NARROWING THE GAP BETWEEN TERMBASES AND CORPORA IN COMMERCIAL ENVIRONMENTS

KARA CORDELIA WARBURTON

DOCTOR OF PHILOSOPHY

CITY UNIVERSITY OF HONG KONG

JULY 2014

CITY UNIVERSITY OF HONG KONG
香港城市大學

Narrowing the Gap Between Termbases and Corpora in Commercial Environments
商用術語庫和語料庫之間的協調問題研究

Submitted to
Department of Linguistics and Translation
翻譯及語言學系
in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy
哲學博士學位

by

Kara Cordelia Warburton

July 2014
二零一四年七月

ABSTRACT

This research investigates the terminological data in terminology databases (termbases) and in corresponding corpora from commercial sources. Four companies in the information technology (IT) sector are used as case studies. Our broad objective is to increase awareness about some of the issues and challenges faced by terminologists in commercial settings. We demonstrate that there are significant gaps between the termbases and the corresponding corpora, that such gaps reduce the effectiveness of the termbases, and that they can be minimised by adopting a corpus-based approach to term identification.

We begin by establishing that the language used in a company contains terminology. After reviewing the conventional theories and methodologies of the field of terminology, we challenge the suitability of some of their precepts for companies that require terminological resources that are both repurposable and production-oriented. We then reveal features in the termbases that depart from established norms. Using a batch concordance technique, we quantify the gap between the termbase terms and the corpora. We then attempt to explain this gap by examining termbase terms that occur in various frequency ranges within the corpora. Using empirical observations, we formulate some guiding principles for selecting terms for termbases with respect to various features including term length, part of speech, term variation, and the use of certain types of modifiers.

We discover that keywords hold potential for discovering multi-word terms that, if documented in termbases, would significantly increase the correspondence between termbases and corpora. We conclude that termbases developed in companies would increase in value if corpus-based approaches to term identification were adopted. This challenges the conventional understanding of what is a term; to open the field of terminology to commercial applications and environments, termhood needs to be established based on communicative purpose and end-use of terminological resources in addition to purely semantic criteria.

Keywords: terminology, terminography, termbases, corpora, keywords, LSP

CITY UNIVERSITY OF HONG KONG

Qualifying Panel and Examination Panel

Surname:	WARBURTON
First Name:	Kara
Degree:	Doctor of Philosophy
College/Department:	Department of Linguistics and Translation

The Qualifying Panel of the above student is composed of:

Supervisor(s)

Dr. FANG Chengyu Alex	Department of Linguistics and Translation City University of Hong Kong

Qualifying Panel Member(s)

Dr. LUN Suen Caesar	Department of Linguistics and Translation City University of Hong Kong
Prof. Jonathan James WEBSTER	Department of Linguistics and Translation City University of Hong Kong

This thesis has been examined and approved by the following examiners:

Dr. FANG Chengyu Alex	Department of Linguistics and Translation City University of Hong Kong
Prof. ZHU Chunshen	Department of Chinese and History City University of Hong Kong
Dr. GUAN Jian Jeff	Department of Computer Information Systems University of Louisville
Prof. L'HOMME Marie-Claude	Département de linguistique et de traduction University of Montreal

ACKNOWLEDGEMENTS

I would like to first and foremost thank the terminologists who participated in this research and their respective companies: Marion Mordenti from Hewlett-Packard, Sue Kocher from SAS, Christina Tolliver from Minitab, and Katrin Drescher from Symantec. They deserve admiration for their willingness to subject their own terminology management practises to scrutiny in order to raise the profile of terminology management in the service of commercial enterprises.

I am indebted to Interverbum Technology AB[1] for donating the TermWeb terminology management software for the purposes of this research. In particular, I wish to thank Bengt Sjogren, CEO, Ioannis Iakovidis, Managing Director, and Mats Granström, Product Director, for their encouragement and technical support. Acknowledgements are also due to Mike Scott, of WordSmith Tools, who promptly responded to my numerous posts on the users' blog about how to resolve problems encountered during batch concordances. His advice on preparing the corpora was invaluable.

I am deeply grateful for the intellectual inspiration and mentoring provided by Dr. Alex Chengyu Fang, Associate Professor, Department of Chinese, Translation and Linguistics at the City University of Hong Kong, who supervised this PhD research. The comments and support extended by fellow researchers in the Dialogue Systems Group provided for a stimulating environment in which to carry out this research. I would like to acknowledge the Hong Kong PhD Fellowship Scheme for the financial support I received in the form of a full scholarship.

Finally, none of this research would have been possible without the unwavering encouragement and support of my husband Mark, and the confidence of our daughters Emma and Madeline. They all made life-changing sacrifices to enable me to undertake this research, for which I am eternally grateful.

1 http://www.interverbumtech.com/

TYPOGRAPHICAL CONVENTIONS

The following typographical conventions are used in this thesis:

- Italics emphasise a term, when discussed meta-linguistically
- Quotation marks are reserved for citations from the literature
- Underlining is used for emphasis

British spelling has been adopted in this thesis, however, American spelling in proper names, citations and references has been maintained.

LIST OF TABLES

LIST OF FIGURES

CONTENTS

CHAPTER 1 INTRODUCTION AND MOTIVATION

1.1 About this thesis

This thesis describes research in the area of terminology management in commercial environments. After completing a Master's degree in Terminology at Université Laval (Québec, Canada), the researcher worked as a terminologist in commercial settings, including 15 years at IBM Corporation. This experience led her to question whether the conventional theories and methodologies of the field are suitable for managing terminology in companies. This PhD research is an attempt to answer that question.

Chapter One (this chapter) includes an introduction to the field of terminology, definitions of key terms used throughout this thesis, some reflections on how terminology relates to commercial interests, and the motivation for this research. In Chapter Two, we review the state-of-the-art, particularly with respect to the key notions of *term* and *language for special purposes*, the predominant theories and methodologies of terminology, and the relationship with corpus linguistics and the role of corpora. In Chapter Three, we critically examine these notions, theories, and methodologies and comment on their suitability for managing terminology in commercial environments. The methodology, objectives and expected outcomes of the research are presented in Chapter Four. Chapter Five is dedicated to describing our data and how it was prepared for analysis. We report the results of our data analysis in Chapter Six, beginning with the terminology databases (termbases) and followed by an investigation of termbase terms found at various frequencies in the corpora. In Chapter Seven we explore the potential of keywords as nodes of frequently-occurring multi-word terms. In the final chapter we draw conclusions, acknowledge limitations of our research, and identify potential areas of future research.

1.2 Key terms and definitions

There are various definitions of terminology as a discipline or as a field of study and practise, reflecting different theoretical views. The definition adopted by the ISO Technical Committee 37, which sets international standards in the area of terminology management, is as follows:

> (Terminology is) the science studying the structure, formation, development, usage and management of terminologies in various subject fields (ISO 1087-1, 2000)[2]

To understand this definition, we refer to TC37's definitions of *terminologies* and *subject fields*:

> (A terminology is a) set of designations belonging to one special language.
>
> (A subject field is) a field of special knowledge.

And also, we need the definition of *special language*:

> (A special language is) a language used in a subject field and characterised by the use of specific linguistic means of expression.

According to these definitions, terminology is concerned with the terms, or designations, confined to distinct subject fields and not with the (general) lexicon of a language, which is the purview of lexicology and lexicography (Sager 1990: 3; Wright et al 1997: 13, 64, 328; Rondeau 1981: 63; Cabré 1999-b: 35; Rey 1995: 119, 130).[3] Lexicology, defined in the Oxford dictionary as "That branch of knowledge which treats of words, their form, history, and meaning," is clearly its closest relative. Sager defines *terminology* as follows (1990: 2):

> Terminology is the study of and the field of activity concerned with the collection, description, processing and presentation of terms, i.e. lexical items belonging to specialised areas of usage of one or more languages.

Considered in all its uses, the word *terminology* is polysemic[4], and therefore, potentially ambiguous. Some even say that it has been the source of considerable confusion and disagreement (Budin 2001: 15). Its original meaning is that of a set of terms, such as the *terminology of sailing*. As theories and methodologies for managing terminology took

2 As quoted from the ISO TC37 termbase, available at: http://iso.i-term.dk
3 ISO TC37 has not proposed any definition for lexicography.
4 See Cabré, 1996, for a detailed description of the meanings of terminology.

shape, the word came to be used to refer to this set of theories and methodologies, and by extension, to the practise of managing terminology, that is, *terminology work* (Rey 1995: 127). To avoid ambiguity, we have adopted distinct terms. We use *terminography* to refer to the collection of activities involved in managing terminology, a position that we will justify in the next section. However, we use *terminology management* and variations thereof (*managing terminology*, *management of terminology*, etc.) when we wish to emphasise the set of activities as an overall program or strategy. We use *terms,* and sometimes other expressions depending on the context, such as *lexical units*, when referring to sets of terms, and *terminological data* when discussing data or information about terms. The aggregate of *terms* and *terminological data* as found in a termbase or in a computer file containing this information (such as an XML file or a spreadsheet) is referred to as a *terminological resource*. In this thesis, we have attempted to reserve the term *terminology* for the set of theories and methodologies.[5]

1.3 Terminology as a discipline and a vocation

Terminology has been accorded the distinction of constituting a scholarly discipline in its own right within the broader field of language studies (Rey 1995: 50; Dubuc 1992: 3; Budin 2001: 16; Cabré 1996: 16). But as Condamines notes (1995: 225), there have been challenges to this notion. Cabré convincingly argues that its disciplinary status has yet to be unequivocally demonstrated (2000), and Sager (1990: 1) and Kageura (1995: 253) deny it altogether. Van Campenhoudt (2001) even challenges the notion that terminology exists separately from LSP lexicography. The question as to whether terminology is a discipline is therefore not yet settled. There has however been a shaping and convergence of theories and methodologies that enable us to talk of terminology as a field of study and as a practise.

In the previous section, we noted that terms, the focus of study of terminology, are confined semantically to a subject field. We will show later that there are differing opinions about

5 Although we have adopted this less ambiguous terminology in the thesis, we cannot change the quotations of other authors. Readers may find the term *terminology* used with any of its possible interpretations in quotations.

what constitutes a subject field[6] (*specialised area* for Sager, above), and in many communicative contexts one cannot easily differentiate between general language and special language (Schubert 2011: 28; Myking 2007:84). Furthermore, in dictionaries, which are lexicographic works, one finds an abundance of specialised terms, some scholars estimating up to 40 percent of the total entries (Bowker 2003: 156; Landau 2001: 34). Thus, the degree of specialisation of the meaning of a lexical unit is not enough to declare whether it falls under the scope of terminology or of lexicology. Maurais (1993: 117) and Lam Kam-mey (2001: xi) even identify computing, the domain of the current research, as being prototypical for witnessing technical terms cross into the general lexicon and vice-versa.

There are also, however, significant differences in methodology which have helped to further distinguish lexicology and terminology.[7] The methods used by terminologists to investigate and manage terminologies are, according to established theory, clearly different from those used by *lexicographers* to describe the lexicon (Cabré 1999-b: 37-38 and 1996: 25; Sager 1990: 3). We will show that in practise, terminologists adopt some methods characteristic of lexicography, and that these methods are particularly relevant to commercial environments.

The original theory of terminology was developed in Vienna by Eugen Wüster in the middle of the twentieth century, culminating in his treatise on the General Theory of Terminology (GTT) which was published in 1979. Several new theories have emerged in recent decades that have broadened our understanding of terminology (see Temmerman 2000, Cabré 2003, L'Homme 2004). These theories will be described later. However, the GTT has dominated, and largely continues to dominate, mainstream knowledge about terminology (Bourigault and Jacquemin 2000: section 9.2.2). This has consequences for our investigation, which concerns the management of terminology in commercial environments.

As lexical units, terms are a part of language. Proponents of the various theories of terminology aim to understand and explain terminology as a feature of language. With this know-

6 ISO TC37 offers no definition for “special knowledge” which is a key term in its definition of subject field.

7 For a deeper discussion of the differences between terminology and lexicology, see: Cabré 1996 and 1999, Rey 1995, Riggs 1989, Dubuc 1997 and L'Homme 2006.

ledge in hand, methods can be developed to create, manage, and use terminological resources for specific aims such as improving communication or managing knowledge resources. In this respect, terminology as a discipline should be able to provide a theoretical and methodological foundation that serves all communication needs. But the needs for terminological resources are changing in response to rapidly evolving technologies used increasingly in commercial environments. We suggest that the theories and methodologies are not keeping apace with these changes.

In information technology, increasingly specialised vocation titles have emerged, such as information developer, localiser, and information architect. Yet the job title terminologist is rarely encountered. In a survey conducted by the Localization Industry Standards Association of users and providers of localisation services, a sector where managing terminology would seem to be important, only 12 percent of respondents claim to employ a terminologist. The remaining 88 percent delegate any terminology management work to staff with other specialisations (Lommel and Ray 2007: 30). Information professionals generally have little knowledge about terminology as a field of study, or about terminological resources as a type of language resource that benefits an organisation. Professional organisations dedicated to technical writing are large, such as Tekom[8] (8,250 members) and the Society for Technical Communication[9] (over 6,000 members), yet terminology and terminologists are hardly mentioned. For example, a search for *terminology* and *terminologist* on the web site of the Society for Technical Communication (STC) [10] produces six and zero results respectively, but a search for *information architect* produces 248, even though, at least to the uninitiated, this appears an unusual job title. In its job bank, as of the writing of this thesis, there are 95 job postings for technical writer and 233 for information architect, but only one for terminologist. 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 employs 40 terminologists to support its 1,300 translators and interpreters[11]. Rey also admits that termino-

8 www.technical-communication.org/
9 www.stc.org/images/stories/pdf/membershipflyer.pdf
10 www.stc.org
11 www.bt-tb.tpsgc-pwgsc.gc.ca/btb.php?lang=eng&cont=826 . This number is currently 1,300 as reported in: http://www.tpsgc-pwgsc.gc.ca/rapports-reports/rpp/2013-2014/rpp-02-eng.html#s2.2.10 . The number 40 for terminologists was provided through direct communication with the Bureau.

logy as a discipline is not widely recognised (1995: 16) and that the disciplinary meaning of the term *terminology* is still unrecorded in many contemporary dictionaries (p. 127).

In summary, one cannot say that there is a unanimous consensus about the nature, scope, principles and methodologies of terminology as a discipline or as a vocation. This field continues to mature.

1.4 Terminology management and terminography

It supports the objectives of this research to clarify the notion of *terminology management.* The term *terminography*, using the suffix *graphy* as motivated by *lexicography*, was proposed by Russian linguist A.D. Hajutin in 1971 to refer to the notion of terminology work. The term *terminographer*, denoting the person who performs terminology work, first appeared in 1975 in the works of the Russian linguist, E. Natanson, and was taken up by Rey in 1976 (Rondeau 1981: 18; L'Homme 2006: 182)[12]. Rey (1995: 125-133) persuasively argued for the separation of terminology and terminography, using the former for theoretical approaches and the latter for "applied descriptive terminology" (p. 23). In an article dedicated to this topic, Sager makes a similar appeal, abhorring the terms *terminology science* and *terminology work* as awkward and imprecise (1994). He maintains that "terminography is a perfectly motivated English term suitable for the applied side of terminology" (p. 379), and that *terminology science* is a tautology given that the suffix *ology* already means science. Thomas (1993: 44) discusses the difference between terminography and lexicography by comparing language for general purposes (LGP) and language for special purposes (LSP), "Terminography is to LSP what lexicography is to LGP." Wright and Budin do the same (1997: 327), as does Cabré (1999-b: 115). However, as we stated earlier, the semantic valence of the linguistic expressions studied is not sufficient for distinguishing the two disciplines, which is why we will also look at methodologies.

The terms *terminography* and *terminographer* have been adopted by many scholars (Cabré 1994: 24 and 1999-b: 115; De Bessé 1997: 65; Rondeau 1981: 18; L'Homme 2004: 21-22;

12 The etymology of these terms was also verified in *Le Robert historique de la langue française*, 2010.

Meyer and Mackintosh 1996: 257; Temmerman 1997: 88 and 2000: 230; Picht and Draskau 1985: 118; Pearson 1998: 209; Schubert 2011: 26; Rogers 2000: 6). Today, the term *terminography* is widely recognised to refer to the work performed by terminologists[13].

There are differing opinions about what is involved in managing terminology. Wright and Budin define *terminology management* broadly as "any deliberate manipulation of terminological information" (1997: 1). In the same volume (p. 327), they cite the ISO 1087 definition of *terminology work* : "any activity concerned with the systematisation and representation of concepts or with the presentation of terminologies on the basis of established principles and methods," and they define *terminography* as "the recording, processing, and presentation of terminological data acquired by terminological research." However, even today, some scholars continue to portray terminology work more narrowly as the application of the principles of the GTT (which shall be explained in more detail later): "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).

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 specialised dictionaries or term banks" (2004: 21, translation). She classifies these activities into seven areas (p. 45, translation):

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

13 As of the writing of this thesis, a Google search on "terminography" produces over 3 million results, whereas a search for "terminology management" produces only 142,000.

Here we can see that terminography comprises more than simply managing terminology in a termbase, the latter corresponding to only one of the seven types of activities, if we adopt L'Homme's perspective. This view gives more formal recognition of the tasks of corpus building, term identification, and terminology research. The LISA study of industry-based terminology management includes a similar range of activities, all oriented towards the goal of improving terminology use in the organisation (Warburton 2001a: 4), with one notable addition: distribution of terms. In this thesis, *terminography* has been adopted with the wider interpretation to include aspects such as manipulation of corpora, term identification, and distribution activities in addition to working directly in a termbase .

1.5 Applications of terminological resources

Two decades ago, it was predicted that applications beyond translation would benefit from richly-structured terminological resources (Knops and Thurmair 1993: 89; Sager 1990: Section 8.4.1; Meyer 1993: 146). Galinski (1994: 142) noted that standardised terminologies are beneficial for indexing, information retrieval, and overall quality management in addition to the conventional uses of translation and technical writing. Sager (1990: 228) predicted that the end-user of terminological resources is not necessarily a human being.

We maintain that there is a wide range of potential applications of terminological resources that are relevant in commercial settings, where, under the pressures of saving costs, increasing productivity, and gaining market share, areas such as authoring, translation and content management, for instance, are naturally subject to increasing automation. We discuss this point further in Section 1.7.3.

1.6 Implications for commercial applications

In this section, we introduce some broad concerns about the implications of conventional wisdom for commercial stakeholders. Later, we will elaborate on many of these points.

