigation that we conducted (Warburton 2014) into the termbases of four companies confirmed Daille’s observation that bigrams are the most frequent. Why would bigrams occur in termbases more frequently than unigrams, especially given that unigrams probably occur more frequently in verbal communication than bigrams? One possibility is that unigrams have a greater tendency to be polysemic and/or too general to be accurately translated, and therefore, difficult to encapsulate in a concept entry. (How would you translate *machine* out of context?) For their part trigrams (e.g. *overlock sewing machine*) are recorded almost as often as unigrams in termbases, according to this same research study. This finding has ramifications for areas such as term extraction, where a tool that only extracts unigrams would be unacceptable, and for setting priorities for the termbase with respect to what kinds of terms it should contain.

These observations have a bearing on commercial terminography. According to research, there needs to be a focus on MWT, particularly bigrams.

Proper nouns

Proper nouns, also called *proper names* and *appellations*,⁴⁴ are very common in commercial content. As defined by ISO, they designate *individual concepts* as

42. Multiword expressions are also the most fruitful sources of new vocabulary in contemporary English (Hanks 2013: 50).

43. Of course, the word *machine* occurs more often in general discourse than *sewing machine*, but that is not the point here. Due to the plurality of meanings, *machine* is imprecise out of context and therefore unlikely to be considered a “term” and more likely attributed to the general lexicon. What Justeson and Katz mean is that unigrams are less frequent as “terms” than bigrams.

44. ISO makes a distinction between appellations and proper names, the former referring to a category of named objects (for example, *iPhone*) and the latter referring to a single named

opposed to *general concepts*, which are designated by (common) nouns.⁴⁵ An individual concept corresponds to one object, for instance, *The White House*, whereas a general concept corresponds to two or more objects, for instance, *white house*.

In commercial content, proper nouns are the names of products, product features and functions, services, programs, companies, and other named entities. They pose significant challenges for the translation process. There is no hard and fast rule, even within a company, about which proper nouns can be translated, and which cannot. In Microsoft, for example, the names of major products such as *Office 365*, and *PowerPoint* remain in English no matter where these products are sold, even though the products themselves (user interface, documentation) are translated. In contrast, the names of the two offerings of Windows 10 (Home and Pro) *are* translated, for example in French they are *Windows 10 Famille* and *Windows 10 Professionnel*. Sometimes the decision to translate a proper noun or to leave it untranslated seems somewhat arbitrary.

The decision to translate, or not to translate, a proper noun is often made by the company's marketing department in the country where the translated version of the information where this name occurs will be distributed. But that decision has to be conveyed to translators, otherwise mistakes occur. Marketing departments typically do not have good communication channels with translation services. This is another function, then, of the company termbase managed by the corporate terminologist.

Variants

In this section, we take a look at the degree of terminological variation, that is, the incidence of terminological *variants*, that occurs in communications in general, and in LSPs in particular. Why discuss this? Because if there is a lot of variation in term usage, it could have a bearing on the motivation for managing terminology and the methods adopted.

Freixa, in her typology of the causes of terminological variation, holds that variation is the phenomenon in which “one and the same concept has different denominations” (2006: 51), and this perspective of semantic equivalence as a basis for identifying *variants* is predominant in the scholarly literature about terminology (e.g. Cabré 2000: 49).

object (for example, *Buckingham Palace*). See ISO 1087:2019. In commercial terminography, this distinction has not proven to be necessary, where the term *proper noun* usually covers all names.

45. See iso.org/obp/ui

As explained in our earlier discussions on theory, the GTT, with its prescriptive focus, aims for univocity and therefore seeks to eliminate terminological variants. It also claims that, in specialized fields, there is a fixed relation between concepts and their designations and therefore, terminological variation as a phenomenon does not exist. More recent theories which recognize the role of cognition and communication dispute this claim. “Even the most cursory examination of specialized language texts shows that terminological variation is quite frequent, and that such variation seems to stem from parameters of specialized communication, such as the knowledge and prestige of the speakers, text function, text content, user group, etc.” (Faber and Rodríguez 2012:13). It is our observation that terminological variants occur frequently in commercial content.

We have previously noted that multiword terms (MWT) are common in commercial content. According to Collet, it would seem that there is a relationship between MWT and terminological variation. Citing Di Sciullo and Williams (1987), she observes that MWT exhibit “syntactic transparency,” as opposed to single-word units which are syntactically “atomistic.” The linear structure of MWT “is accessible to the rules of phrasal syntax, for instance to the rules governing ellipsis or reduction.” This complex linear structure gives them a propensity towards terminological variation: synonyms, hyperonyms, ellipses and reduced forms (2004: 101, 107, 108).

Different typologies of terminological variation have been developed which depend on the application. Daille (2005) presents a typology based on four major application areas: information retrieval (including term extraction), machine-aided text indexing, scientific and technical watch, and controlled terminology for CAT systems.

As mentioned in *Terminology problems and challenges*, a *variant* is usually considered to share properties with its so-called main counterpart term at the surface level; it is in some manner lexically derived from the latter. Variants therefore include abbreviations, acronyms, shorter forms (of multiword terms), spelling variants (such as British and American), and terms with minor adjustments such as the presence or absence of spaces (e.g. *check box* versus *checkbox*), hyphenations (*e-mail* versus *email*), and morphological features (*application program interface* versus *application programming interface*). This interpretation of variants excludes synonyms which do not share any surface form characteristics, such as *lorry* and its American equivalent *truck*.

Reasons why variants occur include a writer’s tendency to avoid repetition, and to be economical, emphatic, creative or expressive. These types of variants are referred to as *discursive* (Freixa 2006:52). The use of synonyms (which include variants) contributes to lexical cohesion (p. 60). Freixa adds that the variants most frequently adopted in specialized texts to avoid repetition, realize economy,

and improve textual cohesion are acronyms and other forms of term reductions. Abbreviations, acronyms, and other abbreviated forms realize economy by shortening the text without loss of semantic precision, at least in context (Kocourek 1982: 142). This may explain why they are so frequent, not only in specialized texts, but also clearly in commercial ones. Abbreviated forms also exhibit a high degree of conceptual equivalence (Freixa 2006: 62). Cabré observes that, once a new term becomes familiar, using an abbreviated form is common in specialized discourse (1999-b: 227). According to Rogers (2007: 29), repeating a full term rather than using an abbreviated form could therefore be disorienting for readers, who assume that they are being given new information. This would result in “over-specification” (p. 30).

Daille (2007: 175) demonstrates that including variants in terminological resources contributes to the improvement of several terminology-oriented applications: information retrieval, machine-aided text indexing, scientific and technological watch, machine translation, computer-assisted translation, and potentially other applications. Other scholars who have described the use of terminologies for indexing indicate that the documentation of variants is essential for this purpose.

While variants are deemed to be terms in all terminology theories, they are perceived and handled differently. Prescriptive-based and preoccupied with standardization, the GTT seeks to eliminate variants from LSPs, or at least to limit their use. Variants (and all synonyms, for that matter) are perceived as a problem that must be eradicated in the interests of clear, unambiguous and objective communication. With their shift towards a descriptive approach based on authentic language, proponents of the remaining theories reject this judgmental position and consider variants, as well as other types of synonyms, as valid expressions with a communicative role and purpose. Sager (1990: 58–59) explains:

The recognition that terms may occur in various linguistic contexts, and that they have variants which are frequently context-conditioned, shatters the idealised view that there can or should be only one designation for a concept and vice versa. (p. 58–59)

There is a need for lexical/terminological variation and this is variously strongly expressed for different text types. Despite the theoretical claims of univocity of reference, there is, in fact, a considerable variation of designation in special languages. (p. 214)

For SEO, it has been shown that lexical variety can actually improve content retrieval. Seeding a text with synonyms increases the coverage of keywords that people may use to search for content on a given topic. Consider the term *USB stick*. According to Wikipedia, this particular form of computer storage goes by

nearly 100 different names (91 to be precise, at the time of writing), including *flash drive*, *jump drive*, *thumb drive*, *pen drive*, and *memory stick*.⁴⁶ Given that it is impossible to anticipate what words people will use when searching for a store that sells USB sticks, a vendor of USB sticks would be well advised to include several of the most popular terms on the company’s web pages.

Thus, terminological variants are a valid form of expression, and play purposeful roles (economy, cohesion, SEO, etc.). Contrary to the decree of the GTT, they will not go away, nor should they. But on the other hand, should they be allowed to occur unfettered, without any controls? For example, in Figure 7, which is a partial view of a terminological entry, there are ten different French terms where Spanish and English only have one. One has to certainly wonder why all the French terms are necessary. So probably no, variants should not be allowed without any kind of controls, as that would have certain negative impacts, such as reducing the leverage of translation memory, and sometimes reducing clarity. The corporate terminologist therefore has to strike a delicate balance when designing the terminology process, resources, and tools: one that allows terminological variation when justified and discourages it when not so.

 English
global environmental issues
 Spanish
cuestiones ambientales de alcance mundial
 French
problèmes environnementaux de portée mondiale
problèmes environnementaux à caractère mondial
problèmes environnementaux à dimension mondiale
problèmes environnementaux de portée planétaire
Problèmes environnementaux à caractère planétaire
problèmes environnementaux à dimension planétaire
problèmes environnementaux à l'échelon mondial
problèmes environnementaux à l'échelon planétaire
problèmes environnementaux à l'échelon de la planète
problèmes environnementaux à l'échelon du globe

Figure 7. Unfettered variants⁴⁷

46. See the Wikipedia article *Pages that link to USB flash drive*.

47. Courtesy Interverbium Technology.

The need to strike this balance is another reason why terminology needs to be managed in corporate environments and why it is so important to adopt a concept-oriented approach.

Applications

In this chapter we describe the various ways that terminology resources can support business processes.

Where can terminology be used?

As described in *The historical confines to translation*, the general impression is that terminology and terminography are confined to the translation industry. Unfortunately, this narrow view has constrained the development of the profession in the information industry. The focus on translation has prevented terminology management from gaining recognition in other areas where it can have great potential, such as authoring, content management and information retrieval.

Three decades ago, it was predicted that applications beyond translation would benefit from richly-structured terminological resources (Knops and Thurmaier 1993: 89; Sager 1990: Section 8.4.1; Meyer 1993: 146). Galinski (1994: 142) noted that standardized terminologies are beneficial for indexing, information retrieval, and overall quality management. Sager (1990: 228) and Meyer (1993) predicted that the end-user of terminological resources is not necessarily a human being, citing spell checkers and machine translation systems as examples.

The application that a terminology resource is used in has a major bearing on what we consider to be a term versus a non-term. L'Homme (2019: 6) states simply, “terminology is deeply rooted in *applications*” and she adds that there are different ways to carry out terminology work according to the needs of those applications. She maintains that the intended application of a terminological resource is the most important parameter to consider when identifying terms, coupled with membership in a subject field (2019: 59–60, 66).

Today there is a wide range of potential applications of terminological resources that are relevant in commercial settings. Preoccupied with saving costs, increasing productivity, and gaining market share, commercial enterprises are constantly seeking ways to further automate areas such as authoring, translation, content retrieval and content management. The pressure to discover ways to use technology to improve our use of language – referred under the umbrella term *Natural Language Processing (NLP)* – is intense.

The scholarly literature abounds with descriptions of how structured terminology resources, i.e. termbases and their digital outputs, benefit applications beyond translation. These applications are described in the sections that follow.

Unfortunately, due to lack of foresight, most existing termbases are developed for the singular use case of translation, but are incapable of meeting the requirements of such applications. They lack the required structure and content. They will need to be re-engineered, and the content cleaned, curated, and extended, at significant cost. No terminologist wants to break the news to upper management that the company termbase is incapable of delivering the required data to support new uses, such as in CA tools, which are often the first language-related technology to be adopted after CAT.

It is therefore critical for corporate terminologists to appreciate the wide range of potential applications of terminological data, and adopt strategies, methods, and structures at the outset that produce terminological resources that are *repurposable*.

In the next sections, we will approach the range of uses in five categories:

- content management
- translation
- authoring
- search engine optimization
- extended applications.

Content management

Terminology is a form of content, and for this reason, terminology management is a form of content management. Let us examine this concept further.

Content management (CM) is the administration of digital content throughout its lifecycle, from creation to publication and finally to archiving or deletion. Global CM is additionally concerned with translation and localization, topics that will be further explored in the next sections.

Content management involves a number of different activities. A breakdown is shown in Figure 8.

Starting at the top left, first we create the content. This may involve a CA tool. Then it is edited for clarity and accuracy according to rules of grammar and style that are meant to boost marketing and are often found in a corporate style guide. There are also standards and best practices for creating certain types of content produced by organizations such as the International Organization for Standardization (ISO). Today, content is digital, so in order to enhance our ability



Figure 8. Activities that are part of content management

to manage the content and optimize it for search, we need to add metadata. Of course, this means that the content is probably encoded in an XML format such as DITA (Darwin Information Typing Architecture).⁴⁸ In a global organization, the content needs to be translated and localized for the target markets. This means that the original content should be internationalized so that it can be translated more easily. To do this we need to infuse the content with markup from another standard, the Internationalization Tag Set (ITS). To translate the content, we use other XML formats such as XLIFF (XML Localization Interchange File Format), TMX (Translation Memory Exchange), and TBX to incorporate terminology in the process. Then the content can be published and made available to our target audience in various delivery media. To be efficient, we need to store the content and reuse it, possibly with some updates, such as for another version of the associated product.

Many people are involved in these various tasks, and they often use tools such as CA software and translation memory (TM) systems. The whole process can be managed with the help of a content management system and central repository.

⁴⁸. DITA is a standard XML format for authoring content that is maintained by OASIS, a standards organization.

Content has increasing levels of granularity. The largest level of content is the entire collection relating to a core asset, such as a product or a service. An information portal is a good example. The next level would be a document, a web page, or a file within that collection. Then we go further into smaller pieces of content: a section within a web page, a paragraph within a section, and a sentence within a paragraph. Within the sentence itself are even smaller bits of content that correspond to concepts. Figure 9 was taken from a web page of Autodesk that describes features of a design software called Notes and Labels. Some important concepts are underlined.



Figure 9. Concepts within a sentence

It is worth noting here that the sentence as a structural unit is traditionally viewed as the smallest level of content that can be reused or recycled via automated means. Translation memory systems, for example, allow translated sentences to be reused for repetitive content in different documents. The idea of reusing translated sentences is now being extended to the authoring stage, where *authoring memory* is beginning to make its way into authoring tools.⁴⁹ However, we maintain that the smaller bits – the individual concepts – are also reusable if they are properly curated, managed, and stored electronically. While this is not an entirely novel idea, as evidenced, for example, by the creation and use of

49. Note that the small print in the figure indicates that this feature applies to different products. Therefore, content reuse at the source, which is possible with an authoring format such as DITA, would be very beneficial here.

termbases in (some) translation companies, the scope of leverage of these conceptual units today falls far below its potential. We will discuss this topic further in *Microcontent*.

Terminology management is a form of (and greatly enhances) content management in various ways. Termbases work in parallel with TM. Terminology data is encapsulated in XLIFF files and TMX files. Terminology markup is a key part of ITS and even of DITA. It is the foundation of search optimization. Lastly, terms themselves are used as keywords to annotate content for classification and categorization purposes.

Translation

It goes without saying that translators have always been and continue to be the primary creators and users of terminology resources. Terminological activities of recent decades have been closely tied to translation (Rey 1995: 50, 129; Pozzi 1996: 69). Historically, the main developers of terminological resources have been public institutions concerned with multilingual communication. And although there are no reasons why terminology work need necessarily be a multilingual endeavor (terminology *does* need to be managed within one language), it almost always is (L'Homme 2004: 21; Rondeau 1981: 33).

As was stated earlier, terminology has close ties with the translation industry (Bowker 2002: 290; Williams 1994: 195). University-level courses that focus on terminology are typically offered under the umbrella of translation studies (Sager 1990: 220; Wright and Budin 1997: 347; Picht and Acuna Partal 1997: 305–306; Korkas and Rogers 2010; Van Campenhoudt 2006: 6). Most of the existing terminology databases were developed to serve the needs of translators (Teubert 2005: 101), and as a consequence the people entrusted with the role of terminologist are almost always primarily translators (Bowker 2002: 291).

Nowadays, many translators access information about terms, such as TL equivalents, from within a CAT tool. In the CAT tool, a small window displays terms from the integrated termbase when they are needed. It is, therefore, vitally important that the central (corporate) termbase is configured in a way that can deliver the required content for the CAT tool. Due to the small window size, this could, for instance, take the form of a limited view of the data, or a filtered subset of the data. It is important to realize, as we will see later, that the termbase that is integrated in the CAT tool may not actually be the central corporate termbase, since terminology management functions in CAT tools are, we maintain, too lightweight to handle corporate requirements of repurposability, among others. There is a high probability that terminology will have to be transferred from a

central termbase to the CAT tool. These and other requirements will be discussed later.

Marta Gómez Palou Allard's PhD Thesis, *Managing Terminology for Translation Using Translation Environment Tools: Towards a Definition of Best Practices*, published in 2012 at the University of Ottawa,⁵⁰ offers a wealth of information about the needs of translators working in CAT tools. Based on a comprehensive survey, she identifies a number of terminographical practices routinely carried out by translators "contrary to what current terminology and terminography literature recommends," mostly due to limitations in their CAT tools. The corporate terminologist should be aware of these limitations and find ways to circumvent them without negatively impacting the repurposability of the central termbase. Lobbying the suppliers of CAT tools to address these limitations serves the interests of corporate terminography.

To serve the goals of repurposability, corporate terminologists must have a view of terminology that extends beyond its use by translators. At the same time, translators remain the primary end user and stakeholder in most cases and therefore deserve a proportional amount of the focus and resources in the terminology program.

The needs of translators must therefore take priority when the requirements of the program are being established. However, adopting methods or technologies that preclude other uses of the terminology resources must be avoided.

A terminology management program that meets the needs of translators has the following main requirements. These and other requirements will be described in more detail later:

- A multilingual termbase
- Access to the termbase by all translators, both in-house and outsourced
- Connectors between the termbase and any used CAT tool
- Automatic search of terms in the termbase from the CAT tool
- Ability to submit terms to the termbase from the CAT tool
- Terminological principles: concept orientation and term autonomy
- Approval workflows.

Authoring

The strong association between terminology and translation, described earlier, has largely prevented terminography from penetrating the authoring stage of

50. Available at <https://ruor.uottawa.ca/handle/10393/22837>

content development, where it is often needed most (see for example Lombard 2006 and Dunne 2007). Still, there is increasing recognition of the importance of managing terminology in the SL for (a) improving the source content itself, (b) rendering the translation process more effective, and (c) reducing costly errors. “Terminology management should be moved upstream in the document cycle to the source and integrated with design and development to prevent the proliferation of problem terms and to shrink the delta between authoring and translation” (Dunne 2007: 37).

We use the term *authoring* to refer to all types of content production in the SL. This includes product interfaces, product documentation, technical reports, marketing material, websites, and internal communications (human resources, policy documents, etc.).

Consistent terminology in the SL has been proven to reduce translation costs, due to the increased leverage of TM (see among others Schmitz and Straub, 2010). However, even without considering translation benefits, it certainly serves the company’s interests to use clear, consistent terms and words in any language. The SL is not only used to communicate with customers in that language, which is typically the largest language market for the company; it also serves as input for translation into dozens of other languages. This makes controlling terminology and word usage in the SL critically important, perhaps even more so than in target languages. And yet, companies continue to allocate disproportionate resources on translation quality control, ignoring the main company language that feeds into translation.

Today, more and more companies are leveraging CA tools to improve authoring. These tools and their particular requirements will be described in *Controlled authoring*.

CA applications do more than check spelling and grammar. They also tell writers when they have used a term that they shouldn’t use, according to company guidelines, and what term to use in its place. This feature requires specially-developed lexical files that contain information about the company’s preferences. Thus, CA requires terminology, but there are many differences when compared with translation-oriented termbases. CA tools require *prescriptive* terminology resources, i.e. ones that indicate preferred and erroneous word usage. They also require words and expressions from the general lexicon, which are often not included in translation-oriented termbases. Companies that have a pre-existing termbase when CA is rolled out often expect to use their termbase “out of the box” in a CA tool. This is simply not possible. There are challenges to incorporating an existing termbase into a CA application.

First, the set of terms required for CA is not the same as the set of terms required for translation. This important distinction will be described in more

detail in *Controlled authoring* and *Inclusion criteria*. Suffice it to say that many of the terms required for CA will be missing from an existing termbase.

The main difference is that existing termbases, which for the most part were developed as an aid for translators, contain few *synsets*, that is, groups of synonyms recorded in the same concept entry. Synsets are essential for CA. Furthermore, any existing synsets in a translation-oriented termbase are unlikely to have usage status values on the SL terms, which are also required for CA. Knowledge about synonyms and their relative importance in the SL is not of great interest to translators. They need to translate the words that appear in the sentence in front of them at any given moment, not, in that moment, to learn about alternate SL words that have the same meaning.

Another consideration is whether all the terminology in an existing termbase should be included in the CA application. While having company-specific terminology in the CA tool is normally beneficial for word recognition and spell checking, experience has shown that a surprising number of conflicts that produce false positives (errors reported that are not errors) are introduced when the whole corporate termbase is included. Faced with the realization that not all terms can be added, the terminologist has the daunting task of figuring out which terms are needed, and then assigning an appropriate marker to those terms so that they can be exported to the CA tool in isolation.

It is therefore necessary for companies to retrofit existing termbases when CA is introduced. This can take a considerable amount of time. Simply importing an existing termbase into a CA tool causes more damage than good. In reality, only a fraction of the total existing entries is beneficial and many terms and words required for CA will be missing in the termbase. This situation can, of course, be avoided if CA is accepted as a potential application of the termbase from the outset.

Aside from improving the quality and consistency of terminology in source content through CA and style guides, the authoring community can, and should, be directly engaged in the process of identifying terms in source content and explaining their meanings. They are more knowledgeable about SL terms and their meanings than most translators. Often, technical writers clarify difficult terms in the content anyways, to help users of the products by way of explanations or definitions. Being unfamiliar with terminography and termbases, writers may not be aware that those types of information can be valuable if encapsulated in a termbase. They may not even be aware that what they are writing *is* a definition.

For example, a formulation like the following one is typical in commercial content: "The Text Encoding Initiative (TEI) is a consortium which collectively develops and maintains a standard for the representation of texts in digital form. Want to become active in the TEI community?" There is a lot of useful information in the first sentence. It contains two terms, *Text Encoding Initiative* and *TEI*,

which are synonyms (full form and acronym), and a definition of those terms (“a consortium which collectively develops and maintains a standard for the representation of texts in digital form”). This is enough information for a terminologist to create an entry in a termbase, mark it as a proper noun, and trigger downstream processes of adding TL equivalents (which in this case are likely to remain in English). One might ask why this term is included in a termbase if it remains in English across all other languages? The answer is that (a) some translators might translate the term if they are not aware that it is a proper noun, and (b) even if translators know that it remains in English, having it in the termbase means that it can be imported into the CAT tool dictionary from where it can be inserted into the translated text with a quick keyboard shortcut, saving the translators the effort of typing nearly 30 characters each time it appears in the text.

Note in the above example that the second sentence includes TEI without its expanded form. This is standard practice in commercial writing: include the full form on first occurrence and use only the acronym thereafter, for economy purposes. Such a practice emphasizes the importance of capturing the information presented in the first sentence, in order to pass the necessary knowledge on to the terminologist. They can then share it with translators and the whole corporate community using the termbase.

Fortunately, if the corporation is using XML as an authoring format, it is possible to mark up content in the source so that this information can easily be found or transferred by using ITS or TEI. Marking up terminological information in source content will be further discussed in *Standards and best practices* and *The authoring community*.

Search

Before we demonstrate how terminology data can be used to enhance search operations and content retrievability, several distinctions need to be made. There are two groups of people who use a web browser to search for information about a company: internal (employees) and external (mostly prospective customers). Internal users are searching for company information on the company’s *intranet*. An intranet is a private network operated by a company which uses internet technologies but is insulated from the global internet. A user id and password issued by the company is required to access the intranet. Employees use the intranet to find information about human resources, company programs and events, policies, staff, departments, and also products and services.

Members of the public typically first discover a company’s website by doing a search on the *internet* with a public search engine such as Google. The words they

enter in the search field are referred to as search *keywords*. If the search keywords appear frequently and prominently on one of the company's web pages, a link to that page should be displayed high up in the list of search results. Thus, in order to achieve a high ranking in internet searches, the company needs to determine which search keywords are used frequently for the types of products or services that the company offers, and ensure that those keywords are also used in the written content of the web page.

On the company's web pages there is usually another search field where visitors can conduct a search within the company's web portal. This type of search is referred to as *enterprise search*.

Figure 10 shows the enterprise search field on the ISO website (in the red box).

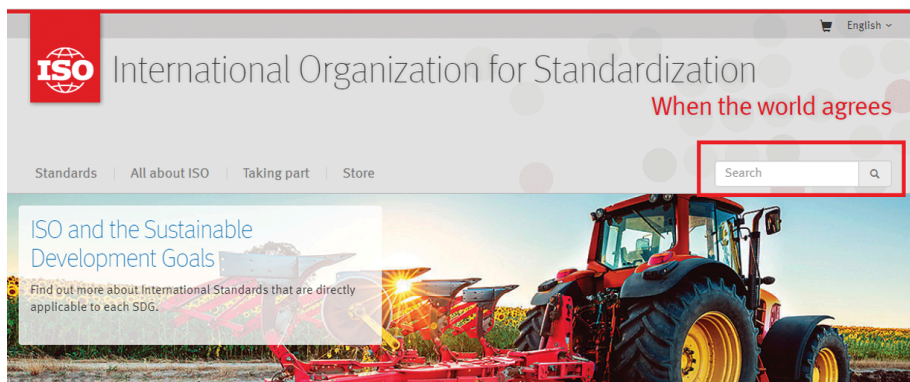


Figure 10. Enterprise search

With global web search engines such as Google, enterprises have little control over how their site ranks in search results. Here, we are referring to *organic* search, as opposed to *paid* search.⁵¹ Google and other web search services do not make their search algorithms public. While there are some best practices for web development that, if followed, can help raise the rankings of a website, there is no precise formula; with the internet constantly evolving, the results can also change overnight. Companies that specialize in search engine optimization (SEO) offer services to help enterprises improve their ranking in global search results, with apparent success. Usually SEO involves determining good search keywords and embedding them in the company's web pages. But global search rankings are also constantly changing, as are the behavior patterns of internet users. So a company may succeed in achieving a high rank one day only to see the results deterio-

51. Paid search is not covered in this book.

rate the next. Nevertheless, achieving a high rank in public searches is extremely important for marketing and sales, such that companies are willing to invest significant resources in SEO. Managing search keywords is the entire focus of global SEO. In this regard, search keywords, which can be viewed as a type of terminology as we will later show, can be managed in a corporate termbase. This application alone will give a lot of weight to the business case.

With enterprise search the company can have more control, since this search engine is, or can be, deployed in-house and the algorithms are known and can often be customized. Many enterprise search systems can integrate structured data such as terminologies and search keywords into the search algorithms in order to improve search results. There are dozens of vendors of enterprise search software. Before purchasing and deploying an enterprise search technology, companies that have a termbase would be well advised to consider products that can use the terminology to improve results, as described in the next sections.

While the use of structured terminological data in enterprise search is still quite rare, it has been successfully deployed in several large companies and others will soon follow. Even if a company is not ready to explore this enhanced search functionality, it should be anticipated that one day it might, and the termbase should therefore be developed with this application in mind.

There are four ways that terminology from a termbase can enhance enterprise search: query expansion, query correction, autocomplete, and faceted search.

Query expansion

Query expansion is a function whereby a user enters a search keyword and a set of alternate search keywords are suggested alongside the results. These alternate search keywords must have the same or a very similar meaning to the original search keyword. For this to work, the search engine must have access to a synonym dictionary, which can be generated from the company's termbase, provided that the termbase has been built correctly.

A synonym dictionary is a file that contains a list of synsets. A synset is one set of synonyms. Usually the dictionary can take the form of a character-delimited text file, with one synset on each row, such as:

```
mobile phone; smart phone; cellular phone; cell phone  
thumb drive; USB stick; flash drive; memory stick; jump drive
```

When a user conducts a search using any of these keywords, the search engine queries the synonym dictionary and presents any alternate keywords found there, and the user can decide whether or not to try one of them. Each alternate keyword

can be presented in the form of a link that when clicked launches the new search automatically.

Query expansion is very useful for acronyms. Because they are so short (usually just a few letters), acronyms are ambiguous even within one company. Does *SUN* refer to Sunday, the star, the company Sun Microsystems, Service User Network, Shipping Unit Number, or any of dozens of other possibilities? When the user types an acronym in the search field, the corresponding full forms that are relevant for the company are presented, and irrelevant ones excluded. From there, the user can click the meaning they intend to search for.

Another form of query expansion involves the suggestion of related terms, i.e. terms that have a similar as opposed to equal meaning. Term candidates for this application include terms denoting a broader concept (hyperonym), a narrower concept (hyponym), or a term with some other associative rather than strictly hierarchical semantic relationship. Bowker and Delsey (2016: 90) describe a “broader information retrieval context” where “more refined searches could be achieved if a greater range of productive lexical knowledge patterns were integrated into search tools.”

There is no reason why the types of terms and their relations required for improving search technologies should not be stored in the central corporate termbase. Of course, for this to be possible the termbase must be concept-oriented. In addition, one of its missions should be to collect synonyms and record semantic relations.

Query correction

Query correction refers to the function whereby the search engine detects a spelling error and either corrects it automatically or suggests a correction to the user. Google has this function with its “Showing results for” prompt.

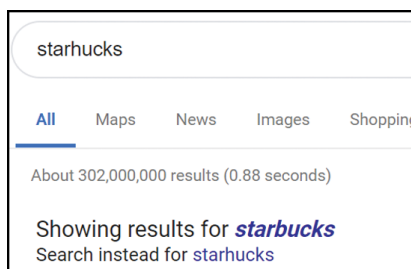


Figure 11. Query correction in Google

Most likely Google determines correct spelling by matching patterns on the fly, based solely on volumes of crawled material. But in order for this to work in enterprise search, the search engine needs to consult a list of correctly-spelled words, since enterprise search does not have access to such large volumes of data.

Most companies have unique names for products, services, and other company-specific concepts, such as *Starbucks* in the example above. These words are unlikely to be in any list of standard English words (which come from language dictionaries such as Oxford). If a company does not include its own internal terminology in the word list, when a user searches a name that is unique to the company, but misspells it, either no suggestion will appear, or one that is completely irrelevant will be offered. If *Starbucks* is in the word list, and the user types *Srarbucks*, the search engine can make the correction.

Autocomplete

Autocomplete, sometimes referred to as *typeahead* or *predictive typing*, refers to the function whereby as a user types a search keyword the system anticipates what the user is typing and offers a list of suggestions. The idea is to help the user by completing the search query. Figure 12 shows this function in Google when the user starts typing “star w.”

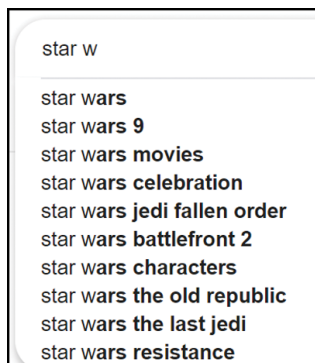


Figure 12. Autocomplete

As with query correction, autocomplete will not be optimized in enterprise search if the company’s own terminology is not integrated into the search engine. None of the company’s own terms will be included in the suggestions. For example, Figure 13 shows the autocomplete suggestions for *Tivoli* when it is not supported by a company-specific autocomplete dictionary. *Tivoli* is an IBM product brand. The

ten suggestions shown here include only two – in fifth and tenth position – that are relevant to IBM.

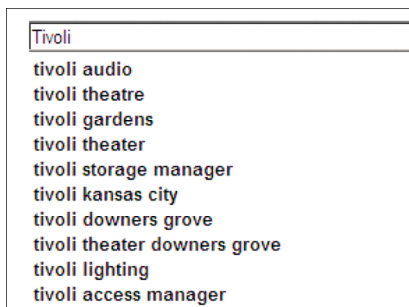


Figure 13. Autocomplete without company terminology

Autocomplete will offer relevant suggestions, and exclude irrelevant ones, only if the enterprise search engine has been configured to use terms from the company's termbase.

Faceted search

Faceted search refers to the function whereby the search engine suggests keywords in increasing order of specificity to help the user find the desired information. Faceted search has become a popular technique in enterprise search deployments, particularly for online retailers. An increasing number of enterprise search vendors provide the technology for implementing faceted search applications.

The following search scenario demonstrates faceted search. In this scenario, electrical engineers are interested in an industrial grade limit switch that is energy efficient. With a Google search, they are directed to the website of a company that offers this type of product. They begin by searching for “switch” using the company's enterprise search field. Through a series of links, they are progressively directed to increasingly specific sub-categories: Note: each of the steps in the sequence above includes a number of choices, only one is shown here for simplicity purposes.

- switch
 - limit switches
 - heavy-duty limit switches
 - heavy-duty low-energy limit switches

With this type of search the engineers have found what they are looking for, even though the resulting product page does not include any of the keywords they had in mind (energy-efficient, industrial-grade).

One way to enable this function is for the search engine to crawl pages on the fly and infer taxonomic relationships based on some automatic text analysis process. However, this does not always produce the desired results. Figure 14 shows a search that has gone wrong. Here, a user selects *limit switches* then *safety limit switches* and then is returned to *limit switches*, which is both illogical and frustrating. And to be offered categories such as *switch family offers full* is simply nonsensical. Such negative search experiences do not reflect a positive company image.

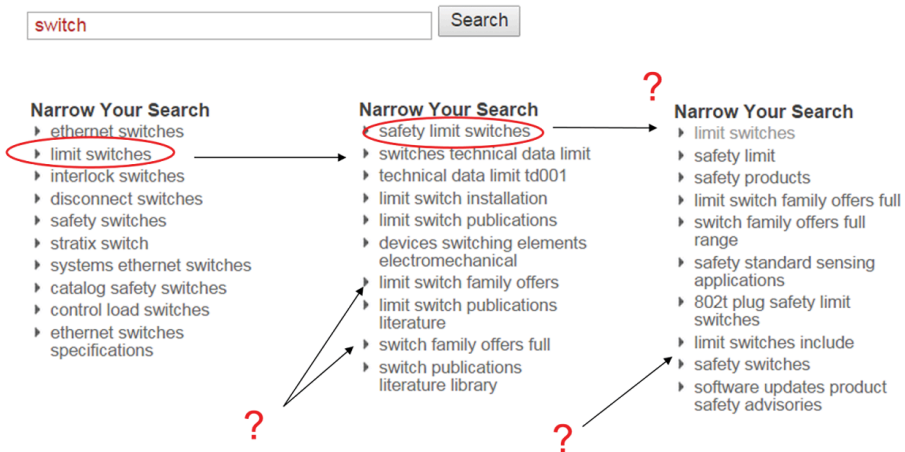


Figure 14. Faceted search without company terminology

A more reliable method is to develop multi-faceted enterprise taxonomies that reflect fixed, logical relationships and then implement them in a properly-designed faceted search function. An enterprise taxonomy is a hierarchically-structured terminology that has been referred to as *intelligent terminology* (Massion 2019), *ontoterminology* (Roche 2012), and *knowledge-rich terminology* (Marshman 2014). This can be done in the termbase if it is properly designed.

Another form of faceted search is when the user is offered a list of categories to narrow the search. This is very common in online retail stores. For example, on a site that sells computers, one might first choose either Desktops or Laptops, then Home or Business, then among various features such as touchscreen, screen size, weight, memory, and price. This also requires a sophisticated multi-faceted product taxonomy.

Intranet search

The company's web pages on the intranet also offer a search field. The enterprise search technology previously described can also be used for intranet search, thus the terminology can be leveraged as explained in earlier sections. However, since the types of content searched by employees on the intranet differ greatly from that searched for by the public, the terminologist needs to ensure that such intranet content (human resources, policies, programs, etc.) is appropriately represented in the termbase. The same resources mentioned before (synset dictionaries, spelling lists, autocomplete lists, etc.) need to be generated from the termbase and deployed in the intranet search engine. This leads to an important realization; the termbase can and should store terminology from both internally-facing and externally-facing content materials, and the two sets of terms should be kept separate in the termbase. If a company takes this approach it will achieve the greatest ROI). This can be done with a picklist field allowing one of four values: internal, external, both, and unspecified.

Search keywords

Search keywords themselves can be effectively managed in the company's termbase. Why would one want to manage search keywords? As mentioned earlier, global search engines such as Google use precise techniques to rank web pages, but these techniques are not made public. Nevertheless, there are some proven methods, which are referred to as Global Search Engine Optimization (SEO). There are companies and suites of technologies that specialize in SEO. It was stated earlier that SEO techniques focus on aligning search keywords used by people on the Internet with words that are used on the company's web pages. For instance, a search for "coffee" using Google will generate links to websites for companies that sell coffee, for coffee houses, and for major coffee brands. According to SEO principles, how close to the top a website appears in the results depends on the frequency and prominence of the search keyword on the web page. There are, of course, other factors that affect SEO, but the selection and placement of high-performing keywords is the main objective. An SEO advisor will therefore recommend a series of best practices for embedding effective search keywords into web content. Thus, it is necessary to determine which search keywords are effective and make them available to the employees who are actually writing web content, and this should be done for multiple languages. It is a form of terminology management and distribution.

Data categories that could support the management of search keywords in a termbase are described in *Search*.

Extended applications

There is plenty of evidence both in the literature and in actual production environments to show that terminology data is used in a range of extended applications (see for example Park et al 2002; Jacquemin 2001; Oakes and Paice 2001; Cabré 1999b; Bourigault and Slodzian 1999; Condamines 2005; Bowker and Delsey 2016). The author of this book has herself deployed terminology resources for CA, indexing, search engine optimization, machine translation, and for term extraction. These experiences demonstrated that many existing company termbases are ill-prepared for such purposes.

Ibekwe-SanJuan et al (2007:2) note that terminological resources are useful for building other types of language resources such as ontologies and aligned corpora. For corpus management, terminological statistics can help to identify comparable corpora, while using terms as pivots can help to align bilingual corpora or even support the federation of heterogeneous corpora. These types of corpus management activities are essential in advanced computer-assisted translation, but they are also useful for managing commercial content in general. As for ontologies, Wettengel and Van de Weyer (2001:458) describe how terminology resources can help build product classification systems and taxonomies (to many there are distinct similarities between the latter and ontologies). Access to a trusted source of standardized terms is very useful for indexing (Buchan 1993; Cabré 1999b:51; Strehlow 2001a; Jacquemin 2001:305; Nazarenko and El Mekki 2007; Greenwald 1994; Condamines 2007:138, 141). And in these modern times, indexing refers not just to back-of-the-book indexes, but also to collections of websites and other digital media. Many readers have probably experienced being prompted to select a topic or subject area from a predefined list to “tag” some content in order to indicate what it is about. This is a form of indexing that can be enhanced by the use of terminology resources.

As already noted, a sound strategy for using terms in strategic areas of content (such as titles and introductory paragraphs) improves information retrieval and search (Strehlow 2001b:433). Cabré even claims that terminology is the key element for successful content retrieval (1996:29; 1995:6). For their part, semantically-structured terminology resources improve the performance of question answering systems (Rinaldi et al 2003). Ahmad (2001) and Massion (2019) describe the role of semantically-structured terminology resources in developing systems for artificial intelligence and knowledge acquisition. Semantically-structured terminology resources contain a rich set of relations between entries. They are key to certain aspects of repurposability and thus they are described in more detail in *Concept relations*. Wright and Budin provide an extensive list of

language engineering implementations that rely on terminology resources (2001: Infobox 31). This list is undoubtedly longer today, some twenty years later.

It would be short-sighted to fail to see the commercial applications of NLP technologies. Call centers rely on text mining to organize, manage, and plan resource allocation. Indeed, Condamines notes that the need for terminology resources that support business processes such as content retrieval and content management is urgent and growing (2007: 134). She describes two types of NLP applications that are used in the workplace, one for information extraction (or retrieval) and the other for knowledge management. Both require company-specific lexicons (2010: 40). According to Bourigault and Jacquemin, the development of corpus-based tools for building terminology resources has been driven by industrial enterprises and institutions in their quest for more efficient content management systems and processes (2000: section 9.1.1). Automatic content classification, document summarization, and document abstraction – all applications that benefit from the infusion of corporate-specific terminology resources – can help to achieve that goal. They add, however, that due to lack of foresight, terminology resources are rarely optimized to effectively realize that efficiency.

Term extraction, which is described in detail in *Term extraction*, is a terminology management process that itself relies on the infusion of company-specific terminology in order to work most effectively. It can be used not only to find terms to add to the termbase but also to find search keywords for SEO.

Bourigault and Slodzian noted long ago that these needs are increasing due to factors such as technological innovation, internationalization, and the growth of the internet and electronic publishing (1999: 29). More recently, Bowker and Delsey (2016) observe that the increasing use of terminological resources in areas such as SEO and indexing are blurring the boundaries between terminology science and information science. Indeed, many of the tasks we describe in this book that corporate terminologists are actually engaged in (or should be engaged in) are closer to information science as they define it than to terminology science, when viewing the latter in its conventional interpretation. Citing Borko (1968), they describe information science as that which is concerned with “the origination, collection, organization, storage, retrieval, interpretation, transmission, transformation, and utilization of information.” Corporate terminologists are doing precisely these tasks, at least with regards to information in its granular form.

It has previously been noted that the translation focus of termbases has affected their ability to serve such extended purposes. As early as 1990, Sager remarked that large institutional termbases only serve a very specific user group (translators), and “they encounter difficulties in adapting to the requirements of the many new user groups who have emerged” (p. 166). Nkwenti-Azeh describes

the various types of data required for different users of a terminology resource. He observes that there is little information in existing termbanks that can be used to support the needs of NLP (2001: 609).

Termbases designed exclusively for translation lack the structure and content needed for other purposes. For example, Condamines (2007a: 136) agrees with Ahmad and Massion (cited above) when she observes that some NLP applications require terminology resources that contain semantic relations. Yet commercial termbases rarely contain semantic relations. And the ones that do often fail to record them in a systematic, “usable” way. In one major commercial termbase that we know of, for example, semantic relations are recorded in a plain text field which makes it impossible to leverage those relations in any automated process. Furthermore, in this same termbase all relations for a given term are recorded in one and the same text field, which violates the principle of data elementarity. Adopting this kind of approach renders the relations almost useless, and therefore, it is a waste of time and effort to record them at all.⁵² We maintain that this is a legacy of the translation focus, which does not emphasize semantic relations. (Semantic relations are discussed in *Concept relations*.)

Another example is the part of speech – noun, verb, adjective, and so forth. According to the *TBX-Basic* specification (TerminOrgs 2014), part of speech is the most important piece of information to include in terminological entries, aside from the terms themselves. Why? A key reason is that the part of speech is the only machine-readable piece of information that contributes to disambiguating homographs. In order to make appropriate suggestions, spell checkers, such as in CA tools, need to know the part of speech of words both in running text and in the spell-checking dictionary that is used to verify spelling. The term *computer*, for example, is a noun while *compute* is a verb. In running text, “the computer” is correct but “he computer” is not. The writer may have simply made a typo. The CA checker *parses* the sentences and determines that in the first case *computer* is a noun (because it follows the article “the”) and in the second it *should be* a verb (because it follows the personal pronoun “he”). It checks the spell-checking dictionary where it finds the word *computer* associated with a noun part of speech value, so it does not flag the first occurrence as an error. But it finds *compute* in the dictionary where it is indicated as a verb only (infinitive form). Through morphological expansion, it should infer that the third person present indicative form of such a verb ends in “s”. It therefore flags the second occurrence as an error and makes a suitable suggestion. The ability for a spell checker to interpret homographs and their inflected forms properly depends on the availability of a part of speech value in the internal spellchecker dictionary.

52. In the termbase in question, this practice continues today.

Most people have experienced the many false positives (errors flagged which are not errors) reported by simple spell checkers; many of these false positives are the result of a lack of part of speech information. Another scenario where term checking requires the part of speech is described in *Controlled authoring*.

In spite of the useful and practical role of part of speech values, few termbases routinely include them. This, again, is due to the historical view that termbases are tools for translators. Translators rarely need part of speech information, and when they do, they can deduce the part of speech from other information in the entry such as context sentences or definitions. This is why even some of the largest and most well-known termbases in the world lack part of speech data. For a commercial termbase, omitting part of speech information will eventually lead to significant problems.

The part of speech is just one example of the importance of capturing certain types of information in termbases. Repurposability of termbases increases with the sophistication of their structure and the scope of their content. Data granularity is key for repurposability. The more detailed and granular, the more repurposable.

Types of information (i.e. data categories) that are useful to ensure repurposability but are often neglected include the following:

- part of speech
- term type (acronym, abbreviation, etc.)
- subject field
- usage information
- process-related information (workflow stages, etc.)
- business-oriented identifiers for subsetting purposes (project, department, source, owner, etc.).

Further information is provided in Data category selection.

A good guideline for corporate terminologists with respect to the data categories to include in a termbase is *TBX-Basic*. Be aware, however, that this guide does not necessarily include all data categories that any particular company will require.

Towards a theoretical framework

In the previous chapters, we described the conventional theories, practices, and principles of terminology management and we then presented the unique challenges it poses in commercial settings. In this chapter we propose some new ways to view terminology work which we feel better support commercial needs.

Statement of the problem

Before proposing a new way to look at terminology management in commercial environments let us summarize the main issues and challenges raised in the previous chapters.

Compared with other practices in the language industry (translation, professional writing, content management, even lexicography), terminography is uncommon. The field of terminology is perceived as an academic discipline whose commercial applications remain largely unrecognized. The established theories and methods have an unbalanced focus on normalization and do not consider commercial needs. The major termbases in the world are developed and maintained by governments and public sector institutions. Finally, due to its close association with translation, terminology has been slow to penetrate into other areas of content development, and existing termbases are, for the most part, not repurposable. Given these factors it is hardly surprising that terminology management has made insufficient headway in the business world. This is troubling considering its potential to benefit that sector of the economy.

The GTT is the theory that is most disconnected from the practical needs of commerce. This is due to its normative focus, prescriptive goals, thematic and onomasiological approach, systematic orientation, objectivist perspective, exclusion of corpora in its methodologies, and failure to consider language in contextual use. However, certain principles derived from the GTT such as concept-orientation and term-autonomy are essential for corporate terminography.

How can we narrow the apparent divide between “classical” terminology and its practical relevance today? As we have previously suggested, answering this question requires a critical review of conventional theories and methodologies

with respect to the wider content management needs of commercial enterprises. The current and anticipated uses of NLP applications in business processes, and the ways terminological resources could enhance those applications, need to be considered. They include automated workflows, information architecture and knowledge engineering, content management, computer-assisted authoring and translation, bi-text alignment, term extraction, content retrieval, and machine translation. Answering that question in its entirety is beyond the scope of this book but should remain an ongoing concern for corporate terminologists.

Nevertheless, we will propose some alternative views, starting with the notions of *termhood* and *unithood*, which are discussed next.

Termhood and unithood

Corporate termbases frequently contain lexical units that do not fit the classical definition of *term*. Thus, in this section we explore the related notions of *termhood* and *unithood* to see if they can help us understand the roles of semantics and syntax in corporate terminography.

It has been noted that the GTT has shaped mainstream ideas and practices of terminology management, and maintains its influence to this day. This is also the case when it comes to defining what a term *is*. Does the GTT-compliant interpretation of “term” apply to commercial terminography? In other words, do terms denote conceptualized objects, are they confined to subject fields, and are they members of a hierarchically-structured concept system? In an enterprise, the main purpose of managing terminology is to drive efficiency and quality in commercial communications, which span a wide diversity of text types and genres. Will managing *only* terms, as they are traditionally defined, realize those objectives?

Terminologists often use the term *termhood* when determining whether or not a particular word or expression can be considered a term. The conventional notion of *termhood* aligns with Kageura and Umino’s definition: it is “the degree to which a linguistic unit is related to a domain-specific concept” (1996, slightly edited). Not surprisingly, the inclusion of *domain-specific concepts* as an essential criterion of termhood reflects the traditional perception of *term*.

Remaining within the confines of this classical definition of *term*, termhood is a matter of degree, which is subjective; a term can have strong termhood, weak termhood and any degree in-between. A term that has strong termhood is highly domain-specific and probably only used by specialists, whereas a term that has weak termhood might have a domain-specific meaning only in certain contexts or communicative situations, or is transitioning to the general lexicon (a process

Meyer and Mackintosh (2000) referred to as *de-terminologization*). Corporate termbases tend to include a larger proportion of terms with weak termhood than most other termbases.

Termhood is therefore based on semantic properties – how *specialized* is the term’s meaning? In contrast, *unithood* has no semantic basis and is determined strictly on the level of syntax. Kageura and Umino define unithood as “the degree of strength or stability of syntagmatic combinations and collocations” (ibid). In other words, how tightly the words in these combinations or collocations “stick” together. Based on the collocational relationship between words, unithood applies thus only to multi-word units, whereas termhood applies to both single-word units and multi-word units (Wong et al 2009). Unithood is also a matter of degree. Strong unithood is reflected in set phrases such as *terms and conditions* and idiomatic expressions such as *rule of thumb*. For example, *book review* has stronger unithood than *review of books*, to cite from Kageura and Umino. The full forms of acronyms and abbreviations, therefore, have strong unithood since the word order has generated a new term (the acronym) and therefore is relatively fixed. Canada has a *goods and services tax* which is abbreviated to *GST*. Although someone could say “services and goods tax” and still convey the same meaning, we are conditioned to say the former due to its stronger unithood.

All terms in the traditional sense have strong unithood, but not all lexical units that have strong unithood are terms (in the traditional sense) due to the lack of the semantic criterion. The transition from unithood to termhood is therefore monodirectional, as shown in Figure 15.

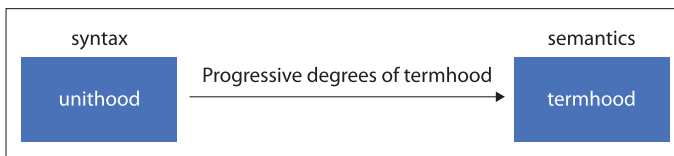


Figure 15. Termhood and unithood

In classical terminology, for multiword term candidates, sufficiently strong unithood is first established to validate the unit syntagmatically, and then sufficiently strong termhood is established to validate the unit semantically. If the unit passes both criteria, it is a term. In corporate terminography, the first criterion can take precedence over the second, and the latter can even be optional. In other words, term candidates that have strong unithood but weak termhood are often included in the termbase. In fact, even candidates that have moderate unithood are frequently desired – to ensure consistency in the SL, to counteract variability

in translations, to increase productivity via shortcut keys for repeated units, and so forth.

In the section *Does commercial content contain terminology?* we argue that commercial language is an LSP and we see the value of a *terminological field* as a pragmatic extension of a *subject field*. At various other points we have explained that commercial terminography is driven by practical needs. Those needs often lead to unithood becoming an important or even sole criterion. Let us now consider some alternate views on termhood and how they might support those needs.

While not denying the role of subject field, several scholars introduce other factors in their views of termhood. Pearson (1998) refers to the relationship between interlocutors such as expert to expert. L'Homme describes translation difficulty (2004) and corpus evidence (context, frequency, etc.) (2005) as key criteria. Temmerman et al (2010) argue that the purpose, application, and target audience of a given communicative act is what establishes termhood. As we have seen earlier, L'Homme accords paramount importance to the application of a terminological resource for determining termhood (2019: 59). Condamines (2005, 2007) states that the "lexical units of investigative interest" (or in other words, the things we decide to put into a termbase) are "observed textual utterances" that warrant our attention by virtue of role and purpose-based criteria. Cabré (1999) views terms as lexical units that, through pragmatic conditions and communication situations, can assume a terminological value. Bourigault and Slodzian produced a treatise in 1999 about the use of corpora in terminology work and coined a name for it: *textual terminology*. All the aforementioned scholars join a growing number who recognize the importance of corpora as the basis for terminological identification and investigation, one that takes into account both semantic and syntactic properties.

Then there is the role of metaphor in the creation of terms. Sánchez et al (2012) provide a comprehensive analysis of this phenomenon, with some examples taken from commercial texts. They note that information technology is "rife" with metaphors. The use of metaphor is a strategy employed to help users understand new concepts by using terms that they are already familiar with. This is also a form of semantic extension. Using examples such as *email* and *desktop*, they note that "being a way of generating new terminology, such metaphors are instrumental in the creation of a user-friendly product" (39). Data is portrayed metaphorically with terms that are normally associated with physical objects: users can *drag and drop* files and folders. The term *folder* and its icon are themselves metaphors. Thus, "metaphor is systematically applied to technology to foster creative thought and enhance user understanding" (40). Metaphor adds a dimension to terms that is culturally-bound and therefore not easily translatable. Transcreating these metaphors in target languages and cultures requires special translation skills, and

the results are carefully stored in termbases. And yet, the classical notion of termhood seems to ignore the role and importance of metaphor.

Thus, there are differing interpretations of termhood (see also *Theories*). What really constitutes a term is subject to continuing debate.

It has been noted already that consistent terms increase the reuse potential of TM, which can lead to significant savings in translation costs. However, even among terminologists, few realize that this principle can also be extended to authoring memory, a technology that is only beginning to make its entry into the content authoring sector. If consistent terms increase the reusability of authoring and TM, and if this objective therefore emerges as a key motivating factor for managing terminology in a commercial setting due to its cost savings potential, this could have wide-reaching ramifications on the scope and methods of commercial terminography. The notion of what constitutes a term in a commercial setting could shift from traditional semantic criteria to statistical measures of frequency and contextual conditions, such as a term's visibility to product users. Setting the semantic basis of termhood aside, these pragmatic criteria go even beyond unithood. We wonder if, for commercial applications, a term could simply be defined as any lexical entity that, when managed according to certain suitable methods, supports the realization of a business goal (efficiency, speed, quality, cost savings, etc.).

Whatever the case may be, it has been indisputably proven that the methods and tools for managing terminology that have developed over the past decades can be effectively applied to commercial content to support those goals, with some adaptations. Corporate terminologists achieve the most success when they embrace the idea that what makes a word or an expression suitable to be proactively managed in a termbase is not whether that word or expression refers to a domain-specific concept or otherwise complies with the traditional definition of termhood. Rather, it is whether or not the effort and costs expended to manage it bring a net benefit to the company.

Companies manage terms to address their language-related issues and support their activities. In commercial settings those issues and activities are driven by pragmatic factors, such as the use of technology and the availability of resources. They are not as linguistically motivated as, for instance, the mission of terminology initiatives in support of minority languages, official language policies in the public sector, or scholarly research. What the corporate terminologist decides is, or is not, a *term*, what should be included in the company termbase and what should not, will necessarily be affected by those priorities.

In commercial environments, therefore, establishing termhood boils down to whether or not including a given term candidate in the termbase will be useful (to any person or to any thing). Lombard, a terminologist at Microsoft, identifies

two broad categories of terms that are managed in software companies: terms important for the marketing of a product (important features and technologies, for instance), and terms that have potential localization (translation) issues (2006: 162) such as translation inconsistencies. Consistency in the SL is also a concern (Warburton 2001b:680), albeit frequently overlooked. While semantics may sometimes be considered when identifying such terms, it is not the only or even the primary consideration.

Ronan Martin (2011), a corporate terminologist at SAS, describes the challenges of deciding what is a useful term in a terminology management strategy intended to meet commercial aims. In his estimation, single-word terms (unigrams) are almost insignificant. This view corroborates Bourigault and Jacquemin (1999) who claim that “single-word terms are too polysemous and too generic.” Martin outlines factors to consider when evaluating term candidates, including the conceptual basis for evaluating termhood, the nature of pre- and post-modifiers, term frequency, context and domain, and term embeddedness (when terms are found embedded in larger terms). Later, term embeddedness as an indication of termhood was also acknowledged by Heylen and De Hertog (2015: 212): “some nested terms occur in many different longer sequences and this also is an indication of termhood.”

Martin points out that the value of a terminological resource is measured according to the reuse of the data: “Building up a set of terms is not a practice that is carried out merely for the purpose of collecting the largest set of terms possible.” Shreve also emphasizes the pragmatic orientation of terminology work, stating that “Terminology management for translators and technical writers is an endeavor with practical goals. It does not exist as an end in itself, but purely to improve the creation of texts and translations” (2001:785).

One of the most important purposes of a termbase is to help translators. Therefore, it makes sense to include in the termbase any word or expression that might pose difficulty for translators or is likely to be translated inconsistently. Certain multiword expressions are challenging, for example:

- *dynamic signal analyzer*
- *direct network measurement*

These two terms are ambiguous due to the lack of explicit markers in English to identify the relationships between the modifiers and the head word. Their unit-hood is rather weak. There are two interpretations of each term and each will be translated differently.

- *analyzer of dynamic signals* OR *dynamic analyzer of signals*
- *measurement of direct networks* OR *direct measurement of networks*

The propensity of English multiword terms to be ambiguous is one reason why they are so prevalent in commercial termbases; there is a perceived need to include them in order to help translators deal with the ambiguity and determine accurate translations. For more on multiword terms, see *Terms considered by length*.

Another challenge for translators is when two different terms occur in the text that might be synonyms. The translator has to decide whether to use one TL equivalent for both (if they are synonyms) or two different ones (if they are not synonyms). Thus, any two terms that could be perceived as synonyms should be documented in the termbase. In this sense, semantic similarity contributes to termhood. Examples of this type of challenge are presented in *Terminology problems and challenges*.

Frequency is an important criterion. A lexical item that occurs frequently in an information set may need to be pre-translated just to ensure consistency in the TL. There is also the productivity factor: sometimes it is desirable to include *repeated fragments* (Bowker 2015) in a terminological resource used in a CAT tool, simply because it is then possible to insert them into the translation with a simple keystroke. This practice of including a wide range of lexical units in translation-oriented terminology resources was first observed by Kenny (1999) and then empirically validated by Allard. The practice in particular was applied to resources integrated in CAT tools where “one-click insertion” (Allard 2012:20; Bowker and Delsey 2016:86) saves time on typing. It seems that translators include a much broader range of terminological units in their termbases than would normally be recommended according to traditional theory.

In addition to domain specific terms the following types of lexical units can be found in commercial termbases:

- Words from the general lexicon, some with a slightly specific meaning, such as *talent* (in HR), *story* (in data modelling), *visit* (a website) and *white paper*.
- Verbs, some specialized such as *rasterize* (an image), some with a slightly specific meaning, such as *sleep* (a computer mode), but others quite ordinary such as *buy* or *copy*.
- Adjectives, such as *agile*, *smart*, and *best-in-class*. They often have a slightly specialized meaning or a marketing connotation.
- Marketing slogans and other key messaging, such as *hire to retire* (from a human resources company), *big ideas for small spaces* (about printers), and *make anything* (pertaining to an architectural design software).
- Set phrases, such as *return for refund*, *terms and conditions*, *point of purchase* and *satisfaction guaranteed*.

- Colloquial and figurative expressions, such as *ballpark estimate*, *gold standard*, and *bottom line*.
- Field and form labels, such as *Dimensions* on a product data sheet and *Family name* on a form.
- Proper nouns (appellations), such as the names of products or companies. For example, *Red Balloon*, *Dear Lucy*, and *Bonjour* (these are indeed names of actual companies or products).
- Some frequently-occurring strings of text.

Each of these types of lexical units needs to be proactively managed throughout the global content pipeline in order to realize goals such as brand consistency, improved content quality, and improved translations. In the termbase of a printer manufacturer, for example, the names of colors were included since their precise (and consistent) translation in all languages is essential. Yet because words denoting colors are members of the general lexicon, they are not terms according to traditional terminology theory. Certain frequently-occurring multiword expressions, such as *return for refund* or *terms and conditions*, need to be translated and managed very carefully. And yet these types of expressions denote more than one concept strictly-speaking (a refund, the action of returning it; terms, conditions). It is uncertain how units like these would or should be handled according to the GTT.

The notion of termhood in commercial terminography must therefore be broad to account for lexical units that are required to achieve goals of efficiency, quality, and economy when producing multilingual content. It needs to focus on the communicative needs and goals of the organization and be based on pragmatic criteria such as translation difficulty, frequency, visibility (such as on packaging), and other factors. The selection of terminological units must also be based on corpus evidence. The use of corpus evidence to establish termhood is further described in *The termbase-corpus gap*.

To summarize, in addition to terms in the semantically-restricted conventional sense, corporate termbases may need to include various other lexical units that are selected according to purpose-based criteria.

The types of lexical units included in commercial termbases do not always fit the classical definition of *terms*. We suggest that even the use of words such as *terms* and *terminology* have contributed to a general misunderstanding of commercial terminography and have made it difficult for objectives such as repurposability to take hold. For this reason, we contemplate introducing a new term that more accurately reflects what corporate terminologists include in termbases: *microcontent*.

Microcontent

Given that in commercial settings the notion of what a term *is* and the criteria that establish *termhood* divert somewhat from their conventional interpretations, the words *term* and *terminology* have been problematic and misleading in representing the types of data that corporate terminologists actually work with. Being generally misunderstood, these words have marginalized (and even undermined) the work.

If we could brand the type of terminology work done in companies differently, we might no longer be shackled by convention, which could change the landscape dramatically. As explained in *Content management* we therefore propose the word *microcontent* as an occasional alternative for *terminology*.

Jakob Nielsen described microcontent in 1998⁵³ as “short content like headlines, understood out of context.” In 2002, blogging pioneer Anil Dash refined the idea, describing it as content that conveys one primary idea or concept.⁵⁴ Although scholars in fields that study content, such as linguistics and lexicology, may not use the term microcontent directly, they generally view this type of “small” content as being univocal (see *Univocity*).

Text that corresponds to one concept is microcontent. Usually it is a single word or a few words. What is interesting for terminologists is that it is the concept level at which content is translated. Contrary to popular belief, translators do not translate words, sentences, or paragraphs, or at least they are not supposed to. They find an equivalent way to express, in the TL, the concepts that are expressed in the SL. Ask any professionally-trained translator and they will agree wholeheartedly. That is why viewing microcontent as content corresponding to a concept and managing content at this microlevel is essential for good translation, as well as for a host of other company applications that depend on lexical knowledge. And this is precisely why managing microcontent is how we should truly brand the terminology work that is carried out in enterprises.

These units of microcontent are, for better or for worse, often called *terms*. Terms express key concepts in an organization. Figure 16 shows, in the form of a word cloud, the key concepts from a company that produces antivirus and security software for mobile devices.

Terms like *malware*, *mobile security*, and *mobile threat* represent significant concepts for this company. One can well imagine that these terms are repeated over and over again in many different documents, websites, videos, and other

53. nngroup.com/articles/microcontent-how-to-write-headlines-page-titles-and-subject-lines/

54. anildash.com/2002/11/13/introducing-microcontent-client/

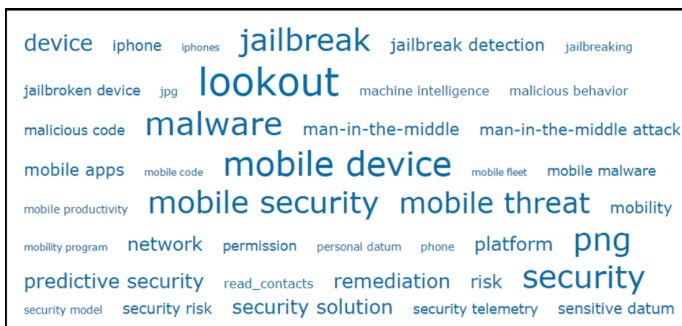


Figure 16. Word cloud

media. Obviously it is important to use these terms consistently, to also avoid using alternate expressions for the same concept, and finally to find appropriate equivalents in other languages and use them consistently as well.

Sometimes a phrase or a slogan can become a fixed piece of content that has significant marketing value. Marketing slogans are examples of phrases that could also be considered microcontent. Consider the following set:

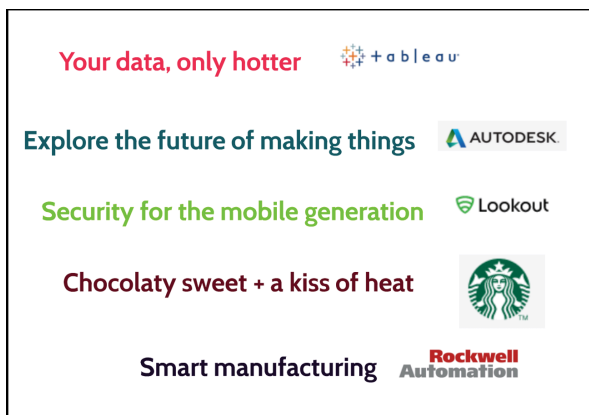


Figure 17. Marketing slogans

The same slogan is repeated throughout the company's content and becomes part of the company's core identity, its brand. At this level, the content cannot be dissected into smaller parts and it behaves like a single concept. Take the example of *smart manufacturing*. We would not normally think of manufacturing as being smart. Intelligence is a property of the human brain and not of a mechanical process. This unusual and unexpected use of language gives the slogan more

impact and increases its retention capacity in our minds. These kinds of fixed phrases can be considered a type of microcontent.

Visuals convey concepts and are therefore also considered a kind of microcontent. Icons that represent products, such as the various symbols that represent the suite of applications, are a good example. With a similar size, shape, color scheme, and so forth, these icons are designed to present a unified visual brand for all the products in the suite.

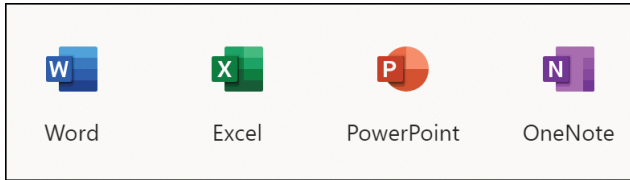


Figure 18. Some Microsoft® applications

Visuals like these represent concepts, and therefore, there is no reason not to include them in termbases.

When key terms are identified, used consistently, translated correctly, and applied to other processes such as SEO, better content is being built from the ground up. This is a bottom-up content management strategy. It minimizes the problems that are inherent in a top-down approach where poor word choices and other problems at the elementary level of content are not discovered until very late in the content management cycle, if at all, at which time they are often widespread and more costly to fix.

Content that is more consistent at the micro-level is more repurposable. More content can be reused or recycled, such as in translation memories, when it is standardized at an elementary level. This actually reduces production costs. With effective search keywords strategically placed in its web pages, the company will rank higher in internet search. And clear, consistent messaging is more engaging, more impactful, and simply better on all fronts.

From a technical standpoint, microcontent has three properties that larger pieces of content, like sentences and paragraphs, do not. Microcontent is univocal, discreet, and infinitely reusable.

As described earlier, univocal means having a single meaning. For example, the concept of a *hybrid cloud* can be clearly defined and even illustrated visually in relation to other concepts.

Figure 19, in addition to providing a good definition for hybrid cloud, is a source of other terms: *cloud*, *cloud computing*, *cloud service provider*, *private cloud*, *public cloud*, and *firewall*. The figure is very effective for conveying the con-

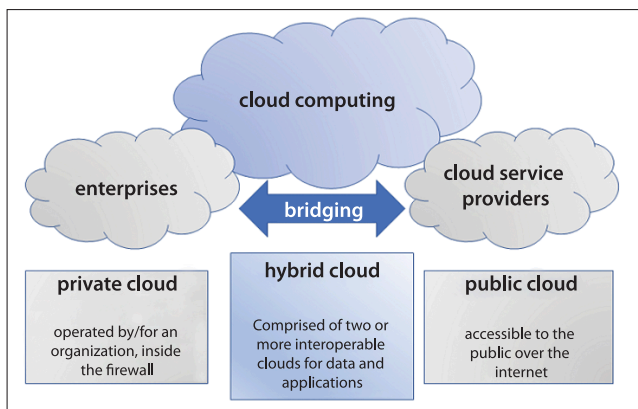


Figure 19. Concept: hybrid cloud

cepts and is ideal for genres such as user documentation. However, in this graphical format, the terms and their explanations are not repurposable. They will not be retrieved by a search and they cannot be used in any other application. By considering these units of information microcontent, and additionally storing them in a termbase, one can unlock their repurposability.

What about words that have more than one meaning? Does that mean that they are *not* univocal? For example, the word *cloud* refers figuratively to the internet, and concretely to the actual clouds in the sky. The property of a word having multiple meanings is called polysemy. Polysemy is very common. Many words, therefore, are not univocal. However, as we stated before about terms, microcontent is also small pieces of text that are based on a single meaning. The word *cloud* when it refers to the internet is one instance of microcontent, and when it refers to clouds in the sky it is yet another instance of microcontent. Each instance of microcontent, considered in the same conceptual frame as terminology with its binary signifier-signified relationship, is univocal, whereas larger forms of content – sentences, paragraphs, documents – are not.

The next property is discreteness. The word *discrete* is often associated with data. Discrete data is minimalist; it cannot be further sub-divided. It has stable properties, and it can be managed as a self-contained unit. Microcontent also has these properties. This means that microcontent can be managed in a database the same way that numerical data can.

An example can help illustrate the nature of microcontent and its repurposing potential. Figure 20 is an excerpt from a SAP product website.

The term *game server* is repeated nine times in a very short text, including twice in the figure. The term *gamification*, also repeated in the figure, could be a neologism or certainly seems so. The term *learning games* suggests that there

Configuring a Learning Game Server

Before you can administer **learning games**, you must have a **game server** in place because the game takes place on the **game server** and reports the results back to SAP SuccessFactors Learning.

Before you can administer the game settings in SAP SuccessFactors Learning, you need to set up a **game server**. We currently support **Bunchball** game servers. Contact **Bunchball** to set up your game design and integration.

After you configure the **game server** you connect to it from the **LMS** through the **GAMIFICATION** system configuration. SAP SuccessFactors Learning sends **points** to the **game server** where your company's game is configured. The **game server** runs the game logic and sends back user achievements (**badges, points, leaders**, and so on).

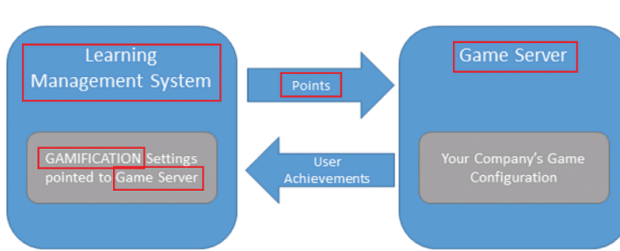


Figure 20. Game server – key concepts

may be other types of games as well. *Bunchball* is a proper noun and probably not translated. The acronym *LMS* must correspond with *Learning Management System* although the only clue is in the figure not in the text. Key concepts for the game are *badges*, *points*, and *leaders*. Obviously, it makes sense to use these terms consistently and to find appropriate, logical, and consistent equivalents in other languages. Taking these words out of their physical location buried in text or in an image as the case may be, recording and managing them as separate entities in a termbase, and making this information available to employees in the technical environment where they work, makes it possible to enforce consistency and good translation practices.

We will further discuss the property of infinite reusability and provide some examples later in this text.

Elements of a new theory and methodology

In this section, we attempt to synthesize previous discussions, debates, and challenges into elements that could be considered as foundational for a theory and methodology that is suitable for commercial terminography, one that considers the needs of the more active and intense production environments (enterprises,

global organizations, language service providers). We do not claim to establish such a new theory and methodology, but only provide observations and raise questions that might support its eventual establishment.

First, terminology management in these environments does not follow any of the existing theories and methodologies, but especially not the ever-dominant GTT. It is not normative, it is not confined to subject fields (in the classical sense), and especially, it is not preoccupied with concept systems. Terminology management in these environments is oriented towards communicative acts (authoring, translation) in widely ranging contexts, genres and media, it leverages technology to the highest degree, and it is repurposed for content management, CA, CAT, SEO, and various other NLP applications.

The GTT has been subject to criticism, even outside of commercial concerns of the type expressed in this book. We maintain that given the significant societal changes of past decades, the basic premises and established methods of terminography in general are subject to critical examination to ensure their applicability to commercial terminography in particular.

What constitutes a term, what defines its *termhood*, is an area where corporate terminologists must frequently deviate from tradition. The field of terminology has witnessed a shift in the notion of termhood from a representation of an objectivist, language-independent concept to a contextually-dependent expression fulfilling a given communicative purpose, an instance of microcontent. In our opinion, adherence to one notion or the other depends on the circumstances and aims of the terminology initiative. In a highly controlled normalization environment, the former interpretation serves a useful purpose. In a more dynamic communicative setting, the latter would be more applicable. Commercial communications fit into the latter category more often than they do the former.

As justified in *Termhood and unithood*, the criteria for termhood should be widened beyond domain specificity to include a series of pragmatic criteria oriented towards the communicative needs of the organization. While not abandoning entirely the conventional interpretations of what a term is, we have to acknowledge that it is equally if not more important to manage the linguistic units that *need* to be managed in order to meet the goals of the company or organization sponsoring the effort. If that means, for instance, occasionally putting into a termbase items that would otherwise not be permitted according to convention, such as full phrases, inflected forms, or other “anomalies,” so be it, provided that doing so is motivated by an express need, does not undermine repurposability, and is not entirely random or accidental.

Aside from univocity, which has been reconciled by constraining its interpretation to specific subject fields, the GTT also recommends that a domain-specific concept should not be denoted by more than one term. This is an idea that we

cannot reconcile since it disallows synonyms. It was meant to support normalization and standardization. In most other communicative situations, including commercial contexts, synonyms are very common.

This brings us to concept orientation, a long-standing core principle. This principle continues to be essential *precisely because* in companies there is indeed a need to know about and manage synonyms. Synonyms are necessary as alternate terms which are more appropriate for given contexts or aims. Economy, for instance, is a stylistic goal that is achieved through the use of acronyms, abbreviations, and shorter forms of multiword terms, all of which are types of synonyms.

The onomasiological approach to terminography prescribed by the GTT is impractical in commercial environments. In fact, with the exception of strictly normative environments, which are quite rare in the terminology field, we might dare to suggest that the onomasiological approach is hardly used anywhere. Nowadays, spurred by the recent availability of large-scale corpora, most terminologists worldwide, regardless of the setting (commercial or otherwise), use the semasiological approach in their daily work. With the semasiological approach, text-based research methods leveraging corpora gain more importance. Contextual information, collocations, and other traits that emerge from corpus evidence are frequently collected and recorded in termbases, and this trend is expected to continue.

To summarize, commercial terminography:

- is semasiological, but still concept-oriented
- adopts a widened interpretation of subject field/domain
- is driven by pragmatic needs (communication-oriented and production-oriented)
- reflects (or should reflect) the organization's corpus
- should develop repurposable resources
- extends the notion of termhood to embrace various types of microcontent
- seeks to discover as many synonyms as possible
- is primarily descriptive, but adopts some prescriptive approaches in particular for controlled authoring and controlled translation.

PART 3

Planning a corporate terminology initiative

Having discussed many of the theoretical and methodological foundations of terminology as a discipline and debated their applicability to commercial terminography, in this part we describe how to propose and get approval for implementing a terminology management strategy and its flagship product, the termbase. This part also covers the key features to be considered when selecting a terminology management system, and how to determine what a termbase should contain.

The proposal

The proposal to develop a terminology management program for the company is prepared by the corporate terminologist, or the would-be corporate terminologist. It is an appeal for the mandate to implement a terminology management process. It should include high-level descriptions of the main products and services to be offered, as well as a business case and budget.

It is a good idea to approach the proposal document as if it were the foundation upon which the program elements will be further defined over time. Many elements of the proposal can be reused in other documents down the road, such as program documentation, training materials, promotional materials, and so forth. Detailed program documentation is required for ISO audits; eventually everything will need to be fully documented. The program documentation is likely to become quite large. Elements of the proposal itself should therefore be *repurposable*, just like we have argued for terminology resources. With this aim in mind, consider writing the proposal in a stable and professional XML format such as DITA.

Organizational position

Large companies that are hierarchically-organized tend to divide authoring⁵⁵ and translation into two distinct departments, if not more (for example if each TL has its own organizational entity). This tendency to view global communications as a tree with two main branches has cultural roots – authoring and translation are still today largely viewed as separate activities and vocations whose workers rarely need to communicate with each other, but it is a view that is gradually changing. This two-sided coin can present difficulties for the terminologist given that the terminology initiative bridges both areas.

55. We use the term *authoring* for all types of content production in the SL. It has been referred to by various names such as *technical writing*, *information development*, *content development*, etc.

As two separate departments, authoring and translation typically have separate budgets. They often have to compete against each other, year after year, to obtain their budgets from the same bucket of funds. This annual fight for funding only aggravates the competitive environment and adds tension to the relationship between the two groups. Under these circumstances obtaining funding for the terminology initiative can be particularly challenging since each group may be reluctant to allocate any of its funds towards it, hoping that the other will take responsibility instead.

There is also a fear factor associated with any new proposed initiative. People may feel threatened and act defensively in response to any change to the status quo. "Introducing terminology management within an organization can meet with some objection" (Cerrella Bauer 2015: 326). The terminologist needs to take careful measures to address concerns and garner support.

Thus, the terminologist's placement and that of any support staff within the organizational structure can have a significant effect on the ability to network with stakeholders and gain their support. In many cases, the practitioners involved do not have a mandate from management for deploying such an initiative. Even with a mandate, too often the terminology unit is positioned in the translation division due to the translation focus described earlier. A formidable divide between authoring and translation communities has become entrenched in corporate culture, as each community has its own rules of operation, reporting structures, budgets, practices, tools, cultural norms, and more.

It is therefore recommended to position the terminology unit organizationally in a way that clearly demonstrates how it serves both communities equally. Positioning it within or subordinate to either the authoring or the translation departments can be a recipe for failure. Figure 21 represents one possible model, where the terminology initiative is subordinate to Information Management which is simultaneously responsible for other information management processes, some of which might be able to leverage the terminology resource, such as SEO.

This structure fosters equitable treatment and stature of the terminology initiative with other information management units while fostering close collaboration between them. Positioning terminology between corporate communications and authoring and translation, while better than positioning it below either of the latter, may have the disadvantage of making it more difficult, logistically or budget-wise, for the terminology unit to partner with other higher-level information management areas.

The overall success of the terminology initiative may depend on the terminologist having direct communication with, if not directly reporting to, an executive-level sponsor who is convinced of the merits of the proposal. This person will have responsibility over both authoring and translation, if not overall information

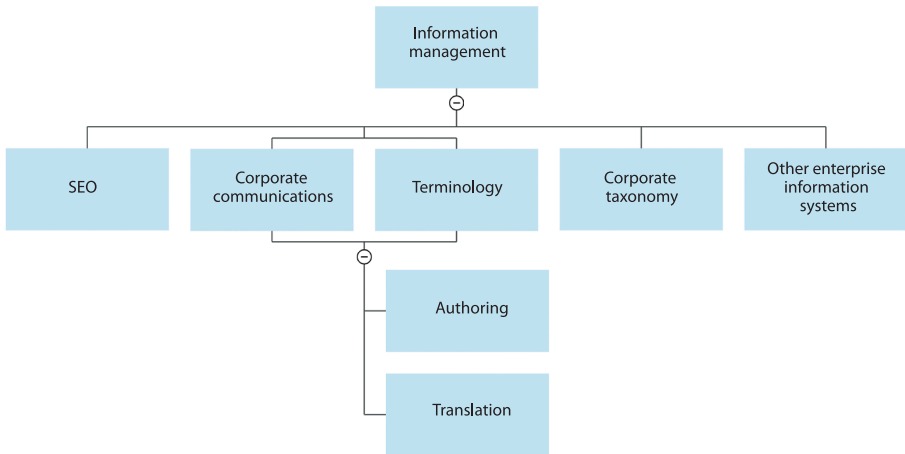


Figure 21. Organizational placement of the terminology initiative

management, such as a VP of Globalization. They can support the terminologist in overcoming the challenges of establishing a framework for cross-department collaboration, bridging authoring and translation, being accorded an officially recognized mandate, obtaining stakeholder buy-in, arranging budgets, and so forth.