We will show that the existing literature about terminology has an academic focus and contains little information pertinent to the practical needs of commercial environments. Terminography continues to be perceived as an academic exercise, and the most notable implementations outside of academic research are undertaken by governments to support language planning policies. In recent years, the growing use of computer-assisted translation (CAT) tools in institutional settings has had some effect in raising awareness about the potential use of terminological resources for commercial purposes. Consequently, some companies have developed in-house termbases[14]. As early as 1996, many medium-sized termbases had already been established by large companies and public institutions in Europe (de Bessé and Pulitano 1996: 35). This trend is likely to continue as natural language processing (NLP) technologies penetrate commercial settings to support the need to efficiently manage content, such as in the areas of controlled authoring and automatic document indexing and classification. A key question raised by this growing presence of termbases in commercial settings is whether the existing theories and methodologies for managing terminology are able to address the needs of commercial implementations.

The aim of the GTT, which will be discussed in detail later, is to standardise, or normalise, terms. This focus on normalisation has had a profound impact on terminography by imposing prescriptivism. For instance, virtually all the guidelines and standards relating to terminography that have been produced by ISO TC37 – widely acknowledged as an authoritative source of knowledge in this field – reflect the prescriptive approach. In its over 50 years of operation, TC37 has published 23 standards, and is in the process of developing at least 20 more, yet none address the commercial applications of terminology. The focus of TC37 and other key organisations in this sector, such as InfoTerm[15], has been on normalisation (Temmerman 2000: 14, 17-18); their approaches to terminography are geared towards language standards and the linguistic programs and services undertaken by governments to support national languages.

14 For example: IBM, Microsoft, SAS, Huawei, Nokia, IMF, and WHO, as well as the four companies in this study.

15 The International Information Centre for Terminology

Is the GTT a suitable foundation for a terminographic methodology that would address the needs of enterprise communication? As early as 1985, it was acknowledged that the works produced by TC37 could not even serve as a theoretical foundation for terminology as a whole, for their goal – standardisation – is too narrow (Picht and Draskau 1985: 249). This view is shared by Temmerman (1997: 54; 2000: 14, 18), L'Homme (2004:27) and even Cabré (2003: 179; 2000: 41). In commercial environments, the inadequacy of these traditional approaches may be pronounced.

For example, in a company, polysemy and synonymy occur due to overlapping concepts in products and technologies and the idiosyncratic linguistic diversity of different writers and translators in its employ. Yet in normative environments these linguistic phenomena are discouraged if not banned altogether. Indeed according to scholars who challenge the GTT, the univocity principle (also known as the isomorphism principle, Temmerman 2000: 126), whereby a term can have one and only one meaning and a concept can be denoted by one and only one term, does not always occur in practise (Temmerman 1997: 88; Cabré 1999-b: 108; Rey 1995: 56; Rondeau 1981: 22), even within a restricted domain (L'Homme 2004: 30; Cabré 1999-b: 40 and 2000: 40). Temmerman even maintains that the univocity principle is "untenable" (2000: 17), a view that is shared by Condamines (2005: 44; 2007b: 45) and Rogers (2007: 15). In commercial texts, terminology consistency is certainly desired, between versions of a product, between related products, and between various communication media. However, at the same time terminology diversification is sometimes necessary for market differentiation (Corbolante and Irmler 2001: 534-535). Furthermore, some search engine optimisation (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 2001-b: 434). Tools have even been developed to find suitable search keywords for a given topic[16]. Since keywords are the terms that represent the topic of a web site or web page, it is evident that terminology is relevant to SEO. All these points suggest that in commercial environments, one needs to balance normalisation interests with freedom of creative expression and marketing priorities.

16 For instance, Google Keyword Tool Box (http://www.googlekeywordtool.com/), Trellian Keyword Discovery tool (http://www.keyworddiscovery.com/search.html), WordStream Keyword Search tool (http://www.wordstream.com/keyword-research-tool)

In 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). Through terminology consistency, translation costs can be reduced by increasing the leverage rate of translation memory (TM), which is one of the metrics used in the localisation industry to justify investment in developing terminological resources (Schmitz and Straub 2010: 18, 20, 25, 26, 29, 57-58, 293). However, even among terminologists, few realise 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 reuse potential of authoring and TM, and if this objective therefore emerges as a key motivating factor for managing terminology in a commercial setting, this could have wide-reaching ramifications on the scope and methods of terminography in commercial environments. 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 the visibility of a term to product users. We wonder if, for commercial applications, a *term* could simply be defined as a lexical entity that, if managed according to certain suitable methods, can bring benefit to the company.

Some scholars maintain that the onomasiological approach to concept description, which is a tenet in the GTT and a methodological difference often cited to distinguish terminography from lexicography, is rarely adopted in practise (Temmerman 2000: 230). According to this approach, which will be further described later, concepts are studied before terms. If it is indeed impractical in commercial settings, a fundamental principle in terminography would need to be revisited because it fails to address the needs of a growing stakeholder in terminology. As Cabré et al noted (2007: 2), even the notion of *term* varies from one application of terminology to another. In the interests of maintaining coherence in the field of terminology, these apparent contradictory positions need to be reconciled. We join Temmerman (2000: 153) in calling for a "diversification in the methodology of terminography" and Cabré (1999-b, p. 114) in recommending that this diversification stress pragmatic aspects.

1.7 Motivation of the current research

We suggest that the established approaches for terminography do not suit needs in commercial settings for developing multipurpose terminological resources. Insofar as we can determine, there are no documented terminographic practises specifically for developing terminological resources for commercial applications.

Unable to find guidelines for managing terminology in commercial settings, companies often turn to sources like ISO TC37, assuming that their prescribed methods apply to them as well. Consequently, they may adopt inappropriate practises, which can result in less-than-optimal technical implementations and costly mis-allocation of resources. For example, failure to document all the variants of a term in a company terminology database (termbase) can lead to lost opportunities for optimising certain applications that need such variants, such as spell checkers and controlled authoring software (terminological variation is discussed in sections 2.3.3 and 3.2.3). And preparing well-formed definitions based on concept systems – a basic tenet of the GTT – is labour intensive and cost-prohibitive; rarely is the investment justified from a business standpoint.

A company may also inadvertently choose a very simplified ad-hoc lexicographic approach (essentially, collecting and minimally describing terms for an immediate need) rather than a systematic terminographic one (further analysing and describing underlying concepts, identifying synonyms and semantic relations). Indeed, this happens quite frequently because of the familiarity of the approach; it produces something that looks similar to a dictionary, with terms as headwords below which their various meanings are described. But because this approach fails to document ontological term relations, if the company subsequently decides to adopt a technology that requires such relations, such as search query expansion or controlled authoring, the termbase may need to be re-engineered, and a lot of new data entered. Failure to adopt a concept-oriented approach to structure terminological resources can result in lost opportunities to leverage those resources.

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.” Aside from the academic and normative focus of the discipline previously described, several additional historical factors may have contributed to the lack of a suitable framework for commercial terminography: (1) the narrow ties of terminology to translation, and (2) the restricted focus of termbases. We comment on these topics in the next sections, and then discuss some key issues for managing terminology in commercial settings.

1.7.1 The close ties to translation

Terminology emerged from the need to name scientific and technical concepts that take shape through innovation. Mass media increased the need to standardise terms in order to facilitate communication. But terminological activities of recent decades have been closely tied to translation (Rey 1995: 50, 129; Pozzi 1996: 69). Eugen Wuster's influential work, which led him to elaborate the tenets of the GTT, is an interlingual dictionary of concepts in the field of machine tools (1967). Since then, the main developers of terminological resources have been plurilingual institutions concerned with multilingual communication. Although there are no theoretical, philosophical, or methodological reasons why terminology need necessarily be a multilingual endeavour (Sager's definition cited earlier states “one or more languages”), it almost always is (L'Homme 2004: 21; Rondeau 1981: 33).

The interlinguistic aspect relates terminology to translation (Rey 1995: 129) and terminology therefore has close ties with the translation industry (Bowker 2002: 290; Williams 1994: 195). University-level courses about 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). Many of the existing terminology databases were developed to serve the needs of translators (Teubert 2005: 101). Consequently, people entrusted with the role of terminologist are almost always primarily translators. Indeed, many employers apparently expect translators to carry out their own terminological research (Bowker 2002: 291). Cabré also observes

that the distinction between these two groups of professionals is often blurred (1999-b: 115). These observations have been confirmed through surveys undertaken with global companies: "Overall, the picture is that terminology is not actually being managed on any systematic basis, but rather, it is treated as a part of the localization process and dealt with on an ad hoc basis" (Lommel and Ray 2007: 31).

While a translation background is beneficial for terminologists, 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). It is also often overlooked in other application areas where structured terminological resources could contribute, such as content management and information retrieval. Indeed, the restricted perception of terminology as an element of the translation process has impeded the recognition of the potential of terminology management for government, industry, and the economy in general (Wright 2001-b: 467). Wright sees this focus shifting gradually with beneficial results (p. 468):

> Forward-thinking enterprises have now integrated terminology generation and documentation into the document and production generation process, a development that has resulted in increased document quality and reliability.

She also notes the impact this has on the vocation:

> The gradual movement of terminology compilation out of the narrow confines of translation-oriented terminology management into the broader focus of technical documentation and information retrieval is accompanied by increased involvement of terminologists in technical communications, information management, and data processing. (p. 469)

Another consequence of this association 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 often have no background in terminology whatsoever. As Cabré states (1999-b: 115), the distinction between these two groups of professionals is often blurred.

The applications of terminological resources that are described in Sections 1.5 and 1.7.3 are not discussed in the normative TC7 standards, which also have a translation focus. Best

practises for managing terminology in commercial environments to address such uses do not seem to exist.

The close ties to translation raise the question whether terminological resources developed according to existing norms of practise, which are designed primarily to serve translation purposes, are able to meet more diverse needs.

1.7.2 The restricted focus of termbases

Historically, termbases have remained outside the realm of commerce. Due to terminology's traditional normative focus, close links to translation, and social dimension supporting national and cultural identity, most terminographic projects have been undertaken in the public sector. Termbases tend to be funded by and designed for public interests such as for language planning (particularly for minority languages) (Temmerman 1997: 53; Rey 1995: 51), for translating government documents, and for 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 (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. Furthermore, termbases designed to serve purposes that are not revenu-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 therefore contain data categories[17] that would be considered impractical or unnecessary in commercial settings. Conversely, they may lack data categories that are important for producing commercial content, such as certain sub-setting values for tracking project-specific terms, and metadata required by controlled authoring and computer-assisted translation software.

17 For information about terminological data categories, see ISO 12620, as well as the ISO TC37 Data Category Registry: www.isocat.org

In 1990, Sager remarked that the large institutional termbases 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). Two decades later, with the advent of NLP applications requiring terminological resources, even in commercial settings (see the next section), 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 which can be used to support the needs of NLP (p. 609). He cites as examples 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 sense disambiguation). He also notes that few termbanks record collocations. Condamines (2007a: 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, it is extremely rare to find a commercial termbase that contains the required semantic relations. We maintain that this is a legacy of the translation focus, which does not emphasise semantic relations.

Rey (1995: 164) notes that large termbases were developed by international organisations like the UN or UNESCO and regional organisations like the European Union and NATO. As with official multilingual states such as Canada, these organisations require substantial translation services and developed large termbases to support this need. (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. Rey feels that this 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, commercial applications of terminological resources were not considered in the design of the major institutional termbases of the world. These termbases, therefore, may not be suitable models for commercial settings.

1.7.3 The need for terminological resources to serve multiple purposes

Terminological resources are not yet widely developed and managed for commercial purposes (Lommel and Ray 2007: 24, 29). Considering the number of global companies and organisations on our planet, few have a termbase (Warburton 2001a; Schmitz and Straub 2010: 87; Lombard 2006: 155). For those that do, the data is often used exclusively to support the translation process. Other potential applications, such as content authoring, are rarely considered (Warburton 2001a; Lombard 2006: 156). A Web site is usually provided where employees can find information such as the definition of a term, or its target language (TL) equivalent. Terminologists in industry refer to this approach as a *pull* approach because it relies on the employee's initiative to enquire about the term. Conversely, source terms and their TL equivalents may be directly integrated in computer-assisted translation (CAT) tools that are used by translators. In this environment, a *push* approach can be used, whereby the terms are automatically shown to the translator when needed; however, considering the industry at large, relatively few companies use this approach. Indeed, spreadsheets are still widely used for developing and storing terminologies (Fidura 2013: 2; Lommel and Ray 2007: 35-38; Lommel 2005: 1, 4).

Hence, the notion of managing terminology in commercial environments has been largely confined to the needs of translation. The tools used for translation are having an impact on terminographic methods in this environment, particularly with respect to what metadata the tools require, what forms the terms themselves must take in order to work properly in the tools, and what selection criteria produce the optimal set of terms for such applications. Our research provides concrete examples of this tool-driven methodology. For example, certain types of variants may not be needed in the autolookup function of a CAT tool if that tool uses fuzzy-matching when looking up terms, whereas other lexical constructs, which are <u>not</u> terms according to conventional theory, may be required to compensate for the limits of TM. Since managing terminology in commercial environments is constantly subject to budgetary pressures and business justification, these types of considerations are fundamental for developing viable terminographic methods that are aligned with business goals.

Yet two decades ago, Meyer (1993: 146) predicted that machines would become a user of terminological data.

> It is predicted that machines may become a category of user for terminology banks; machine translation tools, expert systems, natural-language interfaces to databases, and spelling checkers are just a few of the most obvious applications. (...) Machines will need very large quantities of explicitly represented conceptual information since they do not possess much of the basic real-world information that humans know implicitly.

Ibekwe-SanJuan et al (2007: 2) consider commercial applications of what they call "terminology engineering."

> Applications of terminology engineering include information retrieval, question-answering, information extraction, text mining, machine translation, science and technology watch, all are external fields that can benefit from the incorporation of terminological knowledge.

They further note that terminological resources are useful for building other types of language resources such as ontologies and aligned corpora. Numerous works describe the role of controlled sets of terms for indexing (Buchan 1993; Cabré 1999-b: 51; Strehlow 2001-a; Jacquemin 2001: 305; Nazarenko and El Mekki 2007; Greenwald 1994; Condamines 2007a: 138, 141). Strehlow (2001-b: 433) further notes that a sound strategy for the use of terms in strategic areas of content (such as titles, abstracts and keywords) can lead to significant improvements in information retrieval. Echoing the SEO experts cited earlier, he even recommends the alternating use of synonyms to maximise retrieval. Such a recommendation is strictly contrary to the GTT, as we will show later. Cabré states quite simply that "terminology is a key element for representing the contents of documents and gaining access to them" (1996: 29) and that "it is through terminology that we retrieve information" (1995: 6). Rinaldi et al demonstrate how semantically-structured terminology resources can improve the performance of question answering systems (2003). Wettengel and Van de Weyer (2001: 458) describe how terminological resources can help build product classification systems. Wright and Budin provide an extensive list of language engineering implementations in which terminology management plays a key role (2001: Infobox 31). Ahmad (2001) describes the role of terminological resources in developing various systems for artificial intelligence and knowledge acquisition.

It would be short-sighted not to recognise that commercial enterprises would be interested in such systems and applications. Indeed, Condamines notes that the need for terminological resources that support business processes such as content retrieval and content management is urgent and growing (2007a: 134). She also notes that two types of NLP applications are used in the workplace, one for information extraction or retrieval and the other for knowledge management, and that both types require company-specific lexicons to work properly (2010: 40). Bourigault and Jacquemin claim that research into developing corpus-based tools for building terminology resources is completely driven by needs for more effficient content management systems in industrial enterprises and institutions (2000: section 9.1.1). They state directly: "pour produire, diffuser, rechercher, exploiter et traduire ces documents, les outils de gestion de l'information ont besoin de ressources terminologiques" (translation: to produce, distribute, research, make use of, and translate documents, information management tools need terminological resources." They add, however, that due to lack of foresight, such terminological resources are rarely available in the form required.

According to Condamines, it would seem that this notion that terminology work should be application-driven is gaining ground:

> Most researchers now agree that the usage and control of terminology vary according to the application (translation, knowledge representation, information extraction...). Moreover, with the development of text engineering, applications are becoming progressively more numerous and varied. (2010: 34)

In his paper on term inclusion criteria, Martin (2011) notes that in commercial environments there are multiple potential consumers of terminological resources:

- A machine translation program or programmers
- A writer
- A human translator
- An information architect creating a knowledge network
- A style manager or editor ensuring correct term usage

Numerous works in the literature describe various applications of terminological resources: automatic book indexing, indexing for search engines, ontology building, content classification, contact record analysis, search engine optimisation (query expansion and document

filtering), federation of heterogeneous document collections, information retrieval, document summarisation or abstraction and keyword extraction, product classifications, and automated construction of domain taxonomies and ontologies (Park et al 2002; Jacquemin 2001; Oakes et al 2001; Cabré 1999-b; Bourigault and Slodzian 1999; Condamines 2005). And Bourigault and Slodzian noted quite some time ago that these needs are increasing due to factors such as technological innovation, internationalisation, and the growth of electronic publishing and the Internet (1999: 29).

While it is beyond the scope of our research to define the terminographic requirements of such a wide range of potential applications, we nevertheless wish to emphasise that terminological resources do have multiple applications, and thus need to be repurposable.

Clearly, approaches to terminography will vary according to the ultimate uses to which the terminological resource will be put (Bourigault and Slodzian 1999: 30). In our research, we consider the applications where terminological resources are used in companies, the people who use them, and how they are being used. Given that terminographic approaches adopted in a company will necessarily be driven by other production-oriented processes (authoring, translation, content optimisation, and so forth), these processes may provide an explanation for the prevalence of certain types of terms and metadata in the termbases in this study.

We will consider how the terminology in our sample data meets the purposes for which it is required. Our investigation will consider two aspects: term types and metadata. What types of terms need to be captured and recorded for what specific types of uses? What properties of terms need to be recorded and managed for specific end-uses? Such information is valuable for establishing term selection criteria and data models for termbases that meet the needs of commercial enterprises.

The metadata (data categories) and the types of terms in commercial termbases[18] may provide clues about the processes at play. For instance, in the case of companies that have deployed a controlled authoring software such as Acrolinx or Tedopres, one might expect

18 We use the term *commercial termbase* to refer to termbases that are developed by companies.

some words from the general lexicon to be included in the termbase, as well as data categories indicating preferred usage. We therefore are interested in examining the data categories and the types of terms in termbases, to:

1. Establish a set of common properties shared by commercial termbases.
2. Identify types of metadata present for specific purposes, such as controlled authoring.