Convincing the executive sponsor of the merits of managing terminology is the first major challenge for a terminologist. This involves preparing a high-level proposal and business case, aspects of which are covered in the next sections.

Standards and best practices

To secure the return on investment in the terminology initiative, it is important to be informed about and to consider existing standards and best practices. Demonstrating that the proposal is based on recognized standards and best practices will strengthen the proposal and raise the confidence level of those entrusted with its approval.

At the same time, the corporate terminologist should not accept all standards without question. As with everything else, standards need to be critically examined to ensure that they align with the organization's goals and needs before it is decided that they should be adopted.

A *terminology standard* should not be confused with a *standardized term* or a *standard terminology*. The former refers to a standard about how to manage terminology, how to structure a termbase, and so forth. The latter two refer to terms that have been declared standardized or approved by an authoritative body. An

example of a standardized terminology is section 3, *Terms and definitions*, that is found in ISO standards, which cover everything from bicycle helmets to textiles. This section includes terms and their definitions, which are standardized within the context of the document itself and are meant to be used in communications that cover standards in that area. For example, Figure 22 shows the term *rechargeable extinguisher*, a synonym *refillable extinguisher*, a definition, and a note. Because it is shown in bold, *rechargeable extinguisher* is the standardized, preferred term for this concept according to ISO.

rechargeable extinguisher

refillable extinguisher

fire extinguisher that is capable of undergoing complete maintenance, including internal inspection of the pressure vessel, replacement of all substandard parts and seals, and hydrostatic testing, and of being recharged with media and propellant and restored to its full operating capability by means of the standard practices used by fire equipment service companies

Note 1 to entry: Rechargeable/refillable extinguishers are marked "Recharge Immediately After Any Use" or with a similar equivalent marking.

Figure 22. A standardized term from ISO

Standardized terminology from ISO can be freely accessed in the ISO Online Browsing platform.⁵⁶

As of this writing, ISO Technical Committee 37 comprises over 60 national members and has published 67 standards in the field of terminology, translation, interpreting, and other related fields, with 27 more under development.⁵⁷ It was established with the involvement of Eugen Wüster himself in 1947. Not surprisingly, the standards focusing on terminology resources tend to reflect GTT principles. As a consequence, some are not entirely applicable for commercial terminography.

At the risk of certain opposition, we would like to suggest a certain ISO TC37 standard that we feel is not entirely suitable as a model for commercial terminography, namely *ISO 704 - Principles and methods*. This standard is highly regarded among terminologists, particularly those in academic circles, and is considered among the founding standards of the TC37 portfolio. Describing fundamental concepts and established principles and methods, it is an important standard that all terminologists should be familiar with. It is particularly relevant for constructing terminological knowledge representations, such as SNOMED, which is a systematically-organized collection of medical terms.

By virtue of its title alone, one could presume that ISO 704 establishes standard principles and methods for terminology management in general. In reality, it

56. iso.org/obp/ui

57. The full list is available on the committee website: iso.org/committee/48104.html

describes practices that are more suitable for prescriptive terminology and reflects the principles of the GTT to achieve the goals of standardization (see *Terminology management*). Temmerman makes a convincing argument for a broader set of practices, stating that “the mistake made by traditional terminology theory was to proclaim the standardisation principles as the general theory of terminology” (1998: 78).

For example, ISO 704 requires definitions to be written according to certain rules based on an analysis of concept properties established through the creation of concept systems. These activities are not routinely carried out in commercial terminography. When definitions are written for a company termbase, strict rules are rarely followed, nor are they even necessarily useful. A more flexible and pragmatic approach not only saves time and effort, but it enables definitions to be constructed in a way that focuses on the target audience. The user-focused approach is advocated by the Socio-cognitive theory: “Depending on the type of unit of understanding and on the level and type of specialisation of sender and receiver in communication, what is more essential or less essential information for a definition will vary.” (Temmerman 1998: 79). Corporate terminologists need to be aware of ISO 704 since the principles it describes form the basis of traditional terminology. However, they need not always comply with those principles, especially if the latter conflict with pragmatic needs.

TerminOrgs has published several standards and guidelines intended for terminologists working in large organizations, including *The Terminology Starter Guide*, the *TBX Basic Specification*, and *Terminology Skills*.

Other organizations have produced terminology standards⁵⁸ that address specific use cases. For instance, the W3C Consortium (w3.org) published the *Internationalization Tag Set (ITS)*, which offers a set of markup elements for identifying terms and providing translation-relevant information directly in source content.

The Terminology data category is used to mark terms and optionally associate them with information, such as definitions. This helps to increase consistency across different parts of the documentation. It is also helpful for translation.⁵⁹

Similarly, the *Text Encoding Initiative (TEI)*, a “consortium that collectively develops and maintains a standard for the representation of texts in digital form,”⁶⁰ defines a set of XML elements for representing glossaries, terms, definitions, and other types of terminological data in documents.

58. Some “standards” are referred to by different names, such as *Recommendation*.

59. w3.org/TR/its20/

60. tei-c.org/

For its part, the *Organization for the Advancement of Structured Information Standards* (OASIS) develops the all-important DITA standard (Darwin Information Typing Architecture), an XML markup language and topic-based information architecture that is increasingly being adopted by companies as their authoring format, replacing the unstructured word-processing formats of the past.⁶¹ Using XML as an authoring format provides the ability to encapsulate information about terms in the source, through ITS. We will show how to do this in *The authoring community*.

The Object Management Group (OMG) has developed a specification, *Semantics of Business Vocabulary and Business Rules* (SBVR), which “defines the vocabulary and rules for documenting the semantics of business vocabularies and business rules for the exchange of business vocabularies and business rules among organizations and between software tools.”⁶² Basically, it leverages terminology in the form of rules that can be used to logically structure and automate certain business processes.

The most important standard for guiding the design and development of a termbase is *ISO 30042 – TermBase eXchange (TBX)*, last published in 2019. As the title indicates, the primary purpose of this standard is to foster exchange (interchange) of terminological data between systems or users, but it also reflects best practices for the design and content of termbases. The TBX-info website (tbx-info.net) provides a wealth of information about this standard, as well as some tools and other resources to facilitate its implementation. TBX is compliant with *ISO 16642 – Terminological Markup Framework (TMF)* which specifies the structure of terminological (concept) entries. By virtue of its association with the Data Category Repository (datcatinfo.net), TBX is also compliant with the *ISO 12620* series of standards that govern data categories.

Two core principles of TBX are *modules* and *dialects*. A module is a list of the data categories and their constraints (for example, the levels of the entry where they are allowed to occur) that are permissible in a particular dialect. A dialect is an XML markup language (expressed in TBX format) that adheres to the core structure of TBX and allows the data categories that are specified by a particular module. To put it simply, to comply with this standard, a termbase must adhere to the core structure of TBX and it must allow a certain set of data categories, which are documented. The combination of the two (core structure and selection of data categories) constitutes a dialect. The module is a mechanism to support interchange by documenting the data categories for a given termbase. The TBX steering committee has defined a standard formalism for expressing a module, the

61. This book is authored in DITA.

62. omg.org/spec/SBVR/About-SBVR/

TBX Module Definition (TBXMD). Thus, information about a termbase can be exchanged in a TBXMD file and used by a receiving termbase or a receiving application in order to permit seamless data flow.

Both dialects and modules can be either public or private. Public dialects and modules are meant to be developed through industry feedback, to reflect certain “types” of termbases. For example, TBX-Basic is a public dialect that was developed by TerminOrgs. Private dialects are developed by individual organizations and are not intended for wider industry adoption. You can develop your own private dialect.

Users and their roles

Before you can define a terminology management process you need to determine who has a need to use the terminology resources and services. Establish a good working relationship with all stakeholders of terminology in the organization. This includes:

- producers of various types of SL content: marketing, product development, legal, support, internal communications, editors, HR, etc.
- members of the translation community: project managers, translators, reviewers, subject-matter experts, product managers in target markets
- managers and strategists involved in planning both content production and translation: product managers, information architects, user-centered design, publishing centers, globalization executives.

The primary users of terminology are translators. They usually need little convincing about the value of terminology work. The terminology program should be tailored to meet their needs. This includes, for example, that the termbase should be multilingual, that it should connect with CAT tools, that there should be suitable approval workflows, and so forth. Some translation requirements were already described in *Translation*.

As described in *Where can terminology be used?*, there are users and potential uses of terminology resources that extend beyond the field of translation, and this is where the terminologist faces great challenges. It can be difficult to convince executive sponsors of the potential for such uses, and later obtain agreement and buy-in from employees in those areas.

First, the corporate terminologist should be acutely aware of the need to manage quality in the SL – and this includes both terms as such and word usage in the general sense. They should be prepared to lobby strongly for terminology management in an organization’s SL community, since this is where terminology as a

discipline, resource, and service, is the least recognized. This community includes producers of various kinds of SL content (documentation, marketing, product interfaces, company policies, legal contracts, etc.), and their managers. If the organization has a style guide (and it should), the termbase should be positioned as a way to improve adherence to its contents.

Depending on the process adopted to review and approve terms, which is often referred to as *vetting*, there will be some people entrusted with making decisions about term usage and recording those decisions in the termbase. These people should have subject matter expertise, or in other words knowledge of the products, services, or technologies concerned. They will need to have write access to the termbase, which can be restricted based on need, for instance, to a specific language.

Second, there are uses *by machines*. As microcontent, terminology is a form of machine-readable data. It is therefore subject to various potential uses in the broader field of information science. The terminologist needs to explore innovations in NLP to see what possible uses for terminology data might be on the horizon. We have already touched upon many of these potential uses in *Where can terminology be used?*.

Once the users have been identified, roles can be set up in the terminology management system (TMS) that is used to manage the termbase. These roles will later help to determine interfaces, views, and access restrictions for the various users. The following are some examples (Note: the *levels* referred to in the list will be described in *The data model*):

- System administrator
 - Create users and groups, and specify access controls
 - Unlock a locked-out user
 - Reset passwords
 - Perform backups.
- Terminologist
 - Create a termbase
 - Delete a termbase
 - Create online help
 - Full write access to all entries
 - Import and export
 - Create filters, views, workflows, and other termbase objects
 - Create users and groups, and specify access controls
 - Unlock a locked-out user
 - Reset passwords.

- Content producer (technical writer, marketing writer, etc.)
 - Look up SL terms (to verify spelling, meaning, acronyms, usage, etc.)
 - Submit proposals for new SL terms
 - Add comments to entries
 - Use term checker in CA tool.
- Editor (SL)
 - Look up SL terms (to verify spelling, meaning, acronyms, usage, etc.)
 - Submit proposals for new SL terms
 - Add comments to entries.
- Terminology approver (SL)
 - Edit concept level and SL part of entries
 - Add SL terms to existing entries
 - Create new entries with concept level and SL level content
 - Write usage information for SL terms
 - Rank SL synonyms according to usage preference (preferred, admitted, prohibited, etc.).
- Translator
 - Look up terms (to find meaning, find translations)
 - Submit proposals for new SL terms and TL terms
 - Add comments to entries
 - Use term lookup in CAT tool.
- Terminology approver (TL)
 - Edit TL part of entries
 - Add TL terms to existing entries
 - Write usage information for TL terms
 - Rank TL synonyms according to usage preference (preferred, admitted, prohibited, etc.).
- Other employees
 - Look up terms
 - Add comments to entries.

Note our proposal that the terminology approver for the SL have additional permissions compared to the terminology approver for the TL, specifically with regards to concept-level information. This is merely a suggestion and may not be necessary or even appropriate in some organizations. In a commercial setting, new concepts reflecting new products, services, technologies, or innovations, typically originate in the SL. It therefore makes sense that employees working in the

SL, and preferably those who have subject-matter expertise, should be responsible for describing concepts. This can include providing definitions, explanations, setting the subject field, and so forth.

Stakeholder engagement

Once the users have been identified, you should initiate contact and explain the plan and proposal. Contact the respective managers first and request their permission to contact their team members. Do this early rather than later; teams can become resentful if they feel that plans have already been made without their input. Emphasize that the purpose of your meeting is to present the proposal and get their feedback, which will be duly considered.

Each contact, even the first one with the department manager, should include a slide presentation. Keep it short, use simple language, and explain the basic concepts (such as even *terminology*). Present the elements of the proposed terminology program. Emphasize the benefits using the key arguments in the business case (described in the next section). Do this only for the company at large, but especially for the group you are talking to. Some people may feel threatened by the proposal; be prepared to address concerns. Typical concerns include the following:

- From all users:
 - What is your mandate? Do you have executive-level support?
 - I don't have time to contribute to this effort. What are you expecting from me exactly?
 - Your business case says that this will save us time. How will this affect our jobs? Will we be expected to demonstrate this time savings? Will we be allocated other work to fill the void?
 - If this terminology program saves time and increases productivity, could it eventually lead to job cuts?
- From content producers in the SL:
 - I think this is something that only translators need. It's a multilingual thing. Why are we involved?
 - Why do we need to be concerned about terminology? We don't have problems.
 - Do we have to follow your rules or advice about terms? What about creative license?

- Who is going to have the ultimate say about terms in our writing? This is our responsibility.
- From translators:
 - We have our own existing terminology (database, glossaries, etc.) which is serving our purposes. Why should we move it into a central termbase?
 - This seems like a lot of work. Will our department have to spend time on this or contribute to the budget?
 - What is the benefit for us compared to the status quo?
 - Our translators know terminology already. We don't see a problem or the need to change anything.
 - Terminology in our language is pretty straight-forward. We don't see the need for this.
 - How will this affect our existing procedures which are serving our needs?
 - Will individual translators be able to submit terms? How will we maintain quality control?

Many of the answers to these and other questions and challenges should be found in your business case (see *Business case*). Unfortunately, even when presented with convincing arguments attesting to the value of the terminology initiative, some may not show their support out of fear that they may be affected personally. There is also the classic “turf war” phenomenon – people protecting their situation and their range of responsibilities and resisting change. Some will resent the perceived intrusion by an outsider who they feel is crossing managerial boundaries. The terminologist should be sensitive to people's concerns, work gradually to gain their confidence, be patient, avoid confrontation, and get help when necessary from the Executive Sponsor.

Take note of all feedback and follow up diligently. Demonstrate that you are listening. Address all concerns, if not during the meeting then soon afterwards.

Conduct follow-up meetings at various milestones during the project. Show the achievements and the positive results. Get their continued feedback as they start to use the resources. Offer your services for training and support.

The authoring community

The benefits and applicability of terminology for translation are fairly well known but the same cannot be said for authoring. As described in *Organizational position*, there is also the tendency for upper-level management to relegate the terminology initiative to the translation arm of the company. This should be avoided

at all costs. Therefore, during the proposal phase special effort should be made to clearly articulate the potential benefits for authoring.

In *Authoring*, we briefly described how developers of SL content (technical writers, marketing writers, etc.) should be actively engaged in ensuring that terminology is correct and consistent in their writing. They should also be inserting markup in their documents to identify terms and concepts. We will now elaborate further on the second idea.

Marking up terminology and related information in source content is very powerful for driving efficiency in the translation and localization process. Consider including this approach in the terminology process proposal and its benefits in the business case. But note that it requires that the source content be written in an XML markup language such as DITA, TeX or DocBook. And if XML is not the authoring format already, the company should seriously consider moving in that direction. The advantages are tremendous.⁶³

Certain standards or best practices already exist to provide guidance on how to mark up terms and related information in source content, in particular, ITS and TEI. (These standards were described in *Standards and best practices*.) Each markup language, i.e. DocBook or DITA, may also have its own markup for these purposes. If the organization does not use an existing industry format but has created its own, TEI can act as a useful model for adapting a proprietary format.

With both the TEI and the ITS, corporate terminologists have all the ammunition they need to propose that terms, definitions, and other information be encapsulated directly in the source documents. Doing so will drive efficiencies in various downstream stages, such as extracting terms and definitions for the termbase, and notifying translators directly in the source content if a term should be translated or not. It also encourages the authoring community to take a more active role in the process. They can identify terms as they are writing and share their own definitions, which they are composing regardless to explain concepts to readers, with the terminologist and translators. It gives them an increased sense of ownership.

Figure 23, taken from the ITS specification, shows how a term and its associated definition are identified:

```
<p>We may define <term def="TDPV">discoursal point of view</term> as  
<gloss xml:id="TDPV">the relationship, expressed through discourse structure, between the  
implied author or some other addresser, and the fiction.</gloss></p>
```

Figure 23. Example of ITS markup

63. It is not within the scope of this book to cover this important topic. See *Further reading and resources*.

The types of information that writers can encapsulate in their content by using XML include:

- terms
- synonyms (for instance, indicating the full form of an acronym)
- definitions
- notes to translators
- part of speech
- translatability (whether the term should be translated or not)
- other attributes (XML is extensible to account for any attribute or type of content desired).

By marking up terms, synonyms, and definitions, the writer can automatically generate a glossary for the document. This same markup enables the terminologist to pull the information into the termbase by various programmatic means. Translatability information and translator notes directly aid the translation process, and reduce questions and answers sent around by email or logged into a separate application.

With DITA (and we assume other document markup languages to be similar) writers are already expected to identify *index terms* for generating an index. Clearly, many of those terms are the same ones needed for the termbase and for translators to pay attention to. But since writers will not be accustomed to using such markup when it is initially rolled out, they will need to be educated about what these different types of information are, and how to use markup to encapsulate them. They will need to be convinced of the benefits of using the markup. It will take some time before the markup is regularly used, and writers will need to be periodically reminded about the importance of their collaboration.

Business case

The reasons why a company would decide to manage terminology were presented in *Motivation for managing terminology*. In this section, we present some ideas for creating a business case, which is the first step to obtaining approval and funding for the terminology initiative. A business case includes two basic components: cost and benefits.

It is important not to overwhelm managers by proposing a massive project. The project is more likely to be approved if it is small in scope, for instance, a pilot project with modest funding. The pilot project can serve to prove the viability of subsequent expansions. Normally a business case covers multiple years, often providing a five-year projection. The typical multi-year pattern is gradual

growth, increasing benefits, and decreasing costs. The costs the first year may be higher than subsequent years due to the initial outlay for software and infrastructure. Benefits often increase as the size of the termbase grows. Often the break even point – when benefits equal and begin to exceed costs – occurs in the third year or even later.

Some sponsors will require a new business case each year in order to reflect changes that have occurred (such as, for example, the growth in the termbase size, the addition of a new stakeholder, the acquisition of a new tool, etc.) and to give management regular opportunities to review the merits of this project against others. Each annual business case should reflect the lessons learned from the previous year. Part of the lessons learned involves validating the business case of the previous year. Did the company realize the benefits that were claimed would be realized? If not, why? What can be done to address any shortfalls? If nothing, then the benefits need to be lowered for the following year. What were the actual costs? Can they be reduced in the future?

During the deployment of terminology processes in the company, whenever a new requirement arises, such as the use of a CA software, or a new stakeholder emerges, such as another department that wants to participate, the business case needs to be revised. Each change to the environment potentially produces new costs and new benefits.

Make it clear that terminology management is a permanent commitment. It is not a project with a time and cost limit. Portray terminology work as a corporate strategy. Promote this strategy aggressively and report successes. It is also important to assume that the decision-makers have little or no knowledge of terminology and terminology management. Remember to explain basic concepts and write the business case in plain language.

Costs

The costs of implementing a terminology management program vary in each case and should be determined with management and colleagues. The following are the main categories of costs:

- costs of software licenses (for the TMS, project management, term extraction, concordancers, etc.)
- costs to develop any required custom applications, such as a company web portal for terminology
- software infrastructure (hardware, technical support)

- staff
 - terminologists
 - the portion of time allocated to terminology by other staff – technical support, translators, writers
 - any other staff, such as external consultants or suppliers
- costs for office space
- costs of training materials (videos, online help, documentation, etc.)
- travel costs (to attend meetings, provide or receive training, conferences, etc.).

Benefits

Avoided costs versus saved costs

Financial benefits are divided into two categories: avoided costs and saved costs. Avoided costs occur when it is estimated that without the terminology program the work being described would not necessarily be performed. For instance, in one scenario, the terminology program enables the company to translate more content than would otherwise be possible, but someone on the approval committee might counter that without the terminology program the company would not translate that extra content anyways. One cannot claim that a cost will be saved on a particular activity if the company is not committed to funding that activity. Most of the business case centers on avoided costs.

Figures for saved costs are more difficult to obtain than those for avoided costs. First, employees are reluctant to acknowledge any cost savings for their work out of fear that their budgets would be cut by that amount. Second, the financial sponsors of the activity often request proof that the cost savings were actually realized and may want to allocate the savings to fund either the terminology program itself, or another program with a higher priority. Thus, if claiming cost savings, terminologists may be burdened with additional financial reporting. They may also be held responsible for demonstrating that the savings were realized, while there is no guarantee that the savings will be reinvested into the terminology program. Therefore cost savings estimates need to be extremely reliable, their realization virtually guaranteed, and ramifications for budgets of the stakeholders involved clearly understood. It can be expected that the cost savings will be under-reported.

Avoided costs, on the other hand, refers to the cost of doing something (in a certain manner) that is not currently being done, and how having the terminology program in place would make doing that activity less costly. The difference between doing this future activity without the terminology program, compared to

with the terminology program, represents avoided costs. This is where the corporate terminologist can demonstrate, for example, that centralizing and automating certain tasks achieves economies of scale compared to doing the same tasks manually in an ad-hoc manner. Cost avoidance can also be demonstrated when larger volumes of work can be completed with terminology resources and support, rather than without.

Measurable benefits versus non-measurable benefits

Measurable benefits are those that can be quantified numerically or have some numerical basis, for instance, time saved, costs saved, market share gained, amount of content produced or translated, the number of languages added, size of the termbase, how often queried, and so forth.

Managers are most impressed when you can show measurable benefits. For this reason, measurable benefits should be prioritized over non-measurable ones in the business case.

Non-measurable benefits include improved quality, increased customer satisfaction and loyalty, stronger brand image, better retrievability of content, and other intangible areas.

Examples of benefits

A fair amount has been said and written about the value of managing terminology. It is out of the scope of this book to provide a comprehensive analysis of the benefits of terminology management, and we have already described the main benefits in *Motivation for managing terminology*. In addition, benefits will vary from one company to another. A list of useful references is provided at the end of this book. Based on those and possibly other sources of information, it should be possible to build some sound arguments that apply to the company's situation.

It is important not to simply quote business case arguments generally without directly relating them to the company. Saying, for example, that terminology management will "improve content quality" is not enough. Those types of arguments are not strong. Describing benefits using examples, scenarios and measurements that are specific to the company is much more effective. As mentioned already, measurable benefits (time saved, costs saved, etc.) are always better for the business case than non-measurable ones.

Showing that terminology mistakes can cost the company money (lost revenue, market share, support, cost of fixing them, etc.), and that they can be avoided, can be effective. Showing examples of mistakes that actually occurred will help to dramatize the business case and raise the level of concern and urgency. Consider the following aspects:

- frequency of the error
- how the error was multiplied in different languages
- how the error was multiplied in different media
- the cost of fixing the error, if it was fixed
- what would have happened (or what did happen) if the error was not fixed
- the impact of the error (on customers, market share, support calls, etc.) if it was not fixed.

In 2007, *Multilingual* dedicated its April/May issue to the theme of the ROI of terminology management. In his article, Mark Childress, a terminologist at SAP, presents a sample business case based on a realistic scenario, comparing the costs of creating a term entry to the costs of mistakes that occur when there are none. His calculation method is a good model for this specific ROI argument. He estimates that “investment in terminology standardization before problems arise avoids a ninefold multiplication of costs later”⁶⁴ (2007: 44).

Include some positive success stories as well, particularly from companies with a similar profile. Even just listing companies that manage terminology is useful. Nobody wants to appear to be missing out, especially among competitors.

Validating the benefits

Validating non-measurable benefits is almost impossible. For quality improvements, for instance, you could conduct a customer survey. However, this is difficult and hardly effective.

Another angle is to survey employees, which is easier than surveying customers. The survey can provide valuable input with respect to employee satisfaction with the terminology resources and services, including their own value statements. Examples of questions to ask include how much time they spend using the terminology resources and services, for instance how often they query the termbase each day, and how much time they estimate was saved by using these resources compared to doing the same work without them. However, for questions about saved time and other queries that validate aspects directly tied to employee performance, do not expect entirely truthful responses. Respondents may underestimate the amount of time they saved, uncertain of how that benefit might be interpreted by management (by cutting jobs or reallocating responsibilities, for instance).

Measurable benefits should be validated yearly, showing positive change. This usually involves tracking using whatever technology is in place that utilizes ter-

64. Here, Childress uses the term *standardization* more loosely to refer to the creation and distribution of approved terminological entries in a centralized company termbase.

minology, for instance, the CAT tool or the CA tool. Gather statistics from these tools *prior* to launching the terminology initiative, so that you will have metrics for comparison after the terminology process has been operational. Regular tracking and reporting are essential to justify continued investment in the initiative.

As an example, consider the CAT tool. Data can be gathered relating to the leverage of translation memories: the proportion of non-matches versus partial matches versus full matches. Any trends showing that non-matches are decreasing in favor of fuzzy matches, and that fuzzy matches are decreasing in favor of full matches, may be attributable to improvements in terminology usage. This trend has a quantifiable value in terms of reduced translation costs. Another CAT statistic that can be useful is the size and usage of integrated CAT dictionaries, which come from the company termbase.

Statistics can be obtained from a CA tool as well, such as:

- the size of any integrated dictionaries, which come from the termbase
- how often the dictionaries are queried
- the number of times terminology suggestions are provided and integrated.

CA tools typically provide reports about writing quality. These reports include the results of terminology checks.

Some company products, especially those that are rather technical in nature, include glossaries with product documentation. Another measurable benefit is the potential of the termbase to automatically generate those glossaries, saving the time and effort required to create them from scratch. This benefit is multiplied for each glossary that a term and definition appear in. For example, if the term we discussed earlier, *hybrid cloud*, is included with its definition in the glossaries of ten products, the savings are nine times the effort to document this term and definition in the product materials.

Implementation plan

As stated earlier, terminology management is not a project with an end but rather an ongoing program that supports the company's content production and global marketing objectives. Implementing the program to the point where it starts to show clear benefits can take several years. An implementation plan needs to be developed at the proposal stage, one that specifies clear milestones, deliverables, and measurable objectives or key performance indicators (KPIs). Ensure that what is promised at the milestones is achievable. It is always better to deliver more than promised and outperform projections than to fail to meet targets.

Of course, the implementation phase starts only after the terminology process has been fully documented and formally accepted, as described in The process.

The implementation plan should describe the following elements and stages:

1. the development environment (initial hardware/software required)
2. defining the data model and obtaining approval
3. creating the termbase in the TMS
4. strategies for populating the termbase; available resources
5. beta testing plan
6. announcement plan
7. promotion plan
8. support
9. online help and documentation
10. training.

Strictly speaking, the terminology program is not a project, however many of the best practices in project management are applicable. They include, for example, the SWOT analysis (Strengths, Weaknesses, Opportunities, and Threats). Identifying potential weaknesses and threats will help reduce risk by triggering a reflection on recovery strategies in the event of problems. Tools for tracking and managing deliverables, milestones, resources, schedules, and so forth are essential.

The Project Management Institute (PMI) provides a wealth of resources about project management including guides and best practices, tools, and training opportunities leading to certification as a Project Management Professional. Although PMP certification is of course not required, the corporate terminologist would benefit from becoming familiar with project management practices.

Approval

All of the elements described in this chapter need to be documented in a formal proposal that is submitted to the funding body. This document is the key to getting the project approved. The more professional it looks, the more confidence it will instill and the greater the chances of approval.

As mentioned in *Motivation for managing terminology*, it can be demonstrated that managing terminology supports ISO 9001 compliance. The value that terminology management delivers to quality and to the objectives of ISO 9001 should be a key focus of the business case and should be highlighted in the proposal because of this standard's importance to overall company objectives. But even if the company is not yet following ISO 9001, aligning the terminology proposal with this reputable international standard will likely impress the funding

body. Two key ISO 9001 requirements are that company processes be fully recorded and repeatable. This is another reason why the proposal must be professionally documented.

If you are invited to present the proposal to the members of the funding body, prepare a visual presentation that highlights what will be most interesting to them. Avoid using technical or linguistic jargon that they may be unfamiliar with. Do not assume that the audience knows anything about terminology or terminology management, or about the areas it supports such as CA or CAT. The following are some focus areas to consider for the presentation:

- Introductions: you, the executive sponsor, etc.
- Why the proposal at this time?
- Key concepts: terminology, CAT, CA, SEO, etc.
- The corporate environment necessitating terminology management – markets, languages, translation vendors, products, production technologies, problems that have occurred in the past, etc.
- Elements of the proposed program: termbase, TMS (and how it will be selected), term extraction, dissemination, workflows, etc.
- Users, stakeholders, and how they were consulted
- The business case – costs and benefits (multi-year)
- Implementation plan
- Ongoing operations and reporting.

Ensure that you invite feedback and that you follow it up.

The process

Defining the terminology management process for a company involves the various stages and tasks that are described in this chapter. Another very important task is the selection of data categories. This topic is large enough that it warrants its own chapter, Data category selection.

Access mechanisms and user interfaces

At the proposal stage, you have already identified the main users of the terminology tools and services, be they people or machines (see *Users and their roles*). Assuming that the central component of the terminology program is the termbase, you now need to define the access mechanisms and user interfaces required to service these users.

A termbase is hosted, managed, and stored in a TMS, which is dedicated software for managing terminology. A detailed description is provided in The terminology management system. Since the terminology is stored in a computer system, we use the term *user interface (UI)*, or *graphical user interface (GUI)* to refer to the layout and objects (buttons, menus, etc.) on the computer screen that people use to work with the termbase. A UI enables interaction between machines and people, or in this case, between the termbase and its users. The term *user interface*, however, does not apply to situations where the terminology is provided to a machine, i.e. another computer system. In these cases we use *access mechanisms* or other similar terms.

Translation being the main use case, most TMS have built-in functions and a UI specially-designed for translators. In fact, most commercially-available TMS are directly integrated into or designed specifically for a CAT tool. Standalone TMS are less common but usually more robust and multipurpose. They should provide a dedicated function (sometimes called *connector*) or an API allowing for connection to other systems such as CA or CAT tools.

It is the terminologist's responsibility to ensure that terminology from the termbase can be easily integrated into the various other tools or applications where it is needed, and that it performs as it should in those situations. This also entails checking how the terminology looks in the target system and how users interact with it. For example, the window that displays terminology in CAT tools

is typically very small. It is therefore important to limit the types of information that it displays, to minimize the need for scrolling and ensure that essential information does not fall outside of the window's perimeter where translators might miss it.

If the terminology is to be used in a CA tool, it will be necessary to define a clear process for how the termbase will supply terminology for the CA tool. This is not a trivial undertaking, as will be explained in *Controlled authoring*. CA tools have been found to be sub-optimal for housing and managing multi-lingual terminology (even though they may be technically capable of doing so) due to their authoring focus. (See *The terminology management system* for more reasons.) Typically in corporate contexts the terminology is housed in a separate system, and subsets of it are exported and then imported into the CA tool. This should be automated to occur on a regular basis.

The terminologist needs to be directly involved in deploying the terminology in all external systems, testing it, and determining what it should contain and how it is structured.

Virtually all companies require a web interface for displaying the content of the termbase on a website. Originally, the purpose of web interfaces was to provide all company employees with access to the terminology in the termbase through a browser. The website should therefore be very user-friendly and require no training to be used effectively. The first web interfaces to be developed were secondary read-only interfaces, because users who were more involved in the terminology such as writers and translators were given access through their productivity tools, and terminologists and termbase administrators used the primary UI of the TMS directly. As a result, those early web interfaces tended to have limited read-only features for viewing the terms and other key information such as definitions. Today, with the growing trend towards cloud computing and away from installable software, there is an increasing demand for the web interface to be the primary, full-featured interface for the TMS. Unfortunately, the heritage of the old client/server architecture means that some of the web interfaces provided by TMS that are part of CAT tools remain disappointingly limited. Prior to purchasing a TMS, it is therefore recommended to determine what limitations its web interface may have. If you choose a TMS whose web interface does not meet your requirements, be prepared to develop an entirely new web interface in-house. It has been the author's experience that TMS suppliers are not willing or perhaps not able to implement major new functionality in these legacy web interfaces.

Companies should also consider developing a UI for mobile devices.

Most company websites dedicated to the terminology are for internal use only, not public. The reasons for this are twofold:

1. Some of the terminology is confidential or sensitive in nature, such as names and descriptions of pre-released products and features, or there is the risk that this *could be* the case
2. Some of the content is “untidy,” it has not been subject to quality controls normally applied to public-facing materials. It may contain errors and omissions which would give an unfavorable impression of the company.

In reality, the so-called sensitive terminology normally constitutes a very small portion of the entire termbase. Companies that keep their terminology information away from the public eye risk being criticized by external vendors and partners; translators who are located outside the company firewall are particularly annoyed by this. It would seem a very good strategy to add metadata to the termbase entries, allowing the sensitive terms to be marked as such. A filter could then be used to exclude those terms from the company’s public website.

Unfortunately, untidy content is a common problem. Most termbases include a lot of untidy content and companies have good reason to not show such content to the public. Poor quality content occurs when the terminology initiative is understaffed and/or when existing staff lack the skills and knowledge specific to corporate terminography (hence one of the reasons for this book). A solution to this problem is to create an approval workflow, and to release the approved entries to the public portal by using a filter.

The user-friendly criterion of all the interfaces is satisfied, among other ways, by providing role-specific views of the terminology which hides fields that are irrelevant to the user (see *Views and filters*). An English technical writer does not need to see any languages other than English. Translators only need to see the languages that they work with. None of the users, with the exception of administrators, need to see administrative details such as dates.

Some TMS are entirely web-based, others are client/server based with a desktop interface and offer a web interface as a secondary option. As mentioned earlier, in the latter case, the web interface typically includes a limited number of features compared to the main interface. Customization of this pre-supplied web interface to provide the user-based views may not be possible at least not to the extent desired. Since a robust web interface is normally a show-stopper requirement for most companies, any vendor-supplied web interfaces should be carefully evaluated before they are accepted.

Interfaces may be required for other tools and applications, some unforeseeable at the present time. To optimize interoperability between systems, it is therefore essential that the TMS support the ISO standard XML format for terminology data: *ISO 30042 - TermBase eXchange*. Another useful import/export format is Excel. Import formats are described in *Initial population*.

Stages and workflows

The terminology management process involves terms passing through various stages for various reasons. Examples include:

1. A new concept (invention, innovation, product, program, etc.) has emerged and it needs to be named. The term chosen must comply with certain requirements. (See also *New concepts*.)
2. Writers of product content (user assistance, user interface, website, other types of content) decide that a sentence they are writing contains a term. They decide to explain this term in the document by writing an explanation or a definition. With markup standards like ITS and TEI, it is possible to capture this knowledge (both the term and the explanation) directly in the content, and then it can be passed downstream to the terminologist and to the translators. For more information, see *Standards and best practices* and *Authoring*.
3. Once a term for a new concept has been determined in the SL, suitable terms need to be determined for the languages that the related content is being translated into.
4. A term in the SL has caused some problems and needs to be changed.
5. A term in a TL has caused some problems and needs to be changed.
6. There is a new area being introduced to the company that will bring with it a substantial amount of terms that do not currently exist in the company's corpus. (This could be the result of a merger or acquisition, an organizational change, a new product line, etc.). Action needs to be taken to support writers and translators.
7. A product is being launched in several identified geographical markets (*locales*) for the first time. Its key terms need to be identified and pre-translated in order to shorten translation time and maximize quality.
8. The name of a key feature is changing from one version of a product to the next.
9. Some terminology inconsistencies have been noticed in a product, website, or other content area, in the SL.
10. Some terminology inconsistencies have been noticed in a product, website, or other content area, in a TL.

Each of these situations demands a different course of action involving different people and workflows. Ideally each should be described in the documentation of the terminology process.

For example, when a substantial quantity of new terms is anticipated ("new" terms meaning those that are not already in the termbase), such as after a merger or acquisition, it is recommended to undertake a complete term extraction process.

This would include corpus compilation, extraction of term candidates, post-processing (cleaning), bulk-import, translation, and reviews and approvals. This could be undertaken and managed as a project with its own budget, resources, milestones, and deliverables.

When the term for a product feature changes from one version of the product to the next (which is something that should be avoided if possible), the people who may be affected need to be informed. This includes customers, suppliers and business partners, service personnel, support, content producers (including marketing) and of course, translators. Screenshots and other media showing the previous term need to be retaken. Indexes must be regenerated. Even search keywords in the product's web pages may need to be changed. Translation memories may require modification to reflect the change as well.

Target language equivalents for key terms should be checked to ensure that they are suitable for the target market. Experienced translators with target market knowledge are best qualified for this task. Many companies require these terms to be approved by a marketing expert in the target market.

Consider the impact these various scenarios have on workflows and other termbase objects such as user groups and views. For the last scenario, one could create a group in the termbase for marketing experts (by locale), set up a dedicated section in the termbase for marketing-sensitive terms, create a view for the group that hides the fields unnecessary for marketing experts to see, and assign the group to be approvers of the terms in that section. Then, create a workflow to channel those terms accordingly for approval.

It is important to realize that these scenarios have a direct impact on how the termbase should be designed; it is therefore essential to have a very clear understanding of who will be using the termbase, what they will be doing, and what their needs will be.

The terminology audit

Various files containing terminology can often be found in companies both before and after it has a termbase. These include glossaries, indices, search keywords, catalogs, taxonomies, and so forth. Prior to creating the termbase, the terminologist should carry out an audit of the organization's existing terminology resources. They are likely to be in different file formats (MSWord, XML, PDF, PowerPoint, Excel, OpenOffice, etc.) with varying degrees of quality, information content, and internal structure.

Some of these files contain information that is suitable for inclusion in the termbase, others do not. The selection of files to import to the termbase can be

made later when the termbase is operational and ready for terminology to be imported. The purpose of the audit at this stage is to:

- collect sources of terminology that could be imported at a future stage
- catalog them according to parameters such as source, languages, domain, age, and other criteria (such as product line)
- list the types of information found in each file
- record a high-level assessment of the file's quality, reliability, and internal structure.

To collect the resources, the terminologist needs to reach out to people who are likely to have created or used such resources. Basically this includes anyone who creates content, such as technical writers, editors, translators, web developers, marketing staff, product managers, software engineers, SEO experts, and others. Obtain a copy of all the resources and store them in a central location such as on a network server or cloud-based repository.

Cataloguing the resources according to their salient features facilitates ongoing management and retrieval. The source of the resource is important. Determine whether it is an internally developed resource or one from outside the company. If the latter, there may be copyright restrictions that need to be considered. Internal sources could be identified by employee name, department, product line, etc. Sometimes whether the resource was developed by an experienced in-house employee versus a temporary freelancer or contractor can be an important indicator of its reliability. What languages does the resource contain? Does it cover a particular domain (subject field)? How old is it? All this information should be recorded in a format that allows sorting, filtering, and other manipulations, such as in a database or a spreadsheet.

The next task is listing the types of information contained in each resource. If an employee felt the need to include certain types of information in a glossary, such as a definition or a product identifier, it is likely that this type of information should be included in the termbase as well because it serves some purpose. As described earlier, when implemented in termbases, these types of information are referred to as *data categories*. After preparing the inventory of data categories found in each resource, you will be able to determine how often each data category is used and separate the common ones from those that are rare. A specific data category may be common for certain types of resources but not others. These differences need to be noted. For example, customer-facing glossaries created by technical writers or product developers often include definitions whereas most bilingual glossaries created by translators do not. The goal is to develop a prioritized list of data categories with information about where or by whom they are

used. This information will be used for designing the termbase and for creating customized views for different user roles.

The nature of the content found in each data category – known as its *content model* – also needs to be noted. The content model determines whether the data category contains free text, as do definitions and comments, or whether it contains a limited set of values, such as part of speech (noun, verb, adjective, adverb). There are also cases where a data category's content model is not restricted to a set of values in the collected files, but it could be. For example, if a glossary notes the products that terms are used in, and there are a reasonable number of products in the company, the Product data category could be restricted to a set of values. For reasons that will be explained later, from a design perspective it is recommended to limit the content of a data category to a set of values whenever possible, rather than leaving it open to free text.

Once this terminology audit is complete, you are ready to select data categories for the termbase. This will be covered in the next chapter, after a brief discussion about inclusion criteria.

Inclusion criteria

In a commercial setting, where each additional application of the terms in a termbase multiplies the ROI, repurposability of the termbase is a primary objective. As we will soon see in Data category selection, each application of the terminology has an impact on the selection of data categories. And as we have also seen in *Termhood and unithood*, the same can be said about the criteria for termhood.

A terminology management process in a company should have clearly defined rules of operation and constraints around data structures. Otherwise, “anything goes” and the termbase will lack coherence and rigor, rendering it less effective for all users. One of the key rules governs what should go into the termbase as far as the terms themselves are concerned, and what should not. This rule is referred to as *term inclusion criteria*.

Term inclusion criteria should be based on how the company defines termhood for its purpose. Termhood should be defined according to the aims and applications of the termbase. It should also guide the process for accepting or rejecting term candidates into the termbase. Exceptions should be made to accommodate unique cases. Termhood criteria should be reviewed on a periodic basis to reflect any new or updated requirements.

One thing is certain, the classical notion of termhood inherited from the GTT is not suitable for termbases that are designed to satisfy the demands of corporate communications. The reasons for this are described in *Termhood and unithood*.

When deciding on the term inclusion criteria for a termbase, consider which types of terms:

- are already found in any existing glossaries
- are needed by translators
- are needed by new employees, to support their onboarding
- are needed for the autolookup function of CAT tools
- are needed for CA
- are needed for any other applications, such as SEO.

Term inclusion criteria should also aim to avoid duplication with any other enterprise application. For example, it would be ideal to include terms from an enterprise taxonomy in the termbase, but if there is already another application housing this information, and adequately so, then attempts should be made to ensure that the two efforts are complementary. The technical and administrative environment of both systems (taxonomy and terminology) should be examined to determine the best long-term solution. Should the two data sets be merged, or not? The answer typically depends on how the data is integrated and used in other enterprise systems. The enterprise taxonomy, for instance, may be used by a content management system (CMS) which may be incompatible or more difficult to integrate with the termbase. This is again another reminder how important it is for the termbase to support TBX, which, as an XML format, provides the structure necessary to enable the data to be used by other enterprise systems.

For anyone who will be proposing or submitting terms to the termbase, it is important to understand that the termbase includes terms that are “not recommended”, i.e. their use is discouraged. This is an approach that most people are not familiar with; most hand-built glossaries do *not* include so-called undesirable terms, i.e. terms that are incorrect or should be avoided. Such terms are needed in the termbase for two principle reasons: first, so that users can be notified about the terms that they should avoid, and second, these terms are essential members of synsets, which some extended applications require, such as SEO. In the termbase, they that are not recommended will be clearly identified as such with a **Usage status** value.

The vocabularies used in a CA application differ from conventional terminology resources in two ways: first, many of the lexical items required for CA are words and expressions from the general lexicon. For example, a company may prefer that writers use the adverb *almost* instead of *nearly*, the latter potentially being confused with a spatial concept. According to convention such items would not be included in a termbase since, by conventional definition, terms are domain-specific. Second, CA applications need more verbs, adjectives, and adverbs than are normally recorded in termbases. Actually, they even need some

prepositions and other function words that are practically never included in termbases. These and other differences are described in more detail in *Controlled authoring*.

Turning now to the terms required for translation purposes, there are two user scenarios: translators who use the TMS directly to find information, and translators who use the autolookup function of the CAT tools. In the CAT tool, where the dictionary window is small, the display should focus on the most essential information, usually the recommended translations for the terms that are in the active translation segment. At the moment when they are translating a text, translators generally do not need to see subject fields, related terms, definitions, usage notes for the SL, and possibly other data categories. Therefore these types of information can be hidden from the view to conserve space. This means that a minimalist view of term entries in CAT tools could be restricted to the terms themselves. In any case, most CAT tools allow the translator to open the full entry when needed. On the other hand, the display space available on the UI of the TMS is larger, so the view can contain more data categories by default. We address these options in more detail in *Views and filters*.

With respect to term inclusion criteria, it is true that translators and SL writers need different types of terms, with some overlap of course. We propose the following breakdown of term types required, starting with the overlap:

- Terms required by both translators and writers:
 - acronyms and abbreviations with their expanded forms
 - technical terms and jargon, with definitions
 - terms that are specific to the company's products, services, and the larger industry sector
 - product names
 - names of features or functions within products.
- Terms required more by translators and less by writers:
 - SL terms that can be translated in more than one way in the TL – this can extend to units from the general lexicon
 - multiword terms or expressions – these can often be translated in multiple ways
 - SL terms that are difficult to translate – this can also extend to units from the general lexicon
 - SL terms that should not be translated
 - fixed phrases, such as copyright statements
 - marketing slogans and expressions.

- Terms required more by writers and less by translators:
 - related terms, with clear explanations of how they differ
 - names of related products and services, with clear explanations of how they differ
 - synsets of SL terms with information to guide correct usage.

The industry sector mentioned in the first list refers not only to the core industry sector that the company is active in, but also industries where the products or services are used. For instance, a software company might develop different software products for a wide range of unrelated industries such as banking, shipping, retail, entertainment, and others. The termbase should include software and computing terms but also terms from those industries.

With respect to multiword terms (MWT), any with acronym forms should be included in the termbase along with the acronyms themselves, of course. But the domain-specificity rule for termhood should be loosened to allow some multiword terms into the termbase whose meaning is strictly compositional (and therefore self-explanatory). This is due to the inherent ambiguity of MWT that results from their complex syntax, which is described in *Termhood and unithood*. This ambiguity makes translation difficult and can lead to inconsistencies. Most corporate termbases include an abundance of MWT for this reason.

As explained in previous sections, synsets and concept relations are important although they are not a focus area for translators. The inclusion of synonyms and concept relations should not be overlooked, as they are needed for CA, SEO, and other potential applications. They render the termbase into a knowledge system. We will discuss this further in *Concept relations*.

In establishing term inclusion criteria, it is also helpful to indicate what should *not* be included. This will vary from company to company, but examples that we have found in various companies include:

- code strings and any other machine-readable information, such as file names
- titles (of websites, publications, etc.)
- words from the general lexicon, unless they are required for a specific purpose which can be articulated (CA, translation difficulty, etc.)
- full phrases
- translation memory segments.

Data category selection

With the information gathered in the previous phases about users, roles, interfaces, workflows, and existing terminological resources, the terminologist has everything needed to begin designing the termbase. The first step is deciding which data categories it should include.

In *The terminology audit*, we describe the process of creating an inventory of data categories based on existing terminological resources. To assume that this list is complete would require another assumption: that the existing terminology resources reflect an ideal set of data categories for the organization, even for meeting its future needs. This is unlikely to be the case. Each individual resource, each glossary, was developed by one or several individuals for a restricted purpose, and did not have larger enterprise-wide needs in mind.

The terminologist needs to once again review the primary user roles and interfaces as described in the previous sections, and then consult external guidelines and standards to fill gaps. Key guiding questions include:

- Will terminology be required for CA systems?
- Will terminology be required for CAT tools?
- Will search keywords be stored in the termbase?
- Is there an interest in using the termbase to store enterprise taxonomies?
- Should the termbase be organized such that terms from different administrative divisions, product areas, etc., can be isolated?

We discuss the data category requirements for each of these areas in the following sections.

Computer-assisted translation

In this section we present some of the types of information and data categories that are needed for translation purposes in general and for CAT tools in particular.

As described in *Relevant literature*, the Localization Industry Standards Association published the results of a survey it carried out about terminology management practices in the localization (translation) industry (LISA 2005). The survey includes questions about what types of data categories are most needed in the translation/localization industry and, since CAT tools are widely used in this industry, by extension which ones are needed by CAT tools for their effective use of terminology. Since there are hundreds of different data categories (according to ISO 12620:1999 and DatCatInfo), it was not feasible to ask respondents to indicate the specific data categories that they use. Rather, data categories were grouped together based on overarching themes, with examples given for each one, and respondents indicated for which of these groups they capture information in their termbases.

- grammatical (part of speech, inflection, gender, etc.)
- semantic (definition, explanation)
- contextual/usage (sentence, usage note, etc.)
- categorical (subject fields, products, etc.)
- administrative (status, date, author, etc.)
- term relations (antonym, related terms, etc.)
- illustrations.

Information	Source language	Target language
Categorical	81%	59%
Semantic	78%	59%
Contextual	78%	56%
Administrative	65%	52%
Relational	50%	37%
Grammatical	45%	34%
Illustrations	18%	11%

Figure 24. Percentage of respondents collecting specific types of information

LISA drew the following conclusion from these results:

Categorical information is the most important in this industry which is heavily organized on a client and project basis. Contextual information (e.g. context sentences) ranks higher than semantic (e.g. definitions), confirming a fact previously demonstrated by LISA that context sentences are frequently replacing definitions in the fast-paced translation environment, especially when they can be extracted by machine. Administrative information ranks higher than related terms, grammar, and illustrations, an indicator that workflow information is more important than some strictly terminological data for the translation industry. (LISA 2005)

These findings confirm what is common knowledge and practice in the translation industry, translation-oriented termbases:

- need to track terminology by product, translation job, client, and other administrative categories
- include context sentences more than definitions
- include related terms less frequently
- include grammatical information less frequently.

While this insight can help the terminologist focus on what is important for translators who use CAT tools, it does not give the terminologist license to exclude the less important items such as definitions, synonyms, related terms, and grammatical information. Doing so would cause problems down the road if the termbase is needed for applications beyond translation. When asked why they do not include grammatical information such as the part of speech, translators often respond that they do not need to record this information because they know intuitively the part of speech of any given term. What this type of reasoning does not take into consideration is that the termbase may eventually be used by an application that is incapable of knowing or inferring the part of speech of a term. We will see later, for example, that CA applications depend heavily on this one piece of data. The exclusion of synonyms and related terms will render the termbase unusable for any application that leverages semantic relations, such as keywords for SEO and other knowledge-based resources.

The argument that translators do not need the part of speech has been contested by Dunne (2007: 34) who describes the issue of “lexical ambiguity.” “Lexical ambiguity occurs when a word is used as more than one part of speech – such as a noun and a verb.” He claims that in “disembodied content with no contextual information” (such as software user interfaces), translators “may be unable to determine whether a given word refers to a process or to a product of that process.” He cites as examples computing terms such as *archive* and *filter*, among

others. It would seem beneficial, therefore, to convey part of speech information both in the termbase and in the source content.

The LISA survey results also revealed that more work is carried out on the SL than on the TL. This is because the source term and the concept it refers to need to be completely understood before the term can be translated. Any particular context that the term is associated with, such as a product, also needs to be known to ensure that the selected translation is appropriate for that context. This task – clarifying SL terminology – is, according to the results of this survey, often being carried out by translators. And yet, translators are typically less knowledgeable and informed about SL terminology than the content developers themselves (see Dunne 2007). This is a common problem in enterprises – the entire terminology management responsibility is handed to the translation department while employees in content development are not involved at all. It would be more sensible to have both stakeholder groups – content developers and content translators – contributing equally to the termbase in their respective areas of expertise. Addressing this issue is the main reason why we have advocated for embedding information about terms directly in the source content through markup formats such as ITS.

Some translation-oriented termbases, particularly those integrated into CAT tools, include inflected forms of terms, such as both the singular and the plural form of a noun, or the infinitive and inflected forms of verbs. Usually these terms are added by translators who work with CAT tools. Having inflected forms in the CAT-integrated termbase enables translators to insert the corresponding inflected TL term into the translation with one click, thereby saving time in typing and editing. This practice was noted by Kenny (1999) and Allard (2012).

While this use case – inserting a term from the termbase into the translation in one click – is certainly recognized as a means of increasing translator productivity, including inflected forms in a corporate termbase should be discouraged. Aside from the high degree of data redundancy and additional ensuing costs that this entails, the presence of hundreds if not thousands of inflected forms in multiple languages undermines the repurposability of the termbase. Translators may add inflected forms to their personal CAT dictionaries, and they do, but this practice should be discouraged for a company-wide termbase.

It also arrants mentioning that inflected forms should not be necessary for matching purposes in the autolookup function of the CAT tool. In other words, if the CAT termbase contains *configure*, and the text being translated contains “The process configures...” or “Configuring the process...”, the CAT tool should be able to determine that the two in-text instances correspond to the verb *configure* and display the entry correctly to the translator. A good CAT tool will *lemmatize* inflected forms found in content before looking for a match in the termbase. Lem-

matizing a word involves reducing it to its *canonical form*, that is, reduce a plural noun to its singular form and reduce an inflected verb to its infinitive form and even, reduce a capitalized term to lower case, and *then* search the termbase for a match.

Fuzzy matching is another technique for matching terms encountered in text with the corresponding entries of a termbase when the terms are written in a slightly different form. In some CAT tools this technique is expanded beyond matching morphological variants (inflected forms) to also account for multiword terms written in different word order, for example, *World Wide Fund for Nature* and *World Wide Nature Fund*.

In summary, the corporate terminologist should fully test the CAT tool's abilities for matching terms in running text with their counterparts in the termbase. If the performance is poor, the tool should be replaced by another. Any problems with the autolookup function of the CAT tool should be reported to the tool's vendor. They should not be fixed by tweaking the termbase.

Controlled authoring

Controlled authoring (CA) is “the process of applying a set of predefined style, grammar, punctuation rules and approved terminology to content (documentation or software) during its development” (Ó Broin 2009). With respect to the “approved terminology” part, the goal is to address the problem where different people are using different words for the same thing. The aim is to adopt a *controlled vocabulary*.

CA is increasingly being recognized as a means to improve content quality and overall communications. Content that is consistent and easy to understand is more effective at increasing sales and improving product usability. It is also easier and less costly to translate. Organizations that produce large quantities of information, especially if they do so in multiple languages, can therefore realize significant benefits by implementing CA.

CA can be either active or passive. Passive CA is when writers refer to a company style guide and glossary as they work. Active CA leverages a computer application to prompt writers to adopt style rules and recommended words as they write. Active CA guarantees, more or less, that the rules are followed whereas passive CA relies on the voluntary initiative of the writers.

Companies may turn to active CA after they witness the relative ineffectiveness of passive CA. At the time of writing, examples of software applications that are used for active CA include Acrolinx, crossAuthor (Across), Congree and HyperSTE (Tedopres).

Terminology resources for CA need to be concept-oriented. Organizing the individual words and expressions into synonym sets, or synsets, is fundamental. For each prohibited word there is a preferred alternate, for each acronym there is an expanded form, and so forth.

English	
almost	
Process status	Finalized
Usage status	Preferred
Part of speech	Adverb
Register	Neutral

nearly	
Process status	Finalized
Usage status	Prohibited
Part of speech	Adverb
Register	Neutral
Usage note	Simplified Technical English not approved word

Figure 25. A synset containing two terms

A fundamental data category for CA is **Usage status**.

Table 4. Usage status values used in CA

Value	Description
Preferred	This term is the preferred one of all choices in the entry.
Admitted	This term is allowed, but not preferred.
Restricted	This term is prohibited in some contexts but allowed in certain other contexts. See the usage note in the entry for further clarification.
Prohibited	This term is always forbidden.

Figure 26 shows a sample entry in a termbase that distinguishes four synonyms according to their recommended usage. One is preferred, one is admitted, and two are prohibited.

English	
unrecoverable error	
Usage status	Preferred
Part of speech	Noun

catastrophic error	
Usage status	Prohibited
Part of speech	Noun

irrecoverable error	
Usage status	Admitted
Part of speech	Noun

fatal error	
Usage status	Prohibited
Part of speech	Noun

Figure 26. English synset showing usage status

The CA application uses these values to prompt writers with guidelines on vocabulary use, a function sometimes referred to as *term checking*. With active CA, if writers use a prohibited word, it will automatically be visibly highlighted in the document they are writing. Writers can then right-click to open a dialog that indicates the preferred alternate. Figure 27 shows the result when a writer right-clicks on a prohibited term, here highlighted in yellow. In this case, *catastrophic error* is prohibited, and *unrecoverable error* should be used instead.⁶⁵

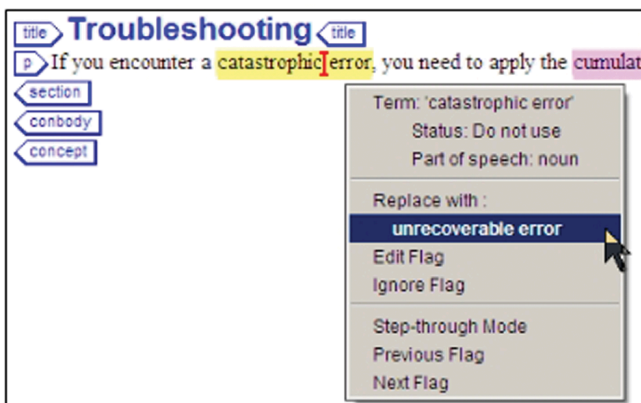


Figure 27. Active CA showing recommended term

⁶⁵. This example uses DITA markup.

Restricted terms and prohibited terms should be easily distinguished in the editing application when a terminology check is carried out, such as with different colors. Writers should be able to quickly see the conditions of restriction, such as if the term is disallowed within the context of a certain product. This requires more information in the back-end database, such as subject field values, product values, and usage notes. To fit into the limited space of the context menu, usage notes must be very short. Longer explanations can be included in an additional dialog, which can be accessed from the context menu.

CA applications require a part of speech value for each term in the termbase (we already demonstrated this by example in *Extended applications*). This is because different usage rules are applied depending on how the term is used syntactically in the sentence. The CA application first parses (analyzes) the sentence by breaking it down into single words and assigning a part of speech value to each. It then checks the internal dictionary to see if any of the words have a usage status value that should be displayed. However, a word can sometimes have two or more part of speech values (these are referred to as *homographs*) with no change in spelling. For example, *sail* could be a noun (the sail on a boat), or a verb (to sail a boat), even when written with an added “s”: “The boat sails with sails.” Without a part of speech value in the respective termbase entries the CA application cannot know which entry corresponds to the noun and which to the verb. And it is not uncommon in corporate communications to constrain the use of words according to the part of speech, for example, only use *port* as a noun, never as a verb. In order to display the correct usage guidance, the CA application therefore needs to know the part of speech value of a word in a given context, which it determines through parsing. It also needs to know the part of speech of the entries in the termbase. The following two examples may help to clarify this use case.

Suppose a software company decides that the documentation should always say “the readme file” and never “the readme.” For example, “Consult the readme file” is correct, but “Consult the readme” is not. One way to approach this is to observe that, by modifying *file*, *readme* in “the readme file” is syntactically playing the role of an adjective whereas in “the readme” it is a noun. The terminologist creates an entry in the termbase for *readme*, assigns the part of speech value *noun*, the usage status *prohibited* and includes a usage note specifying to use “readme file.” When the CA application encounters “readme file” in a text, it knows through parsing that *file* is a noun, and that *readme* modifies this noun and is therefore considered equivalent to an adjective. It checks the internal dictionary for *readme* marked as an adjective and accompanied by a restricted or prohibited Usage status, and does not find it. No usage error is reported. However, when the CA application comes across a sentence such as “Check the readme” it knows through parsing that in this case *readme* is a noun. It checks the dictionary for

readme marked as a noun, finds one that indicates that it is prohibited, colors the term yellow in the text and displays the usage information accordingly.

Another way to handle cases like this is to have one entry in the termbase that contains both terms, *readme file* and *readme*, both marked as nouns, with the former given *preferred* Usage status and the latter *prohibited* Usage status. This will flag as incorrect *readme* when used as a noun, and the correct alternative *readme file* will be presented to the writer. However, the terminologist needs to verify that the CA parser correctly identifies *readme* as an adjective, not a noun, when it is followed by *file*. When it reads “the readme file,” parsing linearly, it may conclude that *readme* here is also a noun because it is preceded by “the.” It will then incorrectly report this usage as an error. Correct term checking behavior can only be determined by extensively testing the terminology checker.

There are many such cases of homographs that have different usage instructions. The only machine readable (programmatic) way to differentiate between homographs is by the part of speech value.⁶⁶ For example, in computing, *port* can be both a noun (the physical ports of a computer) and a verb (to port a software program from one operating system to another). The verb usage is less common and could be considered jargon. It may be advisable to avoid the verb in user documentation and to suggest alternative ways to express this idea. It must, therefore, be possible for the CA application to tell the writer when he has used *port* incorrectly *as a verb* and to display alternatives. This is only possible if the termbase entries contain part of speech values.

The previous examples demonstrate that the part of speech assigned to a word by the CA parser has to align with the part of speech assigned to the termbase entry that needs to be displayed. Because the output of the parser is not always predictable, a lot of testing is required. Sample sentences are *seeded* with the different patterns that require checking. Commonly, several rounds of adjustments to the termbase entries or to the parsing resources used by the CA application (rules, internal dictionaries, etc.) are required before the desired behavior can be achieved.

The use of terminology for CA purposes also has an effect on the company’s definition of *termhood* as was mentioned earlier. Recall that termhood refers to the property of a term candidate that makes the term candidate acceptable for inclusion in the termbase. Traditionally, termhood was restricted to term candidates that have a domain-specific meaning, which is sometimes described as a

66. This applies, of course, only for homographs that do have a different part of speech value, such as *file* the noun and *file* the verb. There are other homographs that share the same part of speech value, and in this case, of course, indicating the part of speech would have no differentiating effect. For example, *port* the strong wine and *port* the place for boats are both nouns.

meaning that falls within language for special purposes (LSP). However, CA often needs to control the use of very common, general words, i.e. words from the general lexicon, which would not meet this traditional criterion.

Corporate termbases that are repurposed for CA typically contain many general lexicon words. Many are non-nouns: verbs, adjectives, adverbs, conjunctions, and so forth. The following set of general lexicon words, for example, was found in the termbase of a company that uses CA: *all, almost, also, can, did not, do, enough, excellent, impossible, minimum, otherwise, previous*. In the termbase each of these words is recorded in a terminological entry that includes another (synonymous) word or expression, with one of the words having a preferred status while the other is prohibited. For instance, as shown in Figure 25 *almost* is preferred, and *nearly* is prohibited. The advice to avoid the word *nearly* comes from *Simplified Technical English (STE)*,⁶⁷ which is a standard for simplified English that was developed for the aerospace industry and has now been adopted in other commercial sectors.

Controlling some general lexicon expressions is typical in corporate style. One will find examples of this in any company style guide. According to a survey conducted by LISA in 2001, 25 percent of commercial termbases include general lexicon words and expressions (Warburton 2001a: 20). With the increasing adoption of CA in corporate environments this figure is likely to be higher today.

It is important to use a dedicated data category to mark terms in the termbase that are needed for CA so that they can be isolated, via a filter, from the rest of the terms. This allows the CA entries to be included in the CA tool as a dedicated set, whether that be by export/import or through an API. For this purpose, we recommend using a custom subsetting data category with picklist values. Considering that CA is an application of the terminology, a termbase field for **Application** with values such as CA, CAT, SEO, and so forth, might be ideal. This is, however, just a preliminary thought; actual values need to be determined based on the needs of the organization or company.

To summarize, the following data categories are necessary to meet the requirements for CA:

- Usage status – picklist with the values preferred, admitted, restricted, and prohibited
- Part of speech – picklist with the values noun, verb, adjective, adverb, proper noun, other
- Usage note – single-line text field
- Additional usage information – multi-line text field

67. www.asd-ste100.org/

- Term type – picklist with the values full form, abbreviation, acronym, short form, and phrase
- A custom subsetting data category for marking terms that are used for CA purposes.

Concept relations

Concept relations (aka semantic relations) are relations between concepts whose meanings are connected in some way. The notion advocated by the GTT that concepts need to be studied and organized in a hierarchical system showing their relations has led some researchers to draw parallels between terminology and ontology. Research in Denmark led by Madsen and Thomsen at the Copenhagen Business School demonstrates methodologies for representing terminological concept systems using Unified Modelling Language (UML), which is a standardized general-purpose modeling language used in object-oriented software engineering.⁶⁸ Using examples from the medical domain among others, they demonstrate that such *terminological ontologies*, as they are referred to, are “vital for the successful development of IT systems” (Madsen and Thomsen 2015).

Most TMS do not support the range of concept relations and features for their representation that can be realized through a modelling language such as UML. Relations usually take the form of a simple link between concept entries, and if there is a typology offered at all, it is usually restricted to three basic types: generic, partitive, and associative. “Most resources that do offer such information merely provide an overview of a specialized field, primarily based on the *is_a* or *type_of* conceptual relation” (Faber 2011: 11). Some more advanced TMS offer the ability to visually display relations in a concept map or concept diagram. Figure 28 shows terms related to forestry. The numbers associated with the terms are the concept identifiers.⁶⁹

Semantic relations in general are recommended for commercial termbases. They transform a termbase from a flat arrangement of terms and term related information into a hierarchically-organized knowledge bank, also referred to as a *terminological knowledge base*. They are useful for a number of applications, including search (see *Search*), indexing, and content classification.

68. These methodologies have been formalized in an ISO standard, *ISO 24156-1:2014 Graphic notations for concept modelling in terminology work and its relationship with UML – Part 1: Guidelines for using UML notation in terminology work*.

69. Source: Interverbium Tech.

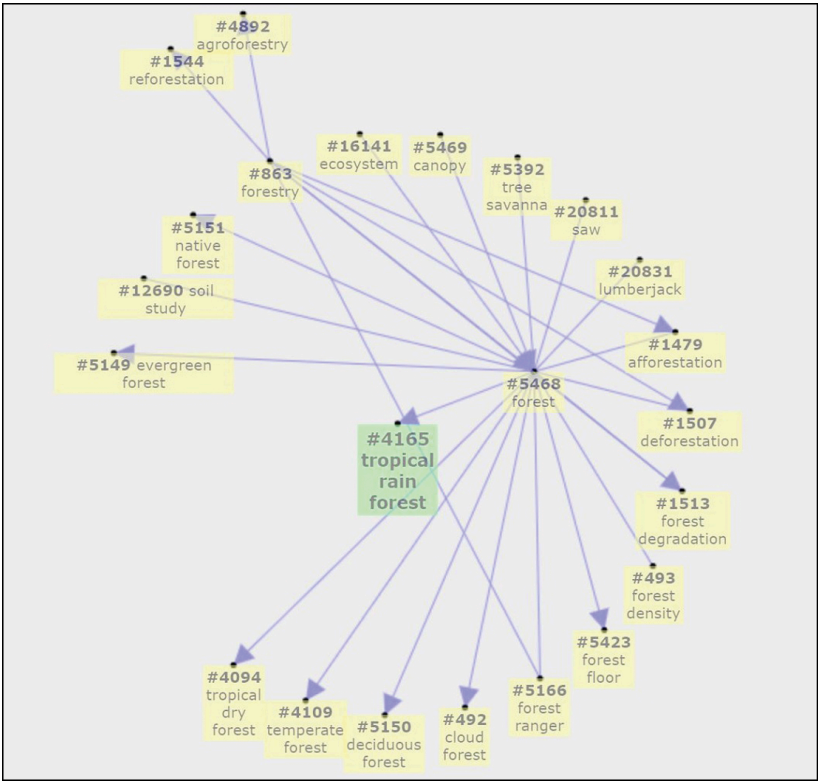


Figure 28. Concept map

There are various types of relations.⁷⁰ Table 5 shows the principal types of relations that would be useful in corporate termbases. The generic, partitive, and associative types are more common than temporal and spatial.

Table 5. Semantic relations

Relation type	Values	Description	Example
Generic	superordinate, subordinate	The “type of” relationship	A pine tree is a type of coniferous tree, which in turn is a type of tree. Therefore, <i>pine</i> has a subordinate generic relationship to <i>coniferous tree</i> which has a subordinate generic relationship to <i>tree</i> .

70. See ISO 704.

Table 5. (continued)

Relation type	Values	Description	Example
Partitive	superordinate, subordinate	The “part of” relationship	Bicycles have wheels which in turn have spokes. Therefore, a <i>spoke</i> has a subordinate partitive relation to <i>wheel</i> which has a subordinate partitive relation to <i>bicycle</i> .
Temporal	before, concurrent, after	A relationship involving time or sequence.	<i>primary school</i> comes before <i>secondary school</i> which comes before <i>college</i> .
Spatial	left, right, above, below, inside, outside, etc.	A relationship involving relative position.	In sewing, <i>interfacing</i> is placed inside two layers of <i>finished fabric</i> .
Associative		A relationship of association, such as cause-effect, product-material, product-user, etc.	A <i>saw</i> is used by a <i>lumberjack</i> who in turn works in a <i>forest</i> . All three terms have an associative relation.

Semantic relations should be bidirectional, and this should be set automatically by the TMS. With bidirectional relations, if term A has a subordinate relation to term B, then term B is automatically given a superordinate relation to term A.

As described in *Data categories*, most off-the-shelf TMS do not handle semantic relations well. Some do not support them at all, while others have faulty linking mechanisms. A common problem is when the link to the target entry is established based on the spelling of a term in the target entry instead of using an unambiguous linking mechanism such as the entry’s concept ID. When there are two or more terms in the termbase that are spelled the same (homographs) the resulting relation may be associated with the wrong term. And due to the principle of univocity described earlier one can expect many homographs in termbases.

Because repurposability is key for maximizing the return on investment of corporate termbases, and semantic relations are required for some applications, corporate terminologists should ensure that the selected TMS supports semantic relations and implements them correctly. We suggest that at a minimum the relation types generic, partitive, and associative should be implemented.

Search

A major goal of any company is to have its web pages rank high in search engines such as Google when people are searching for products or services that the company offers. Google is the most widely used search engine in the world. Thus in the following paragraphs we will focus on Google, although similar techniques apply to other search engines.

How Google ranks web pages, the intelligence behind that ranking, is referred to as its *search algorithms*. In the early days of the internet the search keywords that were to be associated with a website for search ranking purposes were indicated in its *meta tags*. Getting a web page to be returned when users searched with certain keywords was a simple matter of putting those keywords in the meta tags. For example, if the page is about hybrid cars, put *hybrid cars* in the meta tags and bingo, people searching for hybrid cars would be directed to your page. Unfortunately, web developers began to abuse this feature by adding all kinds of irrelevant but popular keywords into the meta tags just to bring people to their site.

The abusive use of the keyword metatags is one of the reasons why Google changed its search algorithms. To prevent this abuse, the new algorithms are not fully disclosed. There are, however, some recognized best practices that can help to raise a website's ranking when certain SEO keywords are used in the search bar.

With SEO techniques constantly changing it is not possible to describe how to manage SEO keywords, nor what data categories are required, in a termbase that would be valid for all companies or even current by the time you read this book. Let us briefly describe, however, one possible scenario where data categories that support global SEO are included in a termbase. The *keyword effectiveness index (KEI)* is a measure of the effectiveness of a search keyword. KEI is calculated based on a formula that uses two figures (1) the number of times that the keyword is used in a search per day (the *search volume*), and (2) the number of pages that are retrieved by that keyword (the *competition*). KEI is also calculated within locale-specific search domains, such as www.google.hk for Hong Kong. Figure 29 shows information recorded in a termbase for the term *MTR*, which is the name of the underground train in Hong Kong.

English	
MTR	
Keyword?	Yes
Market	Hong Kong
KEI	0.6
KEI date	2020-05-01
Competition	50,300,000
Search volume	5,500
Ranking	1
Term type	Acronym
Part of speech	Noun

Figure 29. SEO metadata in a term entry

Essentially one of the principles of SEO is to find effective keywords through calculations such as the KEI and then embed those keywords into certain prominent places of the website, such as the title, headings, and alt text for images.⁷¹ Since there is undoubtedly more than one effective keyword for a given topic it may be necessary to place more than one keyword on a web page and, depending on the similarity of their meanings, they may be synonyms or quasi synonyms and therefore suitable for storage in a concept-oriented termbase.

For example, Figure 30 shows the SEO-related information for *mass transit railway*, which is the full form of *MTR* and is therefore recorded in the same entry. Note, however, the much lower KEI value compared to that of *MTR*. This indicates that it is a much less effective keyword, reflecting the fact that most people search for “MTR”.

mass transit railway	
Keyword?	Yes
Market	Hong Kong
KEI	0.00009
Competition	328,000
Search volume	5.7
Ranking	1
Term type	Full form
Part of speech	Noun

Figure 30. SEO metadata in a term entry

Every time you consider adding a set of terms to the termbase that will be used for a specific purpose, you should also anticipate the necessity of isolating

71. The number of web pages that link to and from the site also has an impact.

this set of terms from other terms in the termbase, such as for generating an export file (this is described in greater detail in the next section). Thus, if search keywords are included, there should be a field allowing these terms to be marked for this purpose. This could be a simple checkbox field, labelled **Search keyword?** which, when checked, indicates that the term is a search keyword. Alternatively you could use a field that allows for a yes or no selection. Another approach that was mentioned in *Controlled authoring* involves a picklist field dedicated to the various purposes of entries in the corporate termbase. This is also further described in the next section.

The following features have been proposed which would allow terminology resources to be developed for search purposes in a termbase:

- Search keyword: yes or no. The default is no.
- Distribution: internal, external, both, unspecified. The default is unspecified.
- Synonyms for query expansion and for ranking keywords
- Hierarchical concept relations for faceted search
- Possibly other data such as the KEI and the variables used to calculate the KEI.

Subsetting

Given the multi-purpose nature of corporate termbases, the ability to isolate parts of the termbase is very important. We refer to this requirement as *subsetting*.

Subsetting can be based on topic area, administrative division, and on purpose. Typically all three types are used. All subsetting fields should be implemented as picklists.

Subsetting based on topic area is needed for isolating all the terms that convey concepts in a certain vertical industry or domain. For example, a company that produces both computer software and computer hardware may need to distinguish between these two sets of terms for various reasons. By providing translators who are translating a document about software with a termbase that includes only software terms, terms from unrelated or irrelevant topic areas do not need do not clutter the small autolookup window of the CAT tool.

Subsetting based on broad topic areas is handled by *subject fields* (sometimes referred to as *domains*). It is often useful to implement a multi-level hierarchy of subject fields. Figure 31 shows a small section of a three-level hierarchy.⁷²

72. Source: Interverbum Technology.

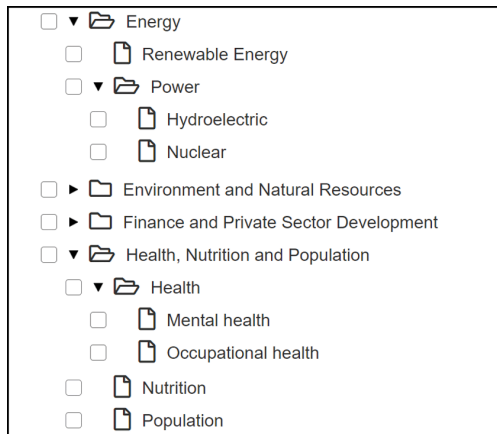


Figure 31. A multi-level hierarchy of subject fields

Topic-based subsetting realized through subject fields may not be finely-grained enough. This is why many corporate termbases feature additional subsetting fields implemented as picklists at appropriate levels in the entry structure. For example, it is not uncommon to find a field such as **Product** or **Product area** to be able to organize terms along product lines. Microsoft's termbase, for example, includes a picklist field at the term level called **Product/technology**, which features values such as *Office 365*, *Skype* and *Sharepoint*. This more granular subsetting capability enables fine-targeted handoffs to translation as well as the publication of product glossaries for end-users.

Subsets based on administrative division enable entries to be associated with various corporate stakeholders, for example, Marketing, Legal, Human Resources, and Engineering.

Subsets based on purpose are critical to serve the goal of repurposing. Subsetting fields can be used for search filters, enabling any subset to be displayed and/or exported separately from the rest of the termbase. Purpose-based subsets can thus be channeled to the applications that require them. If the termbase houses terms and other lexical units that are needed by a CA tool, it is essential to be able to separate those terms from the rest. First, most of the terms needed for CA are not needed by translators. Second, these terms require certain metadata that are seldom required by other terms in the termbase such as usage notes, usage status, and possibly even grammatical information such as verb transitivity. A purpose-based subset picklist field might include, therefore, values such as *CA*, *SEO*, *CAT*, *Corporate taxonomy*, and more.

Finally, it should be noted that subsets of the termbase can be associated with role-based user views for various purposes, such as in a review and approval workflow.

Data category proposal

To propose a set of data categories that would be a good foundation for a multi-purpose termbase in a global enterprise, let us begin by considering the 2003 LISA report which focused on this type of termbase. According to this report, the following types of data categories are common in termbases in global companies:

- subsetting data categories, to organize the terms into logical groups, such as subject field, source, purpose/application, product, etc.
- semantic relations: cross-references and various hierarchical relations
- textual descriptions: definitions, context sentences, usage notes
- term type: abbreviation, acronym, full form, short form, variant
- part of speech: noun, verb, adjective, adverb, other.

Given the information found in the existing literature (see *Relevant literature*), the requirements laid out in the previous sections of this chapter, and the LISA surveys, we propose the set of data categories shown in the following tables as a possible foundation for a termbase serving a large company. However, one should not assume that this set will be suitable in its entirety for any specific industry, company or organization. Some of the data categories listed in the following tables, for instance, may not be needed while others that might be needed could be missing.⁷³

The tables are arranged to indicate data categories according to where they would occur in the hierarchical structure of a terminological entry, as described in ISO 16642:2017 and in *The data model*.

The concept-level contains information that pertains to the concept that the terms in the entry convey. It applies to all terms in the entry.

73. For instance, termbases developed for software development companies will need data categories designed for indicating the UI location of terms, the type of UI object designated, among others. See Dirk-Schmitz, 2015.

Table 6. Concept-level data categories

DC name	Content model	Note
Subject field	Picklist	Values to be determined based on an enterprise taxonomy.
Section	Picklist	Optional additional subsetting DC. For example: marketing, support, product documentation, etc.
Distribution	Picklist	Internal, external, both, unassigned.
Applications	Picklist	A subsetting field intended to allow terminology to be directed to various enterprise applications. Possible values include SEO, CA, CAT, etc.
Definition	Text	A written description of the concept's meaning. ISO 704 provides excellent guidelines for writing definitions, however, in corporate termbases definitions may deviate from those guidelines for various application-oriented reasons.
Source of definition	Text	The source of the definition, i.e. the name of the person who wrote it, a citation from a print resource, a website, etc.
Note	Text	Any note relating to the concept.
Related concepts	Relational	A field that allows unambiguous links to other entries. It should support multiple relation types as described in <i>Concept relations</i> .