1.7.4 Improving term identification

To summarise, compared with other vocations in the language industry, terminography is uncommon. The field of terminology is perceived as an academic discipline whose commercial applications remain largely unrecognised. The established theories and methodologies have an unbalanced focus on normalisation, 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.

Given these factors, it is hardly surprising that terminology management has made little headway into the business world. Despite studies showing that in specialised domains, the most frequent mistakes in translated content are terminology-related (Woyde 2005, Wright 2001a: 492), few companies manage their terminology (Warburton 2001a; LISA 2007: 24).

How can we narrow the apparent divide between terminology and its relevance to commerce? Answering this question requires a critical review of the 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 how 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. The question in its entirety is beyond the scope of the current research.

A good starting point may be to improve our understanding of the nature of terms found in commercial corpora as determined through empirical observations of these terms in their natural contextual environment. Further, if we can observe differences between the terms found in the corpus and the terms recorded in the termbase of the same company, we may be able to identify inappropriate techniques of term identification and collection, which reduce the effectiveness of the terminological resources

CHAPTER 2 LITERATURE REVIEW

Our literature review focusses on two fundamental questions for our research: What is an LSP? and What is a term? We explain why these questions are relevant in the next sections. Because our research investigates the link between commercial termbases and corpora, we also describe how corpora have been portrayed in the literature about terminology.

2.1 Terminology and LSP

Our literature review begins with the question of how the concept of LSP (language for special purposes)[19] is perceived in the field of terminology. We are seeking to justify our claim that the language used in a commercial setting frequently qualifies as an LSP. It is fundamental to justify this claim because terms are typically described as existing within the confines of LSPs (for example: 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). Indeed, as noted earlier, *terminology* is defined by ISO TC37 as a "set of designations belonging to one special language"[20] (our emphasis). As Picht and Draskau (1985: 21) say, "It becomes increasingly clear in any discussion of LSP that one is in many instances referring to terminology. It is equally clear that terminology is part of LSP." If the language used in a commercial enterprise is not an LSP, then one could conclude that it is not a source of terms. We will dispel this potential challenge.

2.1.1 The role of subject field

LSPs 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

19 The expansion of *LSP* is sometimes *language for specific purposes* (Flowerdew 2011, Fuertes-Olivera and Arribas-Bano 2008) and sometimes *language for special purposes*. Pearson (1998) and Kittrredge and Lehrberger (1982) use *sublanguage* with the same meaning. ISO TC37 as well as some scholars (Cabré 1999, Sager 1990, for example) use *special language* and *special subject language*. In more recent LSP studies the term s*pecialised communication* is often preferred to emphasise communicative and cognitive aspects as opposed to purely structural linguistic aspects (see for example Schubert 2011, Fuertes-Olivera and Arribas-Bano 2008).

20 From the ISO TC37 termbase: http://iso.i-term.dk

and Pearson 2002: 25) Very often, subject field (also called *subject domain* or simply *domain*) is a key criterion that differentiates LSP from general language (Rondeau 1981: 30; Sager 1990: 18; Kocourek 1982: 26; Kittredge and Lehrberger 1982: 2; Rogers 2000: 8; Bowker and Pearson 2002: 25). Indeed, the notion of subject field is prevalent in the literature about LSPs. As noted earlier, it is also included in ISO TC37's definition of *special language*: "language used in a subject field..."

General language is the collection of words and expressions that, in the context in which they are used, do not refer to a specialised activity (Rondeau 1981: 26, translated)[21]. 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 define *everyday language* simply as the language that does not satisfy the necessary conditions of scientific texts. However, they also add that it contains indexical signs (pronouns, tense markers, and some adverbs) the interpretation of which depends on exra-linguistic context and that the background of the interlocutor requires only logical and "commonplace knowledge" (1982: 229).

Rondeau's definition seems to reflect the position taken by some linguists that 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 (translation of "usuelle" and "quotidienne" in French)[22], special language, and aesthetic language.

Rondeau adopts the term *communications scientifiques ou techniques (CST)* (technical or scientific communication) to refer to the corpora in which terms are found (1981: 16). He adds, however, that the notion of CST extends to the full range of both pure and applied sciences, and to all techniques, technologies and specialised activities carried out by humans (crafts, professions, trades, occupations, hobbies, leisure activities, etc.). For Rondeau, an LSP is that language used in a CST, in other words, in any communication that can

21 In the field of terminology, definitions formulated with negative structures are discouraged (Bowman et al 1997: 217)

22 Guy Rondeau uses "common," Georges Mounin uses "general" and "ordinary."

be characterised as specialised in a broad sense. Cabré (1996: 22) shares this view, including professional domains and industry in the specialised fields where terms are found.

Terms express concepts that can be classified into subject fields, and the collection of terms and other linguistic means of expression 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). Cabré (1999-b: 81) states, "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." For Pearson also, membership in a subject field is an essential characteristic of termhood (1998: 36). L'Homme (2004: 64) agrees: "Le statut terminologique d'une unité lexicale se définit en fonction du lien qu'on peut établir entre son sens et un domaine de spécialité." (Translation: The terminological status of a lexical unit is defined based on the relationship that can be established between its meaning and a special subject field.) Dubuc (1997: 38) defines the notion of *term* in relation to subject fields, "A term is a word or expression that designates a concept specific to a subject field and the corresponding object in the world." Wright (1997: 13) takes a similar view, "terms are words that are assigned to concepts used in the special languages that occur in subject-field or domain-related texts." And Sager (1990: 19) uses "special reference within a discipline" to distinguish *terms* from *words*. Thus there is a consensus that subject field association is a defining feature of terms.

However, the notion of subject field is somewhat imprecise (Condamines 1995: 227). 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). Condamines extends the notion to a "kind of reasoning" (1995: 227). Rey adheres to the classical criterion of subject field for terminology, but he also acknowledges the existence of "empirically structured terminologies for special cases and practical situations" and mentions the "terminology of a firm" as a typical case (p. 144).

> Systematic description requires the definition of a theoretical or practical subject field. But terminology may also envisage a complex empirical

> object. There are many examples of mixed descriptions on the basis of text corpora, for the purpose of e.g. collecting neologisms, or feeding databases.

Dubuc (1992: 43) makes a distinction between both the exclusivity of a term to a domain, and the term's meaning or usage within that domain:

> L'unité terminologique, ou terme, est l'appellation d'une notion propre au domaine étudié soit parce qu'elle appartient exclusivement à ce domaine et qu'elle ne se retrouve dans aucun autre, soit qu'elle fait l'object d'une utilisation particulière à ce domaine.
>
> (Translation: The terminological unit, or term, is the designation of a concept from the domain being studied, either because the term is exclusive to this domain and occurs in no other, or because it has a specific usage in this domain.)

The principle that a term's meaning is exclusive to a domain leads us to realise that the notion of *term* warrants some clarification. This will be provided in Section 2.3.1.

2.1.2 A closed set of linguistic properties

LSPs are also considered as having restricted linguistic properties of various sorts. Note the part "specific means of linguistic expression" in the ISO TC37 definition cited earlier. 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 formalised and codified variety of language." Rondeau (1981: 29-30) also uses linguistic properties to characterise LSPs, including textual characteristics (concision, precision, depersonalisation), lexical patterns (preponderance of nominal structures), dominance of written form over verbal, frequency of figures, and so forth. Sager (1990: 107, 109, 111, 19, 123) likewise refers to properties such as economy, precision, appropriateness, and referentiality, all of which are achieved through certain stylistic norms.

2.1.3 Communicative context and communicative function

Other scholarly works emphasise the importance of the communicative context for LSPs. LSPs occur "within a definite sphere of communication" (Hoffman 1979: 16). Cabré (1999-

b: 63) refers to "interlocutors in a communicative situation." Over time, the communicative aspect appears to be increasingly recognised. Cabré (2003: 188, 190) points out that terms (and by extension LSPs if we acknowledge the intrinsic interdependency between terms and LSPs) have to be studied in the framework of specialised communication, and indeed terms do not even exist "prior to their usage in a specific communicative context." By stating that terminology "essentially belongs to the sphere of parole," as opposed to the sphere of "langue," Kageura (2002: 14, 251) takes the same position[23].

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

Cabré (1999-b: 63) 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. "Taking general language texts as our point of reference, any type of text that varies from this norm can be considered a special language text." She identifies three common characteristics of special languages: limited number of users (which are defined by profession or expertise, see the next section), a formal or professional communicative situation, and an informative function (p. 68).

Pavel (1993: 21) observes, however, that LSPs are becoming less confined to communicative settings that are restricted to specialists and extend to the private sector and industry:

> LSP communication in any field is less and less confined to specialists in the same domain. It now reaches academia across previously disparate disciplines, extends to public administration and the private sector, permeates industry, and sends its message through the mass media to the general public. (our emphasis)

2.1.4 Conscious acquisition

Another differentiating factor identified in the literature is how language competence is acquired. Some scholars believe that competence in an LSP requires effort over and above

23 Here, the terms *parole* and *langue* are equivalent to the terms *performance* and *competence* used by Noam Chomsky (1965).

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). Education as a criterion for users of LSPs is taken up again by Sager in 1990 (p. 105). Picht and Draskau (1985: 11) support this view but, to account for communication acts of a more didactic nature between specialists and initiates, which are also rich in terminology, they simply state that users acquire the LSP "voluntarily."

2.2 Terminology and genre

A few scholars have begun to investigate the relationship between terminology and genre, as an alternative dimension to LSP. There is a wealth of scholarship about linguistic genre over the last 30 years. In this section we review some of the scholarly record about genre as it applies to terminology and LSP. We discuss its relevance for the current research in section 3.5.

Scholars in genre analysis define genre in varying ways, but a frequently-cited definition is that of Swales: "A genre comprises a class of communicative events, the members of which share some set of communicative purposes" (1990: 58). The class of sub-languages referred to as languages for special purposes in general is not a genre, but a particular instantiation of LSP communication could be construed as one, a classic example being research articles. Other genre researchers identify sets of genres, or over-arching genres that are associated with a number of different information types and media, such as Bhatia's promotional genre (1993: 156), which includes the sub-genres of sales promotion letters and job applications.

The notion of genre has evolved from a simple categorisation of text types to a framework that connects kinds of texts to social actions. Bawarshi and Reiff describe five different perspectives of genre in literary traditions which variably emphasise the classificatory role of genre, its role as an agent for shaping, or constraining, literary production and interpretation, and its influence on cultural and social practises (2010: chapter 2). They then move on to linguistic traditions, including genre analysis for English for Specific Purposes (ESP).

ESP is the English-specific notion of LSP. In Bawarshi and Reiff, ESP generally refers to the studying and teaching of specialised varieties of English, rather than to those specialised varieties themselves. The focus of genre research in ESP has been on academic and research English, and on applying genres for the purpose of language instruction. Inspired by genre analysis in corpus linguistics, genre for ESP involves quantifying the linguistic properties of language varieties. Further, however, ESP genre analysis applies knowledge about linguistic properties to elucidate the role of genre in social context and communicative function. In ESP, a genre is a collection of discourse acts sharing certain properties, occurring within a given context, having a communicative purpose, and involving a discourse community.

Swales' influential work on the genre of academic English (1990) established discourse community as a defining element of genre. He proposes six defining characteristics of discourse community, summarised as follows: a discourse community has (1) a set of common goals and (2) mechanisms of intercommunication; (3) it uses these mechanisms and (4) one or more genres to communicate; (5) it has acquired a specific lexis, and (6) it has members with varying degrees of expert knowledge (1990: 24-27). Requirement number 5 is of particular interest for this research. Swales even cites the information technology discourse community as exemplary with respect to criterion five (p. 26), and acronyms and abbreviations -- common in commercial texts -- as exemplary members of a specific lexis.

Genres are frequently characterised by differences in the use of various linguistic devices. Some of these devices can involve the lexicon. Bhatia, for instance, notes an above-average use in the advertising genre of what he calls complex nominal phrases (a head noun preceeded by a series of adjectives, such as *cordless, lightweight, durable drill*), whereas in the academic scientific genre, the compound nominal phrase (head noun preceeded by other nouns and possibly adjectives, such as *nozzle gas ejection*) is more predominant (1993: 148). He is not clear, however, about whether the same concept can be expressed by these different structures, in other words, whether the different lexico-grammatical choices adopted in different genres constitute semantically-equivalent terminological variants or

whether they reflect shifts in meaning or focus akin to those described by Sanchez, Bowker and Meyer under the notion of multidimensionality (see section 2.3.3). We suspect that both phenomena occur.

Motivated by the need to automatically build ontologies or knowledge bases from LSP corpora, works by Anne Condamines focus on conceptual relation patterns and their variability in different genres (for example: 2007b, 2008a, 2008b). Conceptual relation patterns are words or sequences of words that indicate a relation between two nouns, for instance, the words "such as" may indicate hyperonymy, as in "A flower such as the rose..." She discovered that certain conceptual relation patterns are more likely to occur in specific genres, and that the meaning of a given conceptual relation pattern can differ in different genres[24] (2008). Consequently, any attempt to automatically build knowledge bases from a corpus should use a corpus that has been carefully and accurately structured by genres.

In 2000, Rogers studied the frequency of a selection of terms in the field of automotive engineering across different genres (book, professional journal, popular science, newspaper, and advertisement). She found that the choice of terms did not vary significantly across genres. She concludes that there is "a relatively broad distribution of terms across genres rather than an exclusive use of particular terms in particular genres" (p. 15). These findings are corroborated by Freixa who observed that among the various causes of terminological variation, functional causes (register, genre) are the least productive (see section 2.3.3).

2.3 What is a term?

As we have shown, compared to genre, the relationship between terminology and LSP is more extensively documented. If we can justify that the language used in a company is an LSP, we feel confident in assuming that it contains terms. The next question is whether or not terms found in commercially-oriented texts have unique properties when compared to the conventional view of the notion of *term* found in the literature. Differences in properties

24 For certain conceptual relation patterns, subject field (domain) also has an impact on the frequency and meaning of the pattern. (Condamines 2008: 134)

are interesting for our research, as they can be used to re-examine conventional theories and methodologies in the field of terminology to determine their suitability for commercial environments.

For example, Pavel (1993: 23) notes that in LSPs, neosemanticisms predominate over formal neologisms for expressing new concepts, and that polysemy is a common phenomenon. Kageura (2002: 254) observed a considerable degree of systematicity in the formation patterns of terms in the field of documentation (library science). These kinds of studies lead us to ask whether any patterns can be observed for terms in the corpora and termbases that we are studying, which were produced to serve commercially-driven processes.

The most frequent task carried out by terminographers in commercial settings is term identification, sometimes performed with the aid of a term extraction tool (Warburton 2001a: 8, 12, 14). Having a clear definition of what constitutes a term of value for the people, processes, and applications that use the terminology enables terminographers to focus on those types of terms, thereby ensuring that their work is cost-effective. Does the conventional notion of what constitutes a term hold for commercial terminography, or does it need redefining? We shall come back to these questions later.

Prior to studying the properties of terms from our sample corpora, we therefore need to review the conventional interpretation of what constitutes a *term*. This interpretation will then be used as a reference point when examining terms from our commercial corpora. To what degree do the terms adhere to this interpretation? How do they depart from it?

2.3.1 Disambiguating *term*

Before reviewing the various interpretations of what constitutes a term, we need to clarify an ambiguity in the use of this word, particularly, how a term relates to a concept. The word *term* is often used ambiguously[25]. Consider the following citation from Dubuc (1997: 40):

> A single *term* can be used in several different fields, but the concept it covers changes in each one. (...) For the terminologist, these *terms* are distinct because

25 For a complete article on this issue, see Riggs, 1994.

> the concepts and realities they designate are distinct. Strictly speaking, a *term* belongs to a single subject field.

A contradiction in the uses of *term* can be observed in this quotation, specifically between the first and last sentences. In the first, it means *word form* without any restrictions on the number of meanings that this word form can have, in other words, it refers to any potentially polysemous lexical unit (from LSP, of course). This is a usage of term that is characteristic in lexicography (Riggs 1994: 65). In the other instances, it means term in compliance with the univocity principle, that is, a single lexical form having a single meaning; this explains the use of the plural form in the second sentence.

Cabré (1999-b: 40) explains that lexicography and terminology adhere to two different views about the term-meaning relationship: "The polysemy of the common lexicon is treated as homonymy in terminology."[26] This difference in perspective is fundamental: lexicography recognises the potential of one word to have multiple meanings, whereas terminology views each instance of meaning as a different word[27]. This difference in perspective may also explain why in terminological resources homonyms are more frequent, and conversely, polysemes are less frequent, as compared to lexicographical resources (Cabré 1999-b: 111; Fuertes-Olivera and Arribas-Bano 2008:21; Meyer and Mackintosh 1996: 263). Where terminology sees homonymy lexicography sees polysemy.

Throughout the literature and in common usage (even among classically trained terminologists), many instance of *term* being used in the sense of *word form* can be found. A classic example is that of Gilreath (1994), who discusses the semantic valence of terms in great detail, as if polysemy were an inherent property of terms. Pearson (1998) also frequently uses the word *term* to refer to a lexical unit in its capacity to assume multiple meanings in different subject fields. She takes the position that "a term may be polysemous in different domains but each of its meanings must refer to only one domain" (p. 25). Adhering more closely to the terminology view, Meyer and Mackintosh articulate the same phenomenon differently, "most terms have only one meaning within a given domain" (1996: 24). To

26 She continues this discussion on page 108-112.

27 Here, for simplicity and consistency purposes in this statement, we are using "word" but we are actually also referring to multi-word terms.

specifically refer to a term having only one meaning in a domain, Pearson proposes "subject-specific term" (p. 25). While we do not dispute polysemy as a valid phenomenon in terminology, we wish to clarify our position with respect to the meaning of *term*.

The notions of *signifier*, *signified*, and *sign*, taken from Saussurian linguistics (De Saussure 1916), can help to disambiguate the concept of *term*. The signifier is the phonetic or graphic representation of the concept, the *lexical form*, so to speak. The signified is the concept that the signifier evokes in our mind. The bound relationship between the two is referred to as the sign; thus a given sign has a given physical (or auditory) representation (signifier) and a given conceptual reference (signified).