The language level is primarily used to identify the language of a particular set of terms. Some termbases do not include any data categories at this level other than to specify the language. It is repeatable within an entry to allow for more than one language.

Table 7. Language-level data categories

DC name	Content model	Note
Language	Picklist	Values to be adopted from IETF BCP 47.
Definition	Text	Optional field to allow for definitions in languages other than the working language of the termbase.
Note	Text	To allow for any note or comment about the language or about the terms in this language.

The term level describes the information pertaining to one term. It is repeatable within a given language level to allow for more than one term for a given language (synonyms).

Table 8. Term-level data categories

DC name	Content model	Note
Term	Text	The term, in canonical form. A term field must contain only one term. The term field itself, or elsewhere in the entry, should also allow for the inclusion of symbols as terms.
Source of term	Text	The origin of the term (person, document, website, etc.)
Part of speech	Picklist	noun, verb, adjective, adverb, proper noun, other.
Gender	Picklist	masculine, feminine, neuter.
Term type	Picklist	full form, abbreviation, acronym, short form, variant.
Usage status	Picklist	preferred, admitted, restricted, prohibited.
Process status	Picklist	proposed, under review, finalized.
Context	Text	A sentence that contains the term. Context sentences should be obtained from real sources.
Source of context	Text	The source of the context sentence.
Association	Picklist	Subsetting field to indicate some area or application that the term is associated with. May be more appropriately named Product , Industry , Sector or any other name. This is particularly valuable when there are multiple terms in the entry in the same language; each term may be preferred for a specific product or in some other context. Note: more than one subsetting field may be required.
Usage	Text (single line)	Brief statement about how the term should be used. (For display in CA context menu.)
Usage note	Text (multi-line)	Further explanation about the conditions of usage.
Note	Text	Any other note or comment.

Regarding the **Process status** values recommended above, it should be noted that they are slightly different than those recommended in the TBX standard, which are *unprocessed*, *provisionally processed* and *finalized*. The three values proposed here are equivalent to those values, i.e. proposed is equivalent to

unprocessed, and under review is equivalent to provisionally processed. It is just our opinion that the values we propose are more intuitive. This leads to an important observation: it is not essential that the names of the fields and field values used in a given termbase match exactly the TBX names for any data category. What is important is that the names adopted should be mappable to data categories that are recommended in standards such as TBX, i.e. they should have an equivalent meaning and use. If a terminologist feels, for instance, that *new* is even easier to understand than *unprocessed* or *proposed*, that name can be adopted in the termbase design.

The terminology management system

A *terminology management system (TMS)*⁷⁴ is a software program specifically designed for managing terms and related information in a database. In this chapter, we describe functions and features that should be considered when you are selecting a TMS. Note that not all the features described in this chapter are required; they are simply provided for consideration purposes.

Standalone or integrated

It is important to understand the difference between an *integrated* and a *stand-alone* TMS, as this tends to be a major dividing line in terms of functionality and repurposability. Most commercially-available TMS are part of, offered alongside of, or designed for an over-arching software suite such as a CAT or CA tool. Examples of these systems, which we categorize as integrated, include MultiTerm (SDL), TermStar (Star Transit), crossTerm Manager (Across), QTerm (MemoQ), and TermBase (MultiTrans). Other tools include terminology management functions without offering them as separate products as such, for instance the Acrolinx CA tool and the XTM CAT tool both have integrated terminology management functions. These types of TMS tend to focus on a specific use, i.e. CAT or CA as the case may be, and understandably so.

In general, due to their singular use-case, integrated TMS are not enterprise-scale in the sense that they do not serve all use cases found or potentially necessary in a global enterprise. Unfortunately, many enterprises do not realize this and choose to adopt the TMS within an already purchased CAT or CA tool only to discover later that their termbase is not optimized for other purposes beyond translation or authoring.

In contrast, a standalone TMS is not tied to any other tool or product. The development priorities tend to be more generic and aligned with the features desired by a broad range of users. They have sufficient functionality and/or do

74. The acronym TMS is also used in the translation industry for translation management system. Since systems to manage terminology and translation are often used in tandem, the term *terminology management tool (TMT)* is sometimes used to avoid confusion.

not present some of the limitations of integrated tools, such that they can be used to produce termbases that are more repurposable than those produced in an integrated TMS. They may offer an API or a connector for integration with CAT and CA tools making it feasible to use them in place of the TMS integrated in that tool. Unfortunately, commercially-available stand-alone TMS are few and far between. TermWeb (Interverbum Tech) and Coreon are the major competitors in this league that we are aware of. Some global enterprises have chosen to develop their own standalone TMS in-house.⁷⁵

Due to their interest in working with a range of other tools, companies that produce a standalone TMS tend to be more motivated to fully support the standard for terminology exchange, TBX (ISO 30042: TermBase eXchange).

Core features

The first thing to determine is the system's *deployment model*. Is the TMS a client/server application or cloud-based? Client/server applications require separate installation of the TMS on every computer. Depending on the number of users, providing technical support and managing software updates can be onerous. Cloud-based systems do not require installation as they are accessed through a web browser. If the TMS is cloud-based, is it offered as a Software as a Service (SaaS), which is hosted by the TMS vendor, or does the purchasing company need to provide hosting services? Is self-hosting preferred or even required by the company for security reasons? Does it work on all browsers? If not, it will be necessary to manage browser versions for all users. What are the restrictions imposed by the company's IT security for approval of the software? If it is offered as a SaaS, what are the ongoing costs compared to self-hosting?

Then there is the *development model*. Is the TMS a standalone software product or is it integrated as a component of another software or software suite, as discussed in the previous section? For example, many TMS are developed and offered as part of a CAT tool. Note that a TMS that was developed as a standalone product from the outset is different from a TMS that was originally integrated in another tool but is now being offered as a separate product. The former is typically more robust and full-featured than the latter.

In the following tables, we itemize some general features that a TMS used for corporate terminography should have. Included in these tables are features that relate to the TMS as a whole. Unless otherwise indicated in the **Explanation**

75. For example, Microsoft and IBM.

column, these features are essential. Subsequent sections are devoted to features in specific areas such as entering terms, searching, and exporting.

Table 9. Core design / Data model

Feature	Explanation
Concept orientation	Each entry allows multiple terms in the same language, to account for synonyms, and multiple languages in the same entry to account for TL equivalents.
Term autonomy	All terms in the entry can be documented with the same full set of fields.
Compliance with TMF	TMF is <i>Terminological Markup Framework – ISO 16642</i> . This ISO standard specifies the structure of terminology entries. There should be three nested levels: concept, language, term.
Data model is customizable	The TMS should not have a fixed data model. The terminologist/administrator should be able to create custom fields and field values and specify: the content model of the fields (text, picklist, etc.), default picklist values, the order of fields in the entry, the order of picklist values, mandatory vs optional fields, and so forth. It should also be possible to make straightforward changes to the data model after the termbase is created, including changing the name of a field, adding or removing a field, changing the name of a picklist value, adding or removing a picklist value, changing which fields are mandatory, default values, etc. Of course, some changes to the data model might affect the termbase content, such as removing a field or a field value. In these cases, the administrator should be given a warning message when attempting the change.
Fields assignable at three levels	It should be possible to specify which level (concept, language or term) a field occurs in. It should also be possible to assign a field to more than one level.
Range of field types	Picklist, free text, date, multimedia, and relational.
Mandatory fields	It should be possible to specify that a field is mandatory. Users cannot save the entry without completing this field. To avoid disrupting the progress of users' work and their ability to save entries, even incomplete ones, mandatory fields should be kept to a minimum and they should be restricted to fields that are easy to complete. A field that is often mandatory is the Part of speech .
Nested fields	This feature is not essential but can make the user interface look more well organized. It refers to having a field that itself has a subordinate (child) field. An example is Definition and Source . The subordinate position of Source below Definition signifies that the

Table 9. (continued)

Feature	Explanation
	information in the Source field is the source of the definition and not of another piece of information. Without the nested structure, a workaround is to use specific field labels, such as Source of definition .
Repeatable fields	This is another non-essential feature and in fact it is quite rarely found. ⁷⁶ It refers to the ability to allow users to include more than one instance of a particular field in an entry, for example, to include more than one context sentence for a term. The repeatability of any given field would be set by the terminologist/administrator during the termbase design. Making fields repeatable avoids the situation where users enter more than one instance of information in a single field, for example, including more than one context sentence. Doing so violates the principle of data elementarity. Another case is definitions. Although a concept entry has only one meaning, and therefore, only one definition is necessary in theory, some termbases allow for multiple definitions so that they can be proposed and reviewed as part of entry development with the ultimate goal that only one of the definitions will eventually be retained. On the other hand, some fields should never be repeatable because only one instance is possible. For example, a term (in one concept entry) can only have one part of speech value. If setting a field as repeatable is not possible, the workaround is for the terminologist to design the termbase at the outset with more than one instance of fields where this feature is desired (such as Context). This means, however, that those extra fields will always be visible even when empty.
Order of fields	It should be possible to specify the order of the fields in the entry. Usually the most frequently used fields are placed at the top to minimize scrolling.
Rename fields	It should be possible to change field labels at any time.
Graphics, attachments	The TMS should allow graphics to be inserted in the entry. Usually they are placed at the concept level. It should also be possible to attach documents to the entry, such as PDF files.
Data categories	Being fully customizable, the TMS should allow the use of any of the data categories that are described in terminology standards such as

⁷⁶. This feature is offered in the iTerm TMS, by DANTERM Technologies. Beside the repeatable field, there is a plus sign icon. Users click the icon to add another instance of the field on-the-fly.

Table 9. (continued)

Feature	Explanation
	ISO 12620:1999, ISO 30042:2008, and the DatCatInfo ⁷⁷ data category repository.
Field for workflows	The TMS should support the creation of automated workflows. This typically requires a dedicated field to trigger workflow stages: Process status , with permissible values such as proposed, under review, and finalized.
Picklist values	It should be possible to specify the values of picklist fields, the order of their appearance, and any default values. It should also be possible to add values, remove values, and change values from a picklist field in the data model at any time. Adding and changing picklist values should not be a problem in a well-designed TMS. Any entries that contain a picklist value that is changed should be automatically updated with the changed value. Removing a value, however, can be difficult because doing so affects entries that contain that value. How should those entries be handled? Should the value just be removed, and no new value re-assigned? Or is it necessary to review each affected entry and determine an appropriate new value?
Sections	Termbases that become quite large benefit from being organized into sub-structures, such as divisions or sections. This can be implemented via a dedicated picklist field at the concept level. But a TMS might also provide a dedicated feature for this purpose.
Reusable data model	Once the data model for a termbase has been created, it should be possible to save it and reuse it for other termbases. Saving the data model to an external file is also a good backup practice. Should the data model become dysfunctional or corrupted, the backup copy can be used to recreate it.
File repository	Files such as graphics and document attachments should be stored in a file repository that is associated with the termbase and is easily accessible.

77. datcatinfo.net

Table 10. Data points

Feature	Explanation
Identifiers	Each entry should have a unique concept ID and each term should have a unique term ID. This is essential for cross-referencing purposes. Some other key elements should also have identifiers, such as subject fields and graphics.
Change history	Changes to the termbase should be automatically recorded and it should be possible to view the history of changes.
Revert to previous point	Using the data from the change history, it should be possible to revert the termbase to a previous state. It should also be possible to revert individual entries to a previous historical state.
Record of deleted entries	The TMS should maintain a record (full copy) of deleted entries so that they can be restored if necessary. It should be possible to purge this record periodically when desired.

Table 11. Integration

Feature	Explanation
Connectors	The TMS should provide APIs to allow integration of the terminology into other enterprise systems. It should be possible to provide a mechanism for employees to add terms to the termbase directly from their own working environment, such as an authoring software or even an email application.
Remote termbase can be installed locally	If the TMS is client-server based and the company has employees who may experience connectivity problems, consider offering employees who need to query the termbase frequently, such as translators, the possibility to install the termbase locally and work offline.

Languages and scripts

The TMS should not require a dedicated source language and dedicated target language(s). Users should be able to select any language as their SL (for searching purposes) and any other language as the TL (for a more prominent display than other languages, such as near the top of the entry). Users can search for terms in any language in the termbase by changing the search language at any time.

Some languages have script variants, such as Japanese which contains a mixture of Kanji, Hiragana and Katakana, or Punjabi, which is written in either the Gurmukhi or Shahmukhi scripts, based largely on geographic distribution. Care should be taken to determine how these scripts should be structured and represented in the termbase. For Japanese, for example, all three scripts are an integral

part of the writing system of the language, whereas for Punjabi the language can be written in either script entirely. For Japanese, therefore, it may be useful to have a field for indicating the script at the term level, whereas for Punjabi one could consider approaching each script variant as a separate language, strictly from a data modelling perspective. Chinese has two scripts: Simplified Chinese and Traditional Chinese. These are largely regionally-based (Hong Kong and Taiwan use Traditional Chinese, whereas mainland China uses Simplified Chinese). So, in this case, a **Geographical variant** picklist field with the two values should work.

Term entry

The TMS should include some functions to facilitate the manual addition of terms.

Table 12. Term entry

Feature	Explanation
Customizable input templates	It should be possible for the administrator to create different templates for the purpose of adding new terms, customized for different users or user groups. For example, an input template for the authoring group could include just a few fields pertaining to the SL, such as the SL term, a source of the term, a definition, and a note.
Create entry by duplicating another	Often a user works on a number of entries where some of the fields contain identical content, such as the product area, section, source of term, etc. While not an essential feature, duplicating an entry and then making the necessary changes can be faster than creating each entry from scratch.
Copy field to all terms in entry	Usually the terms in an entry have common features, such as part of speech, source, product area, etc. While not an essential feature, it can be convenient to be able to copy a field to all terms in an entry.
Batch editing	Sometimes the user needs to work on a set of entries. Having a way to work on them together increases productivity substantially. Terminologists often do this by exporting the set (using a search filter) to Excel (or TBX) format, updating the information in the exported file, then reimporting the file to the termbase while using the Concept ID for synchronization with the existing entries. While not essential, it would be helpful if the TMS had a feature that simulates this kind of batch work, so that exporting and importing would not be necessary.

Import and export

Import functions enable existing terminological resources, such as glossaries, to be added to the termbase very quickly. Export functions support repurposability; they enable the terminology in the termbase to be shared with outside organizations and to be integrated into other applications.

Table 13. Import

Feature	Explanation
Supported file formats	The import function should support the following file formats: TBX, Excel (or CSV or Tab-delimited), plus any format that is proprietary to the TMS itself.
Assignment of values	During import, it should be possible to assign a field value to all entries that are being imported. For example, include the client's name in a particular field for all imported terms that come from a particular client.
Field mapping	It should be possible to map a field name in the import file to a specific field in the termbase. For example, if the import file contains the TBX element <descrip type="explanation"> and most of these explanations are actually definitions, it would make sense to move the content of these elements into the Definition field of the termbase. As a workaround, this can also be done by changing the TBX element name in the import file.
Doublette management	The TMS should provide functions to enable the efficient handling of terms in the import file that already exist in the termbase in order to minimize the occurrence of <i>doublettes</i> (duplicates). This feature, which is sometimes referred to as <i>synchronizing</i> , is difficult to implement and is therefore missing from all but the most advanced TMS. Necessary functions include some variation of the following actions: (a) ignore the incoming entry, (b) merge the incoming entry with the existing entry, (c) replace the existing entry with the incoming entry. Synchronizing incoming entries with matching ones in the termbase can be based on matching terms, but this can be dangerous due to the existence of homographs. The most precise synchronizing is based on the concept ID. However, this is only possible for an import file that was previously exported from the same termbase. More primitive TMS simply import all the entries then provide a view of the doublettes post import to allow the terminologist to work on them. Doublettes are further described in <i>Workflows</i> .

Table 13. *(continued)*

Feature	Explanation
Merge entries	After import, even if the TMS has sophisticated import actions as described above, there will probably be some doublettes that need to be merged into one entry. The TMS should provide an easy way to merge two entries.
Logging	The import process should produce a log file that provides detailed information about the import: how many entries were successfully imported, how many failed, how many were merged, skipped, overwritten, etc. Entries that failed to be imported should be identified and the reason given for the failure. Any failures due to non-compliant information in the imported file should be clearly documented.
Reusable import settings	It should be possible to save the settings used to import the file for future reuse.

Table 14. Export

Feature	Explanation
Supported file formats	The export function should support the following file formats: TBX, Excel (or CSV or Tab-delimited), plus any format that is proprietary to the TMS itself. Exporting to a printable format such as PDF with a suitable layout is also desirable.
Field mapping	It should be possible to map a field name in the termbase to another name in the export file. For example, if the termbase contains a field named Note and most of these notes are definitions, it would make sense to export the Notes into the TBX element <descrip type="definition">. As a workaround, this can also be done by changing the TBX element name in the exported file.
Apply a filter	It should be possible to export only a portion of the termbase by applying a search filter to the export.
Choose fields	It should be possible to export only a selection of fields for each entry. For example, you could export only the terms and definitions and send the exported file to someone who is assigned to review the definitions.
Choose languages	It should be possible to export a subset of the languages in the termbase.
Choose entries	This optional feature refers to the ability to export a set of entries that have been manually selected by the exporter (as opposed to a set of entries based on a search filter).

Table 14. *(continued)*

Feature	Explanation
Logging	The export process should produce a log file that provides detailed information about the export: how many entries were successfully exported, how many were not exported, which ones were not exported, and the reason for the failure.
Reusable export settings	It should be possible to save the settings used to export the file for future reuse.

Views

Different users have different needs and therefore require customized views of the termbase entries. Providing custom views according to user needs helps to increase user satisfaction with the termbase. For example, translators of a certain language pair need only see those languages; other languages can be hidden. Writers are interested in definitions, synonyms and usage notes. Translators also need context sentences, whereas writers do not.

Table 15. Views

Feature	Explanation
Customizable views	It should be possible to create customized views, save them, and assign them to user groups. Types of view customizations include hiding specific fields, hiding specific languages, and hiding subsets of the termbase (through a search filter).
Select SL and TL	It should be possible for users to set their preferred source (search) language and target language. The selected TL should appear immediately below the selected SL in the entries (to minimize scrolling).
Language display	Users should be able to customize their own display with respect to the languages that appear and the order of their appearance.
Layout styles	It should be possible for the termbase administrator to customize the style of the layout (colors, fonts, etc.)

Search

Providing advanced search functions is essential not only to find terms, but also to focus the work on specific subsets of the terminology.

Table 16. Search

Feature	Explanation
Wildcards	The search function should support wildcards in any position of the search string. It should be possible to use multiple wildcards.
Fuzzy search	Fuzzy search is an extension of wildcard search which may include different pattern matching algorithms such as re-arranging the words in a multi-word search string, changing the order of letters in a word, etc.
Full-text search	Full-text search is a search that is carried out in one, several, or all text fields of the termbase. It should be possible for users to carry out a full-text search and to select which text fields they wish to search in. For example, a user may want to search for a particular sequence of words in the definitions.
Search in more than one termbase	If users have access to multiple termbases, they should be able to carry out a search in any combination thereof.
Batch search	It should be possible for users to conduct a batch search, meaning that they upload a text file that contains a list of search terms, or put the terms into a provided text box, and the system searches for all the terms and provides a report of findings (which terms are found, which are missing).
Filters	It should be possible for all users to create search filters based on any combination of fields in the termbase. For example, a user might want to search for all acronyms (which is a Term type value) that were created during a certain time period or by a certain user. Boolean operations (AND, OR, NOT, CONTAINS, etc.) should be supported.
Incomplete records	Among the possible search filters one that displays entries that are <i>incomplete</i> based on criteria specified by the user should be included. For example, a user may want to search for entries that are lacking a definition.
Entries without translation	It should be possible to search for entries that are lacking a term in a specified language. This is necessary to fill language gaps in the termbase.
Filters can be reused	It should be possible to save all filters for reuse.

Table 16. *(continued)*

Feature	Explanation
Filters can be shared	It should be possible to keep a saved filter private or to share it with other users.
Statistics	Search filters should display summary statistical information: how many concepts (entries) and how many terms satisfy the filter.

Access controls

Access controls help preserve the quality of the termbase. They enable the corporate terminologist to assign responsibility for certain parts of the termbase to specific users.

Table 17. Access controls

Feature	Explanation
Restrictable rights	It should be possible to restrict termbase access at different levels for users and user groups.
Restrictions by function	It should be possible to restrict functions such as creating, modifying, deleting, importing, and exporting entries for users and user groups.
Restrictions by language	It should be possible to restrict access for users and user groups based on language, for example, to allow translators to modify only information in their working languages.
Restrictions by subset/section	It should be possible to restrict access for users and user groups based on a subset or a section of the termbase (set by a filter, for example).
Restrictions by field	It should be possible to restrict access for users and user groups to specific fields. For example, one might want only certain users to write definitions.
Restrictions by role	User groups are often created according to roles (writer, translator, reviewer, for specific languages, etc.) and restrictions applied according to roles.

Relations

The capability to record concept relations enables the termbase to develop into a knowledge system. Relations are also essential for implementing faceted search.

Table 18. Relations

Feature	Explanation
Entailed terms	<i>Entailed</i> terms are terms found in text fields (such as in definitions) that are hyperlinked to another entry in the termbase. It should be possible to create entailed terms in any text field.
Between entries	There should be dedicated fields at the concept level to link concepts.
Precise link mechanism	A relation pointing to another entry should be based on the concept ID of the target entry or on the term ID of the target term, not based on the term's written form. Links based on the written form are ambiguous and problematic due to the presence of homographs.
Bidirectional	A relation established from one entry to another should automatically produce a corresponding relation in the opposing direction. For example, in concept entry A if there is a superordinate generic relation to concept entry B, then in concept entry B a subordinate generic relation to concept entry A should be automatically established.
Different types	Various types of relations should be supported, such as generic, partitive, and associative. See <i>Concept relations</i> .
External links	It should be possible to create hyperlinks to external sources.
Relation graphs	Sometimes called concept maps. Visual representations of linked concepts are a nice feature particularly for termbases that develop into knowledge systems. For an example, see Figure 28.
Avoidance of broken relations	If a user attempts to delete a concept entry or a term, and doing so would break a relation pointing to that entry or term from another entry or term, a warning should be displayed. An alternative approach would be to automatically delete the relation.
Manage broken relations	A feature should be provided to assist the terminologist in finding and correcting broken relations.

Workflows, community input

Table 19. Workflows, community input

Feature	Explanation
Users can suggest terms	There should be a very simple way for users to suggest terms. This can be either directly in the TMS or through another channel. The key requirement is that it is very simple and intuitive and requires no training or knowledge of terminology or of terminology tools.
Commenting	Users should be able to add comments to an entry.
Feedback to admin	Users should be able to send feedback to the administrator.
Permalinks	It should be possible to obtain a direct link to an entry, for example, to copy/paste this link into an email so that the recipient can click the link to access the entry directly.
Role-based workflows	It should be possible to create workflows for specific users and user groups. For example, a Spanish terminologist could receive automatic notifications when Spanish translators add new terms. The new terms are automatically set to a proposed status until the terminologist approves them.
Workflow triggers	Workflows should be triggered when certain changes occur in the termbase, and this should be fully customizable. For example, when terms are assigned a particular product value and the translation of the product needs to be started for a specific set of languages, once a certain number of those terms are created in the termbase or for example a start date is reached, a job is sent to those TL terminologists to fill the TL terms. The job could take the form of, for example, a TMS view of the entries that need to be worked on, or even an email with links pointing to those entries.
Workflow notifications	Similar to the previous requirement, when a workflow includes a notification that is sent to a user, it should be possible to configure when these notifications are sent, for example, after a certain volume of entries meet the workflow conditions, after a certain time period (daily, weekly), etc.
Workflow scheduling	It should be possible to schedule when workflows start as well as set target dates for completion.
Customizable workflows	Workflows should be fully customizable. It should be possible to set pre-conditions and post-actions to be taken based on termbase fields.

Administrative functions

Various administrative functions are required to provide statistics for reporting purposes, to manage users and groups, to make global changes, and so forth. These should be restricted to users with administrator access, i.e. the corporate terminologist.

Table 20. Administrative functions

Feature	Explanation
User management	Administrators should be able to create, modify, and delete users, and reset passwords.
Group management	It should be possible to create groups, assign users to groups, and assign various termbase objects to groups (workflows, views, templates, filters, etc.)
Job scheduling	It should be possible to define certain automated processes, define conditions for their activation, and schedule when they run. For example, send notifications to users, purge deleted entries, hand-off to translation, etc.
Projects	Being able to define a <i>project</i> , which is a collection of settings, filters, tasks, users or groups, workflows, etc., provides a framework for completing work in a particular area of the termbase that needs attention.
Automated backups	It should be possible to schedule automated backups.
Restore from backup	It should be possible to fully restore a termbase from a backup.
Global changes	It should be possible to make global changes across the entire termbase, for example, change a word in all definitions, etc.
Enhance performance	If performance slows down, there should be some debugging features to attempt to resolve the problem such as through re-indexing (if applicable).
Statistics	There should be an administrative dashboard that provides detailed statistics: number of entries, number of terms per language, number of entries per section, number of definitions, number of search queries, how often a user accessed the system, etc.
Search log	There should be a log of searches. This log can be used, for instance, to discover terms that people are searching for but are missing from the termbase.

Implementing and operating the termbase

In this part we describe how to create the termbase and populate it with terms. The topics of customization, quality evaluation, and promotion and training are also addressed.

Terminology management is an ongoing process; it is not a “project,” which by definition has a beginning and an end. In this part we therefore also discuss ongoing operation of the terminology initiative.

Create the termbase

In this chapter we describe how to create the termbase starting with the data model and data categories.

As described in *Users and their roles*, *Access mechanisms and user interfaces* and *Data category selection*, different users need different types of information. A robust TMS should be customizable to cater to the needs of different types of users. In this chapter we also discuss areas of the TMS that frequently need to be customized such as filters and views.

The data model

A data model is a representation of the structure and content of a comprehensive terminological entry in the termbase. It includes:

- the structure of a terminological entry
- a list of the data categories
- the content model of each data category
- the placement of the data category in the entry structure
- an indication of whether the data category is mandatory or optional
- an indication of whether the data category is repeatable
- an indication of whether the data category has a default value.

The structure of a terminological entry should adhere to *ISO 16642 – Terminological Markup Framework*. This means that the entry should have three nesting levels or sections: concept, language, and term. For demonstration purposes, in Figure 32 the English language section contains two term sections, and each term section contains one English term. The French language section contains one term section. This is to demonstrate the principles of both concept orientation (multiple terms in the same entry) and term autonomy (each term having its own section). Having more than one term section within a language allows for documenting synonyms.

Data categories, which are further described in *Data categories*, *Data category selection* and *Data category proposal*, should be selected according to user needs. Before beginning to develop the termbase, ensure that you have a final list of the

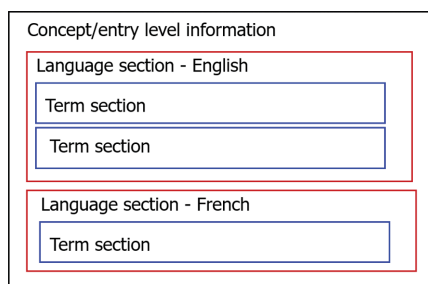


Figure 32. Structure of a terminological entry

data categories needed. Most TMS allow data categories to be added to a termbase after its initial creation, but changing or deleting data categories from an existing termbase can be difficult.

Determine the content model for each selected data category. Content models are described in *Data categories* and in *The terminology audit*.

The placement of each data category in the entry structure needs to be determined, i.e. either at the concept, language, or term level. Some data categories can occur at multiple levels. For instance, it may be desirable to allow definitions at both the concept level and the language level, the latter to allow for definitions in different languages or for slight variations of the concept definition for specific socio-linguistic cultures. Notes are commonly allowed at multiple levels as well. A proposal was provided in *Data category proposal*.

To avoid imposing restrictions in the creation of terminological entries, which can cause bottlenecking and slow down the growth of the termbase, mandatory data categories should be kept to a minimum. In practice, many termbases do not have any mandatory fields except that each entry must have at least one term in one language.⁷⁸ However, according to best practice it is now recommended (by TerminOrgs in particular) to make the part of speech mandatory. Another data category that is frequently mandatory is the source of the term.

Some new initiates to terminology work, who are led to believe by proponents of the classical approaches that definitions are of utmost importance, make definitions mandatory. This has proven to be misguided for commercial environments where definitions are actually not needed by some users nor for some purposes of corporate termbases. If they are not always required, why make them mandatory? That would be an inefficient use of resources, as researching and writing definitions is time-consuming. Definitions should only be created for terms whose meaning would not be known by many users and if those users actually need to

⁷⁸. Coreon does not even impose this requirement.

know the meaning in order to do their jobs. An example would be a technical or product-specific term that translators are unfamiliar with but must fully understand in order to produce an equivalent in the TL. In practice, translators are often adequately served by a context sentence, which is easier to obtain.

Some TMS offer the ability to design templates for creating terminological entries which can be assigned to different users or user groups. A template could include mandatory fields and default values. This enables the assignment of specific mandatory fields and default values for certain users or user groups. For instance, if one user works on terminology for a specific product area, the **Product** field could be pre-filled with a default value for that user.

Similarly, some TMS use workflows to assign mandatory fields and default values, and those workflows are assigned to specific users or user groups. For instance, a workflow could be assigned to a specific user group which automatically sets a field to be mandatory, or which assigns a default value, when members of the group create entries.

Assigning mandatory fields by user, through templates or workflows, enables certain requirements to be selectively imposed, such as requiring certain users to include definitions but leaving this optional for others.

Assigning default values to data categories can be an effective way to increase productivity. For example, if the **Part of speech** is a mandatory field as suggested, making *noun* the default value saves time because the vast majority of terms in termbases are nouns. It saves the effort of manually selecting this value over and over again for many entries. Default values that are enabled for specific users or user groups reduce the amount of information that those users need to input manually.

The user should always be able to override default values.

Using a spreadsheet to gather and organize all this information is an effective means of preparing, reviewing, and finalizing the data model with stakeholders.

Controlling access

To avoid accidental, uncontrolled, and incorrect tampering of data in the termbase access controls should be established which restrict users' access to only those areas of the termbase where they should be authorized to make changes. To set this up efficiently, typically user groups are created in the termbase access control functions based on user role. Individual users are then assigned to these groups. The section *Users and their roles* describes typical role-based groups and the types of tasks they are expected to perform.

It should be possible, for example, to restrict a group's write access to

- specific languages
- specific levels of the entry (concept, language, term)
- specific fields
- specific functions (such as disabling import or disabling the ability to delete an entry).

Field-level access control can help to monitor actions at a very granular level. Write access to the **Definition** field, for instance, could be restricted to certain individuals who have been given training on how to write definitions. The **Process status** field should be restricted to authorized approvers.

Views and filters

The termbase will be more user friendly and consequently you will benefit from greater buy-in for the entire initiative if you provide views of the data that align with user needs. A view is a display of the content of the termbase that includes some fields while hiding others, and possibly also restricts the viewable content through a filter. Keep in mind that screen size is limited, especially for users who access the terminology from within the window of another application. For example, translators using a CAT tool will see the terminology in a small window within the CAT tool editor. A typical range of different views would include:

- a full view of all the fields for terminologists system administrators and other advanced users
- a restricted view for translators
- a restricted view for content creators in the SL.

What is hidden or displayed for each view should be determined in consultation with each group. Examples of typical scenarios include:

- For the translator view, hiding:
 - all languages except the SL and the TL for the given translator
 - all administrative fields (person names, dates, etc.)
 - term relations
 - usage status and usage information for the SL.
- For the content creator view, hiding:
 - all languages except the SL
 - all administrative fields (person names, dates, etc.)
 - context sentences.

Views should be set up so that the hidden parts can be re-displayed by the user during a given session.

Filters are used to search for terminology and to export terminology based on certain criteria. The TMS should support search functions with multiple conditions using Boolean operators. For example, a filter can be configured to display all the terms created by a certain person between certain dates within a specific subject field. It should also be possible to use conditions such as equals, contains, or does not contain, and for dates operators such as equals, before, or after. As mentioned, it should also be possible to use filters in views.

Search filters are also used by terminologists to create business analytics for the purpose of reporting progress to upper management and for identifying areas of the termbase that need additional work. Similarly, search filters can be used to focus the work for a contributor. For instance, a search filter could be created that identifies all entries that lack a French term. This filter is then provided to the French translator, who accesses the entries that require attention. If that filter produces too many entries, it could be further restricted by another condition, such as a specific subject field, so as not to overwhelm the translator. This also allows the work to be focused on terms that are more likely to be semantically related.

General, casual users of the termbase should not be expected to use search filters as this would be considered an advanced function. However, certain filters could be very practical for some users. These filters can therefore be pre-created by the terminologist and made available to users through a link or other means, such as a drop-down list. For example, the home page of the termbase could have a link named “New terms added this week,” to draw attention to the latest company terminology.

Workflows

A workflow is one or more actions that are completed based on the presence of certain conditions. The most common workflow involves the progression of an entry or of a term in an entry through various stages. Other types of workflows may be set up to complete various other actions on entries, which allows the administrator to have a high degree of control over the termbase and its use. Workflows vary greatly from one organization to the other. Most TMS have very limited workflow functions, and some none at all.

Workflows are very helpful for managing the content of the termbase semi-automatically. The following are a few examples of actions that could be done automatically through workflows:

- Routing all new entries to a certain section, user group, etc.
- Routing terms created in a specific language to the user group responsible for that language.
- Giving terms that satisfy a certain condition a specific field value, such as giving all new terms the *unprocessed* Process status.
- Making certain fields read-only for certain user groups.
- Setting the color and style of terms that have certain field values, for instance, displaying *prohibited* terms in bold red.
- Sending email notifications of various sorts.

Regarding the workflow involving the progression of entries, there are some basic principles that apply to all cases. Terminology management normally starts in the source language and then continues on to target languages. Different user groups are responsible for each part, but all work together in the same termbase. Work on SL terms should be done by SL writers and editors, and work on the TL terms should be done by experienced translators. This may seem obvious, but it is surprising how often translators are expected to add SL content to termbases. They may be required to do so on occasion of course, such as when they want to add a TL term to the termbase, but the corresponding SL term is missing, but creators of SL content should be primarily responsible for the SL in the termbase.

There are three different workflows that are common in company termbases: project driven, process-oriented, and usage-oriented.

A project-driven workflow identifies the terms necessary to complete the creation and translation of a given project (document, product, website, etc.) in all relevant languages and ensures that the necessary terminological resources are made available to the employees involved. A typical project-driven workflow might look like the following:

1. SL term is identified prior to translation kickoff (this could involve a term extraction process).
2. SL terminologist adds the SL term to the termbase with additional information including project ID.
3. Languages into which the project will be translated are identified and the entry is marked for translation to those languages. This could involve the use of a special field for this purpose.
4. TL terminologists for those languages are notified through a workflow.
5. TL terminologists add a TL term.
6. Bilingual terminology for the project is made available to translators using a filter on the project ID. This involves either an export/import to the CAT tool or a direct connector to the CAT tool.
7. The project content (document, website, etc.) is sent for translation.

8. Translators can add new terms to the bilingual terminology during translation.
9. Newly added terms are included in the termbase (either through export/import or a direct connector).

Process-oriented workflows indicate how complete an entry is, when it is considered to be finalized, and sometimes, if it should be deleted. The DatCatInfo data category repository⁷⁹ includes a picklist field called **Process status** with the values *unprocessed*, *provisionally processed* and *finalized*. Note that a value for deletion is missing, yet a mechanism to mark entries for deletion is required since the incidence of doublettes (duplicate entries) is quite common. We suggest either adding this value to Process status or creating another field for this purpose.

The criteria used to determine when an entry is considered finalized must be clearly defined. It is preferable to automate the assignment of the finalized process status value when the criteria are met, if the TMS supports this. Avoid imposing too many conditions for this to occur. Excessive conditions can create bottlenecking, with many entries in the review queue, to the point that much of the termbase is in a pre-finalized state for long periods of time. This situation will be negatively perceived by both users and executive sponsors. The conditions that must be met for entries to be considered finalized should be determined based on what is deemed to be essential information for an entry. It is not necessary, for example, for an entry to include a term in every supported language; one might require only the SL, or the SL and one or two of the most prevalent languages required for company materials. It is also not standard practice in commercial terminography to require a definition.

The following describes a basic process-oriented workflow (Figure 33):

1. An employee submits a proposal for a new term. This can be done in one of two ways: (1) use a feedback mechanism that bypasses the termbase completely such as an email link on the user interface, or (2) allow employees to add new terms themselves directly in the termbase. In the latter case, the term is automatically given the process status *unprocessed* and if desired, through a filter and custom view, it can be hidden from general view until the process status has been upgraded to *provisionally processed* or *finalized*.
2. The proposal is routed to a user group whose members have the authority to create, edit, and finalize new terms. For demonstration purposes, this group will here be referred to as the Terminologist group and its members referred to as Terminologists. If option (1) is used in the previous step the routing is via an email or other mechanism. If option (2) is used the routing is done

79. datcatinfo.net

automatically by a workflow in the TMS. Ideally, routing is directed to specific Terminologists based on certain conditions, for instance, by subject field, language, product area, or some other value. This requires the submitter to specify that information in the original proposal.

3. The Terminologist creates the entry in the termbase (or updates it, if it was created by the proposer), adds information, and changes the **Process status** to *provisionally processed* (if it requires more work) or *finalized*.
4. The Terminologist routes *provisionally processed* terms to another Terminologist who can finalize the entry. Routing should be done via a workflow in the TMS that allows the Terminologist to select the next person.
5. This second Terminologist adds the necessary information and sets the status to *finalized*.

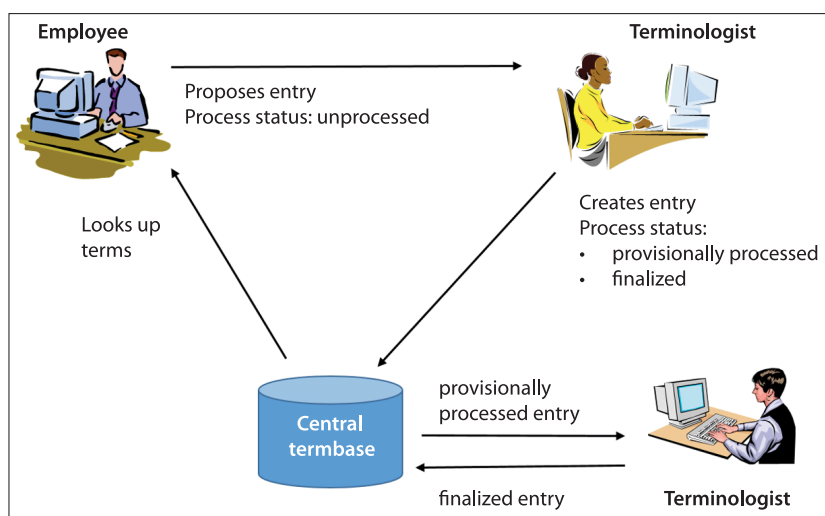


Figure 33. Basic process-oriented workflow

A usage-oriented workflow involves a review and approval process that focuses on the usage rating of an entry or a term. It is used to rank terms based on preferred usage and it involves users who are tasked with approving terms. For this purpose, *TBX-Basic*⁸⁰ includes a picklist field called **Usage status** with the values *preferred*, *admitted*, *not recommended*, and *obsolete*. The *obsolete* value is not commonly used in commercial termbases, where any undesirable term is often assigned the value *not recommended*. There is usually no need to have a separate value for terms that are not recommended, specifically because they are obsolete,

80. Available from TerminOrgs: www.terminorgs.net

as opposed to other reasons. Another concern is that a value *not recommended* could be perceived ambiguously: does it suggest that the term is not permitted at all, or simply that it is not “recommended”? Some company termbases include the value *prohibited* either as a replacement for *not recommended* (to remove any ambiguity), or in addition to it. These values are essential for CA applications and should be set up according to the functions of the CA application that require them. This is why, in *Controlled authoring* and *Data category proposal*, we have suggested the values *preferred*, *admitted*, *restricted*, and *prohibited*.

The usage-oriented workflow is only (but not always) applied on entries that have more than one term in a given language. In such cases there may or may not be a need to decide which of the two or more terms, all of which have the same meaning, should be “preferred.” This ranking of synonymous terms is carried out when there is a desire to enforce consistency of term usage. This is important for some concepts, such as the name of a product feature, but less so for others, such as certain general lexicon expressions. Almost all entries that are used in CA are synsets where each term has a usage status value.

The names of data categories used in the previous paragraphs can be replaced by others if desired, but the fundamental meaning and way that they are used should align with these descriptions. For example, *provisionally processed* seems awkward and unduly long; an acceptable alternative could be *in progress* or even *draft*. However, ensure that the words chosen are intuitively understood. One problem that has occurred in the past is using the value *approved* for either the process-oriented or the usage-oriented workflows because it is frequently misinterpreted. Some termbases erroneously included *approved* as a value for **Process status** to indicate that the entry has been checked for accuracy and completeness, and therefore “approved.” The real meaning, however, is *finalized*. Users can mistaken “approved” to mean that the terms in the entry are approved for usage, when in fact the contrary may be true (a finalized entry can include terms that are not recommended). Likewise, *approved* as a value for **Usage status** is not as clear as *preferred*.

Launch the termbase

In this chapter we discuss how to populate the termbase with initial data and officially launch the system. Prior to and during the launch, keeping the communication lines with stakeholders open is important to ensure their continued support.

Initial population

One of the goals of any terminology program is to increase the number of valuable terms in the termbase as quickly as possible. In commercial settings, there is less of a need to “vet” or check the information before it appears in the termbase compared to more prescriptive termbases operated by governments and other public institutions. Termbases in companies are viewed as work in progress and there is a higher tolerance for errors and omissions which are viewed as temporary issues to be fixed eventually.

An efficient means of quickly adding terms is to collect terminology from across the company and prepare it in the form of import files. Importing terms is faster than adding them one at a time. **Always make a backup of the termbase (even if it is empty) before importing any data.** In *Interchange*, it was mentioned that Excel files are often used as an interchange format. In the same manner, they can be used as an import format provided that the content to be imported is not excessively complex (with many data categories, many languages, etc.), in which case TBX is preferred.

During the terminology audit, if you have collected terminology resources in various file formats, such as MSWord, it will be necessary to convert those files to plain text format with a separator (delimiter) character, and from that format you can generate an Excel file. This is often not straightforward at all and in some cases the data almost needs to be retyped (or copy/pasted) completely, in which case you can just work with Excel directly from the start. But if you do work with plain text, for the *separator* character, use either a tab character or an unusual character such as the tilde (~) or vertical bar (|). The separator character is needed to separate the information so that it will align correctly in distinct columns in the spreadsheet. Using a comma as a separator is not recommended

since commas occurring inside text-based fields such as definitions will also be interpreted as separators, producing an unusable spreadsheet.

Most spreadsheet programs allow the separator character to be specified when the file is opened so that the columns can be aligned properly. Note that if a file has the .txt extension it may be necessary to change it to .csv. This way, the spreadsheet program knows that it is opening a character delimited file.

Figure 34 shows a plain text file (named with a .csv extension as just mentioned) with the vertical bar separator.

```
subject field|definition|en term|part of speech|en term|part of speech|fr term|part of speech
banking|an account for businesses|current account|noun|business account|noun|compte commercial|noun
environment||greenhouse gas|noun||gaz à effet de serre|noun
```

Figure 34. CSV file

The first row corresponds to the header row of a spreadsheet; it indicates the column headings which correspond to the fields in the termbase. The actual names in the header row will vary from one TMS to another depending on the data model, the internal (system recognized) names of fields, and how the import and export systems function. In order to determine the correct values to put in the header row, a useful technique is to perform an export of the termbase to Excel and examine the header row that is produced by the export routine. The concept level fields should occur first, followed by language level and term level.

The next row is the terminological entry for *current account* which includes a subject field, definition, the term, part of speech, a synonym (*business account*), part of speech, the French equivalent *compte commercial*, and part of speech. The third row is for the term *greenhouse gas*.

Note that each row in the above example has the same number of separator characters (7). The third row, for example, does not include a definition or a second English term but the separators required for that data are inserted.

When this file is opened in Open Office, a dialog appears allowing the user to specify that the separator is the vertical bar (Figure 35), which produces a properly-formatted spreadsheet.

Import

Character set: Unicode (UTF-8)

Language: Default - English (Canada)

From row: 1

Separator options

☐ Fixed width

☒ Separated by

☐ Tab ☐ Comma ☒ Other

☐ Semicolon ☐ Space

☐ Merge delimiters

Text delimiter: "

Other options

☐ Quoted field as text

☐ Detect special numbers

Fields

Column type: [dropdown]

	Standard	Standard	Standard	Standard
1	subject field	definition	en term	part of speech
2	banking	an account for businesses	current account	noun
3	environment	greenhouse gas	greenhouse gas	noun

Figure 35. Specifying the separator

	A	B	C	D	E	F	G	H
1	subject field	definition	en term	part of speech	en term	part of speech	fr term	part of speech
2	banking	an account for businesses	current account	noun	business account	noun	compte commercial	noun
3	environment		greenhouse gas	noun			gaz à effet de serre	noun
4								

Figure 36. CSV file after import

It was stated in *Interchange* that the format of an import or export file needs to be sufficiently robust to handle all the types of information in the termbase. Spreadsheet formats often are not powerful enough. Furthermore, they are awkward for handling entries that include synonyms, data in many fields, or multiple languages. With so many columns required to represent such data, the user is constantly scrolling horizontally, making the spreadsheet very difficult to work with.

As a format for representing terminological data in a file, XML is more robust and stable than spreadsheets. For large termbases that have the aforementioned properties (many fields, many languages, etc.), XML is a more suitable export and import format. *TermBase eXchange (TBX)* is the XML markup language specifically designed for terminological data (ISO 30042: 2019). Figure 37 shows the first entry.

A major challenge is how to prepare files for import that lack internal structure. For example, a glossary in MSWord format may contain paragraphs for terms and related information including definitions, comments, usage notes,

```

<conceptEntry>
<descrip type="subjectField">banking</descrip>
<descrip type="definition">an account for businesses</descrip>
<langSec xml:lang="en-US">
  <termSec>
    <term>current account</term>
    <termNote type="partOfSpeech">noun</termNote>
  </termSec>
  <termSec>
    <term>business account</term>
    <termNote type="partOfSpeech">noun</termNote>
  </termSec>
</langSec>
<langSec xml:lang="fr-FR">
  <termSec>
    <term>compte commercial</term>
    <termNote type="partOfSpeech">noun</termNote>
  </termSec>
</langSec>
</conceptEntry>

```

Figure 37. Entry in TBX format

synonyms, and equivalents in other languages, all together without any consistent formatting to separate the unique information types.

agrobiodiversity = agricultural biodiversity

biodiversité du/des paysage(s) agricole(s), valorisation agricole de la biodiversité, diversité des espèces élevées/cultivées, biodiversité agricole, biodiversité des cultures, agrobiodiversité
 agrobiodiversidad, biodiversidad agrícola

One of the program priorities (q.v.) established by the Conference of the Parties to the Convention on Biological Diversity. Also an operational program (q.v.).

Figure 38. An unstructured glossary entry

Various formatting conventions are used in Figure 38, such as the equal sign (=) indicating English synonyms, italics for French terms with a comma separating each one, unmarked font for Spanish terms, a definition following the Spanish terms, and the lexical marker “Also an” suggesting another English synonym or perhaps a hypernym. Indeed, this example is actually more structured, through formatting, than most glossaries. Prior to importing this information into a termbase, to respect the principle of data elementarity, each distinct piece of information needs to be physically separated into its own structure or container (just as they appear in separate cells in a spreadsheet). If stylistic conventions such as italics or symbols are used consistently throughout the glossary it may be possible to utilize them with advanced search/replace functions to reorganize the information in a tabular layout. This could eventually be converted to or copy/pasted into

a spreadsheet. However, usually the stylistic conventions are not used consistently, and a lot of manual effort is required to structure the information properly.

Depending on the structural properties of the original file a decision is made to restructure it into either a tabular or character delimited format, which can then be opened in a spreadsheet application, or into TBX. As stated earlier, if the glossary contains many languages, many different types of information, or many synonyms, TBX is recommended.

When the glossary is available in either the spreadsheet or TBX format it can be imported, assuming these file formats are supported for import. However, at this point it is probably too early to do so. The file likely needs further enhancements to improve the content and add missing information. For example, it is highly recommended, as per the TBX-Basic standard, that a part of speech value be included for every term. Most terms are nouns, so the noun part of speech value can be added in batch to all terms and then the exceptions (verbs, adjectives, etc.) manually reset. Terms should be checked for correct spelling and case. Poor definitions should be improved.

It is important to ensure that there are no cells (or <term> elements in TBX) that contain two terms, such as a full form and an acronym, which is a common style used in glossaries:

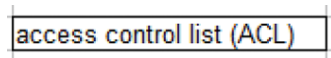


Figure 39. Cell containing two terms

Making modifications to a spreadsheet in batch requires only basic spreadsheet manipulation skills. Making modifications to a TBX file in batch requires more advanced text manipulation skills and the use of a text editor that supports regular expressions such as Notepad++ or UltraEdit. Another useful tool for manipulating TBX is an XML editor, which in addition to regular expressions offers the ability to make even more complex manipulations such as with XPath expressions. In our estimation, these text manipulation skills are among the most useful that a terminologist can have.

Beta test

Prior to its official launch the termbase needs to be beta-tested. Beta testing refers to having all the system's functions tested by individuals who were not involved in its development. Testers will need to be selected for the various role-based user groups and accounts set up accordingly.

Each tester will require precise step-by-step instructions describing how to complete the test; these are referred to as *test cases*. Different test cases need to be developed for different usage scenarios and also to test the various access levels. All the functions of the system need to be tested, for example:

- Create a profile, set and change password, etc.
- Browse through the list of terms
- Search for a term (exact and with wildcards)
- Create a new entry with one term
- Add a term to an existing entry
- Delete a term from an entry (or mark a term for deletion)
- Delete an entry (by users with sufficient access rights)
- Prevent the deletion of an entry (by users with insufficient access rights)
- Add, modify and delete concept level information in an existing entry
- Add, modify and delete language level information in an existing entry
- Add, modify and delete term level information in an existing entry
- Create, apply and save a search filter
- Customize the view
- Create and test workflows.

Any other available functions not listed above also need testing. For example, run a batch search, add a comment to an entry, print an entry, import/export (by those with sufficient rights), etc. Any problems found during the test must be fixed, and the function retested to verify the fix, before the launch.

Launch

The official launch of the termbase and any other major deliverable should be a well-publicized event. Make use of the various internal communication channels which may be available for publicizing major company projects and developments.

The terminology initiative itself should have its own web portal to serve as an information hub and a gateway for accessing the termbase and other resources. This should be available simultaneously with the official launch of the program and its components. Ensure that there is a feedback mechanism, such as a discussion board, in the web portal.

In addition to the web portal, prior to officially launching the termbase the following should also be completed:

- Beta test
- Online help and documentation
- Training programs available
- Terminology process documentation.

After the official launch and first announcements the terminology initiative will need to be promoted again periodically. This is necessary to reach employees who were recently hired or who may have missed previous announcements.

Documentation and training

If the termbase and the TMS are ergonomically designed with appropriate user-based views accessed from a web browser most users should not need any training. Users who have more advanced responsibilities such as reviewing and approving terms should attend a short demonstration of the tasks and associated workflows. The demonstration should be recorded, and the recording made available for later viewing.

All the components of the terminology program should be fully documented, and the documents made available on the program's web portal. A brief termbase user's guide should also be available. It should include topics such as:

1. Accessing the termbase
2. Views
3. Fields in the term entry
4. Search filters
5. Submitting terms
6. Providing feedback.

The user's guide can also include more advanced topics such as inclusion criteria and workflows. However, avoid providing too much information as this can trigger negative impressions. It is not essential, for instance, that users memorize a long list of criteria that their submissions of new terms must meet in order to be included in the termbase. Complicated "acceptance" criteria will discourage their feedback. The terminologist and other reviewers can take responsibility for ensuring that new submissions meet those criteria as part of the normal vetting process.

Community outreach

It is important to maintain open communication lines with all stakeholders to help ensure their continued support and engagement.

There should be a simple yet efficient mechanism to collect feedback about the termbase and other components of the terminology program on a continuous basis. Preferably this mechanism should be more sophisticated than simply providing the terminologist's email address.

For collecting feedback about terms, the best method is to have a commenting feature that is built into the TMS. This could be, therefore, one of the required features when conducting the competitive evaluation. It should be possible for users to see all comments. It should also be possible for the terminologist and others who have manager access to search for comments based on search filter criteria such as dates, submitters, etc. Consider the option of categorizing comments with labels such as new, in progress, rejected, duplicate, and resolved.

A discussion board on the terminology program's web portal is a good way to collect feedback on topics relating to the overall program.

For staff who are accessing terminology from within applications other than the TMS directly, such as from a CA tool or a CAT tool, it should be possible to submit feedback while working within that tool. Avoid requiring users to open an application outside of their working environment.

Some companies operate in fields that include very specialized terminology. In this case it will be necessary for the terminologist and other people involved in managing terminology (editors, writers, translators, etc.) to occasionally consult with subject matter experts, allowing them to verify the appropriateness of a term or to clarify its meaning.

Thus, it would be proactive to identify experts with knowledge in the various specialized areas that are needed and obtain their agreement to act as advisors. These individuals are not likely to be interested in playing an active role in the terminology management process. Rather, their involvement should be limited to providing expert advice when called upon.

Regular reporting to management is essential for maintaining executive support. The corporate terminologist should develop a series of measurable KPI (key performance indicators) and track their change on a regular basis (quarterly, annually, etc.). It should be possible to obtain the data by using search filters. Examples of KPI include:

- Number of concept entries in the termbase
- Number of terms in the termbase, per language
- Number of concept entries containing synsets, per language

- Number of comments, in total and by status (new, in progress, rejected, duplicate, resolved)
- Number of terms that have been reviewed and approved
- Number of active users.

The amount and types of feedback from the user base is also valuable information to report to management.

KPI can also be used to demonstrate areas that are under-developed and need additional resources, such as when there is an insufficient number of terms in the termbase in certain languages or when a high-profile product area has a disproportionately low number of terms. Use graphs to show these KPI whenever possible.

Examples where the terminology program solved a problem or issue should be collected and the best ones relayed to management. It is especially valuable to showcase extended applications, beyond translation, where terminology from the termbase is being used. Any such extended application should also have its own tracked KPI. These stories can also be featured on the terminology web portal to raise awareness and support.

Expand the termbase

In this chapter we explain various ways to expand and improve the termbase after its initial launch. The importance of using corpora to discover and validate terms is emphasized.

Term extraction

Automatic term extraction (ATE), also known as *term harvesting*, *term mining*, *term recognition*, *glossary extraction*, *term identification* and *term acquisition* (Heylen and de Hertog 2015), refers to the process of identifying the key terms in a set of documents. It requires a software program (term extraction tool) and a terminologist to run the program and refine the output.

What is considered to be a key term depends on how the list of extracted terms will be used, which is discussed in the next section. Generally speaking, key terms are often words that express important concepts, i.e. they reflect the topic area of the text. For instance, in an automobile user manual the names of the car's parts, functions, and features, but also general driving and operating expressions are important terms. On the other hand, on the car's website the colorful language crafted to influence potential buyers will contain other interesting but possibly less technical terms.

One can extract terms manually by reading the document and highlighting the important ones. However often the text is too large for manual extraction to be feasible. (The information set for most products comprises hundreds if not thousands of individual files.) In this case, we need to use a *term extraction tool*.

Why would we want to extract terms?

Term extraction is useful for a number of reasons. The first supports the entire terminology management program. It enables terms to be identified fairly quickly across the entire corporate corpus, or a targeted subset of it, and then imported to the termbase. It is an efficient way to increase the size of the termbase with terms that represent the company's collective communications.

The second supports individual translation projects. It is a recognized best practice to determine TL equivalents of key terms found in a text (or more often, a

collection of texts) that is scheduled to be translated before the actual translation starts and make those TL terms available to the translators working on the project. Term extraction supports this process. First, terms are extracted from the source texts. The extracted terms are then researched and TL equivalents are determined (in a controlled way where quality is guaranteed, such as by an experienced translator who is familiar with the topic area). Then, the list of terms (which is now a bilingual glossary) can be provided to the translators who will be translating the text. If the translators use a CAT tool, the bilingual glossary can be loaded into the CAT tool where the terms are automatically shown to the translators. This avoids having to rely on translators to look up terms. This process guarantees that the translated versions of a text will meet expected standards, at least with respect to the terminology, even when translators with different qualifications or backgrounds are involved.

This approach is particularly useful when multiple translators contribute to a translation project, which is often the case in large enterprises. Providing translators with pre-determined TL terms ensures that the terminology will be consistent throughout. Failing to do so means that translators have no guidance on how to render the important terms and their choices will vary, leading to inconsistencies in the translated text.

Companies that operate on a multinational or a global scale have to translate the information about their products and services into multiple languages. Usually the information is spread in multiple files. The information needs to be translated quickly so that the products can be released to market as closely to (and preferably at the same time as) the release of the SL version. This objective that is referred to as *simultaneous shipment* or *simship*. Typically, a translation company is engaged and to meet tight deadlines the job is divided into parts and given to different translators. Extracting the key terms and determining their TL equivalents is an effective strategy for ensuring that the quality of the final translation will be acceptable. The process also helps to shorten revision time and reduces the number of errors that need to be corrected.

Another use of ATE is to identify the key words in a document for building an index or a search engine lexicon. The key words can also be used to tag a document as to its main content such as for a content management system or an automatic content categorization tool. Some websites invite viewers to tag the content being viewed as a means of crowd-based content classification. One can imagine a term extraction tool running in the background and offering the viewer a list of potential candidates for the tag categories. Selecting from a list of predetermined choices would make the tagging much more effective by reducing variation among responses. The technology necessary to do this is available today, it is only a matter of implementation. These more advanced information technologies are

gaining momentum in commercial settings to help deal with the exploding volumes of information in electronic form.

The output of a term extraction tool is a list of *term candidates*, called so because some of the items in the list are not terms (i.e. they do not meet the termhood criteria set for the company). After a terminologist has gone through the list and has removed unwanted items, what is left are now *terms* because they are deemed to be “interesting” for downstream stages such as addition to the termbase.

Term extraction tools

A *term extraction tool*, or *term extractor*, is a software program that scans a text and outputs a list of term candidates that were found in that text. There are a number of commercial term extraction tools on the market. There are also a number of tools that have been developed in research settings such as universities,⁸¹ but these tools are generally not meant for production purposes (support services may be unreliable or unavailable, there may be disclaimers and no guarantees, and so forth). Furthermore, due to IT security controls, many companies do not allow the use of experimental or non-commercial software.

Most term extraction tools extract term candidates in one language from files in that language. There are also tools that can extract terms from two *parallel texts* (sometimes called *bitexts*) in two languages. Mostly this works as follows: The two files are first segmented and aligned sentence by sentence. If the parallel texts are from a TM then they are already segmented and aligned, albeit sometimes improperly.⁸² Working on one paired segment at a time, they extract SL terms from the SL text then compare the SL segment with the corresponding TL segment to identify possible TL equivalents. This bilingual term extraction is also called *term alignment* (Heylen and De Hertog 2015). The results are not reliable, and consequently a translator will have to review the output and make corrections. Depending on the quality of the raw output, validating the output may take more or less effort than it does to determine equivalent TL terms manually. Also, when translating terms manually the translator should be checking the company’s TM anyways to see how the terms might have been rendered in the past. This suggests that companies with large repositories of translation memories might benefit from a bilingual term extraction process that is run on those memories even if all the output needs reviewing. The corporate terminologist needs to weigh the two

81. For example, Termostat from the University of Montreal: termostat.ling.umontreal.ca/

82. This is an example that demonstrates the importance of high-quality segmentation in translation memories.

options and determine which is most cost effective and produces the most reliable bilingual terminology.

Considering the technical approach, term extraction tools fall into one of three categories: statistical, rules-based or hybrid. Statistical tools are the least effective; some only extract single words (*unigrams*). As we have seen in *Termhood and unithood*, there is evidence that many important terms are comprised of more than one word. Therefore it is essential that the term extraction tool be capable of extracting multiword terms. Tools that use a rules-based (grammatical) approach tend to produce better output than statistical ones. With this approach, the part of speech (noun, verb, etc.) of the words in the text is considered. This allows terms that follow certain patterns to be given priority, such as:

- noun, e.g. *laptop*
- noun + noun, e.g. *laptop computer*
- adjective + noun, e.g. *smart phone*
- adjective + noun + noun, e.g. *incandescent light bulb*

Patterns

Common_Noun= 440 (33 %)
 Common_Noun Common_Noun= 361 (27 %)
 Adjective Common_Noun= 336 (26 %)
 Adjective Common_Noun Common_Noun= 60 (5 %)
 Common_Noun Common_Noun Common_Noun= 52 (4 %)
 Adjective Adjective Common_Noun= 23 (2 %)
 Common_Noun Preposition Common_Noun= 11 (1 %)
 Common_Noun Preposition Adjective Common_Noun= 6 (0.5 %)
 Adjective Common_Noun Common_Noun Common_Noun= 5 (0.4 %)
 Common_Noun Adjective Common_Noun= 4 (0.3 %)
 Adjective Coord_Conjunction Adjective Common_Noun= 4 (0.3 %)

Figure 40. Partial list of the word patterns extracted by TermoStat from a sample corpus

Heylen and De Hertog (2015) refer to these patterns as *syntactic templates*.

Such an approach therefore requires more sophisticated algorithms involving part-of-speech tagging and syntactic parsing. Some tools also perform morphological stemming to convert inflected forms to their base form. The most advanced tools (and these are usually those developed in a research setting) also calculate the prevalence of a term candidate in the analyzed corpus and compare that figure to its prevalence in a secondary *reference corpus*. If the former figure is greater than the latter, then there is a good chance that the term candidate is domain-specific and therefore, a “term” of interest. Heylen and De Hertog (2015) refer to this as *contrastive term extraction*.

In addition to extracting term candidates, some tools can extract additional information, such as:

- One or more context sentences
- Name of the file or files where the term was found
- Part of speech.

Cleaning the term candidates

The list of term candidates contains unwanted items, i.e. term candidates which are ultimately rejected; these are called *noise*. The greater the amount of noise the less *precision* in the output (Bowker and Delsey 2016). Term extraction tools generally produce a lot of noise; it often comprises more than 60 percent of the output. Most tools allow you to reduce the noise by using an *exclusion list* (sometimes called a *stopword list*). But this does not solve the problem entirely. Most of the noise has to be removed manually through a process referred to as *cleaning*. Cleaning involves not only removing unwanted term candidates but also consolidating families of terms into their key members and adding new terms by resetting the boundaries of some multiword term candidates. If the effort to remove the noise exceeds the effort of identifying terms manually from the start, then the tool is not useful at all. Unfortunately, this is often the case.

When a terminologist tries a term extraction tool for the first time, the experience is often negative, and the process is soon abandoned. But the process can be effective if sufficient time, resources and patience are dedicated to finding a tool that performs reasonably well, in addition to learning how to use stopwords lists. Terminologists also get better and faster at cleaning the raw output over time.

Aside from the problem of excessive noise there is also the matter of *silence* to be concerned about. Silence refers to the important terms that were not extracted by the tool. The more valid terms that are missed, the less *recall*. All term extraction tools fail to identify some important terms.

To be effective, a term extraction tool must produce low levels of noise and silence (i.e. perform with high precision and recall). However, there will always be a degree of noise and silence, since term extraction tools are not perfect and never will be. The terminologist needs to modify and enhance the output before it can be used.

As described in *Termhood and unithood*, the notion of termhood (i.e. what makes a term candidate valuable enough to be selected for inclusion in the company's termbase) is different in commercial terminography compared to the conventional interpretation of termhood inherited from classical theory. Terminologists must keep this in mind when cleaning the output. They should establish a set of parameters or guidelines for cleaning that will result in terms being retained that meet and support the company's needs and objectives and align with the company's own definition of termhood. Of course the parameters

should also align with the term inclusion criteria that are established for the termbase (see *Inclusion criteria*).

Karsch (2015) describes a series of practical selection criteria. They are reproduced here, slightly reworded, with some brief explanations:

1. abbreviations, acronyms and their long forms
2. homographs
3. new or unfamiliar terms (e.g. *social distancing*, *app*)
4. terms that could be confusing or misinterpreted
5. terms that result from the process of terminologization – when a general word assumes a specialized meaning (e.g. *cloud* and *crowd*)
6. terms that result from the process of transdisciplinary borrowing – when a term from one discipline takes on a new meaning for another discipline (e.g. *bricks and mortar*)
7. terms that reflect a degree of specialization (domain specificity)
8. terms that occur frequently or are widely distributed
9. terms that are highly visible – on packaging, legal notices, user interfaces, etc.
10. terms that are members of a concept system – if the term obviously is part of a larger set of terms
11. terms that need standardization – presence of inconsistencies, undesired variants, etc.

Concordancing

A *concordance* is an alphabetical index of the principal words in a book or the works of an author with their immediate contexts (Merriam-Webster). It is also known as *key word in context* (KWIC). Figure 41 shows a concordance of the French term *changement climatique* (*climate change* in English) obtained from TermoStat,⁸³ a concordancing and term extraction tool developed by the Université de Montréal (Canada).

les plus susceptibles de faciliter l'atténuation du stratégiquement les plus prometteuses pour atténuer le rôle de la technologie dans l'atténuation du prévoit le transfert de technologies d'atténuation du les technologies d'adaptation aux effets négatifs du de TER utilisables dans le domaine du transfert de technologies dans le contexte du du FEM sur l'atténuation du risque de stratégiquement les plus prometteuses pour atténuer le	changement climatique changement climatique changement climatique changement climatique changement climatique changement climatique changement climatique changement climatique changement climatique	ou l'adaptation à ce changement et d'intéresser (ou s'y adapter) . étant plus clair et mieux compris , et de technologies d'adaptation , mais privilégient . , donnant ainsi un caractère vraiment mondial en nous appuyant sur les travaux du . (aussi s'y adapter) .
---	--	--

Figure 41. A concordance from TermoStat

83. termostat.ling.umontreal.ca

A concordancing tool or software, also called a *concordancer*, scans a corpus, extracts words and arranges them in alphabetical order showing each with a certain number of words or characters to the left and right of the word. Concordancers are used by linguists and researchers to study the vocabulary used by a particular socio-linguistic community or author,⁸⁴ compare different usages of the same word, determine the most frequent words in the corpus, find phrases and idioms, and create indices among other uses. Large scale corpora of language variants, such as American English and British English, are used by lexicographers to identify words for dictionaries.

Concordancers are very useful for finding multiword terms and compound words that are formed from a given single word. For instance, using TermStat again as above, a concordance of *changement* by itself would easily reveal that the word *climatique* often occurs to its right.

Concordancers are evolving rapidly due to advances in NLP research. Some incorporate a technology called part of speech tagging, which assigns a part of speech value to each word using a combination of techniques including looking the word up in internal dictionaries and analyzing the structure of the sentence. For example, the normal structure of a simple English sentence in the active voice is subject-verb-object (“John threw the ball”). Closed word classes such as articles and prepositions are more easily identified and offer clues about the words around them. An article is followed by a noun and a noun may be preceded by adjectives and other determiners (“The big bad wolf”). Concordancers that perform part of speech tagging can search for combinations of words based on grammatical patterns (such as adjective+adjective+noun). They are therefore effective for identifying multiword terms.

The most advanced concordancers offer a variety of additional functions which can be very useful. One such function involves comparing the frequency of occurrence of the words in the analyzed corpus (sometimes referred to as the *working corpus*) with the frequency of occurrence of the same words in a larger, general language corpus, which is referred to as a *reference corpus* (e.g. the British National Corpus, the Open American National Corpus, etc.). The frequencies are normalized, that is, the frequency count is adjusted taking into account the total word count of the corpus so that they can be compared (apples to apples, so to speak), and if the normalized frequency of a word in the analyzed corpus is significantly higher than its normalized frequency in the reference corpus, then the probability that the word has a domain-specific meaning or usage in the analyzed corpus is quite high. The word is statistically distinctive. In other words, a higher normalized frequency is a good indicator of termhood.

84. For example, the Shakespeare concordancer: opensource-shakespeare.org/concordance/

An example will help to demonstrate this principle. Figure 42 shows two lists of words that occur frequently in a corpus from a software company. The lists were produced by WordSmith Tools, which performs various text analysis functions in addition to concordancing. The list on the left shows high-ranking words after comparison with a reference corpus whereas the list on the right shows high-ranking words without comparison with a reference corpus (therefore, based on internal frequency alone). It is clear that the list on the left includes a high proportion of domain-specific unigram terms whereas the list on the right is much less interesting.

1	SYMANTEC
2	NORTON
3	BACKUP
4	SECURITY
5	YOUR
6	SERVER
7	WINDOWS
8	PROTECTION
9	CLICK
10	#
11	DATA
12	INTERNET
13	EXEC
14	PRODUCT
15	FILE
16	FILES
17	ANTIVIRUS
18	EMAIL
19	SOFTWARE
20	TM
21	INFORMATION
22	STORAGE
23	RECOVERY
24	ONLINE
25	COMPUTER
26	WEB
27	SYSTEM
28	NETWORK
29	MICROSOFT
30	ENTERPRISE
31	SUPPORT
32	PROGRAM
33	SERVERS

1	THE
2	#
3	AND
4	TO
5	OF
6	A
7	FOR
8	YOU
9	IN
10	SYMANTEC
11	YOUR
12	IS
13	OR
14	ON
15	THAT
16	WITH
17	NORTON
18	SECURITY
19	THIS
20	BACKUP
21	ARE
22	BE
23	FROM
24	SERVER
25	IT
26	WINDOWS
27	AS
28	CAN
29	DATA
30	IF
31	PROTECTION
32	BY

Figure 42. High ranking unigrams with and without comparison to a reference corpus

In WordSmith Tools, to distinguish between the output of these two different processes, the words on the left (which are ranked in comparison to a reference corpus) are referred to as *keywords* whereas the words on the right (which are ranked only according to internal frequency) are simply *words*.

In *Term extraction*, the problem of silence produced by term extraction tools was described. Silence refers to terms (that is, useful ones) that the term extraction tool did not identify and extract. Since many of the terms of interest for corporate terminography are bigrams and trigrams (see *Terms considered by length* and *Termhood and unithood*), one could assume that unidentified bigrams and trigrams make up a significant part of the silence. The use of concordancers can help find these missing terms. We therefore suggest a novel approach: after using a term extraction tool and cleaning the output to remove the noise use a concordancer to identify more terms, thereby reducing the silence.

The procedure⁸⁵ is as follows:

1. Using the concordancer's word list function, make a word list from the working corpus.
2. Make a word list from a reference corpus (WordSmith Tools includes a selection of different reference corpora).
3. Make a keyword list by using the two word lists as input.
4. Select some interesting domain-specific terms from the keyword list (at this point they are all unigrams) and record them as a list in a plain text file.
5. Run a batch concordance using the text file as input and the cleaned list of terms from the term extraction tool's output as an exclusion list (this will prevent terms you have already identified from appearing in the concordance).
6. Check the resulting concordance for interesting bigrams and trigrams.

Collocations also reveal words that tend to *collocate*, or in other words occur together. From the concordance shown in Figure 41, for example, the terminologist or translator can notice that the noun *atténuation* and its verb form *atténuer* are frequently used with *changement climatique* to convey *reduction* and *reduce* in English. Without this observation, the translator might be inclined to use the more direct translations *réduction* and *réduire*. Translations using those words would be understood but would not be as idiomatic as translations that adopt the frequent collocates.

To summarize, concordancers can complement term extraction tools by finding terms that are missed by the latter and by providing views of multiword terms in context to validate termhood.

85. The procedure described here reflects the WordSmith Tools concordancer, however, it should be possible to complete a similar process in other concordancers.

Target language terms

If your company uses CAT technology for its translations it is essential that the termbase be accessible to translators directly within the CAT tool. Translators must also be able to submit terms (both SL and TL) to the termbase directly from within the CAT tool while they are translating documents. Most CAT tools provide this functionality by allowing the submitted terms to be recorded in the terminology module of the CAT tool. However, since the terminology module that is part of the CAT tool is frequently inadequate for large-scale corporate terminology management (see *Standalone or integrated*), this module may not store the company's central termbase. If that is the case, then the terminology module in the CAT tool is likely used as a temporary location for bilingual terminology that the translator needs for the task at hand, but the company's central termbase is stored elsewhere in a more robust environment.

The corporate terminologist needs to ensure that these two systems (CAT terminology module and central termbase) are synchronized and that terminology can flow between them appropriately. This can be done through an import/export process or by developing a direct connector between the two systems. Both methods are bidirectional: terminology flows from the central termbase to the CAT module and from the CAT module back to the termbase.

The first method is only feasible if both systems support the TBX XML standard; spreadsheets will not likely support the range of data required. A terminologist with basic XML skills should be able to utilize the existing export/import functions in both systems to facilitate the transfer of data. The round trip (export from termbase, import to CAT tool, export from CAT tool, import to termbase) needs to be set up to occur at regular intervals. Search filters can be utilized to export only the data that the translator needs. (Remember that the screen space for viewing the terminology in the CAT tool is very small.) With some engineering support the process can even be automated.

The second method involves using an API (application programming interface) or another communications protocol and therefore will require more advanced computer programming skills beyond knowledge of XML. The suppliers of the CAT tool and/or the TMS may be able to provide the necessary programming support. Consider discussing these options when you are negotiating the purchase of these tools. The advantage of this method is that the two systems are synchronized in real time.

Finally, another method to obtain translated terms for the termbase involves mining TL terms from translation memories. This method is recommended to fill gaps in the termbase for a particular area of the company, provided that a TM or another parallel text (a SL text and its translated version) exists for that area.

For example, consider the following scenario. After a merger or acquisition a new product line is added to the company's offerings. The acquired company has a TM but no termbase. Using a bilingual term extraction tool, for each language pair, the terminologist can discover terms that are already used in the acquired company's documentation and add them to the termbase. If the company does not have a TM but can provide a parallel text the same procedure can be followed after the documents are aligned by using an aligning tool. Most CAT tools include alignment functions for this purpose. However, none are flawless, and some require a lot of manual adjustments to correct misalignments. Terminologists should be aware that there are technology suppliers that specialize in alignment tools. Their technology is sometimes superior to that of CAT tools where alignment is one feature among many. Mining terms from TMs provides terms for the termbase that help translators ensure that future translations are consistent with previous ones.

New concepts

As a scholarly discipline, terminology has devoted significant attention to developing principles and methods for creating new terms, which are known as *neologisms*.⁸⁶ New terms need to be created when a concept has been recently introduced to society and has not yet been named. This occurs in the case of new inventions, such as the term *smart phone*, which according to Google NGram viewer first appeared around the year 2000. Typically, a new term is first created in the language of the community where the concept or innovation originated, and then corresponding terms are created for other languages. The challenge in the original or primary language is to name the new concept appropriately, but the other, secondary languages have the additional challenge of avoiding unwanted linguistic influence from the primary language. Terms that are borrowed directly into a language are often unwelcome by those who desire to keep the language pure. The term *marketing* used in French is a famous example of a term that is disliked by many defenders of the French language, notably because it violates French morphology (the suffix "ing" does not exist in French). Another type of influence that is often perceived negatively is a *calque*, which is a direct translation. For example, the term *flea market* has been literally translated into *marché aux puces* for French and likewise into over a dozen other languages.

The desire to avoid over-contamination by a foreign language has led some linguistic communities to develop proactive approaches for dealing with

86. See Dubuc (1992) and Sager (1997), as well as the vast scholarly publication record of Jean-Claude Boulanger.

neologisms. Authorities in Canada, for instance, have implemented a language management policy to reduce the incidence of anglicisms in French. They have developed guidelines on how to create neologisms in the French language, they have expert linguists in charge of creating neologisms, and they disseminate the new terms through a broad public awareness campaign. The result is a high success rate in entrenching the new terms into general use. It has been anecdotally said that the French language in Canada contains fewer anglicisms than the French language in France. If that is true, this policy may be responsible. Moreover, both the Canadian federal government and the Quebec provincial government maintain extensive terminology databases which are widely used not only by translators but by the public at large.⁸⁷ The Quebec termbase proposes two alternate French terms to replace *marketing*: *commercialisation* and *mercatique*.

Indeed, in these secondary languages neologisms are needed not just to name new concepts or innovations but also to replace words and terms that have penetrated the language and that demonstrate characteristics of undesired foreign influence.

In a company, terminologists in fact are rarely called upon to create terms. More often, the employees involved in developing a new product, feature, or other new offering will name it themselves and by the time the terminologist becomes aware of a new term it may be too late to make any changes, with documentation and other records often already created. The terminologist should, however, have some level of oversight on the incidence of new terms so that several things can take place, as necessary. First, employees should be informed of the new term as quickly as possible to avoid alternate terms being adopted ad-hoc. Second, the terminologist should ensure that the new term adheres to some best practices that apply to the naming of new concepts, such as:

- the term should not have any negative connotations or potential negative connotations in any culture
- the term should be at the appropriate register (technical, formal, jargon, colloquial, etc.)
- the term's meaning should be easily apprehended from the term itself, which is a property referred to as *transparency*. For example, *mobile phone* is more transparent than *handset*.
- the term should not be tied to a specific culture. For example, *given name* and *family name* are more culturally neutral than *first name* and *last name*.

87. Termium Plus: btb.termiumplus.gc.ca, and the Grand Dictionnaire Terminologique (GDT): granddictionnaire.com/

A sound set of principles for creating new terms, including examples, is provided in ISO 704.

An example of the first type of problem, negative connotations, is the term *cheat sheet*. This term was adopted by a software manufacturer to name a type of interactive online help. Apparently, it does not have a negative connotation in American English where it signifies any kind of quick aid. However, in other English communities, such as Canada and Britain, the term retains its original meaning of a sheet of paper used for cheating on a test or exam. By the time the terminologist discovered that this term was appearing in the software product it was too late to change. When the product was sent for translation it was necessary to provide additional support to the translators to ensure that they could find TL equivalents that did not have the negative stigma associated with cheating. This is where raising awareness of the terminology process among all employees can bear fruit, as a more informed production team may have submitted the term to the terminologist for his or her input.

Creating new words to denote new concepts is rather rare (examples: *selfie*, *staycation*). It is much more common to use existing words. And since, as noted before, bigrams and trigrams are very common to denote concepts prevalent in commercial language, one can expect such compound nouns to be very productive for naming new concepts (a process referred to as *compounding*). An example is *smart TV*,⁸⁸ by analogy with *smartphone*, which is a television that has internet and computing capabilities.