For the purposes of clarity, we prefer to adopt a purist usage of the word *term*, that is, a term is a *sign* in the Saussurian sense. A term has only one meaning. Two identical signifiers that have two different signifieds correspond to two distinct terms (assuming, of course, that other criteria for terms are satisfied, such as subject-field specificity). This interpretation is in line with that of Rondeau, Dubuc (1992: 26) and other classic theorists. Rondeau uses the term *denomination* to refer to the signifier and *concept* (translation of *notion* in French) to refer to the signified. He states (1981: 21-23):

> Ce qui distingue le terme des autres signes linguistiques, c'est d'abord que son extension sémantique se définit par rapport au signifié plutôt que par rapport au signifiant. (...) Pour une notion donnée, il y a, théoriquement, une dénomination et une seule. (...) Un terme constitue un couple dénomination-notion clairement identifié par le contexte. (...) Le rapport qui s'établit entre une dénomination et une notion est monoréférentiel, c'est-à-dire que *pour un terme donné*, à une dénomination correspond une notion et une seule.
>
> (Translation: What distinguishes terms from other linguistic signs, first and foremost, is that a term's semantic extension is defined in relation to the signified rather than to the signifier. (...) A term instantiates a binary relationship between a denomination and a concept which is clearly identifiable by the context.(...) The relationship between a denomination and a concept is monoreferential; *for a given term*, a denomination has one and only one meaning.)

What is also noteworthy in this interpretation is that a term's meaning is only discernible in context. Dubuc (1992: 26) describes this condition as follows:

> Un même terme peut-il appartenir à plusieurs domaines? Bien sûr, on peut

> trouver formellement un même terme dans plusieurs domaines. (...) Cependant pour le terminologue, il s'agit là de termes distincts sans rapport direct entre eux. Strictement, donc, un terme n'appartient qu'à un seul domaine d'emploi.
>
> (Translation: Can a given term belong to several subject fields? Yes of course, one finds the same term, formally speaking, in several subject fields. However, for the terminologist, what we have in this case are distinct terms that are not directly related. Strictly speaking, a term belongs to only one subject field.)

Recognising the ambiguity of *term* that we have just identified, Riggs (1989, 1994, 1997) proposed some new terminology to clarify these concepts. He uses *vocable* to refer to the *lexical form* (signifier). “A vocable can be used to designate two or more concepts in the same subject field” (1994: 69). A vocable that has more than one signification is “equivocal” (1994: 71). Alternatively, an *unequivocal term* is any expression used to represent only one concept within any given subject field, i.e. it is monosemic (p. 72). Thus, Rigg's *unequivocal term* corresponds to Rondeau's *term* and also to the notion of term in traditional terminology. We find Rigg's proposals of vocable and unequivocal to be appealing, as they overcome the awkwardness of using *word form* when in fact many terms comprise more than one word, and the need to use the more technical and archaically-perceived *signifier*. However, Rigg's terminology has never taken hold in the field.

The terminological perspective, whereby each different meaning of a given lexical form corresponds to a distinct term, explains and justifies concept orientation as a fundamental design principle for termbases. We will later demonstrate that this principle, although not always adopted, is critical for developing multi-purpose terminological resources for commercial applications. In this interest, a distinction needs to be made between the univocity principle as a criterion for so-called valid terms, and its value as a guiding principle for termbase modelling. We have previously shown that the univocity principle in terminology has been challenged (we will provide further evidence of this in the following sections). We support those challenges because polysemy and homonymy are observed in commercial texts and are rarely seen as a problem (provided that they are not excessively widespread). However, in conjunction with the focus of terminology on concepts, the univocity principle motivated the terminographic practise of concept orientation for terminological resources.

Thus, while the univocity principle cannot be artificially imposed to prevent a lexical form from having multiple meanings, thereby resulting in different terms that share the same surface form, at least, each of these identical-looking terms must be treated in an independent structure (entry) in a termbase.

2.3.2 Theoretical interpretations

As was the case with the concept of LSP, there are different opinions as to what constitutes a term. These differences reflect the respective theories of terminology that the scholars espouse. Our exploration of the meaning of *term* needs to therefore include an overview of the various theories of terminology. Although the views differ, some basic tenets supported by all theories can be identified. Let us briefly review the key theories of terminology and their interpretations of *term*.

We first describe the original theory for terminology, and then theories that emerged in reaction to the original theory. The latter theories are more corpus-based partially due to advances in NLP technologies. They emphasise communicative, cognitive, and lexical aspects (notably by Cabré, Temmerman, and L'Homme respectively).

2.3.2.1 General Theory of Terminology

The original theory is known as the General Theory of Terminology (GTT)[28] (Picht and Draskau 1985: 29; Felber 1984: 96-97). It is also referred to by various scholars as the Vienna school (Temmerman 2000: 3; Cabré 1999-b: 2; Picht and Draskau 1985: 31), the Wusterian theory or approach (Lara 1998, Roche 2012, Temmerman 2000: 231), traditional terminology (Temmerman 2000: 1) and the traditional theory or classical theory (L'Homme 2005). Eugen Wüster, an engineer, developed this theory while preparing a multilingual dictionary of machine tools (Wüster 1967). According to the GTT, objective communication is achieved by fixing the relationship between terms and concepts. This theory therefore favours biunivocity, whereby a linguistic form corresponds to one and only one concept, and a concept is expressed by one and only one linguistic form (L'Homme 2004:

28 It has also been called the objectivist theory and the conceptual theory.

27). The focus of study is the concept, to which terms are secondarily assigned as designators (Cabré 2003: 166-167; L'Homme 2005: 1114), however theoretically difficult that may be (Rey 1995: 145-146). Concepts occupy fixed positions in a language-independent concept system, where they are hierarchically related to other concepts. The approach is onomasiological (delimit the concept first and foremost, then find terms to denote it), and the goal is normalisation (L'Homme 2005: 1114-1115; Cabré 1996: 25).

Until recent years, the GTT dominated the field of terminology, and it continues to do so in educational settings (L'Homme 2005: Note 2; Alcina 2009: 7; Cabré 2000: 40) and in normalisation settings (such as for language planning and standards development). It is the theory espoused by ISO TC37. Works based on the GTT include Felber (1984), Rondeau (1981), Dubuc (1992) and Picht and Draskau (1985), and to some degree Cabré (1999-b)[29] and Rey (1995)[30]. The following citations about the concept of *term* are taken from several of these works.

Dubuc (1992: 25):

> Le terme est l'élément constitutif de toute nomenclature terminologique liée à une langue de spécialité. On peut donc le définir comme l'appellation d'un objet propre à un domaine donné.
>
> (Translation: A term is an element of any terminological nomenclature that is connected to an LSP. One can therefore define it as the designation of an object from a given domain.)

What is noteworthy here is the reference to a nomenclature, which is a highly structured and invariable set of terms reflecting a concept system, and that a term is a designation of an object.

Rondeau (1981: 21), alluding to Saussurian linguistics (De Saussure 1916):

29 Cabré devotes a large portion of her monograph to describing and justifying many of the tenets of the GTT, but concludes with arguments for a new orientation which stresses social, pragmatic, and communicative aspects. Some recognise these arguments as a new Communicative Theory of Terminology (CTT).

30 Although Rey states (1995, p. 116) "I am not a true Wusterian," his monograph *Essays on Terminology* adheres largely to the principles of the GTT. Rey does not explicitly state where his position diverges from that theory. However, being a lexicographer himself, he does acknowledge the lexical dimension of terminology more than a classical General theorist would.

> (…) la partie signifié qui compose le terme se définit par rapport à un ensemble de signifiés appartenant au même domaine.
>
> (Translation: The signified part of the term is defined in relation to a collection of signifieds belonging to the same domain.)
>
> (…) sur le plan logique, le terme trouve sa place dans une structure hiérarchique notionnelle à l'intérieur d'un domain.
>
> (Translation: At the logical level, a term is located in a hierarchical conceptual structure within a domain.)

Terms thus exist only in being conceptually related to and distinguishable from other terms within the same concept hierarchy.

Picht and Draskau (1985: 95):

> The characteristics of the term which distinguish it from the non-term are precision and the fact that it belongs to a system of terms (reflecting a system of concepts).

Thus, according to the GTT, terms are designations of objects the conceptualisations of which can be classified systematically. A terminology (set of terms) must correspond to a conceptual system (Rey 1995: 140).

2.3.2.1.1 Criticisms of the General Theory

Since the mid 1990's, the GTT has been subject to significant criticism (L'Homme 2005: 1115). The main critique is that it does not take into account language in use (see Pearson 1998, Temmerman 1997 and 2000, L'Homme 2004, Cabré 1999-b, Collet 2004). Temmerman (2000: 21) puts it simply: "Traditional Terminology does not have a theory of communication." She criticises the GTT for not recognising the role of natural language (p. 60): "In traditional Terminology, natural language is treated as a necessary evil which needs to be constricted." Collet concurs, "The objective (of Traditional terminology) has its roots in the positivist belief that natural language possesses characteristics which are likely to constitute an impediment to clear and precise communication" (p. 100). Concepts are studied outside of their use in communicative settings. Terms are considered at the level of langue,

and not of parole. This criticism is also shared by Kageura (2002: 251), among others. Indeed, long before recent challenges, Rondeau declared quite emphatically that terms are actualised at the level of parole, not of langue, and that only in contextualised discourse can one determine whether a given linguistic unit is a member of an LSP or of LGP (1981: 28).

Pearson (1998: Section 1.9) notes that the relationship between the interlocutors in the communicative act has a bearing on termhood, i.e. whether they are both experts in the subject matter, or whether the text is written by an expert for a learner and thus is more didactic in nature. L'Homme (2004: 54) further observes that the notion of *term* differs according to the needs of users. For a translator it takes on a pragmatic meaning:

> Pour le traducteur, le terme se confond avec la notion d'unité de traduction faisant problème, c'est-à-dire une unité dont le sens n'est pas clair ou dont l'équivalent n'est pas connu.
>
> (Translation: For the translator, the notion of term extends to any problematic translation unit, that is, any unit that is not clear or for which the target language equivalent is not known.)

L'Homme (2004: 27) goes on to assert that the GTT is not a theory at all, but rather a specific objective that determines a set of methodologies:

> La perspective de normalisation est si ancrée dans la démarche classique qu'on a même cherché à la théoriser, mais il s'agit bel et bien d'un objectif que s'est donné la terminologie classique et non d'un principe théorique véritable.
>
> (Translation: The standardisation perspective is so entrenched in the traditional approach that there have even been attempts to formalise it into a theory. In reality, standardisation is simply an objective of traditional terminology, rather than a true theoretical principle.)

Temmerman (2000) makes similar observations:

> The mistake made by traditional Terminology was to proclaim the standardisation principles as the general theory of terminology. (p. 220)
>
> 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. (p. 37)

Cabré (2003: 167) puts it succinctly:

> Wüster developed a theory about what terminology should be in order to ensure unambiguous plurilingual communication, and not about what terminology actually is in its great variety and plurality. (our emphasis)

The advent of NLP technologies such as term extraction tools, the availability of large machine-readable corpora, and the shift from structuralist linguistics to corpus linguistics and cognitive linguistics, influenced scholars to critically examine the GTT with an interest in what discoveries the use of corpora could bring. Indeed, some scholars deem that the GTT lacks empirical foundation and that these shifts in perspective brought about by technological innovation and theoretical maturation call for a new empirically-inspired theory.

> Since the tenets put forward by traditional terminology do not equip the term with features that result in the required behaviour in discourse, it can be concluded that these tenets are not borne out by empirical data. Consequently, there is a need for a theory of the term that respects the facts of special language communication, thus for a theory that is empirically adequate. (Collet 2004: 102)

In questioning some of the basic tenets of the GTT, a number of alternative views on terminology emerged. Several of these theories are briefly presented in the next sections.

2.3.2.2 Socio-cognitive Theory

Following Sager's discussion about the communicative, cognitive, and linguistic dimensions of terminology (1990), Temmerman (1997, 2000) developed the Socio-cognitive Theory in reaction against the GTT. Based on cognitive semantics, this theory emphasises the role of human experience in the formulation of concepts, and thus, claims that many concepts are, in fact, language dependent. LSPs are viewed as a part of language as a whole complete with its functional paradigm. Hence, in line with Halliday's theory of Systemic-Functional Linguistics, LSPs have both experiential and logical ideational metafunctions as well as textual and communicative (interpersonal) ones (Halliday and Webster 2009: 253).

In Temmerman, Kerremans, De Baer (2010: 187), a term is a natural language representation of a unit of understanding, considered relevant to given purposes, applications, or groups of users. Here, the use of “unit of understanding” instead of *concept* is a reaction against the GTT. And by using “natural language representation,” they emphasise the

communicative aspect. Finally, “relevant to given purposes” suggests that termhood is purpose-dependent. Note that any reference to subject field or LSP is curiously omitted.

Thus, the Socio-cognitive Theory views terms as expressions of meaning that are dependent on the context of communication. The communicative situation distinguishes terms from units of the general lexicon.

2.3.2.3 Lexico-semantic Theory and Textual Terminology

A lexico-semantic approach to terminology emerged in Canada, led by Marie-Claude L'Homme. Inspired by the Meaning-Text Theory developed by Igor Mel’čuk (1995), this approach considers terms first and foremost as lexical units. Terms are studied in their linguistic environment to determine their lexical properties and behaviours particularly in relation to other lexical items with which they co-occur in corpora. The focus is on lexical structures rather than conceptual ones.

In her treatise on what is a term (2005: 1123), L'Homme adheres to the principles of a “textual terminology,” based on the ideas of Bourigault and Slodzian (1999) and Slodzian (2000), whereby a term is a construct that takes shape through an analysis which gives consideration to corpus evidence, validation by subject-matter experts, and the purpose of the terminographical product (L'Homme 2004: 25). The text is the starting point for all terminological investigation (Roche 2012). The works of Anne Condamines relating terminology closely to corpus linguistics, and specifically to corpus semantics, also take up this notion of textual terminology (2005, 2007a, 2007b: 44). For Condamines, textual terminology is similar to discourse analysis in that analysing a text results in a lexical “construct” that is itself subject to interpretation (2005: 43). One is in effect constructing a system of lexical units from a set of observed textual utterances, and it is difficult to affirm that this lexical system corresponds to the terminology of the domain, purely speaking. Rather, the system of lexical units established through textual observation is first and foremost associated with the interlocuteurs and the purpose of the communicative act. Condamines thus suggests that the lexical units of investigative interest in textual terminology are determined by role and purpose-based criteria.

Adhering to similar viewpoints, Collet redefines the notion of term within the theoretical framework of text linguistics (2004), which she calls the "text-linguistic approach" (to terminology) (p. 109). Adopting the terminology of Halliday and Hasan (1976), she argues that terms contribute to a text's "texture," i.e. the properties of coherence and cohesion that allow the text to function as a unit (p. 103). Terms bring coherence as carriers of "meaning content," which can vary laterally (from language user to language user) and vertically (through time) (p. 106). These differences in meaning can be contextually and situationally dependent, or they can lead to polysemy. This capacity of a term's meaning to change is another challenge to the univocity principle. Terms contribute to cohesion through various lexical means including repetition and terminological variation (terminological variation is discussed in sections 2.3.3 and 3.2.3). These two instances of variation -- semantic and syntactic -- are contrary to the univocity principle of the GTT.

> In texts for specific purposes, the term, a semantically charged linear structure, exhibits variability on the level of both its meaning content and of its linear structure. Traditional terminology finds itself at odds with these phenomena. (p. 108)

Ibekwe-SanJuan et al (2007: 1-2) also view terms essentially as text units, but consider a term's relevance to an application of utmost importance: "In an application-oriented framework, a term designates the meaningful text unit in a specialised discourse considered useful for an application" (our emphasis). Note the similarity here to Temmerman's "relevant to given purposes."

The literature affirming *textual terminology* as an independent theory or set of methodologies is not yet abundant. Condamines dates its beginnings to the mid 1990's (2010: 30; 2007b: 44), precipitated by advances in corpus linguistics. Bourigault and Slodzian's paper that appeared in 1999 is cited by other scholars as setting the stage for further enquiry. Textual terminology appears thus to be in its early stages of development as a recognised approach to terminology. The focus is on bringing terminology into the sphere of corpus linguistics and on demonstrating the validity of corpus-based discovery and validation of terminological units. We have not found sufficient evidence in the literature to confirm the existence of a consolidated textual terminology theory and methodological framework[31].

31 This is why we have not dedicated a separate section for this approach.

2.3.2.4 Communicative Theory

In an attempt to ground the postulates of the GTT to the dynamic realities of communication, Cabré (1999-a) formulated a series of reflections that are the foundation of the Communicative Theory of Terminology (CTT). The CTT adopts a linguistic approach to identifying concepts; the objects of study are terminological units as part of natural language. Terms are first and foremost lexical units that, through pragmatic conditions and communication situations, can assume a terminological value. While the CTT recognises the existence of conceptual structures as a framework where terms can be formalised as designators of concepts, it allows these structures to be more loosely defined according to the criteria established in a terminology project.

Rather than limit the object of its study to the purely scientific or technical lexicon, as the GTT does, the CTT recognises different levels of specialisation of the lexicon for serving different communicative purposes. The CTT adopts a descriptive approach to terminology, with an emphasis on the linguistic properties of terms, in order to develop different kinds of terminological resources for different purposes.

2.3.3 Views on variation

The notion of *variation* in the scholarly record relating to terminology is pluralistic (Cabré et al 2005: 12; Daille 2005: 182). For Condamines (2008b, 2010), it encompasses a range of phenomena: variation of meaning, of conceptual relation markers relating to genre, of terminological resources for different applications, of terms diachronically, of language conditioned socially, in other words, virtually any change affecting terminology both linguistic and otherwise. For most scholars, terminological variation refers to the production of linguistic *variants* of terms. However, the definitions of *variant* itself vary according to the end use or application of the terminological resource (Cabré et al 2005: 12).

On the one hand, there is the viewpoint that a variant need not necessarily have the same meaning as the term of which it is considered to be a variant. Ibekwe-Sanjuan defines

variation broadly as "changes affecting the structure and the form of a term producing another textual unit close to the initial one" (1998: 1). Her research focusses on syntactic variants which, according to this definition, can differ in meaning (for example: *root hair deformation* and *deformed root hair*). Citing evidence from previous research, Collet observes a predominance in LSPs of multi-word terms that exhibit "syntactic transparency," and points out that their complex linear structure gives them a propensity towards terminological variation: synonyms, hyperonyms, ellipses and reduced forms. She calls each set of such related forms a "paradigm" (2004: 101, 107, 108). Daille (2007: 164) defines *variant* as "an utterance which is semantically and conceptually related to an original term." According to this definition, any term that has even a similar meaning as another could be called a variant. Daille also notes that most researchers investigating terminological variation avoid defining this phenomenon, preferring instead to identify the specific kinds of variation they study (2007: 165), as she herself does in an earlier study: "a variant is a term derived from an existing term through insertion, juxtaposition, permutation and coordination" (Daille et al 1996: 205).

Some scholars describe this partial (as opposed to total) semantic equivalency of variants as a cognitively-motivated phenomenon called *multidimensionality* (Sanchez 2011: 184, citing Bowker and Meyer 1993), whereby the concept denoted by the variant and the main term is essentially the same, but the variant reflects a different perspective or feature. In her typology of the causes of terminological variation, Freixa holds that variation is the phenomenon in which "one and the same concept has different denominations" (2006: 51), and this perspective of semantic equivalence is predominant in terminology (Cabré 2000: 49). "What is required by terminological resources or for translation is that the variant of a term belong to the same semantic class as that term" (Cabré et al 2005: 12).

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

In research oriented towards term extraction, Daille et al (1996) identify the following types of variation: inflectional, graphical (e.g. *air flow* and *airflow*) and orthographic, syntactic, and morpho-syntactic. These categories can be further broken down, for instance, syntactic variation includes insertion and juxtaposition, coordination, and permutation.

We suggest that some of the types of syntactic variation that are described in the context of term extraction would not be considered term variants by a terminologist building a corporate termbase. This again reflects the assertion by Daille and others that the notion of variation is application-dependent. Some syntactic variants that can be observed in corpora reflect stylistic choices at the writing stage, such as avoiding the strict repetition of a term. However, the produced variant may not be sufficiently lexicalised to constitute a distinct term in the eyes of a terminologist. For example, if the term *transmission mode* is in the termbase, it is unlikely that *mode of transmission* would also be necessary.

In terminography, a variant is usually considered to also 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 multi-word 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*), hyphenation (*e-mail* versus *email*), and morphological features (*application program interface* versus *application programming interface*). Variants may also result from differences in case, however, one must distinguish true variants of this type from other uses of case such as for beginning a sentence or a list and distinguishing common nouns from proper nouns. These uses of case are not variants. This interpretation of variants excludes synonyms which do not share any surface form characteristics, such as *football* and its American equivalent *soccer*. Since the term *synonym* refers to <u>any</u> term that has the same meaning as another, and thus includes variants, for clarity purposes we have chosen *lexical synonym* to refer to the latter type.

Freixa observes five basic motivations for variation: dialectal (the authors have different origins), functional (due to different registers or genres), discursive (reflecting stylistic and

expressive differences), interlinguistic (resulting from contact between languages) and cognitive (reflecting different conceptualisations) (2006: 52). She also asserts that discursive causes generate the most variation and functional causes the least.

This is an interesting point, since discursive motivations as described by Freixa would seem to naturally occur in commercial settings. These include a writer's tendency to avoid repetition, and to be economical, emphatic, creative or expressive. Furthermore, the use of synonyms (which includes variants) contributes to lexical cohesion (2006: 60). Freixa adds that the variants most frequently adopted to avoid repetition, realise economy and improve textual coherence in specialised texts are acronyms and other forms of term reductions. As a form of variation, abbreviations, or abbreviated forms, realise economy by shortening the text, without loss of semantic precision, at least in context (Kocourek 1982: 142). This may explain why they are “irresistible” among communicators of specialised texts. Abbreviated forms also exhibit a high degree of conceptual equivalence (Freixa 2006: 62). Cabré observes that abbreviation is common in specialised discourse once a new term becomes familiar (1999-b: 227). According to Rogers (2007: 29), repeating a full term rather than using an abbreviated form could therefore be disorienting for the reader, who assumes that he or she is being given new information. This would result in “overspecification” (p. 30).

In her article, Daille 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 (p. 175). All the aforementioned scholars who 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 the aforementioned theories, they are perceived and handled differently. Prescriptive-based and preoccupied with standardisation, 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 real language, proponents of the remaining theories reject this judgemental position and consider variants, as well as lexical 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 in 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)

Sager even suggests that variants are more prevalent in special language than in general language:

> The means of alternate designation do not differ markedly between general and special languages; because of the higher concentration of reference terms, there may, however, be a higher density of alternate forms in special language discourse. (p. 214)

For the CTT, variation is a fundamental phenomenon in terminology. It plays an essentially pragmatic role in retrieving information, providing writers and translators with real contexts, and demonstrating which terms are actually used in specialised texts in order to allow decisions to be made about term standardisation (Campo et al 2005: Section 4).

The purposeful existence of synonymic variants in LSPs can no longer be denied. For Rogers (1997: 219), they are common in special-language texts "despite the best efforts of standardising bodies." Jacquemin (2001: 6, 215) found that variants account for about one third of term occurrences in English medical corpora, and concludes that variation "is a crucial characteristic of terms." He states (p. 115): "Variation is not a peripheral symptom, it is a pervasive phenomenon in terminological linguistics. Syntactic and morphosyntactic variants represent respectively 25 and 15 percent of term occurrences." Daille (2007:163) reached similar findings: 15 to 35 percent, a figure that is also recognised by Cabré et al (2005: 12). Kerremans' research accepts as "fact" that terminological variation is a widespread phenomenon in specialised discourse (2010: 5). Cabré asserts that variation is "in-

herent in both general and special communication" (2000: 42). Bourigault and Jacquemin view variation as a "massive yet long underestimated phenomenon" in technical and scientific corpora (2000: section 9.3.2.5). Slodzian maintains that variation has been empirically proven as a frequent phenomenon (2000: 69). Marshman et al (2008: 43) conclude that the analysis of synonyms and variants is an important part of terminological description in a specialised field. Variation is thus a common linguistic device in the use of terminologies.

Cabré (2003: 179) notes that the use of variants is dependent on the degree of text specialisation:

> A text of lesser degree of specialisation and didactic function is conceptually more redundant and consequently will contain more variation of designation than a highly specialised text intended for conveying scientific innovations to colleagues at the same level.

Shreve (2001: 782) attempts to explain why the normative view (elimination of variants) has not taken hold in practise, and points out the value of corpora for documenting variants:

> Terminology standardization attempts to reduce the variation in the linguistic forms of terms. But until standardization is complete and universally accepted, terminology, because it appears in texts, will be affected by a variety of cohesive mechanisms such as the use of synonyms, near synonyms, substitution by hyponym or hypernym, etc. Translation-oriented terminology management systems need to document these variations by collecting and commenting on their appearance in texts.

2.3.4 Predominance of nominal forms

It is generally agreed in the literature that, with respect to word class, terms are predominantly nouns (Cabré 1994: 23 and 1999-b: 36, 70 and 112; Rey 1995: 29, 136; Kocourek 1982: 71; Condamines 2005: 44; Daille et al 1996: 2077). These include single-word nouns and multi-word nouns, referred to as *noun phrases* or *nominal phrases* by some linguists (*syntagme lexical nominal* or *syntagme nominal* in French). Condamines attributes this to the notion that nominal forms are associated with a high degree of stability and designatory power (2005: 44). Indeed, in a study of specialised dictionaries carried out by L'Homme (2003), between 84 and 98 percent of the entries were nouns, and in two terminological

dictionaries, Rey found an even higher proportion (1995: 137). Cabré (1999-b: 112) claims that two-thirds of terms are nouns, whereas Sager (1990: 58) suggests that terms are almost exclusively nouns:

> Concepts which are linguistically expressed as adjectives and verbs in technical languages are frequently found only in the corresponding noun form and some theorists deny the existence of adjective and verb concepts.

Rey, who emphasises denomination of extra-linguistic concepts and objects as an essential motivation of terms (1995: 29), may be one of the theorists Sager was referring to. Rey often prefers the term *sign* to *term* when discussing the object of terminological study, borrowed from the field of semiotics. He defines signs as "words and units larger than the word... that function as names, denoting objects, and as indicators of concepts" (p. 29). Emphasising denotation results in the near exclusion of non-nominal forms from terminology, and indeed, Rey only admits "some" verbs and adjectives, "the conceptual content of which cannot be reduced to a noun," into the sphere of terminology. However, even 30 years ago, at least one prominent terminologist, Heribert Picht, acknowledged that verbs also contribute to the specificity of LSPs (1985: 5).

Defenders of the Lexico-semantic Theory and other text-based approaches regret the near exclusive focus on nouns in terminology (for example, Bourigault and Slodzian 1999: 31).

> La sélection exclusive de noms est incompatible avec ce qui peut être observé dans les textes spécialisés (L'Homme 2004: 60).
>
> (Translation: The exclusive selection of nouns (as terms) is incompatible with what can be observed in specialised texts.)

L'Homme (2005: 1122) even suggests that the "conceptual perspective"[32] is partly responsible for the common assumption that most terms are nouns because, being objectivist, its methodology actually favours nominal structures.

> The theoretical models of terminology still exclusively accommodate the description of nouns and are not well suited to take other parts of speech into account. (2002: 66)

32 Our translation of "optique conceptuelle," a term L'Homme uses to refer to the General Theory of Terminology.

Bourigault and Jacquemin share this view, but attribute it to GTT's emphasis of the denotational function of terms (2000: section 9.4). They encourage the extension of terminological description to other word classes, particularly 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*. L'Homme devoted an entire research paper on the "terminological verb" (1998). The increasing recognition of the terminological status of verbs has given rise to research devoted to the automatic extraction of verbs (Kubler and Frerot, 2003).

The Lexico-semantic Theory challenges the traditional view that concepts can only be described in reference to our conceptualisation of objects in the real world (L'Homme 2004: 62). Real-world objects are typically expressed by nouns, and by objectivising the nature of concepts, the traditional view denies that concepts – at least those of interest to terminology – can be expressed by other word classes. In contrast, the Lexico-semantic Theory acknowledges that some concepts can only be described in relation to other concepts, and specifically to concepts whose lexical expression establishes a predicative relationship. Such concepts are referred to as "semantic predicates" and the concepts in relation to which they are explained as "semantic actants" (L'Homme 2004: 62). This view is used to account for non-nominal structures that have a terminological interest. L'Homme uses the verb *bequeath* to illustrate this point. This verb can hardly be defined without making reference to the fact that this action can only be carried out by a *person*, who upon his or her death leaves *something* to *a beneficiary* (person, museum, etc.). A term whose meaning is understood with reference to such semantic actants is called a *predicative term* ("terme à sens prédicatif" – literally, a term with a predicative meaning). Terms whose meaning is more intrinsically bound, such as those referring to real-world objects, are non-predicative (p. 63).

This lexico-semantic approach to meaning description can explain why some verbs, adjectives, and adverbs seem to acquire a terminological characteristic in LSPs. But it is also relevant for describing many nouns that refer to non-concrete concepts, including those that express verbal notions (such as *displacement* and *configuration*) and properties (such as

feasibility and *magnitude*). These types of non-objectivist concepts are difficult to account for according to the GTT.

Jacquemin (2001: 273) notes that "studies in terminology have generally focused on noun phrases, but other categories convey important concepts in documents." He observes (p. 313) that "through variations, verbal and adjectival phrases can be recognised as conceptually equivalent to nominal terms, and thus extend the domain of terminological knowledge beyond the frontiers of noun phrases."

L'Homme (2004: 102) refers to relations of semantic equivalence between words of different classes as "dérivation syntaxique" (syntactic derivation). She describes how verbs and adjectives can be "nominalised" and hence are terminologically interesting (for example, *to treat* (v) vs. *treatment* (n), and *compatible* (adj) vs. *compatibility* (n)).

Kubler and Frérot (2003) conducted a study of verbs in the computing domain with the aim to develop an algorithm for extracting verbs from parallel corpora. In their study, verbs posed a challenge for learners of the LSP, and for translators. They also maintain that the need for documenting verbs increases with the use of electronic corpora for term identification and description (p. 429). In her study of semi-technical vocabulary in computer science texts, Lam Kam-mei (2001: 73) observed that verbs pose significant challenges in comprehension to non-English readers of the texts and by extension, also to users of the corresponding software, particularly those that are used figuratively (for example: *run*, *call*, *interrupt*), which is a common feature of computer-science texts.

At IBM, an enterprise in the computing domain that manages its terminology, verbs were found to be important because of their prevalence in software user interfaces[33]. Following in the steps of Kubler and Frerot (2003), IBM's term extraction tool, originally designed to extract only nouns, was enhanced to extract verbs. It also records grammatical properties such as transitivity, collocations such as prepositions, and deals with discontiguous verb-preposition structures (for example: *check the file in*).

33 Based on the researcher's observations while working as terminologist for the company.

Indeed Thomas specifically identifies the computing domain as one that is rich in domain-specific verbs (1993: 59):

> In a highly-restricted domain, such as virology or computing, which may be considered 'terminologically autonomous', that is domains which barely overlap with others, there is a high number of verbs which belong solely to their domain, or else appear only rarely in LGP.

In spite of the acknowledgement in the literature that non-nouns as well as non-*concrete* nouns can evoke terminological meanings and display terminological behaviours, they have been the subject of relatively few dedicated studies (see for example Li 2011 and Cao 2011, in addition to Kubler and Frérot just cited). To what degree do commercial texts contain lexical items that fit our definition of *term* but which are not nouns? We will examine terms in commercial corpora and in the termbases in an attempt to answer this question.

2.3.5 Predominance of multi-word terms

Terms comprising more than one word are abundant in terminological resources (Meyer and Mackintosh 1996: 259; Nagao 1994: 406; Cabré 1999-b: 36; Cabré et al 2005: 1; Maynard and Ananiadou 2001: 265; Knops and Thurmair 1993: 87; Daille et al 1996: 207)[34]. Following other scholars, we refer to these types of terms as *multi-word terms*, abbreviated to *MWT* (L'Homme and Marshman 2006: 73). Other designations are found in the literature including: *complex term* (Dubuc 1997: 38), *terminological phrase* (Dubuc 1997: 38; Cabré 1999-b: 90), *phrasal term* (Cabré 1999-b: 91), and *compound* (Picht and Draskau 1985: 108; Sager 1990: 61; Dubuc 1997: 140). Dubuc (1997: 38-39) makes a distinction between complex term and terminological phrase, the latter having a higher syntactic function. This is taken up by Cabré (1999-b: 90-92), who discusses degrees of freedom of multi-word constructions. When the terminological nature of the lexical unit is de-emphasised, the expression *multi-word unit* (*MWU*) or *phraseological unit* (Cabré 1999-b: 229) is used.

The predominance of MWTs in terminology has also been observed empirically. Research conducted by Maynard and Ananiadou demonstrates that the average length of terms is 1.91

34 Multi-word expressions are also the most fruitful sources of new vocabulary in contemporary English (Hanks 2013: 50)

words (2001: 25). For their part, Justeson and Katz (1995) observe that, overall, as the length of a term candidate increases, the likelihood of it being a valid term decreases. They also note that terms in the form of bigrams are more frequent than unigram terms. Daille et al (1996: 204) take this further and demonstrate that terms comprising two words are actually the most frequent of all. Our research, therefore, will focus on MWTs.

2.4 Methodologies

In this section, we briefly describe the conventional methodologies for terminography.

2.4.1 Onomasiological vs semasiological approaches

Since knowledge about the nature of terms found in commercial corpora is a precondition for managing terminological resources in commercial settings, we need to consider the task of term delimitation, i.e. the setting of term boundaries. According to the GTT, concepts are delimited first, through an analysis of the concept system. Then and only then can terms be identified to denote those concepts. Term delimitation is therefore based on concept delimitation. This methodology is onomasiological. In such an approach, it would seem that the delimited concept would lead to the corresponding term, and therefore, morphosyntactic rules for term delimitation would be irrelevant.

The onomasiological approach is one of the basic tenets of the GTT (Cabré 1999-b: 7; Rey 1995: 127), and Rey attributes it to German semanticists who adhere to the Wusterian school (p. 137). It contrasts with the semasiological approach, which is used in lexicography (Cabré 1999-b: 8, 38; Rey 1995: 120). In the literature, these two approaches are fundamental to distinguishing terminology from lexicography (Bowker 2003: 155). Lexicographers describe words and their various meanings. Terminologists describe concepts and then, secondarily, adorn these concepts with words (designators, terms).

Some scholars claim, however, that the approach frequently adopted by terminographers is semasiological, more in line with lexicography (Temmerman 2000: 230; L'Homme 2005:

1117; Rey 1995: 137; Bowker 2003: 155; Martin and van der Vliet 2003: 335).

> 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 Terminology required them to (artificially) pretend that they were starting from concepts. (Temmerman 2000: 230).

Sager (2001: 761) positions terminography even more in the direction of lexicography: "The principles and methods of terminology compilation now have a greater commonality with lexicography than ever before." He also recognises the predominance of the semasiological approach (p. 765): "Terminology compilation is therefore becoming increasingly text-oriented and less governed by the desire to construct separate concept systems." Later, L'Homme (2004: 30), states:

> Même s'il est vrai que la délimitation précise d'un concept le guide lorsqu'il sélectionne des termes et qu'il les définit, le terminographe procède généralement à un repérage des termes dans des textes. Une fois qu'il les a identifiés, il en appréhende le sens. Il adopte donc la démarche inverse de celle préconisée par la terminologie classique.
>
> (Translation: Even if it is true that a precise delimitation of concepts assists in selecting and defining terms, the terminographer typically collects terms from actual texts. First he identifies a term, then he establishes its meaning. This approach is opposite to that prescribed by traditional terminology.)

The approach described by L'Homme is semasiological. Temmerman and Cabré share the view that the semasiological approach is more common in practise (Temmerman 2000: 230; Cabré 1999-b: 162). L'Homme continues to challenge the traditional dividing line between terminography and lexicography in subsequent publications (2005; 2006: 184), stating:

> Les fiches terminologiques sont souvent décrites comme le résultat d'une démarche onomasiologique. Toutefois, dans les faits, le terminologue fait rarement de l'onomasiologie stricte, pusiqu'il relève les formes linguistiques dans les textes (2005: 1117).
>
> (Translation: Terminological entries are often portrayed as though they are the result of an onomasiological process. However, in reality, terminologists rarely adopt an onomasiological approach, since they are called upon to identify linguistic forms in their textual environment.)

Myking discusses semasiology and onomasiology as a classic dichotomy in the field of terminology (2007:86). In his view, this dichotomy can be reformulated as an opposition of

lexeme-based vs concept-based terminography. Citing Pearson,Temmerman, and Cabré, he notes that the lexeme-based approach advocates the use of corpus-based methods.