Care must be taken to avoid adopting an existing term for a new concept when this could lead to misunderstanding, due to, for example, the two concepts being used in proximity (if they are used simultaneously in the same product or other set of related information) or the two concepts having some kind of conflicting nature. The earlier example of *smart* being used to convey the meaning of “computing and internet capable” is quite different from its use in *smart manufacturing* which, while it shares this meaning, also includes other properties such as high levels of adaptability, rapid design changes, and flexible workforce training. While it is not always possible, nor necessarily recommended, to avoid using these types of terms, one should ensure that consumers and other readers of company materials are informed as to their meaning. Everyone has experienced the frustration of coming across a term whose meaning is unclear and being unable to find an explanation anywhere. Acronyms without an accompanying full form are annoying. Documentation writers should never assume that readers know key terms, acronyms, abbreviations, and other important and/or cryptic terms.

88. TVs that do not have networking capabilities are now sometimes called *dumb TVs*.

Methods for naming new concepts include the following:

- Terminologization⁸⁹ – the use of general words for technical concepts. For example, *cookie* on websites, *ribbon* in software user interfaces, *crowd* in the context of social networking, and *cloud* referring to remote storage.
- Compounding – the use of two or more words to form a new term. For example, *sound bar* and *mountain bike*. Can involve truncation, such as *cannabusiness*.
- Transdisciplinary borrowing – the use of a term from another discipline. For example, *real estate* referring to the available space on a computer screen for an application, or *dashboard* referring to a collection of functions on a computer screen.
- Derivation – creating a new word from an existing one, often by adding a prefix or suffix. Example: *luddite*, used today to refer to people who are uncomfortable with technology, from Ned Ludd⁹⁰ and the suffix *ite* (denoting followers of a movement or doctrine).
- Conversion – when a word changes grammatical category (part of speech). For instance, *google* is now used as a verb meaning to search on the internet (even when using a search engine other than Google itself).
- Blending – forming a new term by joining parts of existing terms. Example: *Brexit* from Britain and Exit, *emoticon* from emotion and icon, and *malware* from malicious and software. This type of word formation is also known as a *portmanteau*.

89. Terminologization is the opposite phenomenon of what Meyer and Mackintosh (2000) describe as *de-terminologization*, where a technical term is adopted in general language often with a change in meaning.

90. A weaver who smashed knitting frames out of frustration in the late 18th century.

Maintain quality

In this chapter, we consider how to ensure the quality of the termbase. Quality is determined on the basis of the termbase's ability to meet its objectives and purpose. Does it contain the right terms? Are the fields used properly so that they deliver the required information? Is it repurposable?

The termbase-corpus gap

The company termbase has two missions which determine what terms it should include:

- Guide usage towards an *ideal* language (how people *should* write and translate) (prescriptive approach).
- Reflect current usage, in all its imperfections (how people *actually* write and translate) (descriptive approach).

The reason for the first mission is self-explanatory, but there are a number of reasons for the second.

The first mission cannot, in fact, be completed without the second. We need to have a good overall picture of the words, terms, and expressions that employees use, whether correct or not, before we can identify areas for improvement.

In practical terms translators remain the main users of the termbase. To deliver the productivity gains afforded by the autolookup function of CAT tools the termbase must include words, terms, and expressions that are used on a frequent basis in the company. It should include even those that do not reflect the *ideal* language, or that contravene the corporate style guide or word usage rules.

Other reasons for the second mission relate to repurposability. Potential uses such as SEO, indexing, automatic content classification, and so forth require as large a collection of terms and expressions found in the company as possible.

The two missions may appear conflicting at first glance. How can the termbase reflect current usage, which at times is incorrect, and at the same time guide writers towards correct language? The answer, as previously explained, is in the principle of concept orientation whereby multiple terms representing the same concept are organized in one concept entry. By marking one of those terms

as *preferred* the first mission is realized, and by including other terms in use for this concept so is the second.

Terms required for the first mission are obtained gradually over time as writers, editors, and translators note errors or inconsistencies and these issues are properly reflected in the termbase. Corporate style guides also often include lists of preferred and prohibited terms, which should be added to the termbase.

Terms required for the second mission are frequently under-represented in termbases. A research study carried out using four companies and their termbases revealed that in all cases there was a significant “gap” between the terms in the termbases and the terms used in the company (Warburton 2014). Two types of problems contribute to this gap: (1) the termbase contains terms that are not used in the company at all (or are used very infrequently), and (2) some terms that are frequently used by employees are missing in the termbase. We refer to the former type as *unoptimized* and the latter type as *undocumented*.

A large corpus-termbase gap (when there are many unoptimized and undocumented terms) undermines the terminology initiative. Our experience suggests that the corpus-termbase gap for corporate termbases in general is very large, and the reason for this is because terminologists working in commercial environments are generally not aware of the need for the termbase to align with the company’s corpus. As a result few adopt a corpus-based approach in their work.

A corpus-based approach to term identification enables termhood to be confirmed with corpus evidence. Every corporate terminologist should learn the fundamentals of corpus linguistics and become proficient in the use of corpus analysis software. By corpus analysis software,⁹¹ we refer to tools that perform the following types of functions:

- corpus management functions, such as crawling directory paths, file encoding, file conversions, markup recognition, etc.
- concordances (KWIC – key word in context), both for terms that are searched individually, and terms submitted in batch
- summary statistics of the concordances
- creation of word lists (frequency based)
- creation of keyword lists (saliency based, by comparison with a reference corpus).

See *Concordancing* for more information about how a concordancing tool can be used to identify terms from corpora.

91. For example, Wordsmith Tools. Concordancing functions are built into some CAT tools.

Unoptimized terms

Unoptimized terms are named as such because they do not contribute to the cost-effectiveness and the goals (increased employee productivity, improved quality, etc.) of the termbase. Research has shown that unoptimized terms in termbases are a major problem that impacts the return-on investment of termbases (Warburton 2014). Unoptimized terms exist in all the languages in the termbase, but their presence in the SL is the most problematic due to the ripple effect that is caused when they are translated (each translation of an unoptimized SL term is also unoptimized). Although this may sound harsh, these terms are useless. They are not supporting any company process or satisfying any need, and including them in the termbase is a waste of time and resources. Unoptimized terms take up space in the termbase and incur costs by adding to the burden of data management. Including terms in the termbase that do not further the goals of the terminology process (which in turn serves the goals of the company) reduces the value of the termbase and diverts resources away from more productive areas. The terminologist should also be concerned about the probability that these terms were selected, vetted, curated, *and translated* at the expense of other more important terms which were overlooked. Consider the wasted cost of translating, often to dozens of languages, a term that is not actually needed.

It has been statistically proven, for example, that a multiword term that includes a non-essential premodifier is less optimized (occurs less frequently in the corpus) than its counterpart with the non-essential modifier removed (Warburton 2014).

If a term in the termbase does not occur in the organization's corpus then it is unlikely that end-users will need to enquire about it. Likewise, if the term occurs rarely in the corpus then it will probably be rarely queried in the termbase as well. Below a certain threshold of use it becomes economically unjustified to include a term in the termbase when users could probably find the information they need elsewhere, such as by conducting an internet or intranet search. This type of unfocused search is not efficient if it is repeated many times, but it is justified if repeated infrequently. On the other hand, a termbase is cost-effective by reducing the time it takes for employees to find information that they require on a frequent basis. This is why frequency of occurrence is an important criterion of termhood for termbases that are developed for production-driven requirements.

Of course, frequency of occurrence is not the only valid termhood criterion; certain other criteria actually justify the inclusion of infrequent terms such as domain specificity, translation difficulty, or legal or marketing importance. When a term currently used in the company needs to be replaced by another term the latter must be added to the termbase even though it may not yet exist in the

corpus. This scenario is typical for CA. However, infrequent terms that have no special status and are present accidentally are unjustified and add undue costs.⁹² The termbase will be more effective if these redundant terms are replaced by more productive ones. But if the corporate terminologist uses a corpus-based research methodology from the outset, the number of unoptimized terms that end up in the termbase will be minimal.

Identifying unoptimized terms in the termbase involves performing a concordance of all the termbase terms in the company's corpus. This requires a tool that supports concordancing in batch (using an input file). The process entails exporting all the SL terms from the termbase into a plain text file (in a list, one term per line), and running a concordance of that file against the company corpus. The summary statistics indicate the total number of times each term was found in the corpus. Terms that have a very low frequency, or even a frequency of zero (which are referred to as *nonextant* terms), fall into the unoptimized category.

There are, of course, exceptions, and therefore the terminologist should not simply remove all the nonextant and infrequent terms from the termbase without some additional consideration. A term could have a frequency of zero or a very low frequency for various reasons. For example, the term could designate a new concept (new product, service, function, etc.) and material that will contain this term has not yet been produced. The term could occur infrequently in the corpus because it is a non-standard variant of another term, and yet this fact alone justifies its inclusion in the termbase (with an appropriate usage value and note) to support extended applications. It is also possible that the corpus is incomplete. It is difficult, sometimes even impossible, to compile a corpus of the entire company's holdings. Missing files could affect the frequency count of some terms.

As stated earlier, the terminologist must also remember that corpus frequency is but one important criterion for termhood. Some low-frequency terms are still important, for example legal terms, regulatory terms, safety warnings, terms that present significant linguistic or cultural challenges for translators, and so forth.

Nevertheless, the list of infrequent and nonextant terms will reveal patterns that suggest reasons for their low frequency, such as compound nouns that are perhaps too long, terms with unnecessary premodifiers or postmodifiers, inflected forms, plurals, and terms that contain numbers or punctuation. Generally speaking, the more words a term contains the less frequently it occurs. Setting the boundaries of a multiword term properly in order to "optimize" its value in the termbase is difficult. Knowing, however, that a long term is likely to occur less frequently than a shorter one, the terminologist should critically examine long compound nouns to see if there is any advantage to breaking them down

92. For some cost scenarios, see Warburton 2014.

into smaller components. Table 21 shows some examples taken from a corporate termbase (Warburton 2014):

Table 21. Frequency of multiword terms before and after boundary adjustments

Infrequent termbase term	Corpus frequency	Adjusted term	Corpus frequency
exponential growth trend model	2	exponential growth	46
global worksheet variable	2	global worksheet	70
proof sheet error	0	proof sheet	221
absolute correlation coefficient	0	correlation coefficient	330

There are other types of minor modifications that can render an infrequent or nonextant term into an optimized one, such as singularizing a plural term or changing the case. The terminologist can verify this by making the modification and then running a concordance on the adjusted term to see how the frequency changes.

Table 22. Frequency of terms before and after adjustments

Infrequent termbase term	Corpus frequency	Adjusted term	Corpus frequency
ghost boot partition	0	Ghost boot partition ⁹³	168
software updates	4	software update	44
antispam	2	anti-spam	25

If the frequency of an adjusted term increases significantly, and if the adjusted term does not yet exist in the termbase, it is an important undocumented term. Replacing the unoptimized term (before adjustment) with the undocumented term (after adjustment) narrows the termbase-corpus gap.

Undocumented terms

Undocumented terms are terms that are needed in the termbase to support the termbase’s mission, but they are missing. They represent a lost opportunity to realize the tangible goals of the terminology process: increase productivity, save costs, and improve quality and customer satisfaction. Empirical research suggests

93. Here, *Ghost* is the name of a product, and should therefore be capitalized.

that the economic impacts of failing to document key terms are greater than those of documenting unoptimized terms (Warburton 2014).

A term extraction tool and cleaning process can reveal some undocumented terms, and this is most effective if the existing termbase terms are used as an exclusion list during processing (see *Term extraction*). However, some undocumented terms will not be found (contributing to the silence described earlier). In fact, the number of important terms that a term extraction tool fails to identify is typically quite high. Relying on a term extraction process alone is therefore not sufficient.

According to tests, another method that has shown promising results is to identify *salient* unigrams, or in other words, statistically prominent single-word terms. Then use them in a batch concordance search to find interesting multiword terms (Warburton 2014). Salient unigrams are referred to as *keywords*. “Keywords are words which are significantly more frequent in one corpus than in another” (Hunston 2002: 68). When used in a search context, as described here, Drouin refers to them as “specialized lexical pivots” (2003).

This approach is based on the assumption that multiword terms, particularly bigrams, are important. It is widely acknowledged in the literature that terminological units frequently comprise more than one word. Not surprisingly, termbases in general contain a high proportion of multiword terms (see *Terms considered by length* for more discussion about term length). A multiword term consists of a headword and modifiers. And so it would make sense that salient unigrams might be among those important headwords, that they might be the “building blocks” for multiword terms, and that searching for those salient unigrams would lead to the discovery of important bigrams and trigrams. A number of researchers adopted similar approaches with varying degrees of success (see Drouin 2003, Drouin et al 2005, Chung 2003, Kit and Liu 2008, Anick 2001, Bowker and Pearson 2002).

The procedure in WordSmith Tools⁹⁴ is as follows:

1. Using the word list function, create a word list from the company corpus.
2. Create a word list from a reference corpus.
3. Using the Keyword function, use the word lists to produce a keyword list.
4. Examine the top and bottom of the keyword list for salient unigrams (the most interesting salient unigrams are at the top, but some are also found at the bottom) and note them down.
5. For each selected keyword, do the following:

94. The process is similar to the procedure described in *Term extraction* for addressing the silence produced by term extraction tools, however there are some differences with respect to leveraging the existing termbase terms.

1. Conduct a search in the termbase for terms that contain this keyword and export the resulting list of terms to a plain text file (or just type the results into a text file if the list is not long).
2. Run a concordance on the keyword, using the list of termbase terms that contain this keyword as an exclusion list. (Thus, concordances containing the termbase terms will not be produced.)
3. Examine the results for interesting new multiword terms. (Focus on bigrams and trigrams that occur frequently.)

An example will help to demonstrate the process. In one company's corpus, using the Keyword function, the unigram *model* was identified as salient. The word was searched in the termbase, and 30 terms were found that contain *model*. They were exported to a plain text file. A concordance was run on *model* using the text file from the termbase as an exclusion list. The following interesting multiword terms were found containing *model* in the concordances:

- *regression model*
- *quadratic model*
- *response surface model*

These are all important terms that are undocumented. They should be added to the termbase.

Determining the optimal boundaries of multiword terms needs to be based on corpus evidence. All terms should be checked against the company's corpus, but this rule holds true particularly for multiword terms that contain three words or more. Any non-essential or incidental premodifier should raise a red flag. When added to a core term, a non-essential or incidental word produces a term that is rarely encountered in the corpus, when compared to the core term without that word. For example, *single exponential smoothing* occurs much less frequently in one corpus we examined than *exponential smoothing*. Including *single* severely inhibits the term's repurposability across applications such as CA and CAT. But also, as a numeric concept, *single* adds no unique semantic meaning that would pose translation difficulty. Moreover, including *exponential smoothing* alone in the termbase enables repurposability for various larger compounds: *single exponential smoothing*, *double exponential smoothing*, *exponential smoothing method*, *single exponential smoothing method*, and so forth.

The previous example was relatively straightforward given that *single* is readily recognized as a non-essential modifier. Setting term boundaries is not always so straightforward. A general guideline could be stated in this way: if a term is "productive" in forming other terms, include it in the termbase.

The frequency of occurrence of termbase terms in the company corpus is a major indicator of their value to the terminology initiative. From a business perspective, there is little justification for incurring the costs of managing (and therefore translating) a term that is rarely used, with of course some exceptions, such as when the concept has critical legal or safety ramifications, as noted earlier.

Field content

On a regular basis a quality assessment of the termbase should be performed to identify problem areas at the level of the termbase fields. This will enable the discovery, for example, of a pattern of incorrect field use by particular users, who can then be provided with additional training. Table 23 provides some examples.

Table 23. Common errors in termbases

Error	Example	Corrective action
Two terms in one Term field	[file transfer protocol (FTP)].	Create two terms in the entry: [file transfer protocol] and [FTP].
Extraneous characters in the Term field	Such as punctuation, parentheses, extra spaces.	Remove all extraneous characters.
Term is not in canonical form	Entering a verb in the present participle form or the past participle form rather than infinitive. For example, <i>roaming</i> (present participle) instead of <i>roam</i> (infinitive). Also, for verbs, do not include the copula ("to").	All terms must be in canonical form: infinitive for verbs, singular for nouns (unless the noun represents a plural concept).
Wrong part of speech	One reason why this occurs is because the source content is not written properly. For instance, an error message may read "The install failed," which is incorrect English, since <i>install</i> is a verb. The correct message would be "The installation failed." Giving <i>install</i> a noun part of speech value in the termbase, reflecting the incorrect usage, is also incorrect. (On the other hand,	Assign the part of speech value that one would find in a dictionary.

Table 23. (continued)

Error	Example	Corrective action
	including <i>install</i> as a prohibited noun in the entry for <i>installation</i> is useful.)	
Wrong term type	For instance, giving all terms in the termbase that are not an acronym or an abbreviation the value <i>full form</i> .	Follow the guidelines in <i>TBX-Basic</i> .
Poorly written definitions	Definitions that are excessively long, contain non-essential information, do not clearly define the concept, contain embedded definitions of other terms, and other problems.	Follow the guidelines in <i>ISO 704</i> . However, avoid spending an excessive amount of time and resources writing excellent definitions as this is not commonly justified in a corporate termbase. Focus on fixing definitions that do not accurately describe the concept or are misleading, which could lead to incorrect translations.
Context sentences that do not contain the term	Context sentences <u>must</u> contain the term. They must also be authentic, which means that they are not created by the terminologist but rather are found in existing documents, websites, or other sources	Find a context sentence that contains the term.
Information in wrong text fields	For instance, putting a note or a comment in the Definition field.	Ensure that all fields contain the content that they are intended for and only that content.
Text fields contain multiple types of information	For instance, the Definition field contains a definition and also the source of that definition.	Ensure that there are dedicated fields for each type of information and move text to those fields where necessary.
Field used for wrong purpose	For instance, putting a person's name (translator, reviewer, etc.) in the Note field.	Ensure that the content in each field aligns with the field's purpose. If a certain type of information appears frequently in the wrong field, perhaps it is necessary to create a new dedicated field for this purpose.
Term is in the wrong case	For example, terms incorrectly written with an initial capital letter. This often occurs in glossaries that were prepared in a word processing software when the autocorrect	Ensure that only proper nouns use upper case characters. All common nouns should be in lower case.

Table 23. (continued)

Error	Example	Corrective action
	function that automatically sets the first letter on a line to upper case has been enabled. All terms in these kinds of glossaries start with an uppercase letter. If they are imported to the termbase without any correction, the terms will be in the wrong case.	
Cross references and other terms in text fields and not in relational fields	For example, mentioning a related term in the Note field or the Definition field without creating a proper concept relation in a relational field.	In some TMS, you can make a link to related terms in text fields such as the Definition field. If the definition contains a term that has an entry in the termbase, it is useful to make this term into a hyperlink pointing directly to that entry. These are called <i>entailed terms</i> . While this method of indicating related terms is acceptable and very helpful for users, who can simply click the link to go to the target entry, it is important that this method not replace the use of dedicated relational fields which support hierarchical taxonomic structures. For more information about relational fields, see <i>Data categories</i> and <i>Relations</i> .

You can often use functions such as search wildcards, filters and views to find problem areas. For instance:

- Set a filter for verbs and verify that all terms returned by the filter are indeed verbs. Repeat for other parts of speech.
- For English, search for terms ending in “ing” and check if some are present participles that should be changed to the infinitive. Repeat for other problematic word patterns such as terms ending in “s” which might be plurals that should be changed to singular form and terms ending in “ed” which might be past participles. Sometimes past participles are acceptable as a canonical form provided that the part of speech is *adjective* to reflect the fact that they modify nouns.

- Create a view that includes only the **Term** and the **Context** fields. Verify that each context contains the term.
- Search for `[*(*]` to find all terms that contain a parenthesis character. Do the same for other extraneous characters.

Another efficient way to check field content is to use a text mining approach. Export the entire termbase to a file and use a global search function to view all the content of a particular field at once. This can be done on either a text file, XML file (such as TBX), or even a spreadsheet. Advanced text editors such as Notepad++ or UltraEdit are particularly effective for this purpose. They allow you to show all instances of a particular string, such as the TBX element `<termNote type="partOfspeech">verb</termNote>` so that, in this case, you can verify that all terms that have the verb part of speech value are indeed verbs. Alternatively, you can search for all definitions and quickly verify the content of that field, not only that the definition is acceptable but also that the field does not contain other types of information. You can do the same with spreadsheets as well by using the sorting capabilities to focus on specific fields and field values.

You can make the corrections in the termbase itself, but if there are a large number of corrections to be made often it is more efficient to make the changes directly in the exported file and then reimport the corrected content to either the existing termbase, using a synchronize option (on the concept ID), or into a new termbase (which has been created using the same data model). The choice depends on how complete an exported file is (some systems do not allow all data in all fields to be exported) and how sophisticated the synchronize options are (some systems do not merge imported data well into an existing termbase). More information about exporting and importing is provided in *Interchange* and *Initial population*. Always make a backup of the existing termbase before importing new content into it.

Backups

Any database should be backed up regularly and termbases are no exception. Ideally this should be automated. The backups should include the entire termbase content and the termbase data model (often the two require separate backup processes). Study the various backup options and file formats available in the TMS and use the most comprehensive and stable one. While the international exchange standard TBX is a reputable format, its implementation in some TMS is defective and therefore it should not be assumed to be the best option. The native export format in a TMS is often the most robust since it was developed specifically for

that TMS. Carefully study all export and backup options in order to identify their limitations and determine the best one.

Some TMS do not allow the content of all fields to be exported or the export may lose or change some data. For instance, relational fields are infamous for not being exportable and reimportable (yet, we hope this will be resolved at some point). Administrative information might change: dates on re-import may automatically update to the current date, and names of people who created entries might be updated to the name of the person doing a subsequent import. Therefore, a data export may not be a full backup. In this case investigate database backups using a database management system or file management option which might be external to the TMS itself. Nevertheless, even if an alternative method to data export is used for backups, perform data exports on a regular basis as an additional security measure.

Leveraging opportunities

The repurposing potential of structured terminological resources is a frequent theme in this publication. It is the corporate terminologist's responsibility to continuously seek new opportunities to leverage the termbase and to extend its scope and use. If CA is being considered, the terminologist should be a key contributor to that initiative. If the company wants to establish, improve, or localize its product taxonomy, the termbase is an excellent supporting resource. If employees are empowered to "tag" content with keywords, those keywords should be fed into the termbase to help it grow. The termbase is a multipurpose corporate knowledge repository.

New uses of the termbase often require changes to the termbase. Additional data categories may be required. The use of an existing data category may need to change (for instance, adding new values to a picklist field). One example that comes to mind is when the termbase is used in a CA application, where part of speech data is sometimes interpreted and used in slightly different ways. For example, in the termbase a verb in past participle form such as *encrypted* can be given the adjective part of speech value if it is frequently found as a noun modifier in the company corpus. However, a CA application, which relies more on linguistic rules, may perform incorrectly with this approach where the adjective part of speech value is reserved for words that are purely adjectives as indicated in a dictionary, such as *technical* and *noxious*. One linguist we consulted provided an interesting tip: true adjectives are words that you can put "very" in front

of, such as “very technical” and “very noxious” (in contrast “very encrypted”⁹⁵ and other similar past participle constructions such as “very downloaded” do not sound right). CA applications also sometimes have difficulties correctly interpreting homographs according to each part of speech value in the termbase. This is an area that requires a lot of testing in the CA application (see *Controlled authoring*.)

Sometimes, therefore, it may be necessary to even modify some of the existing content to adapt to new use cases. During mergers and acquisitions, for example, one product line may become subsumed into another and this may have ramifications to corresponding product line values in the termbase. Before making any widespread changes to the termbase, again, always ensure that you create a backup.

95. Something is either encrypted or not encrypted. It can't be "very" encrypted.

Conclusion and future prospects

The main motivation for managing terminology in a commercial setting is to reduce costs for content authoring and translation. To this we add that thought-leaders and policy makers in enterprises are interested in developing multipurpose terminological resources that can be leveraged in various extended NLP-based applications. They are seeing the writing on the wall. “The future of terminological resources is evidently related to their potential interoperability and exploitation in new applications and resources” (León-Araúz et al 2019: 223).

Bourigault and Slodzian point out that these new applications are primarily “textual” which means that terminological resources intended to serve them must also be corpus-based in order to reflect those texts. We like the simple way that they express this bi-directional dependency on corpora:

La terminologie doit venir des textes pour mieux y retourner (1999: 30). (Translation) Terminology must come from texts so that it can better return to texts.

In other words, terminological resources (i.e. termbases) must reflect terms in active use in order to enable productive reuse. In an article dealing with workplace applications Condamines also stated this very directly: “Terminology has to be drawn from texts written in the workplace” (2010: 45). This perspective contrasts with other environments such as public institutions where the motivation centers around language preservation or conceptual standardization. It shifts the focus of what terminology actually is away from semantic criteria towards authentic discourse, purpose-driven communication needs and in particular to degrees of repurposability. The more a term is used, the more it will be required in various applications and situations, and thus the more it should be recorded and managed in a structured digital format, so that the information necessary for these uses can be leveraged in various production-oriented NLP technologies. Linguistic and semantic properties of various sorts, while important, are secondary to this pragmatic criterion. This may be difficult for some to accept, but this is reality for the corporate terminologist.

In commercial terminography, application-oriented needs rather than strict semantic criteria should be the determining factor in defining termhood. And adopting a corpus-based approach to term identification will reduce the gap described in *The termbase-corpus gap*. The use of corpora for selecting terms would greatly increase the value and repurposability of commercial termbases.

We have suggested that among the theories of terminology, the GTT is the least relevant for addressing the needs of commercial applications of terminological resources. With commercial terminography, we see a departure from purely semantic criteria towards a model for term selection that is purpose-driven, that values repurposability above all, and is based on corpus relevancy. The various theoretical perspectives on the notion of *term* that evolved post GTT give greater importance to the communicative intent of interlocutors, to the application of the terminological resource, and to the role of corpora for providing empirical linguistic evidence. We find that these perspectives resonate for commercial terminography. Condamines has already claimed that textual terminology “constitutes an important part of linguistics of the workplace” (2010: 46). There is definitely a place for terminology management in the private sector, for *corporate terminography*, among the modern emerging theories of terminology.

The theoretical foundations of terminology need to adapt to modern applications; an application-oriented terminology theory and methodology is needed. A new paradigm for terminological resources needs to take shape, one that is less constrained by fixed semantic models and is sufficiently flexible to serve different linguistic contexts, communicative goals, and end users of terminological resources.

We propose that a methodological framework for commercial terminography would include the following elements:

- adopting more statistically-based criteria for term selection
- using the organization’s corpus as the primary source of terms
- using corpus analysis technologies such as concordancers, keyword identifiers, and collocate relationship calculators
- adopting a termbase data model that ensures that the terminological resource can be repurposed in a range of NLP applications.

Corporate terminologists are in an extraordinary position. They have fantastic opportunities for professional development, for engaging in innovation, and for being part of the digital evolution on the leading edge of language technology. These opportunities must be recognized and seized. A corporate terminologist needs to leverage terminology in extended applications, and prove the value of the termbase for supporting the company’s strategic objectives in all matters that involve language.

Commercial terminography is not terminography in the classical sense. Corporate terminologists are working in uncharted territory. The aim of this book is to raise awareness of the terminology discipline, as it is officially conceived, falling short of meeting modern demands. Corporate terminologists can shape the development of a new theory and methodology for commercial terminography. In fact, they even have a responsibility to do so. Hopefully this book has triggered some reflections in this direction.

Further reading and resources

The following works are recommended here because they focus on aspects related to terminology management. Works covering these topics more generally are not listed due to their potentially large number. A search in any university library catalog will provide ample suggestions for further reading on these topics.

General principles

- Dubuc, Robert. 1997. *Terminology: a practical approach*, Brossard, Québec: Linguattech éditeur inc.
- ISO Technical Committee 37: Language and Terminology. *ISO 704 – Terminology work – Principles and methods*. Geneva: International Organization for Standardization. The forthcoming version of this standard (after 2019) will contain a detailed typology of concept relations.
- Kockaert, Hendrik and Frieda Steurs(eds). 2015. *Handbook of Terminology*, V.1. Amsterdam: John Benjamins.
- Pavel, Silvia. *The Pavel Tutorial*. Originally developed by the Terminology Standardization Directorate, Translation Bureau, Public Works and Government Services Canada. Available online: linguistech.ca/pavel/
- Rondeau, Guy. 1981. *Introduction à la terminologie*, Montreal: Centre éducatif et culturel inc.
- Sager, Juan. 1990. *A Practical Course in Terminology Processing*, Amsterdam: John Benjamins.
- TerminOrgs. 2016. *Terminology Starter Guide*. Available from: terminorgs.net.
- Wright, Sue Ellen and Gerhard Budin(eds). 1997. *Handbook of Terminology Management*, V.1. Amsterdam: John Benjamins.
- Wright, Sue Ellen and Gerhard Budin(eds). 2001. *Handbook of Terminology Management*, V.2. Amsterdam: John Benjamins.

Terminology in commercial environments

- LISA Terminology SIG. 2001. *Terminology Management in the Localization Industry*. Localization Industry Standards Association.
- LISA Terminology SIG. 2003. *Terminology Management: A study of costs, data categories, tools, and organizational structure*. Localization Industry Standards Association.
- LISA Terminology SIG. 2005. *Terminology Management practices and trends*. Localization Industry Standards Association.

- Schmitz, Klaus-Dirk and Daniela Straub. 2010. *Successful terminology management in companies*. Stuttgart: TC and More GmbH.
- Warburton, Kara. 2008. "Terminology: A New Challenge for the Information Industry." *American Translators Association Journal*.
- Warburton, Kara. 2014. "Narrowing the Gap Between Termbases and Corpora in Commercial Environments." In *LREC Proceedings*, 2014. Reykjavik.
- Warburton, Kara. 2014. "Terminology as a Knowledge Asset." *MultiLingual*, June 2014, 48–51.
- Warburton, Kara. 2014. *Narrowing the gap between termbases and corpora in commercial environments*. PhD Thesis. Hong Kong: City University of Hong Kong. Available from: <http://termologic.com/resources/>
- Warburton, Kara. 2015. "Terminology Management." In *Routledge Encyclopedia of Translation Technology*, ed. by Chan Sin-wai, 644–661. Oxfordshire, UK: Routledge.
- Warburton, Kara. 2015. "Managing Terminology in Commercial Environments." In *Handbook of Terminology*, V.1., ed. by Hendrik J. Kockaert and Frieda Steurs, 361–392. Amsterdam: John Benjamins.
- Warburton, Kara. 2018. "Terminology Resources in Support of Global Communication." In *The Human Factor in Machine Translation*, ed. by ChanSin-wai. Routledge Studies in Translation Technology. Oxfordshire, UK: Routledge.

Terminology management systems

- Warburton, Kara and Arle Lommel. 2017. *Terminology Management Tools*. Common Sense Advisory. Burlington, MA: CSA Research.

Termbases

- ISO Technical Committee 37: Language and Terminology. 2017. *ISO 16642:2017 Computer applications in terminology – Terminological markup framework (TMF)*. Geneva: International Organization for Standardization.
- ISO Technical Committee 37: Language and Terminology. 2019. *ISO 26162-1 - Management of Terminology Resources – Terminology databases – Part 1: Design*. Geneva: International Organization for Standardization.
- ISO Technical Committee 37: Language and Terminology. 2019. *ISO 26162-2 - Management of Terminology Resources – Terminology databases – Part 2: Software*. Geneva: International Organization for Standardization. Note: publication of *ISO 26162-3 – Part 3: Content* is forthcoming as of this writing. It will provide guidance on the quality of termbase content.
- ISO Technical Committee 37: Language and Terminology. 2019. *ISO 30042:2019 Management of terminology resources – TermBase eXchange (TBX)*. Geneva: International Organization for Standardization.
- TerminOrgs. 2014. *TBX-Basic Specification*. Available from: terminorgs.net

Data categories

Information about data categories for languages resources has been collected in a centralized website, the Data Category Repository (DCR) at www.datcatinfo.net.

Controlled authoring

Warburton, Kara. 2014. "Developing Lexical Resources for Controlled Authoring Purposes." In *LREC Proceedings*. Reykjavik.

Search engine optimization

Warburton, Kara and Barbara Karsch. 2012. *Optimizing global content in Internet search*. Available from Research Gate.

Term extraction

Bernth, Arendse, Michael McCord, and Kara Warburton. 2003. "Terminology Extraction for Global Content Management." *Terminology*, 9(1): 51–69.

Karsch, Barbara. 2015. "Term extraction: 10,000 term candidates. Now what?" *ATA Chronicle*, Feb 2015: 19–21. American Translators Association.

Warburton, Kara. 2010. "Extracting, preparing, and evaluating terminology for large translation jobs." In *LREC proceedings*, Malta.

Warburton, Kara. 2013. "Processing terminology for the translation pipeline." *Terminology*, 19(1): 93–111.

Term variants

Daille, Béatrice. 2017. *Term Variation in Specialised Corpora. Characterisation, automatic discovery and applications*. Amsterdam: John Benjamins.

Drouin, Patrick, Aline Francœur, John Humbley, Aurélie Picton(eds). 2017. *Multiple Perspectives on Terminological Variation*. Amsterdam: John Benjamins.

Freixa, Judit. 2006. "Causes of denominative variation in terminology. A typology proposal." *Terminology*, 12(1): 51–77.

Workflows and project management

- Cerrella Bauer, Silvia. 2015. "Managing terminology projects." In *Handbook of Terminology*, V.1, ed. by Hendrik J. Kockaert and Frieda Steurs, 324–340. Amsterdam: John Benjamins.
- Dobrina, Claudia. 2015. "Getting to the core of a terminological project." In *Handbook of Terminology*, V.1, ed. by Hendrik J. Kockaert and Frieda Steurs, 180–199. Amsterdam: John Benjamins.
- Dunne, Keiran and Elena Dunne(eds). 2011. *Translation and Localization Project Management*. American Translators Association Scholarly Monograph Series, XVI. Amsterdam: John Benjamins.
- Karsch, Barbara. 2006. "Terminology workflow in the localization process." In *Perspectives on Localization*, ed. by Keiran Dunne, 173–191. American Translators Association Scholarly Monograph Series, XIII. Amsterdam: John Benjamins.
- Project Management Institute, Inc. 2017. *A Guide to the Project Management Body of Knowledge (Guide)*. Sixth edition.