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 (Cabré 1999-b: 163; Sager 1990: 136). Rey seems to agree: "In the electronic medium the observance of terminological requirements is even more endangered than in the printed version" (1995: 156)[35].

Piet van Sterkenburg (2003) appears to even equate specialised lexicography and terminology by defining them in virtually the same terms:

> specialized lexicography: the branch of lexicography concerned with design, production and evaluation of specialized dictionaries (p. 414)
>
> terminography: the branch of lexicography concerned with the theory and practice of designing and compiling specialist dictionaries in fields like physics, medicine, law, etc. (p. 417).

However, we attribute this to a failure to acknowledge the database medium in terminography and the concept-orientation of terminographical resources, among other differences, which are widely recognised in the literature. In our opinion these differences are significant enough to challenge this claim that terminography is a branch of lexicography. Van Campenhoudt, however, maintains that the resources produced through terminology or through LSP lexicography can be derived from the same deep structure, and can take the necessary different surface forms through the application of technologies such as XSL stylesheets (2002: 102). We wonder about that.

The exclusivity of the onomasiological approach in terminography is certainly being challenged in the literature. And scholars are increasingly recognising the value of some lexicographic methods for certain types of terminology work. Van Campenhoudt attributes this methodological contradiction to a discord between theory and practise:

35 Rey doesn't elaborate on which "requirements" he is referring to and in what manner they are "endangered."

> Qu'importe si l'auteur du dictionnaire a le sentiment d'adopter une démarche plutôt onomasiologique ou plutôt sémasiologique: il n'y a que les purs théoriciens pour penser que l'une ou l'autre suffirait à la tâche. (2006: 7)
>
> (Translation: Whether the dictionary creator feels inclined to adopt an onomasiological or a semasiological approach matters little. Only a pure theorist would believe that either approach alone would suffice.)

2.4.2 Thematic vs ad-hoc methodologies

Another methodological point of divergence between how terms are managed according to the GTT as compared to practical requirements relates to the importance of the dependency relation between terms (or concepts, to be more precise for the former). According to the GTT, 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). This approach to terminography is referred to as *thematic* (L'Homme 2004: 45; Meyer and Mackintosh 1996: 280)[36] or *systematic* (Cabré 1999-b: 129, 151). In practical situations, such as when a translator or a writer needs immediate help with a terminological problem, this approach is rarely followed. "Every practitioner knows that such a method is highly artificial" (Rey 1995: 45). A task- and text-driven approach is adopted, whereby an individual term is studied in context. This approach is referred to as *ad-hoc* (Wright et al 1997: 147; Cabré 1999-b: 129, 152), *punctual* (Cabré 2003: 175; Picht and Draskau 1985: 162; Alcina 2009: 7)[37] or *term-oriented* (Meyer and Mackintosh 1996: 280). Broadly speaking, thematic terminography is onomasiological, and ad-hoc semasiological. The sequence of steps in the ad-hoc approach is roughly the opposite of that recommended for thematic terminography (Wright et al 1997: 150).

These different methodologies have different interpretations about the notion of term. In the thematic approach, a term only exists as a designator of a conceptual node in a structured concept system. Real contexts are then studied to confirm the term. The ad-hoc approach, in contrast, accepts the existence of a term based solely on observations of text.

36 Dubuc (1997, p. 55) calls it "subject-field research."
37 Dubuc (1997, p. 47) calls it "term research" (as opposed to "subject-field research"). In the French literature, it is referred to as "démarche ponctuelle" (L'Homme 2004, p. 46).

2.5 The contributions of corpus linguistics

Since we use corpora in the current research, we wish to briefly cover the contributions of corpus linguistics to the evolving discipline of terminology.

A corpus (plural: corpora) is any body of text collected with the aim of analysing its features (Landau 2001: 273). Corpus linguistics has been broadly defined as "the study of language using collections of text in a computerised file that can be analysed by applying statistical procedures" (Landau 2001: 277). According to this definition, the current research falls into the realm of corpus linguistics. In the previous sections, we noted that terminography has more in common with lexicography than conventionally acknowledged. Understanding the contributions of corpus linguistics to lexicography might therefore lead to a deeper appreciation of how it benefits terminography. For these two reasons, we now consider the scholarly record with respect to the use of corpora in lexicography and its subsequent emerging influence on terminography.

2.5.1 Corpus linguistics and lexicography

The body of knowledge relating corpora to lexicography is extensive. Lexicographers have been using corpora as statistical evidence of word usage since at least Thorndike's Teacher's Word Book in 1921. Today, using corpora to compile dictionaries is standard practise (Cermak 2003: 20). "The corpus is generally acknowledged as an indispensable resource for the creation of dictionaries and lexicographic tools" (Prinsloo 2009: 182)[38].

Corpora provide the means to measure the frequency of use of a given lexical entity (word, term, expression). In lexicography, frequency of use is a key criterion for deciding which lexical items should be included in a dictionary.

> A low frequency for a lexical item is sound justification for omitting it from one's dictionary. Conversely, a high frequency argues strongly for inclusion. (Sinclair 2001: 297)

38 The Cobuild Dictionary is frequently cited as exemplary for using corpus evidence systematically and faithfully (Sinclair 2003: 167; Varantola 2003: 231)

Beyond frequency information, corpora provide a virtually unlimited pool of contexts, or concordances, from which one can observe collocations and other patterns of use. Computational linguists joined corpus linguists in developing techniques for lemmatising inflected forms, tagging words with their part-of-speech value, measuring the strength of the relationship between specific words, and parsing sentences to establish syntactic relations such as whether a noun is a subject, object, or indirect object of a verb. File-level annotations provide lexicographers with insights into the text genre, source, and other details, whereas word-level annotations handle lexical properties such as part-of-speech. These techniques can be applied to infuse raw corpora with additional grammatical, syntactic, and semantic information rendering them even more powerful as reference material for lexicographers.

2.5.2 Corpus linguistics and terminology

The relationship between corpus linguistics and terminology is less pronounced in the literature compared to lexicography. None of the aforementioned theories of terminology is specifically and explicitly linked to corpus linguistics. Nevertheless, all except the GTT place emphasis on the importance of studying terms in context. Over 30 years ago, before electronic corpora became widely accessible, Rondeau (1981: 79) emphasised that terms need to be studied in the context of their actual use. Indeed, a data category called *context*, which is a sentence containing the term that is obtained from an existing corpus, has been a key component of termbases since the earliest times (for example, TERMIUM®, the Grand Dictionnaire Terminologique, and IATE[39], all decades old). The authenticity of a context sentence, i.e. that it is a citation of an actual spoken or written source (as opposed to being fabricated), is a property considered essential to provide some empirical evidence of the existence of the term. This view is also adopted in lexicography (for example, John Sinclair, as cited in Landau 2001: 282). The value of the context to shed semantic, syntactic, and morphological information is recognised by all the theories of terminology.

39 www.termiumplus.gc.ca/, www.granddictionnaire.com/, iate.europa.eu/

Dubuc (1992: 45) even developed a typology of contexts that has been adopted by other scholars (Dubuc and Lauriston 1997: 80-87; Cabré 1999-b: 138) and has even appeared in ISO standards (12620:1999, ISO TC37 Data Category Registry). Dubuc points out that the context can lead to the identification of collocates that occur with significant frequency in a domain, which could in turn offer evidence of lexicalisation. These are important aids for determining term boundaries.

Meyer and Mackintosh discuss the use of corpora for terminology work, and coin the term "corpus terminography" (1996: 258). They point out, however, that the tools and techniques used in corpus lexicography are not always applied in the same way by terminographers. For example, lexicographers have an intuitive knowledge of the concepts conveyed by general language and therefore use the corpus primarily as a source of linguistic information, whereas to a terminologist, a corpus is a source of both linguistic and domain knowledge. Lexicographers use more "introspection" than terminographers (p. 265). There are, however, similarities. Like lexicographers, terminographers require corpus evidence about the full range of terms used (including variants and synonyms), with usage contexts to elucidate their conditions of use (p. 267). Indeed Temmerman and Van Campenhoudt note that it is now quite feasible to study special language in large corpora (2001: 2).

The use of corpora has entered into the debate about the differences between lexicography and terminography. Fuertes-Olivera and Arribas-Bano describe the synergies between specialised lexicography (or LSP lexicography) and modern descriptive terminography, where "words and terms are mostly being differentiated in terms of functional and pragmatic approaches, leaving aside established views which focus on the conceptual component of terms" (2008: 7). They claim that there is a methodological confluence between LSP lexicography and terminography that has resulted from the increasing use of text (corpora) in terminology work (p. 8). This view is shared by Van Campenhoudt (2002: 102), who as noted earlier even appears to equate the two disciplines.

Frequency, which as previously mentioned is a key criterion for selecting words for a dictionary, is also a key criterion in automatic term extraction. Frequency is also helpful

information for deciding whether or not a term candidate is a true term (Kit and Liu 2008: 218). Even so, due to its focus on conceptual relevance, the GTT has paid little attention to the importance of frequency for term selection. Similarly, while agreeing that frequency is a good measure of the relative importance of words and thus helps to guide decisions on the lexicographical entries in dictionaries, Nagao (1994: 406) admitted that he is not sure whether frequency would be equally valid for terminography:

> Nobody knows whether this statistical approach is applicable to the specialised fields of natural science and engineering. We do not have any reliable statistics of word occurrence in a large text corpus from such a specific subject field as computer science.

In the literature, it seems that the use of corpora is more frequently mentioned for terminology work when the focus of discussion shifts towards dictionaries, a product more often associated with LSP lexicography than with terminology where termbases are dominant. In an article about "terminological dictionaries," Martin and Van der Vliet state, "Terminography, much like lexicography, proceeds *corpus-based* nowadays" (2003: 339, their emphasis). Likewise, Bowker describes corpus selection, processing and analysis for compiling a specialised dictionary (2003: 160-163).

Pearson (1998) wrote the only monograph dedicated specifically to the application of the principles of corpus linguistics to terminology. This is not to say that corpus linguistics has not made an impact on terminology. On the contrary, as stated earlier, all the aforementioned theories, except the GTT, place emphasis on studying terms in their natural linguistic environment. One can say that the use of corpora has dominated terminology for at least the past decade. Term extraction, a technology developed to automatically identify and extract terms from corpora, has been the subject of considerable research (Ahmad and Rogers (2001), Daille (1994), Drouin (2002), Kit and Liu (2008), to mention just a few).

Corpora have been the focus of some research projects carried out in Canada, particularly as a resource for the automatic discovery of terminological relations (L'Homme and Marshman 2006: 67), and the semi-automatic construction of domain-specific corpora for terminological research purposes (Barriere 2006: 81). Rogers (1997) used a genetic engineering corpus to study synonymic variation in LSPs, and identifies contextually-dependent

triggers of the use of variants that could only be discovered through direct observations of text. The use of corpora to conduct terminology research is, at least in academic settings, now commonplace, and there are too many examples to list.

All that considered, Ahmad and Rogers (2001: 729) note that, compared to lexicography, terminography as a practice has been slow to recognise the value of corpora, and they attribute this delay to the very principles inherent in the GTT:

> The use of corpora in terminology management has been accepted both theoretically and practically much more slowly than in lexicography, largely as a result of its generally more prescriptive orientation and onomasiological, i.e. concept-based approach.

In spite of this comparative slowness, the value that corpus linguistics can bring to terminology is recognised. Sager (2001) explains:

> Terminology is now adopting a corpus-based approach to lexical data collection. By being studied in the context of communicative situations, terms are no longer seen as separate items in dictionaries or part of a semi-artificial language deliberately devoid of any of the functions of other lexical items. The increasing tendency to analyse terminology in its communicative, i.e. linguistic, context leads to a number of new theoretical assumptions and also to new methods of compilation and representation. (p. 761)

> The production of glossaries or terminological dictionaries is now quite unthinkable without the basis of textual corpora that circumscribe the range and scope of the material on which the collection is based. (…) Terminology compilation is therefore becoming increasingly text-oriented and less governed by the desire to construct separate conceptual systems. (p. 762 and 765)

Exploring the relationships between terms and texts, Shreve (2001: 772-787) advocates for an increased use of corpora in terminography. He prefers to see this achieved through a direct linkage between the meta-terminological resource such as a termbase (he uses the term “glossary”) and the textual corpus, as he finds the capacity for encoding contexts in the termbase itself to be too limiting. What he is describing is concordancing software.

> Translators and technical writers should move toward the collection and annotation of their source and target texts as a parallel to their terminology management activities. If this is not done, then valuable information about the effects that textual content has on the actual usage and appearance of terms in texts will be lost. This information is too complex and varied to be imported into the glossary, rather, the glossary must be linked or associated

> with a corpus of texts so that for each term entry in a glossary there are a range of textual contexts in which the term appears. This textual information is precisely the translation-strategic or writing-strategic information that is needed to actually use the terms in texts. (p. 786)

The use of corpora in terminology was recognised early and is evidently now well established. Rey states unequivocally: "The set of linguistic units, words, and phrases which serve as terms is extracted from a corpus" (1995: 146). Sager adds: "Serious terminology compilation is now firmly corpus-based" (1990: 154). More recently, Teubert describes how "new instruments have been developed (from corpus linguistics) which can be applied to descriptive terminology work" (2005: 103). He maintains that "modern text-based terminology work is becoming an economic factor" because it "paves the way for the development of cutting-edge technology" by providing near instantaneous access to knowledge.

2.6 Summary

There are differing views on what defines an LSP. While there is consensus that LSPs are restricted to a subject field, the notion of subject field itself is not clear. More recent theories tend to broaden the definition of subject field beyond its traditional interpretation of a field of knowledge within a highly structured taxonomy of sciences and technologies, to include various activities carried out by humans as well as criteria such as text purpose, text type, and the communicative setting. There is also consensus that LSPs elicit certain stylistic textual features. Pearson (1998: 7) summarises it well: "Terminologists, LSP and sub-language researchers contend that what distinguishes LSP from LGP are restrictions on vocabulary and syntax." We will later explore how these and other features of LSPs can justify whether or not the language used in a company can be considered an LSP.

The terminology literature has so far not paid a great deal of attention to the role of genre in term discovery or characterisation or, conversely, the role of terminology in the characterisation of genres. Cabré attributes the defocalisation on genre to the reductionism of the GTT: "By neglecting the communicative aspects of terminological units -- another consequence of the concentration on the denominative function of terms -- research has been

prevented into the contribution terminology can make to the differentiation of text types at various levels" (2000: 42). The scholarly record has focused on genre as a determining factor in the use and interpretation of linguistic markers that establish relationships between terms, rather than on aspects such as genre-motivated terminological variation or termhood itself (see for example, Condamines 2007b, 2008a, 2008b). Further research may prove fruitful in these areas.

The theories described in this chapter (General, Socio-cognitive, Lexico-semantic, and Communicative), as well as the textual terminology movement, diverge considerably in their perspective of what constitutes a term, emphasising respectively, the concept, cognition, lexical behaviour, and contextual factors such as the role of the interlocuteur and communicative intent. We could not summarise this better than Cabré (2003: 187):

> At the core of the knowledge field of terminology we, therefore, find the terminological unit seen as a polyhedron with three viewpoints: the cognitive (the concept), the linguistic (the term) and the communicative (the situation).

For the GTT, the criterion for termhood is that the concept denoted by the term necessarily is a member of an objectivist, structured system of concepts. Subsequent theories place more emphasis on a range of linguistic properties (morphological, syntactic, paradigmatic, etc.) of terms and contextual factors, much of which can be observed in the corpus. Indeed, in these theories, the notion of term is intrinsically linked to the text in which it occurs.

Scholars who recognise the communicative aspect consider that termhood is relative to the purpose or application of the terminological resource. Van Campenhoudt extends this application-oriented perspective to broad missions such as language planning, harmonising terminology within a discipline, describing language variation, and facilitating inter-linguistic communication: each would have a different notion of what is a term (2006: 4).

While terminology has lagged behind lexicology in using corpora as the basis for research, likely due to its historical foundations in the onomasiological perspective, the value of corpus-based investigation to identify and validate terms is now widely recognised by researchers and scholars in the field of terminology.

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. In our opinion, adherence to one notion or the other depends on the circumstances and aims of the terminology initiative. In a highly controlled normalisation environment the former interpretation serves a useful purpose. In a more dynamic communicative setting the latter would be more applicable. We maintain that commercial communications fit into the latter category more often than they do the former.

Variants, which the GTT seeks to eliminate, are today recognised as valid means of expression in LSPs. Several empirical studies have validated this claim. Likewise, the premise that terms are almost always nouns has been challenged, again with empirical evidence.

While earlier scholars defended the onomasiological approach and only reluctantly admitted the existence of the semasiological[40], there is now increasing acknowledgement that terminographers often adopt a semasiological approach to term identification and description. Due to advances in computerised text processing capabilities, terminography is increasingly being driven by corpora. These two factors are interdependent; the more corpora that terminographers have easy access to, the more their investigative processes will be driven by text. They have ramifications on the notion of what constitutes a term, as this notion would necessarily be text-bound under these conditions. Condamines cites a particularly problematic case in point directly relevant to commercial settings, that of selecting terms from a list of term candidates produced by a term extraction tool:

> Il faut avoir fait l'expérience qui consiste à choisir les termes parmi la liste de tous les syntagmes nominaux d'un corpus (fournis par un extracteur de termes) pour se rendre compte de la difficulté qu'il peut y avoir à décider ce qu'est un terme. (2005: 44)
>
> (Translation: One needs to have experienced the task of choosing terms from among a list of noun and noun phrases produced by a term extraction tool from a corpus to realise how difficult it is to decide what is a term.)

As L'Homme (2005: 1130) explains, perhaps the differences in perspectives are to be expected:

40 For instance, Meyer and Mackintosh stated in 1996 (p. 280) that the semasiological approach is “rarer”

Il est extrêmement difficile, voire impossible, d'aborder le terme en faisant abstraction de l'application dans laquelle il est mis à contribution. La diversité des applications terminologiques (construction d'ontologies, élaboration de dictionnaires, documentation, traduction, etc.) et les postulats théoriques qu'elles appellent (description ou normalisation, cadre conceptuel ou lexico-sémantique, onomasiologie ou sémasiologie, etc.) mènent à des conceptions diversifiées de l'objet terminologique.

(Translation: It is extremely difficult, if not impossible, to define what a term is without considering the applications in which it is being used. The range of uses of terminology (ontology construction, dictionary publishing, content authoring, translation, etc.) with their respective theoretical postulates (descriptive vs prescriptive, conceptual vs lexico-semantic, onomasiology vs semasiology, etc.) lead to diverse interpretations of what constitutes the object of terminological study.)

CHAPTER 3 CRITICAL DISCUSSION OF THE LITERATURE

In this chapter, we comment on the notions expressed in the literature, with respect to terminology used and managed in commercial environments. We make assumptions that we will subsequently attempt to validate in our empirical research.

3.1 Company-specific language as an LSP

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-recognised properties, the language used in most companies does indeed constitute a type of LSP, although few authors state this explicitly[41]. Even according to some of the classic definitions in the literature, this type of language qualifies as an LSP. For instance, while the notion that an LSP is restricted to a subject field is widely accepted, the definition of *subject field* has broadened from a highly-structured objectivist hierarchy of science and technology to an experiential delimitation that is context- and application-dependent. Even thirty 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 a range of different degrees of specialisation; some are obviously more specialised than others. The texts produced in most companies describe tangible products, services, and activities, often within one vertical industrial or economic sector, which could be viewed as a subject field according to the broader interpretations we have presented. They often adhere to specific linguistic rules and styles;

41 Cabré (1999-b, p. 22 and 63) and Pavel (1993, p. 21) identify commercial texts as a type of LSP.

many companies have a style guide, and some are automatically implementing the style rules through controlled authoring[42] software. The written form predominates. Aside from certain types of marketing material, most of the information companies produce is strictly informative, and often somewhat didactic. While the target audience of this information – often consumers – may not have undertaken special education in order to read and use it, they are actively engaging with the goal of acquiring knowledge. Depending on its level of specialisation, a company's informational content could therefore fit into three communicative contexts that Pearson (1998: Section 1.9) describes for sub-languages: expert-to-initiates, relative expert-to-uninitiated, or teacher-pupil.

Even though Rey recognises the importance of clearly delineated scientific and technical subject fields to terminology 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, our emphasis). We suggest that this term resonates well for terminography in commercial settings.

3.2 The notion of term, in commercial environments

> The language services of private businesses address all language-related issues (text writing, translation, terminology) in the company and deal with the specialised terminology necessary for the activities of the company. (Cabré 1999-b: 23)

Companies manage terms to address their language-related issues and support their activities, as Cabré notes. In this sense, the motivation for managing terminology in a commercial setting is no different than in other environments. However, in commercial settings those issues and activities are driven by pragmatic factors, such as the use of technology and the availability of resources, and are not as linguistically motivated as, for instance, terminology initiatives in support of minority languages or official language policies. The

42 For example, Xerox, SAP, Eastman-Kodak, IBM, Digital, ITT, Ericsson, CISCO, DELL, Adobe, PayPal, Hitachi, eBay, Microsoft, Oracle, Philips, Boeing, Toyota, to name just a few. (Gopferich 2000:237 and www.acrolinx.com)

notion of *term* will necessarily be affected by these factors. In this section we mention some potential commercial applications of terminological resources and we describe three areas that could be affected: non-nouns, variants, and semi-technical vocabulary.

3.2.1 Purpose or application of terms

As noted in the literature review, some scholars maintain that termhood is relative to the ultimate purpose of the terminological resource or application where it is used. This point is particularly relevant to commercial environments, where virtually every task undertaken is project-driven.

Martin (2011), a terminologist at the company SAS, describes the challenges of deciding what is a useful term in a terminology management strategy intended to meet commercial aims. He outlines factors to consider when evaluating term candidates, including the conceptual basis for evaluating termhood, the nature of pre- and post-modifiers, term frequency, the context and domain, and term embeddedness (when terms are found embedded in larger terms). In his estimation, single-word terms are almost insignificant. He points out that the value of a terminological resource is measured in terms of the reuse of the data: "Building up a set of terms is not a practise that is carried out merely for the purpose of collecting the largest set of terms possible." Shreve also emphasises the pragmatic orientation of terminology work, stating that "Terminology management for translators and technical writers is an endeavour with practical goals. It does not exist as an end in itself, but purely to improve the creation of texts and translations" (2001: 785).

We maintain that in commercial environments, establishing termhood entails deciding that managing a given term candidate serves the communicative needs of the company. Lombard, for instance, 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 localisation (translation) issues (2006: 162), such as translation inconsistency. We would like to add that consistency in the source language (SL) is also a concern (Warburton 2001b: 680). Note that these categories are

purpose-based not semantically-based. Through our research, we will establish those communicative needs, which in a modern context are based on the end-use applications of the terminological data. We then seek a co-relation between the terms chosen to be managed and those communicative needs. In this respect, we expect that a terminographic methodology for commercial environments will need to be solidly based on pragmatic requirements.

3.2.2 Importance of non-nouns

As in other text types where terminology is found, the terms in commercial texts 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 in commercial corpora. We suggest that in commercial communications, domain-specific verbs, and possibly adjectives and adverbs to a lesser degree, exist and need to be proactively managed as well as nouns.

Indeed, one only needs to consult a few company Web sites to remark that in commercial texts, domain-specific concepts are also expressed by non-nouns, and many of the nouns themselves express non-tangible concepts, such as those referring to services, programs, activities, and events. Consider the following paragraph from a document about Microsoft SQL Server 2008[43], for which we have underlined some terms[44]:

> 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 candidate terms, 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*

43 Available from: http://download.microsoft.com/download/B/F/6/BF66161B-3804-49DA-AB95-1D8E4F3BA14E/SQLServer2008R2_DW_DataSheet_12_10.pdf

44 The determination of term boundaries shown here is one possible interpretation. Using a larger corpus can help to more precisely determine term boundaries.

schema). What is further interesting is that the document in question is a data sheet – a particular type of product document in which one would expect a high incidence of tangible concepts because it describes specific features of a product.

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 its general use; 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, verbs such as *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 and translation.

As stated earlier, at IBM[45], verbs are significant because of their prevalence in the user interface of software. Verbs like *open*, *save*, *view*, *export*, and *print*, are common in software interfaces. Considered as belonging to the general lexicon, they would have no terminological interest according to traditional theory and practise. But in software products – which are often translated into dozens of languages – where consistency between the user interface and the help system is critical, these *are* terms of interest for pragmatic purposes. This scenario represents exactly what L'Homme was referring to when she described how terminology is used by translators, cited in section 2.3.2.1.1.

Thus, the IBM term extraction tool, initially designed to only extract nouns, was subsequently modified to extract verbs and record properties like transitivity, collocations such as prepositions, and discontiguous verb-preposition structures (for example: *check the file in*).

45 Based on personal experience.

Likewise, FASTR, the term extraction tool described in Jacquemin (2001: 273), was extended to non-nominal phrase structures such as verbal, adjectival, and adverbial phrases.

3.2.3 Prevalence of variants

Recent theories have recognised that variants have a valid expressive and semantic role in LSPs. We suggest that variants are common in commercial texts, and that they have an explicit purpose such as to realise economy. Our research will attempt to quantify the scope of variants in commercial texts. In this section, we clarify how variants are perceived in practical terminography.

We have cited references in the literature that identify abbreviation as a popular form of terminological variation in LSPs because it achieves economy without loss of semantic precision. There are various forms of abbreviation which go by various names, and this has led to some confusion among terminologists. We wish to clarify how we categorise abbreviations as a best practise for terminography in commercial settings and for the purpose of this research. We adopt the perspective of the TBX editorial committee, which recommends data categories for abbreviated forms in terminography. This perspective is based on the TBX-Basic specification, which was developed by commercial terminologists in under the auspices of the Localization Industry Standards Association, and is now maintained by TerminOrgs[46], a think-tank of terminologists working in commercial settings.

Abbreviation is just one form of terminological variation, as noted previously in the literature review. For the purposes of terminographical marking of term variants, ISO TC37 produced a standard set of 34 variant types, called term types, in ISO 12620:1999. The complete list from the standard is reproduced in the following table.

46 www.terminorgs.net

entry term synonym quasi-synonym international scientific term common name internationalism full form abbreviated form of term • abbreviation • short form • initialism • acronym • clipped term • contraction	variant transliterated form transcribed form romanized form symbol formula equation logical expression sku part number	appellation idiom phraseological unit • collocation • set phrase • synonymous phrase standard text string string category product name

Table 1: ISO 12620 Term type values

Several of these values are not relevant to terminography, such as *synonym* and *entry term* (all terms in an entry are synonyms and at the same time entry terms), while others are rare if not non-existent entirely in terminological resources, such as *string category* and *logical expression*. Several represent linguistic items that do not ordinarily belong in a termbase because they are not *terms* according to conventional theories and methodologies, such as *idiom* and *phraseological unit* along with its sub-types. Some values appear to overlap, such as *appellation* and *common name*. Others appear to be an element of an incomplete set, such as *product name*; why are other kinds of names omitted? Finally, a few might be misplaced, for instance, idioms are often if not always phrases, why therefore does this value not appear as subordinate to *phraseological unit*? Overall, this list is too complex, large, and incoherent to be practical for terminographers.

The Terminology Special Interest Group (SIG) of the Localization Industry Standards Association recognised this problem and proposed a smaller set of values, based on input that had been received from practising terminographers through surveys. Through this consultation process, it was confirmed that in commercial texts, abbreviation is a linguistic device commonly used to achieve economy. Two types of abbreviation dominate these texts: acronyms, and truncated multi-word terms (MWTs). The SIG recommended the following term type values, to be used with the meaning given below, for TBX-Basic:

full form

The complete representation of a term for which there is an abbreviated form. For example, in an entry that contains both *ACL* and *access control list*, the term *access control list* is the full form and ACL is the acronym.

abbreviation

An abbreviated form formed by omitting letters from a longer form. Example:

full form: volume

abbreviation: vol.

short form

An abbreviated form that includes fewer words than the full form. Example:

full form: Intergovernmental Group of Twenty-four on Monetary Affairs

short form: Group of Twenty-four

acronym

An abbreviated form made up of the initial letters of the components of the full form or from the syllables of the full form. Examples:

full form: United Nations

acronym: UN

full form: Extensible Markup Language

acronym: XML

variant

An alternative form of a term other than an abbreviated form. Variants can include words that have an alternative spelling, punctuation, capitalisation, word formation, or even a numeric representation. Example:

term: soft switch

variant: softswitch

phrase

Any group of two or more words that are frequently expressed together and that denote more than one concept. The individual words in a phrase usually function in more than one grammatical category (part of speech) within the syntax of a sentence. Examples:

send feedback

work offline

Here, the value *variant* has a more specific meaning than used elsewhere in this thesis, as it does not include abbreviated forms. This is because abbreviated forms are so common in commercial texts that there is a need to categorise them more granularly into three types: abbreviation, acronym, and short form. But aside from abbreviated forms, there is no need to granularly categorise other instances of terminological variation, such as spelling variants and hyphenated forms. For practical purposes, terminographers need only one value to label such other cases of variation, and this is the purpose of the value *variant*.

TBX-Basic was produced by terminologists employed in commercial enterprises. It provides some evidence that terminological variation, and particularly various devices of abbreviation, are common in commercial texts.

3.2.4 Semi-technical vocabulary

Previous research has shown that non-native readers of English texts in scientific and technical fields have more difficulty with "semi-technical vocabulary" than they do with strictly technical vocabulary (Kennedy and Bolitho 1984, Trimble 1985, Nation 1990). These findings apply not only to learners of English, but also to translators, since English is typically their second language.

Semi-technical vocabulary, sometimes referred to as sub-technical vocabulary (Trimble 1985) or non-technical (Nation 1990), has been the subject of considerable discussion and debate among researchers in English for special purposes (ESP), resulting in various definitions of this category of lexical items[47]. Lam Kam-mei refers to semi-technical vocabulary as those words and expressions that fall into the "hazier range" between general and basic English vocabulary and domain-specific terminology (2001: 6). A common thread is that these units are either shared across multiple disciplines (for example, *method* and *function*), or are basic English words that have assumed a domain-specific meaning (for example, *view* and *field*). But translations are meaning-based not form-based. One cannot assume that

47 Lam Kam-mei (2001) reviews the perspectives of 23 scholars on this topic.

the TL equivalent of a given English word or expression will be the same in different disciplines just because it is the same in English, nor that the TL equivalent of an English term that has shifted in meaning will remain the same. However, translators can easily assume that they are the same, especially if they are unfamiliar with the subject field. If translators have difficulty with this category of terms in any significant measure, this would explain why most termbases developed for translators as primary end-users contain many such terms and expressions which, due to their lack of domain specificity, would not be considered "terms" according to conventional theory.

3.3 Views on theory and methodology

Various views about terminology theory and methodology are expressed in the literature, but none address the uses of terminological resources in commercial settings. This leads us to consider the possibility that the mainstream terminology theories and terminographic methods may not be suited to the pragmatic conditions in commercial enterprises. Important features may be missing that are needed to effectively deploy terminology in business processes. The foundation of a theory and a methodology for managing terminology in commercial settings needs to be based on empirical evidence from commercial corpora.

The GTT, with its normative focus, prescriptive goals, thematic approach, systematic orientation, denial of the interference potential of linguistic formance, objectivist perspective, and exclusion of corpora in its methodologies, is the theory most disconnected from the practical needs of commerce. A descriptive, ad-hoc approach is more effective in content production environments (Wright 1997: 147-148). Teubert further notes that the descriptive approach is particularly necessary in domains marked by rapid change (2005: 103).

Aspects of the subsequent theories could contribute towards a model for managing terminology in commercial environments; but they all fail to define such a model completely, and some of their tenets that may be considered inappropriate for business applications. For instance, the Socio-cognitive Theory exemplified by Temmerman focuses on the role of cognition in forming concepts. In her 2000 monograph, she uses examples from the life

sciences to prove her postulates. She herself acknowledges that this is a limitation that could be the subject of criticism (2000: 234). While her views hold merit and can increase our understanding of the forces at play behind concept formation, terminographers working in commercial environments are not preoccupied with how concepts are formed. The socio-cognitive aspects of terminology, while their existence is not being challenged here, rarely intervene in terminographic practise in a commercial setting. This fact was acknowledged by Rey who stated (1995: 58), "The cognitive function, especially its creative aspect, however essential and primordial in itself, is less important for the practising terminologist than the other functions[48]." Nevertheless, this theory raises the importance of the communicative role of context, and subsequently, of the need to study terms in use.

The Lexico-semantic Theory, which emphasises the status of terms as lexical units, characterises the relationships between terms and the behaviour of terms in a highly granular way that likely exceeds the needs for terminology description in pragmatic settings. This approach provides a framework for describing the lexical properties and relations of terms in great detail. Some of its methods could prove useful for delimiting and describing terms from commercial corpora. However, overall the lexico-semantic approach does not provide practical methods for managing terminology in a company.

In the literature, it is also apparent that the terms typically used as exemplary models for the precepts of the various theories are either considered as isolated units out of their communicative context, or are taken from domains presenting characteristics that justify those precepts. Terminology in content produced by companies in the private sector has largely been ignored. The theories of terminology have yet to be critically examined in a modern context of content production for commercial purposes.

3.4 The role of corpora

We maintain that the semasiological approach is almost exclusively used in commercial environments. This suggests, however, that corpora need necessarily figure prominently as

48 The other functions Rey is referring to are the linguistic functions and the socio-cultural functions.

sources of research materials in these environments. However, while there is growing recognition that corpora are useful for selecting terms and obtaining information about terms, the actual use of corpora remains low among terminologists in the private sector. Having frequently more background in translation, as previously mentioned, terminologists in general lack awareness about corpus linguistics, large-scale corpora, and corpus-analysis tools. Furthermore, in many companies, especially large ones, it may not even be possible to gain access to the company corpus in its totality; by virtue of its size and organisational structure, the corpus is typically distributed across different content repositories. Even the company's Web site on the Intranet will lack a significant portion of the total corpus. When terminologists do check a corpus to validate a term, due to these factors typically they will 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 analysed, validating terms based on a sub-section of the full corpus will result in some terms being selected that are not optimised for large-scale repurposability.

Myking raised questions as to how corpus-based approaches should be reflected in methodology, but left them unanswered:

> To what extent can lexeme- (corpus-) based methodology provide a genuinely new method of terminography in its own right? To what extent is it intended as a supplement to the concept-based method? To what extent is fruitful coexistence possible within terminography? Can the lexeme-based methods provide practical solutions beyond the identification of term candidates? (2007: 86).

3.5 Genre as a deterministic factor for terminology

We noted in section 2.2 that the notion of discourse community is fundamental to genre. According to Swales description (1990: 24-27), producers of commercial texts and their target audiences do constitute a discourse community (or set of communities), and the IT field is cited as exemplary in this regard. Also according to Swales, a discourse community communicates in one or more genres (our emphasis). Therefore we can anticipate that commercial content is delivered in the form of one or more genres. We also noted that genres involve choices at the level of the lexis.

Inspired by the exemplars of genre provided in the literature (Bawarshi and Reiff 2010: 138; Rogers 2000: 9; Gopferich 2000: 229; Ditlevsen 2011: 199; Bhatia 1993: chapter 3), we suggest that typical genres in public-facing materials in a business setting would include annual reports, white papers, marketing Web sites, product manuals, maintenance manuals, operating instructions, frequently-asked questions, trouble-shooting guides, packaging material, licenses, and so forth. For the IT industry, which is the subject of our research, we would add on-line help, Web logs, and user interfaces for various media such as software and mobile applications. Internal-facing genres would include meeting minutes, employee performance reports, product design specifications, project proposals, and so forth. Commercial content comprises multiple genres.

One would expect such genres to utilise different linguistic techniques to achieve their *communicative purpose*, which scholars maintain is the most important factor in genre identification (Bhatia 1993: 45). A marketing web site uses persuasive language and a concise style, a maintenance manual active voice, simple syntax and a controlled vocabulary such as recommended by STE, and a white paper a style similar to that of an academic research article. But the choices of vocabulary representing the same key specialised concepts in these different genres should not, in theory, differ significantly. This assumption is confirmed by Rogers' study described in section 2.2 and by Freixa's in section 2.3.3. This is not to say that terminology does not differ between certain genres, for instance, the use of *oncology* in a scientific medical journal article and *cancer research* in a brochure designed to raise funds for a hospital from the general public seems perfectly suitable for their respective target audiences. There could be different conventions with respect to genre-based lexical variation in different domains. Indeed, Condamine's research shows that the relation between genre and domain is complex, and she recommends that we “ponder the definition of genre and, more specifically, its link with the notion of domain” (2008: 134).