Bibliography

- Ahmad, Khurshid. 2001. "The Role of Specialist Terminology in Artificial Intelligence and Knowledge Acquisition." In *Handbook of Terminology Management*, V.2, ed. by Sue Ellen Wright and Gerhard Budin. 809–844. Amsterdam: John Benjamins. <https://doi.org/10.1075/z.htm2.32ahm>
- Alcina, Amparo. 2009. "Teaching and Learning Terminology. New Strategies and Methods." *Terminology* 15(1): 1–9. <https://doi.org/10.1075/term.15.1.01alc>
- Allard, Marta Gómez Palou. 2012. "Managing Terminology for Translation Using Translation Environment Tools: Towards a Definition of Best Practices," PhD Thesis. Ottawa: University of Ottawa. Available from: <https://ruor.uottawa.ca/handle/10393/22837>
- Anick, Peter. 2001. "The Automatic Construction of Faceted Terminological Feedback for Interactive Document Retrieval." In *Recent Advances in Computational Terminology*, ed. by Didier Bourigault, Christian Jacquemin, Marie-Claude L'Homme. 29–52. Amsterdam: John Benjamins. <https://doi.org/10.1075/nlp.2.03ani>
- Bellert, Irena and Paul Weingartner. 1982. *Sublanguage. Studies of Language in Restricted Semantic Domains*, ed. by Richard Kittredge and John Lehrberger. Berlin: Walter de Gruyter.
- Bourigault, Didier and Monique Slodzian. 1999. "Pour une terminologie textuelle." *Terminologies nouvelles* 19: 29–32.
- Bourigault, Didier and Christian Jacquemin. 1999. "Term Extraction and Term Clustering: An integrated platform for computer-aided terminology." In *Proceedings of the ninth Conference on European Chapter of the Association for Computational Linguistics (EACL)*, 15–22. Stroudsburg, PA, USA: Association for Computational Linguistics. <https://doi.org/10.3115/977035.977039>
- Bourigault, Didier and Christian Jacquemin. 2000. "Construction de ressources terminologiques." In *Ingénierie des langues*, ed. by J.M. Pierrel. 215–234. Paris: Hermès.
- Bowker, Lynne and Jennifer Pearson. 2002. *Working with Specialized Language. A practical guide to using corpora*. London: Routledge. <https://doi.org/10.4324/9780203469255>
- Bowker, Lynne. 2002. "An Empirical Investigation of the Terminology Profession in Canada in the 21st century." *Terminology*, 8(2): 283–308. <https://doi.org/10.1075/term.8.2.06bow>
- Bowker, Lynne. 2003. "Specialized Lexicography and Specialized Dictionaries." In *A Practical Guide to Lexicography*, ed. by Piet van Sterkenburg. 154–164. Amsterdam: John Benjamins. <https://doi.org/10.1075/tlrp.6.18bow>
- Bowker, Lynne. 2015. "Terminology and Translation." In *Handbook of Terminology* V. 1, ed. by Hendrik J. Kockaert and Frieda Steurs. 304–323. Amsterdam: John Benjamins. <https://doi.org/10.1075/hot.1.16ter5>
- Bowker, Lynne and Tom Delsey. 2016. "Information Science, Terminology and Translation Studies – Adaptation, collaboration, integration." In *Border Crossings: Translation Studies and Other Disciplines*, ed. by Yves Gambier and Luc van Doorslaer. 73–96. Amsterdam: John Benjamins. <https://doi.org/10.1075/btl.126.04bow>

- Buchan, Ronald. 1993. "Quality Indexing with Computer-aided Lexicography." In *Terminology: Applications in Interdisciplinary Communication*, ed. by Helmi B. Sonneveld and Kurt L. Loening. 69–78. Amsterdam: John Benjamins. <https://doi.org/10.1075/z.70.06buc>
- Budin, Gerhard. 2001. "A Critical Evaluation of the State-of-the-art of Terminology Theory." *Terminology Science and Research: Journal of the International Institute for Terminology Research*, IITF, 12(1–2): 7–23.
- Cabré, Maria Teresa. 1995. "On Diversity and Terminology." *Terminology*, 2(1): 1–16. <https://doi.org/10.1075/term.2.1.02cab>
- Cabré, Maria Teresa. 1996. "Terminology Today." In *Terminology, LSP and Translation. Studies in Language Engineering in Honour of Juan C. Sager*, ed. by Harold Somers. 15–33. Amsterdam: John Benjamins. <https://doi.org/10.1075/btl.18.04cab>
- Cabré, Maria Teresa. 1999. *La terminología: Representación y comunicación. Elementos para una teoría de base comunicativa y otros artículos*. Barcelona: Institut Universitari de Lingüística Aplicada.
- Cabré, Maria Teresa. 1999b. *Terminology – Theory, Methods and Applications*. Amsterdam: John Benjamins. <https://doi.org/10.1075/tlrp.1>
- Cabré, Maria Teresa. 2000. "Elements for a Theory of Terminology: Towards an Alternative Paradigm." *Terminology*, 6(2): 35–57. <https://doi.org/10.1075/term.6.1.03cab>
- Cabré, Maria Teresa. 2003. "Theories of Terminology. Their Description, Prescription and Explanation." *Terminology*, 9(2): 163–199. <https://doi.org/10.1075/term.9.2.03cab>
- Cerrella Bauer, Silvia. 2015. "Managing Terminology Projects." In *Handbook of Terminology*, V. 1, ed. by Hendrik J. Kockaert and Frieda Steurs. 324–330. Amsterdam: John Benjamins. <https://doi.org/10.1075/hot.1.17man1>
- Champagne, Guy. 2004. *The Economic Value of Terminology. An Exploratory Study*. Ottawa: Translation Bureau of Canada.
- Childress, Mark. 2007. "Terminology work saves more than it costs." *MultiLingual*, April/May 2007, 43–46.
- Chung, Teresa Mihwa. 2003. "A Corpus Comparison Approach for Terminology Extraction." *Terminology*, 9(9): 221–245. <https://doi.org/10.1075/term.9.2.05chu>
- Collet, Tanja. 2004. "What's a term? An attempt to define the term within the theoretical framework of text linguistics." *Linguistica Antverpiensia, New Series, NS3 – The Translation of Domain Specific Languages and Multilingual Terminology Management*, V. 3: 99–111.
- Condamines, Anne. 1995. "Terminology: New Needs, New Perspectives." *Terminology*, 2(2): 219–238. <https://doi.org/10.1075/term.2.2.03con>
- Condamines, Anne. 2005. "Linguistique de corpus et terminologie." *Languages*, 157(1): 36–47. <https://doi.org/10.3917/lang.157.0036>
- Condamines, Anne. 2007. "Corpus et terminologie." In *La redocumentarisation du monde*, ed. by R. T. Pédaque. 131–147.
- Condamines, Anne. 2010. "Variations in Terminology. Application to the Management of Risks Related to Language use in the Workplace." *Terminology*, 16(1): 30–50. <https://doi.org/10.1075/term.16.1.02con>
- Corbolante, Licia and Ulrike Irmeler. 2001. "Software Terminology and Localization." In *Handbook of Terminology Management*, V. 2, ed. by Sue Ellen Wright and Gerhard Budin. 516–535. Amsterdam: John Benjamins. <https://doi.org/10.1075/z.htm2.14cor>
- Daille, Béatrice, Benoit Habert, Christian Jacquemin and Jean Royauté. 1996. "Empirical Observation of Term Variations and Principles for their Description." *Terminology*, 3(2): 197–257. <https://doi.org/10.1075/term.3.2.02dai>

- Daille, Béatrice. 2005. "Variations and Application-oriented Terminology Engineering." *Terminology*, 11(1): 181–197. <https://doi.org/10.1075/term.11.1.o8dai>
- Daille, Béatrice. 2007. "Variations and Application-oriented Terminology Engineering." In *Application-Driven Terminology Engineering*, ed. by Fidelia Ibekwe-SanJuan, Anne Condamines and M. Teresa Cabré Castellvi. 163–177. Amsterdam: John Benjamins. <https://doi.org/10.1075/bct.2.o9dai>
- De Saussure, Ferdinand. 1916. *Cours de linguistique générale*. Paris: Éditions Payot et Rivages. (Republished in 1995).
- Depecker, Loïc. 2015. "How to Build Terminology Science?" In *Handbook of Terminology*, V. 1, ed. by Hendrik J. Kockaert and Frieda Steurs. 34–44. Amsterdam: John Benjamins. <https://doi.org/10.1075/hot.1.o3how1>
- Dobrina, Claudia. 2015. "Getting to the Core of a Terminological Project." In *Handbook of Terminology*, V. 1, ed. by Hendrik J. Kockaert and Frieda Steurs. 180–199. Amsterdam: John Benjamins. <https://doi.org/10.1075/hot.1.1oet1>
- Drouin, Patrick. 2003. "Term Extraction using Non-technical Corpora as a Point of Leverage." *Terminology*, 9(1): 99–115. <https://doi.org/10.1075/term.9.1.o6dro>
- Drouin, Patrick, M. C. L'Homme and C. Lemay. 2005. "Two Methods for Extracting Specific Single-word Terms from Specialized Corpora: Experimentation and Evaluation." *International Journal of Corpus Linguistics*, 10(2): 227–255. <https://doi.org/10.1075/ijcl.10.2.o5lem>
- Dubuc, Robert. 1992. *Manuel pratique de terminologie*. Quebec: Linguattech.
- Dubuc, Robert. 1997. *Terminology: A Practical Approach*. Quebec: Linguattech.
- Dunne, Keiran. 2007. "Terminology: ignore it at your peril." *MultiLingual*, April/May 2007: 32–38.
- Faber, Pamela Benítez, Carlos Márquez Linares and Miguel Vega Expósito. 2005. "Framing Terminology: A Process-Oriented Approach." *Meta*, 50(4). <https://doi.org/10.7202/019916ar>
- Faber, Pamela. 2011. "The Dynamics of Specialized Knowledge Representation: Simulational Reconstruction or the Perception–action Interface." *Terminology*, 17(1): 9–29. <https://doi.org/10.1075/term.17.1.o2fab>
- Faber, Pamela and Clara Inés López Rodríguez. 2012. "Terminology and specialized language." In *A Cognitive Linguistics View of Terminology and Specialized Language*, ed. by Pamela Faber. 9–32. Berlin/New York: Mouton De Gruyter. <https://doi.org/10.1515/9783110277203>
- Fidura, Christie. 2013. *Terminology Matters*. White paper. SDL Inc. Available from: sdl.com/download/terminology-matters-whitepaper/76365/
- Freixa, Judit. 2006. "Causes of Denominative Variation in Terminology. A Typology Proposal." *Terminology*, 12(1): 51–77. <https://doi.org/10.1075/term.12.1.o4fre>
- Galinski, Christian. 1994. "Exchange of Standardized Terminologies within the Framework of the Standardized Terminology Exchange Network." In *Standardizing and Harmonizing Terminology: Theory and Practice*, ASTM STP 1223, ed. by Sue Ellen Wright and Richard A. Strehlow. 141–149. Philadelphia: American Society for Testing and Materials. <https://doi.org/10.1520/STP137525>
- Galisson, Robert and Daniel Coste. 1976. *Dictionnaire de didactique des langues*. Paris: Hachette.
- Greenwald, Susan. 1994. "A Construction Industry Terminology Database Developed for use with a Periodicals Index." In *Standardizing and Harmonizing Terminology: Theory and Practice*, ASTM STP 1223, ed. by Sue Ellen Wright and Richard A. Strehlow. 115–125. Philadelphia: American Society for Testing and Materials. <https://doi.org/10.1520/STP137505>

- Hanks, Patrick. 2013. *Lexical Analysis – Norms and Exploitations*. London: The MIT Press. <https://doi.org/10.7551/mitpress/9780262018579.001.0001>
- Heylen, Chris and Dirk de Hertog. 2015. “Automatic Term Extraction.” In *Handbook of Terminology*, V. 1, ed. by Hendrik J. Kockaert and Frieda Steurs. 203–221. Amsterdam: John Benjamins. <https://doi.org/10.1075/hot.1.11aut1>
- Hoffman, Lothar. 1979. “Towards a Theory of LSP. Elements of a Methodology of LSP Analysis.” *International Journal of Specialized Communication*, 1(2): 12–17.
- Hunston, Susan. 2002. *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press. <https://doi.org/10.1017/CBO9781139524773>
- Hurst, Sophie. 2009. “Wake up to terminology management.” *Communicator*, Spring 2009. Croydon: Quarterly journal of the Institute of Scientific and Technical Communicators.
- Ibekwe-SanJuan, Fidelia, Anne Condamines and M. T. Cabré Castellvi (eds). 2007. *Application-Driven Terminology Engineering*. Amsterdam: John Benjamins. <https://doi.org/10.1075/bct.2>
- ISO Technical Committee 37: Language and Terminology. 2000. *ISO 1087-1:2000 – Terminology work – Vocabulary – Part 1: Theory and application*. Geneva: International Organization for Standardization.
- ISO Technical Committee 37: Language and Terminology. 2007. *ISO/TR 22134:2007 – Practical Guidelines for Socioterminology*. Geneva: International Organization for Standardization.
- ISO Technical Committee 37: Language and Terminology. 2019. *ISO 26162: Management of terminology resources – Terminology databases – Part 1: Design, and Part 2: Software*. Geneva: International Organization for Standardization. Note: Publication of Part 3: *Content* is forthcoming as of this writing.
- ISO Technical Committee 37: Language and Terminology. 2014. *ISO 24156-1:2014 Graphic notations for concept modelling in terminology work and its relationship with UML – Part 1: Guidelines for using UML notation in terminology work*. Geneva: International Organization for Standardization.
- ISO Technical Committee 37: Language and Terminology. 2017. *ISO 16642:2017 Computer applications in terminology – Terminological markup framework (TMF)*. Geneva: International Organization for Standardization.
- ISO Technical Committee 37: Language and Terminology. 2019. *ISO 1087:2019 – Terminology work and terminology science – Vocabulary*. Geneva: International Organization for Standardization.
- ISO Technical Committee 37: Language and Terminology. 2019. *ISO 30042:2019 Management of terminology resources – TermBase eXchange (TBX)*. Geneva: International Organization for Standardization.
- ISO Technical Committee 176: Quality Systems. 2015. *ISO 9001: Quality management systems – Requirements*. Geneva: International Organization for Standardization
- Jacquemin, Christian. 2001. *Spotting and Discovering Terms through Natural Language Processing*. Cambridge: The MIT Press.
- Justeson, John and Slava Katz. 1995. “Technical Terminology: Some Linguistic Properties and an Algorithm for Identification in Text.” *Natural Language Engineering*, 1(1): 9–27. <https://doi.org/10.1017/S1351324900000048>
- Kageura, Kyo. 1995. “Toward the Theoretical Study of Terms.” *Terminology*, 2(2): 239–257. <https://doi.org/10.1075/term.2.2.o4kag>
- Kageura, Kyo. 2002. *The Dynamics of Terminology. A Descriptive Theory of Term Formation and Terminological Growth*. Amsterdam: John Benjamins. <https://doi.org/10.1075/tlrp.5>

- Kageura, Kyo. 2015. "Terminology and Lexicography." In *Handbook of Terminology*, V. 1, ed. by Hendrik J. Kockaert and Frieda Steurs. 45–59. Amsterdam: John Benjamins. <https://doi.org/10.1075/hot.1.04ter2>
- Kageura, Kyo and Bin Umino. 1996. "Methods of automatic term recognition." *Terminology*, 3(2): 259–289. <https://doi.org/10.1075/term.3.2.03kag>
- Karsch, Barbara. 2015. "Term Extraction: 10,000 Term Candidates – Now What?" *ATA Chronicle*, Feb 2015: 19–21. American Translators Association.
- Kelly, Natalie and Donald DePalma. 2009. *The Case for Terminology Management*. Common Sense Advisory. Burlington, MA: CSA Research.
- Kenny, Dorothy. 1999. "CAT Tools in an Academic Environment: What are They Good for?" *Target: International Journal of Translation Studies*, 11(1): 65–82.
- Kit, Chunyu and Xiaoyue Liu. 2008. "Measuring Mono-word Termhood by Rank Difference via Corpus Comparison." *Terminology*, 14(2): 204–229. <https://doi.org/10.1075/term.14.2.05kit>
- Kittredge, Richard and John Lehrberger. 1982. *Sublanguage. Studies of Language in Restricted Semantic Domains*. Berlin: Walter de Gruyter. <https://doi.org/10.1515/9783110844818>
- Knops, Eugenia and Gregor Thurmair. 1993. "Design of a Multifunctional Lexicon." In *Terminology: Applications in Interdisciplinary Communication*, ed. by Sonneveld, Helmi B. and Kurt L. Loening. 87–109. Amsterdam: John Benjamins. <https://doi.org/10.1075/z.70.08kno>
- Kocourek, Rostislav. 1982. *La langue française de la technique et de la science*. La Documentation Française, Paris. Weisbaden: Oscar Brandstetter Verlag GmbH & Co.
- Korkas, Vassilis and Margaret Rogers. 2010. "How much terminological theory do we need for practice? An old pedagogical dilemma in a new field." In *Terminology in Everyday Life*, ed. by Marcel Thelen and Frieda Steurs. 123–136. Amsterdam: John Benjamins. <https://doi.org/10.1075/tlrp.13.09kor>
- L'Homme, Marie-Claude. 2002. "What can verbs and adjectives tell us about terms?" In *Proceedings of the Terminology and Knowledge Engineering conference*, 65–70. Nancy, France.
- L'Homme, Marie-Claude. 2004. *La terminologie: principes et techniques*. Montreal: Les Presses de l'Université de Montréal.
- L'Homme, Marie-Claude. 2005. "Sur la notion de "terme"." *Meta: Translators' Journal*, 50(4): 1112–1132. <https://doi.org/10.7202/012064ar>
- L'Homme, Marie-Claude. 2006. "The Processing of Terms in Dictionaries: New Models and Techniques." *Terminology*, 12(2): 181–188. <https://doi.org/10.1075/term.12.2.02hom>
- L'Homme, Marie-Claude. 2019. *Lexical semantics for terminology: an introduction*. Philadelphia: John Benjamins. <https://doi.org/10.1075/tlrp.20>
- Leon-Arauz, Pilar, Arianne Reimerink and Pamela Faber. 2019. "EcoLexicon and By-products – Integrating and Reusing Terminological Resources." *Terminology*, 25(2): 222–258. <https://doi.org/10.1075/term.00037.leo>
- Lombard, Robin. 2006. "A Practical Case for Managing Source Language Terminology." In *Perspectives on Localization*, ed. by Keiran J. Dunne. 155–171. Amsterdam: John Benjamins. <https://doi.org/10.1075/ata.xiii.13lom>
- Madsen, Bodil Nistrup and Hanne Erdman Thomsen. 2015. "Concept Modeling vs. Data Modeling in Practice." In *Handbook of Terminology*, V. 1, ed. by Hendrik J. Kockaert and Frieda Steurs. 250–275. Amsterdam: John Benjamins. <https://doi.org/10.1075/hot.1.13con1>
- Marshman, Elizabeth. 2014. "Enriching Terminology Resources with Knowledge-rich Contexts: A Case Study." *Terminology*, 20(2): 225–249. <https://doi.org/10.1075/term.20.2.05mar>
- Martin, Ronan. 2011. "Term Inclusion Criteria." *Internal SAS document*, SAS Inc., Cary, N.C.
- Massion, Francois. 2019. "Intelligent Terminology." *MultiLingual*, 30(5): 30–34.

- Maynard, Diana and Sophia Ananiadou. 2001. "Term Extraction Using a Similarity-based Approach." In *Recent Advances in Computational Terminology*, ed. by Didier Bourigault, Christian Jacquemin, and Marie-Claude L'Homme. 261–278. Amsterdam: John Benjamins. <https://doi.org/10.1075/nlp.2.14may>
- Meyer, Ingrid. 1993. "Concept Management for Terminology: A Knowledge Engineering Approach." In *Standardizing Terminology for Better Communication: Practice, Applied Theory, and Results*, ASTM STP 1166, ed. by Richard Alan Strehlow and Sue Ellen Wright. 140–151. Philadelphia: American Society for Testing and Materials. <https://doi.org/10.1520/STP18002S>
- Meyer, Ingrid and Kristen Mackintosh. 1996. "The Corpus from a Terminographer's Viewpoint." *International Journal of Corpus Linguistics*, 6(2): 257–285. <https://doi.org/10.1075/ijcl.1.2.05mey>
- Meyer, Ingrid and Kristen Mackintosh. 2000. "When Terms Move into our Everyday Lives: An Overview of De-terminologization." *Terminology*, 6(1): 11–138. <https://doi.org/10.1075/term.6.1.07mey>
- Nagao, Makoto. 1994. "A Methodology for the Construction of a Terminology Dictionary." In *Computational Approaches to the Lexicon*, ed. by B.T.S. Atkins and A. Zampolli. 397–411. Oxford: Oxford University Press.
- Nakagawa, Hiroshi and Tatsunori Mori. 1998. "Nested Collocation and Compound Noun for Term Recognition." In *Proceedings of the First Workshop on Computational Terminology*, ed. by Didier Bourigault, Christian Jacquemin, and Marie-Claude L'Homme. 64–70. Montreal: Université de Montréal.
- Nakagawa, Hiroshi and Tatsunori Mori. 2002. "A Simple but Powerful Automatic Term Extraction Method." In *Proceedings of the Second International Workshop on Computational Terminology*. Stroudsburg, PA: Association of Computational Linguistics. <https://doi.org/10.3115/1118771.1118778>
- Nazarenko, Adeline and Touria Ait El Mekki. 2007. "Building Back-of-the-book Indexes?" In *Application-Driven Terminology Engineering*, ed. by Fidelia Ibekwe-SanJuan, Anne Condamines and M. Teresa Cabré Castellvi. 199–224. Amsterdam: John Benjamins. <https://doi.org/10.1075/bct.2.10naz>
- Nkwenzi-Azeh, Blaise. 2001. "User-specific Terminological Data Retrieval." In *Handbook of Terminology Management*, V. 2, ed. by Sue Ellen Wright and Gerhard Budin. 600–613. Amsterdam: John Benjamins. <https://doi.org/10.1075/z.htm2.20nkw>
- Oakes, Michael and Chris Paice. 2001. "Term Extraction for Automatic Abstracting." In *Recent Advances in Computational Terminology*, ed. by Didier Bourigault, Christian Jacquemin, and Marie-Claude L'Homme. 353–370. Amsterdam: John Benjamins. <https://doi.org/10.1075/nlp.2.18oak>
- Ó Broin, Ultan. 2009. "Controlled Authoring to Improve Localization." *MultiLingual*, Oct/Nov 2009.
- Packeiser, Kirsten. 2009. "The General Theory of Terminology: A Literature Review and a Critical discussion," Masters Thesis, Copenhagen Business School. Available from academia .edu
- Park, Youngja, Roy J. Byrd and Branimir K. Boguraev. 2002. "Automatic Glossary Extraction: Beyond Terminology Identification." In *Proceedings of the 19th international conference on computational linguistics*, V.1. Pennsylvania: Association for Computational Linguistics. <https://doi.org/10.3115/1072228.1072370>

- Pavel, Silvia. 1993. "Neology and Phraseology as Terminology-in-the-Making." In *Terminology: Applications in Interdisciplinary Communication*, ed. by Helmi B. Sonneveld and Kurt L. Loening. 21–33. Amsterdam: John Benjamins. <https://doi.org/10.1075/z.70.03pav>
- Pearson, Jennifer. 1998. *Terms in Context – Studies in Corpus Linguistics*. Amsterdam: John Benjamins. <https://doi.org/10.1075/scl.1>
- Picht, Heribert and Jennifer Draskau. 1985. *Terminology: An Introduction*. Copenhagen: Denmark LSP Centre, Copenhagen Business School.
- Picht, Heribert and Carmen Acuna Partal. 1997. "Aspects of Terminology Training." In *Handbook of Terminology Management*, V. 1, ed. by Sue Ellen Wright and Gerhard Budin. 63–74. Amsterdam: John Benjamins. <https://doi.org/10.1075/z.htm1.35pic>
- Pozzi, Maria. 1996. "Quality Assurance of Terminology Available on the International Computer Networks." In *Terminology, LSP and Translation. Studies in Language Engineering in Honour of Juan C. Sager*, ed. by Harold Somers. 67–82. Amsterdam: John Benjamins. <https://doi.org/10.1075/btl.18.07poz>
- Rey, Alain. 1995. *Essays on Terminology*. Amsterdam: John Benjamins. <https://doi.org/10.1075/btl.9>
- Riggs, Fred. 1989. "Terminology and Lexicography: Their Complementarity." *International Journal of Lexicography*, 22:89–110. <https://doi.org/10.1093/ijl/2.2.89>
- Rinaldi, Fabio, James Dowdall, Michael Hess, Kaarel Kaljurand, and Magnus Karlsson. 2003. "The Role of Technical Terminology in Question Answering." In *Proceedings of TIA – 2003 – Terminologie et Intelligence Artificielle*, Strasbourg.
- Roche, Christophe. 2012. "Ontoterminology: How to Unify Terminology and Ontology into a Single Paradigm" In *Proceedings of LREC conference*, 2012. Available from academia.edu
- Rogers, Margaret. 2000. "Genre and Terminology." In *Analysing Professional Genres*, ed. by Anna Trosborg. 3–21. Amsterdam: John Benjamins. <https://doi.org/10.1075/pbns.74.03rog>
- Rogers, Margaret. 2007. "Lexical Chains in Technical Translation. A Case Study in Indeterminacy." In *Indeterminacy in Terminology and LSP*, ed. by Bassey E. Antia. 15–35. Amsterdam: John Benjamins. <https://doi.org/10.1075/tlrp.8.05rog>
- Rondeau, Guy. 1981. *Introduction à la terminologie*. Montreal: Centre éducatif et culturel Inc.
- Sager, Juan, David Dungworth and Peter F. McDonald. 1980. *English Special Languages. Principles and Practice in Science and Technology*. Wiesbaden: Brandstetter-Verlag.
- Sager, Juan. 1990. *A Practical Course in Terminology Processing*. Amsterdam: John Benjamins. <https://doi.org/10.1075/z.44>
- Sager, Juan. 2001. "Term Formation." In *Handbook of Terminology Management*, V. 1, ed. by Sue Ellen Wright and Gerhard Budin. 25–41. Amsterdam: John Benjamins. <https://doi.org/10.1075/z.htm1.06sag>
- Sager, Juan. 2001. "Terminology Compilation: Consequences and Aspects of Automation." In *Handbook of Terminology Management*, V. 2, ed. by Sue Ellen Wright and Gerhard Budin. 761–771. Amsterdam: John Benjamins. <https://doi.org/10.1075/z.htm2.29sag>
- Sánchez, Maribel Tercedor, Clara Inés López Rodríguez, Carlos Márquez Linares, Pamela Faber. 2012. "Metaphor and metonymy in specialized language." In *A Cognitive Linguistics View of Terminology and Specialized Language*, ed. by Pamela Faber. 33–72. De Gruyter Mouton.
- Santos, Claudia and Rute Costa. 2015. "Domain Specificity." In *Handbook of Terminology*, V. 1, ed. by Hendrik J. Kockaert and Frieda Steurs. 153–179. Amsterdam: John Benjamins. <https://doi.org/10.1075/hot.1.09dom1>

- Schmitz, Klaus-Dirk and Daniela Straub. 2010. *Successful Terminology Management in Companies*. Stuttgart: TC and more GmbH.
- Schmitz, Klaus-Dirk. 2015. "Terminology and Localization." In *Handbook of Terminology*, V. 1, ed. by Hendrik J. Kockaert and Frieda Steurs. 452–464. Amsterdam: John Benjamins. <https://doi.org/10.1075/hot.1.ter7>
- Seomoz. 2012. *The Beginner's Guide to SEO*. Available at: <http://www.seomoz.org/beginners-guide-to-seo>
- Shreve, Gregory. 2001. "Terminological Aspects of Text Production." In *Handbook of Terminology Management*, V. 2, ed. by Sue Ellen Wright and Gerhard Budin. 772–787. Amsterdam: John Benjamins. <https://doi.org/10.1075/z.htm2.30shr>
- Strehlow, Richard. 2001a. "Terminology and Indexing." In *Handbook of Terminology Management*, V. 2, ed. by Sue Ellen Wright and Gerhard Budin. 419–425. Amsterdam: John Benjamins. <https://doi.org/10.1075/z.htm2.05str>
- Strehlow, Richard. 2001b. "The Role of Terminology in Retrieving Information." In *Handbook of Terminology Management*, V. 2, ed. by Sue Ellen Wright and Gerhard Budin. 426–444. Amsterdam: John Benjamins. <https://doi.org/10.1075/z.htm2.06str>
- Temmerman, Rita. 1997. "Questioning the Univocity Ideal. The Difference between Sociocognitive Terminology and Traditional Terminology." *Hermes, Journal of Linguistics*, 18: 51–90.
- Temmerman, Rita. 1998. "Why Traditional Terminology Theory Impedes a Realistic Description of Categories and Terms in the Life Sciences." *Terminology*, 5(1): 77–92. <https://doi.org/10.1075/term.5.1.07tem>
- Temmerman, Rita. 2000. *Towards New Ways of Terminology Description: The Sociocognitive Approach*. Amsterdam: John Benjamins. <https://doi.org/10.1075/tlrp.3>
- Temmerman, Rita, Peter De Baer, and Koen Kerremans. 2010. "Competency-based Job Descriptions and Termontography. The Case of Terminological Variation." In *Terminology in Everyday Life*, ed. by Marcel Thelen and Frieda Steurs. 179–191. Amsterdam: John Benjamins. <https://doi.org/10.1075/tlrp.13.13ker>
- TerminOrgs. 2014. *TBX-Basic Specification*. Available from: terminorgs.net
- TerminOrgs. 2016. *Terminology Starter Guide*. Available from: terminorgs.net
- Teubert, Wolfgang. 2005. "Language as an Economic Factor: The Importance of Terminology." In *Meaningful Texts*, ed. by Geoff Barnbrook, Pernilla Danielsson and Michaela Mahlberg. 96–106. London: Continuum.
- Thomas, Patricia. 1993. "Choosing Headwords from Language-for-special-purposes (LSP) Collocations for Entry into a Terminology Data Bank (Term Bank)." In *Terminology: Applications in Interdisciplinary Communication*, ed. by Helmi B. Sonneveld and Kurt L. Loening. 43–68. Amsterdam: John Benjamins. <https://doi.org/10.1075/z.70.05tho>
- Thurrow, Shari. 2006. *The Most Important SEO Strategy*. Available from: <http://www.clickz.com/clickz/column/1717475/the-most-important-seo-strategy>
- Van Campenhoudt, Marc. 2006. "Que nous reste-t-il d'Eugen Wüster?" In *Intervention dans le cadre du colloque international Eugen Wüster et la terminologie de l'École de Vienne*. Paris: Université de Paris 7.
- Warburton, Kara. 2001a. *Terminology Management in the Localization Industry – Results of the LISA Terminology Survey*. Geneva: Localization Industry Standards Association. Available from: terminorgs.net/downloads/LISAtermsurveyanalysis.pdf
- Warburton, Kara. 2001b. "Globalization and Terminology Management." In *Handbook of Terminology Management*, V. 2, ed. by Sue Ellen Wright and Gerhard Budin. 677–696. Amsterdam: John Benjamins. <https://doi.org/10.1075/z.htm2.25war>

- Warburton, Kara. 2014. "Narrowing the Gap between Termbases and Corpora in Commercial Environments." Doctoral thesis. Hong Kong: City University of Hong Kong. Available from: termologic.com/resource-area/
- Warburton, Kara. 2015. "Managing Terminology in Commercial Environments." In *Handbook of Terminology*, V. 1, ed. by Hendrik J. Kockaert and Frieda Steurs. 360–392. Amsterdam: John Benjamins. <https://doi.org/10.1075/hot.1.19man2>
- Wettengel, Tanguy and Aidan Van de Weyer. 2001. "Terminology in Technical Writing." In *Handbook of Terminology Management*, V. 2, ed. by Sue Ellen Wright and Gerhard Budin. 445–466. Amsterdam: John Benjamins. <https://doi.org/10.1075/z.htm2.o8wet>
- Williams, Malcolm. 1994. "Terminology in Canada." *Terminology*, 1(1): 195–201. <https://doi.org/10.1075/term.1.1.18wil>
- Wong, Wilson, Wei Liu and Mohammed Bennamoun. 2009. "Determination of Unithood and Termhood for Term Recognition." In *Handbook of Research on Text and Web Mining Technologies*, ed. by Min Song and Yi-Fang Brook Wu. 500–529. Hershey, PA: IGI Global. <https://doi.org/10.4018/978-1-59904-990-8.ch030>
- Wright, Sue Ellen. 1997. "Term Selection: The Initial Phase of Terminology Management." In *Handbook of Terminology Management*, V. 1, ed. by Sue Ellen Wright and Gerhard Budin. 13–23. Amsterdam: John Benjamins. <https://doi.org/10.1075/z.htm1.o4wri>
- Wright, Sue Ellen and Gerhard Budin. 1997. "Infobox No. 2: Terminology Activities." In *Handbook of Terminology Management*, V. 1, ed. by Sue Ellen Wright and Gerhard Budin. 327. Amsterdam: John Benjamins. <https://doi.org/10.1075/z.htm1.o2wri>
- Wright, Sue Ellen and Leland Wright. 1997. "Terminology Management for Technical Translation." In *Handbook of Terminology Management*, V. 1, ed. by Sue Ellen Wright and Gerhard Budin. 147–159. Amsterdam: John Benjamins. <https://doi.org/10.1075/z.htm1.19wri>
- Wüster, Eugen. 1968. *The Machine Tool*. London: Technical Press.
- Wüster, Eugen. 1979. *Einführung in die allgemeine Terminologielehre und terminologische Lexikographie*. Translation: Introduction to the General Theory of Terminology and Terminological Lexicography. Vienna: Springer.

Index

A

access controls 174, 183
acronyms 43, 54, 56, 68, 80, 84, 139
active controlled authoring 145
ad-hoc terminography 17
adding terms 169, 191
adjectives 63
administrative data categories 30
administrative functions 177
admitted terms 30, 146
adverbs 63
alignment 210
appellations 66
applications of terminology 73, 89
approved terms 185
approving terms 117
ATE *see* term extraction
authoring 78, 121
authoring memory 74, 97
autocomplete 81
automatic term extraction *see* term extraction
avoided costs 125

B

backups 225
benefits 42, 57, 125
beta testing 195
bigrams 65
Boolean operators 173
business case 42, 57, 123

C

CA *see* controlled authoring
CAT *see* computer-assisted translation
cleaning term candidates 201
collecting feedback 198
collocations 51, 206
commercial environment 35
Communicative Theory 11
community input 176
complex terms *see* multiword terms
compounds *see* multiword terms
computer-assisted translation 18, 30, 42, 77, 131,
142, 210
concept diagram 14

concept entry 3, 5, 23, 181
concept level 158
concept orientation 5, 20, 23, 23, 57, 67, 105, 181
concept relations *see* relations
concept systems 11, 14
concepts
 naming 211
 relations 151
 universality 15
conceptual data categories 30
concordancing 206
content management 74
content models 31, 135, 181
controlled authoring 18, 23, 25, 42, 47, 78, 131, 137,
145, 185, 226
corpora 8, 52, 89, 105, 206, 215
corpus analysis tools 215
cost avoidance 125
cost savings 125
costs 60, 124
cross references 30, 175

D

Darwin Information Typing Architecture

see DITA

data categories

 administrative 30
 concept level 158
 conceptual 30
 content models 31, 181
 for authoring 145
 for search 81, 154
 for translation 31, 142
 KEI 154
 language level 158
 part of speech 92, 142, 145
 picklists 27
 process status 185
 proposed set 158
 relations 31, 151
 selecting 141
 subject fields 92, 156
 subsetting 92, 156, 158
 term level 158
 term type 92, 146, 158

- terminological 30
- usage status 27, 30, 145, 185
- data elementarity 27
- data granularity 27
- data integrity 27
- data model
 - concept orientation 23
 - content models 181
 - data elementarity 27
 - data granularity 27
 - data integrity 27
 - default values 27, 181
 - levels 181
 - mandatory fields 181
- DatCatInfo 23, 30, 113
- de Saussure 21
- de-terminologization 94, 214
- default values 27, 164, 181
- defective terminology 45, 57
- definitions 80, 182
- deleting terms 185
- delimiting characteristics 14
- deprecated terms *see* prohibited terms
- see also* restricted terms
- descriptive terminography 18, 69, 215
- dialects 113
- DITA 74, 113, 121
- DocBook 121
- documentation 197
- domains *see* subject fields
- doublettes 170, 185

E

- embeddedness 98
- entailed terms 175, 224
- enterprise search 81
- entry *see* concept entry
- errors 53, 126, 222
- Eugen Wüster 11
- exclusion list 205, 215
- executive sponsorship 111
- export 29, 170
- extended applications 89
- external terminology 81

F

- faceted search 81
- feedback 198
- filters 184
- Frame-based Terminology 11
- fuzzy search 173

G

- general lexicon 4, 79, 138, 150
- General Theory of Terminology xxi, 8, 11, 94, 105, 229
- globalization 47
- glossaries 135
- GTT *see* General Theory of Terminology

H

- homographs 31, 91, 148, 153, 226
- homonymy 23

I

- identifiers 164
- import 29, 170
- importing terms 191
- inclusion criteria 137
- inflected forms 144
- input models 181
- integrated TMS 163
- interchange 29
- internal terminology 81, 131
- internationalization 47
- Internationalization Tag Set 74, 78, 113, 121, 144
- intranet 81
- ISO xxi, 113
- ISO 16642 *see* Terminological Markup Framework
- ISO 704 4, 8, 63, 113, 213, 223
- ITS *see* Internationalization Tag Set

K

- KEI 81, 154
- key performance indicators 128, 198
- Keyword Effectiveness Index *see* KEI
- keywords 81, 154, 206, 220
- KWIC 206

L

- language for general purposes 36
- language for special purposes 21, 36, 94
- language level 158
- language planning 35
- languages 168
- Lexico-Semantic Theory 11
- lexicographer 3
- lexicography 3
- lexicological entry 5
- lexicologist 3
- lexicology 3
- LGP *see* language for general purposes
- limited-value fields *see* picklists
- localization 47, 57, 142

Localization Industry Standards Association 57
 LSP *see* language for special purposes

M

machine translation 89
 mandatory fields 182
 microcontent 47, 74, 101
 modules 113
 multiword terms 54, 65, 94, 137, 215
 MWT *see* multiword terms

N

naming new concepts 211
 Natural Language Processing 15, 35, 47, 50, 73, 89, 93, 105, 117
 neologisms 211
 NLP *see* Natural Language Processing
 noise 205
 nonextant terms 218
 normalization 11
 normative terminology *see* prescriptive terminology
 nouns 63

O

OASIS 113
 Object Management Group 113
 onomasiology 11, 14, 23, 24, 105
 ontologies 89, 151
 organic search 81

P

parallel texts 210
 part of speech 63, 91, 92, 142, 148, 226
 passive controlled authoring 145
 phrasal terms *see* multiword terms
 picklists 27, 31, 31
 polysemy 24
 precision 205
 predictive typing 81
 preferred terms 30, 146
 prescriptive terminology 18, 79, 215
 process status 185
 prohibited terms 30, 146
 project management 128
 proper nouns 56, 66
 proposal 129
 punctual terminology *see* ad-hoc terminology

Q

quality assurance 222

query correction 81
 query expansion 81

R

recall 205
 reference corpus 204, 206
 relations 31, 51, 57, 89, 151, 175
 reports to management 198
 repurposability 28, 50, 73, 77, 89, 226
 restricted terms 30, 146
 ROI 123
 roles 117

S

saved costs 125
 SBVR 113
 search engine optimization 47, 53, 67, 73, 81, 154
 search keywords *see* keywords
 searching terms 173
 semasiology 14, 105
 SEO *see* search engine optimization
 sign 21
 signified 21
 signifier 21
 silence 205, 206
 Simplified Technical English 150
 simship 44, 202
 Socio-cognitive Theory 11, 14
 socioterminology XXI, 50
 sponsorship 111
 spreadsheets 29, 191
 stakeholders 117, 120
 standalone TMS 163
 standardization 8, 11
 standardized terms 113
 standards 113
 STE *see* Simplified Technical English
 stopword list 205
 style guide 117, 150
 subject fields 3, 21, 36, 51, 92, 94, 156
 subject matter experts 117, 198
 subsetting 92, 156
 synonyms 23, 27, 53
 synsets 57, 78, 81, 146
 systematic terminology 17

T

target-language terms 98, 210
 TBX *see* TermBase eXchange
 TBX-Basic 30, 92
 technical writing 78, 121
 TEI *see* Text Encoding Initiative

- term autonomy 26, 181
- term candidates 201
- term entry *see* concept entry
- term extraction 18, 52, 89, 201
- term harvesting *see* term extraction
- term level 158
- term mining *see* term extraction
- term type 92, 146, 158
- termbase
 - approval 129
 - backups 225
 - collecting feedback 198
 - content models 181
 - data categories 30, 135, 141, 181
 - data model 181
 - default values 181
 - designing 27, 28
 - documentation 197
 - implementation 128
 - inclusion criteria 137
 - input model 181
 - key performance indicators 198
 - launching 196
 - mandatory fields 181
 - proposal 129
 - quality assurance 222
 - reporting to management 198
 - roles 117
 - stakeholders 117
 - testing 195
 - training 197
 - user interface 131
 - users 117
 - uses 73
 - web interface 131
- see also* terminology management systems
- TermBase eXchange 29, 30, 113, 131, 163, 191
- termhood 36, 94, 101, 105, 137, 145, 201
- terminographer 3
- terminography
 - ad-hoc 17
 - descriptive 18, 69, 215
 - onomasiological 14
 - prescriptive 18, 79, 215
 - semasiological 14
 - systematic 17
- terminological data categories 30
- terminological entry *see* concept entry
- Terminological Markup Framework 116, 158, 165, 181
- terminological phrases *see* multiword terms
- terminologist 47

- terminologization 214
- terminology
 - applications 89
 - difference with lexicology 3, 11
 - meanings of 3
 - problems 45, 53, 126
 - standardization 11
 - theories 11
 - uses 73
- terminology audit 135
- terminology database *see* termbase
- terminology management systems
 - access controls 174, 183
 - adding terms 169, 191
 - administrative functions 177
 - community input 176
 - core features 164
 - cross references 175
 - data model 164
 - default values 164
 - doublettes 170
 - entailed terms 175
 - export 170
 - filters 184
 - identifiers 164
 - import 170
 - integrated 163
 - languages 168
 - overview 163
 - relations 175
 - search 173
 - standalone 163
 - views 172, 184
 - workflows 176, 185
- see also* termbase
- terminology problems 53
- terminology standards 113
- TerminOrgs 30, 58, 89, 113
- terms
 - acronyms 43
 - adding 169, 191
 - admitted 30, 146
 - approving 185
 - deleting 185
 - doublettes 185
 - entailed 224
 - errors 53
 - formation 211
 - from corpora 215
 - importing 191
 - marking in source 121
 - nonexant 215

- preferred 30, 146
- problems 53
- prohibited 30, 146
- restricted 30, 146
- undocumented 215
- unoptimized 215
- usage 145
- testing a termbase 195
- TeX 121
- Text Encoding Initiative 113, 121
- text mining 89
- Textual Theory 11
- thematic terminography *see* systematic terminography
- theories
 - Communicative Theory 11
 - Frame-based Terminology 11
 - General Theory 11, 94, 105
 - Lexico-Semantic Theory 11
 - Socio-cognitive Theory 11, 14
 - Textual Theory 11
- TM *see* translation memory
- TMF *see* Terminological Markup Framework
- TMS *see* terminology management systems
- TMX 74
- Traditional Theory of Terminology *see* General Theory of Terminology
- training 197
- transcreation 15
- transfer comment 26
- translated terms *see* target-language terms
- translation 46, 77, 78
- translation memory 43, 75, 97, 210
- transparency 211
- trigrams 65
- typeahead 81
- U
 - undocumented terms 216
 - unithood 94
 - univocity 11, 21, 23
 - unoptimized terms 216
 - usage status 27, 30, 138, 146, 185
 - use cases 134
 - uses of terminology 73
- V
 - variants 51, 53, 67
 - verbs 63
 - Vienna School *see* General Theory of Terminology
 - views 131, 172, 184
- W
 - W₃C 113
 - web interface 131
 - wildcards 173
 - word class *see* part of speech
 - workflows 134, 176, 185
 - working corpus 206
 - writing 78, 121
 - Wüster 11
- X
 - XLIFF 74
 - XML 49, 74, 121

The Corporate Terminologist is the first monograph that addresses the principles and methods for managing terminology in content production environments that are both demanding and multilingual, such as those found in global companies and institutions. It describes the needs of large corporations and how those needs demand a new, pragmatic approach to terminology management. The repurposability of terminology resources is a fundamental criterion that motivates the design, selection, and use of terminology management tools, and has a bearing on the definition of termhood itself. *The Corporate Terminologist* describes and critiques the theories and methods informing terminology management today, and practical considerations such as preparing an executive proposal, designing a termbase, and extracting terms from corpora are also covered. This book is intended for readers tasked with managing terminology in today's challenging production environments, for those studying translation and business communication, and indeed for anyone interested in terminology as a discipline and practice.

ISBN 978 90 272 0849 1



John Benjamins Publishing Company