We note that Rogers (2000: 11) emphasises the relevance of genre for comparing the use of terminology in different genres within an LSP, for instance, between learned articles in mechanical engineering and advertisements about the products produced. Extending this

notion to a commercial setting, constructing a company corpus according to genre by separating, for instance, online help from print documentation, would enable the role of genre in terminological variation to be empirically observed. Indeed, Condamines notes that applications (of linguistic resources) could be grouped into classes to which one could attempt to associate types of terminological resources. In content of a commercial nature, these application classes could correspond to application-oriented genres. After investigation, it may be possible to link choices of terms, relations and their representations to the various applications (2010: 35). Information workers in companies frequently point out that maintaining terminology consistency in different text types is challenging, often because the texts are produced by different people in the organisation (Warburton 2001b: 678, 680). Taking up Swales again, these different people could belong to different discourse communities depending on their background and communicative intent. A genre-based comparison of terminology may help to elucidate this situation.

The language used in a commercial setting comprises a set of different genres across which the lexical stock of a company is distributed. Genre may be a deterministic factor for terminological variation within a company corpus. Whether there are clear divisions in terminology use in these different genres can only be determined through an investigation of terminology in business corpora subdivided into genres. The current research is, however, interested in the lexis of companies taken as a whole and how well terminology resources developed in companies are adapted to meet their broad communication needs.

3.6 Summary

We have demonstrated that the scholarly literature has neglected to examine the specific needs for terminography in commercial environments. We suggest that among the theories of terminology, the GTT serves these needs the least. According to Alcina (2009: 7) the GTT is no longer the most appropriate for training communicators and translators and it should be complemented by the more recent theories such as the communicative and socio-cognitive. We see the relevance of certain precepts of the other theories for commercial environments, such as communicative intent, application of the terminological resource,

and text-driven criteria in determining termhood. We aim to validate the relevance of these precepts through empirical evidence.

The next chapters are dedicated to the research component of this study. We begin by formulating the research objectives. Then we describe the commercial data and how it was prepared for analysis. We then produce the results of our data analysis, focusing on the gap between commercial termbases and corpora. This analysis leads us to consider the value of keywords in identifying the most important terms that a company should manage.

CHAPTER 4 RESEARCH OBJECTIVES

4.1 Research questions

The object of terminography is terms. To consider terminography in the context of commercial settings, it is fundamentally important to establish what we consider to be a term in this context. In our literature review, we described how the various theories of terminology define the notion of *term.* We suggest that in commercial settings, this notion has to widen to include pragmatic criteria. Given that, in such settings, all terminology work needs to be economically justified[49], the question: What is a term in commercial settings? can be more appropriately expressed as: What types of linguistic units need to be actively managed to support the communicative needs of commercial enterprises?

This research examines terms in commercial texts and terms in corresponding termbases to determine the effectiveness of the terminographic procedures used. We investigate the soundness of the assumption that terms selected for management by terminographers in commercial environments are in fact the terms that need to be managed. To this end, of special interest is how well the terms found in the corpora correspond to the terms found in the termbase of a given company. Any significant gap between the two suggests that there are weaknesses in the term selection process used by the terminographer. We seek to answer the following questions:

1. Is there a significant gap between commercial termbases and corpora?
2. What are the causes of this gap?
3. Does the gap reduce the effectiveness of the termbase?
4. How can the gap be narrowed?

Answers to such questions can then motivate theoretical and methodological reflections. How can mainstream terminology theory and methodology account for an effective

49 Commercial justification involves a cost-benefit analysis, sometimes referred to as a return-on-investment analysis, or a business case.

commercial terminography? What are the theoretical and methodological implications of the empirical evidence?

4.2 Research methodology

In this study, we compare the termbases and corpora of four companies as case studies: Minitab, SAS, Symantec, and Hewlett Packard (HP). The companies range in size and all are active in the field of computing and information technology (IT).

As English is the primary language of business and also the SL for translation purposes for all four companies, the corpora are comprised of English texts. Likewise, our investigations on the termbases – all but one of which are multilingual – will be on the English terms.

First we needed to be granted permission to obtain copies of the termbases and corpora and use them for research purposes. This required a legal framework which had no precedent either in the four companies or in the City University of Hong Kong. The companies had never before received a request to provide their linguistic assets for research purposes, and the university was unaccustomed to research projects involving the intellectual property of commercial enterprises. It was necessary to consult with the legal department of each company, and a legal agreement was prepared and signed by both parties (a company representative and the researcher). A separate agreement was required for each company, reflecting its particular conditions. A copy of one of these agreements is provided in Appendix G. The university's legal department was also consulted to ensure that there would be no violation of its policies. Finally, due to the large sizes of the termbases and corpora, it was necessary to install a special FTP server at the university to allow the files to be transferred. The entire process took several months.

The termbases were studied first to determine the nature of the terms they contain, the range of data categories used, and the data model and file formats. As it was not possible to gain direct access to in-house company resources for security reasons, each termbase was provided in the form of an exported file. The termbases were then analysed using the

following methods:

1. On the exported files directly, by using Xpath queries and other statistical collection methods
2. By importing the export files into the TermWeb terminology management system[50], and using its filtering and statistical reporting functions.

Using a terminology management system (TMS) was deemed necessary after it was determined that the structure of the export files was not conducive to carrying out Xpath queries necessary to perform certain investigations on the data. (This limitation will be further explained later.) A TMS enables a wider range of filters to be applied to the data to gather statistical information than Xpath queries, at least, it was the case for our data.

Two software tools were used to examine the exported files directly: oXygen, an XML editor, and UltraEdit, a text editor. oXygen provides features such as XML validation of the source file and running of Xpath queries. UltraEdit is an advanced text editor that allows searches using regular expressions, and is very practical for making global changes and for gathering statistics. UltraEdit also has a basic concordancing function which was used to double-check some of the concordance statistics provided by a dedicated concordancing software (described below). It was also used to prepare the exported termbase files for import into the TMS.

The following observations are of particular interest as they present opportunities to discover better methods of term identification:

a) the number of terms in the termbases that occur rarely or not at all in the corpora, and the nature of those terms
b) the number of terms in the termbases that occur frequently in the corpora, and the nature of those terms
c) the terms in the corpora that demonstrate characteristics of interest, such as unusual frequency or domain specificity, and yet are missing from the termbases.

50 TermWeb was generously supplied by Interverbum Technologies Inc.

To measure the correspondence between the termbases and the corpora it was necessary to establish the frequency by which the terms from the termbase occur in the corpora, for each of the four companies. Concordances of the termbase terms, and various variations thereof, are then generated and examined in an effort to find ways to increase the correspondence. We therefore required a concordancing software that takes an input list of terms (i.e. the terms from each company's termbase), searches that entire list in the corpus, and provides global statistics. WordSmith was selected after examining the features of five major concordancing software tools: AdTat, AntConc, MonoConc Pro, KH Coder and WordSmith.

WordSmith has more of the desired features than the other tools, and it benefits from a respected reputation among natural language processing experts. In particular, WordSmith is the only tool that runs a concordance in batch against an input list of terms; the remaining tools only generate concordances for one term at a time, which the user has to submit manually. The batch concordance function was essential for our research.

For each company, the corpus had to be prepared for the concordancing software. The size of the corpus and of the list of termbase terms submitted for the batch concordance was too large for WordSmith to handle, even with our powerful laptop computer[51]. We experienced significant performance issues such as system freezing and crashing. Several months of effort and time were spent discovering ways to reduce the processing load, such as by converting files into a different format, removing markup, and merging files to reduce their total number. These tasks are explained in more detail later.

Likewise, for each company, the terms from the termbase needed to be prepared; they must be extracted from the termbase, which was provided in an XML format. If general lexicon words are present, they must be removed for the concordancing step; otherwise, too many concordances would be generated. The presence of general lexicon words in a termbase is atypical but not necessarily unjustified (for example, they may be required for controlled authoring purposes). We will comment further on this situation later.

51 2.8 GHz processor, 8 GB RAM, 280 GB hard disk storage

We are interested in examining the parts of company termbases which are possibly not optimised because the terms they contain are not members of the corresponding corpus. We therefore generated concordances of the termbase terms using the corpus as input to identify those terms that do not occur in the corpus or occur very infrequently. We then attempted to draw conclusions about the nature of these under-optimised terms and sought approaches to reduce their occurrence in the termbase.

We also examined the termbase terms themselves to determine if there are any features common to several or all of the companies under study, such as the prevalence of certain word classes and the term length (number of tokens in a term). Given our interest in closing the gap between termbases and corpora, we also explored the potential of keywords as nodes for identifying MWTs that are frequent in the corpus.

To explain the existence of potentially redundant terms in the termbase (terms that do not occur, or occur very infrequently, in the corresponding corpus), we adopted the following procedure:

1. Run a concordance against the corpus with the termbase terms, for each company. The concordance is case sensitive to retrieve only exact matches. The aim is to establish a baseline statistic measure of the gap between the termbase and the corpus.
2. Identify the termbase terms that do not occur, or occur very infrequently, in the corpus, based on a set of comparable frequency ranges.
3. Present the findings obtained from the previous step to company terminologists to seek any possible explanations for these terms based on company-specific terminology requirements or processes. For instance, the supplied corpus may be incomplete. Make any necessary adjustments to the process and repeat it to establish an accurate baseline statistic.
4. Analyse the nature of the terms that are not found in the corpus, as well as terms occurring in various frequency ranges, to reveal any patterns of interest.

We used functions provided by the WordSmith tools to help discover reasons for the unusual infrequency of certain terms, such as errors in the setting of term boundaries, with

a focus on relationship measures, clusters, collocates, and keywords. For some of these functions, we needed to generate a list of words (tokens) from our corpus, and use a word list from a reference corpus for comparison.

The purpose of these investigations is to discover ways to improve term selection by narrowing the gap between termbases and corpora. In addition, the value of corpus-based terminology identification may become apparent.

4.3 Expected outcomes

To our knowledge, no empirical study of the relationship between termbases and corpora from commercial enterprises has as yet been carried out, and in this sense our research is novel. However, the basic assumptions that motivate this research are shared by other terminologists and scholars. Sager, for instance, pointed out that advances in information retrieval systems and the availability of large machine-readable corpora have affected the motivation and methods of terminology compilation (1990: 131). Technical texts can be converted into a format for terminological analysis using techniques developed by computational linguists. Consequently, texts can be analysed and compared with current terminology holdings in order to discover missing terms, and, we add further, under-optimised terms. He also observes that statistical analyses of large volumes of text can reveal changes in the frequency or usage of a term (p. 132). Finally, Sager raises the prospect of using the comparative termbase-corpus investigation that we have adopted for "semi-automated control over existing term collections" and for "machine-assisted terminology identification and compilation." While in very early stages of consideration at the time he made these statements, these applications, he predicted, would increase dramatically in importance in the future. We believe that these types of observations can do more than improve the quality of the holdings of a given termbase by filling gaps and reducing redundancy; they can also contribute to developing effective strategies for term identification as a whole.

Because terminologists in companies typically use an ad-hoc approach to term identification that rarely takes into consideration large-scale corpora, we expect to discover a signifi-

cant gap between the termbases and the corpora. We also expect to find lexical constructs in the termbases that do not adhere to the traditional semantic-based notion of termhood, but rather are present for pragmatic purposes driven by the end-uses of the terminology.

Through this investigation we expect to clarify the notion of termhood from the perspective of production-oriented motivations. We also expect to observe that determining the optimal boundaries of MWTs is problematic and could be improved through a corpus-driven approach to term selection. The knowledge gained through this research can lead to the elaboration of terminographic practises that are effective in commercial environments. Terminological resources developed according to those practises will be better able to meet needs in areas of language production and language processing (see sections 1.5 and 1.7.3), due to their greater correspondence to the language actually used in commercial enterprises.

CHAPTER 5 DESCRIPTION AND PREPARATION OF THE DATA

Our research examines the terms, terminological data, corpora, terminology uses, and terminographic practises in four companies: Minitab, SAS, Symantec, and Hewlett Packard. All the companies produce products in the information technology and computing fields. Products are produced in English as the SL, and are then translated into multiple target languages. Each company has a termbase. In this chapter, we describe the four companies, their termbases, and their corpora.

5.1 Description of the data

In this section we describe the four companies that participated in this research and their data (termbase and corpus).

5.1.1 Minitab

Minitab Inc. develops software and provides services for quality improvement and statistics education. More than 450 Fortune 500 companies use Minitab's statistical software to analyse their data and manage quality improvement projects. Minitab also develops the world's leading software for teaching statistics, which is used by 4,000 colleges and universities. Head-quartered in State College, Pennsylvania, the company has subsidiaries in Australia, France and the United Kingdom.

With 330 employees, Minitab is the smallest of the four companies studied[52]. Its content is produced by a staff of 12 writers.

Minitab began to actively manage its terminology in 2009. One part-time terminographer manages the termbase using crossTerm, the TMS that is included in the Across Language

52 As a privately-owned company, Minitab does not release revenue figures.

Server CAT tool. The database was initially populated with existing company glossaries from translators, most of which were in spreadsheet form.

A relatively new undertaking, the termbase is currently used only by technical writers and translators (meaning that no other employees have access to it). Technical writers use XMetaL as their authoring tool. crossTerm provides a plug-in for XMetaL, called cross-Author, which allows the writers to access the termbase directly from XMetaL. Translators can access the termbase in their CAT tool, crossDesk. Translators can also access the termbase directly in crossTerm and have the ability to add or modify the data, whereas technical writers must contact the terminologist to suggest additions or modifications.

In the following screen capture of the XMetaL plug-in, the right frame shows a list of the terms from the termbase that occur in the sentence that the writer is currently editing. For each term in the list, the writer can view a pop-up window that contains further information about the term, such as a definition or a usage note. Terms marked with a green check mark are approved, while terms with a red circle with a line through it are prohibited. Writers use this information to ensure that the terms they use adhere to company standards. This is a controlled authoring implementation.

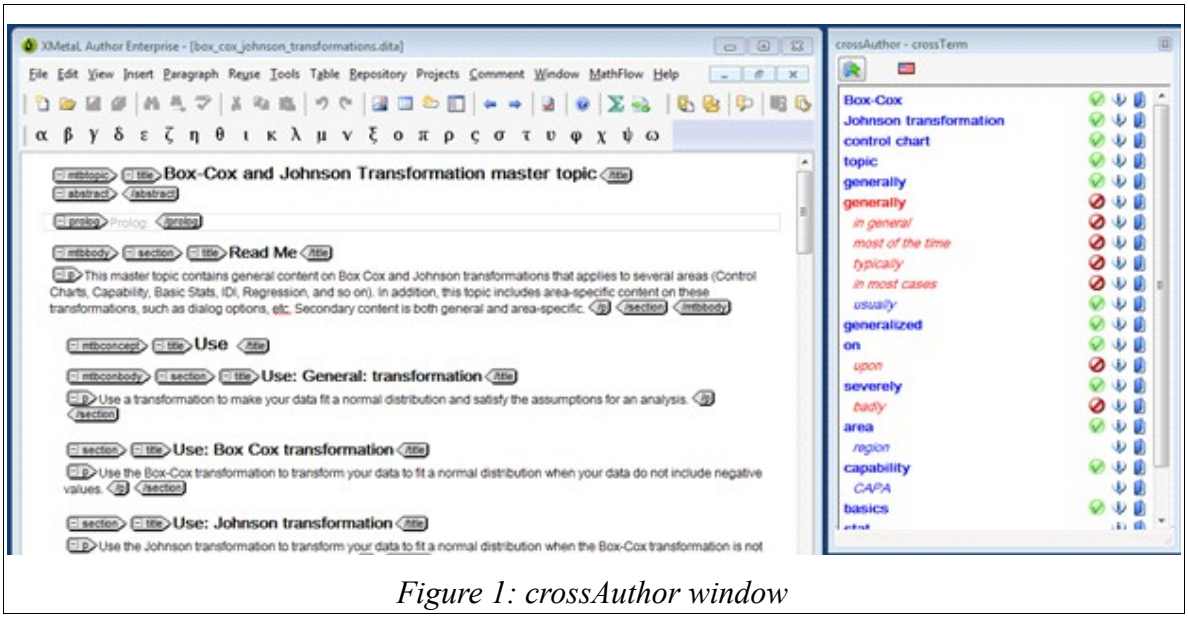


Figure 1: crossAuthor window

The use of the termbase for controlled authoring has implications with respect to the types of terms and other lexical units that the terminologist accepts into the termbase, which some refer to as term inclusion criteria. Any lexical construct that needs to be used by content producers in a certain way in order to achieve the company's goals of consistency, style and technical accuracy qualifies for inclusion. Controlled authoring software provides the opportunity to proactively change writers' choice of words. When a change is required to existing usage, such as to impose the use of a long form in place of an abbreviation, or an industry-standard term rather than a company-specific synonym, the terminologist adds the preferred term to the termbase and indicates that the existing term is to be avoided. In some cases, the preferred term has never or rarely been used by the writers at this point, and therefore, it may not be reflected in the corpus.

Minitab outsources its translation work to one multi-language vendor which uses SDL Trados CAT tools. However, like many companies Minitab maintains intellectual property rights over its TM and manages its TM and terminology resources in-house. Minitab also supports business partners who undertake the translation of its products into languages that are not translated by the vendor, by sharing any necessary resources and tools. This is the main reason why Minitab purchased Across as an in-house localisation tool.

According to Minitab's staff localisation specialist, one of the greatest difficulties when translating software documentation is making sure that the translations in documents describing the user interface (UI) exactly match the translations of the UI itself (sometimes referred to as software strings). TM should take care of this, but in practise it does not always, since software strings are often sub-segment level text. Minitab has adopted an innovative solution to this problem. Rather than translating the software strings in the documentation, translations are directly sourced from the translated UI files through the use of XML referencing mechanisms. This means that when the software is translated and the documentation is built, the strings will be identical. One reason why Minitab is able to achieve this is because DITA[53] is its strategic authoring format.

53 Darwin Information Typing Architecture, an XML authoring standard.

The termbase comprises 2,311 concept entries containing 12,500 terms in 14 languages, 4,121 of which are English terms. Over half the entries (1,420) contain more than one English term, indicating the presence of lexical synonyms and variants. This reflects its use for controlled authoring. For the same reason, the termbase includes terms that are not technical in meaning. For example, one concept entry (that means, basically, *difference*) includes the following terms with the accompanying usage indicator:

- difference - Preferred
- discrepancy - Rejected
- distinction - Rejected

These types of terms range from very general or common in meaning, to the semi-technical vocabulary described in section 3.2.4.

The contents of the termbase were exported from crossTerm by the terminologist and provided as a TBX file. A description of the data categories is reproduced in Appendix C. The termbase is concept oriented, as shown in the following figure.

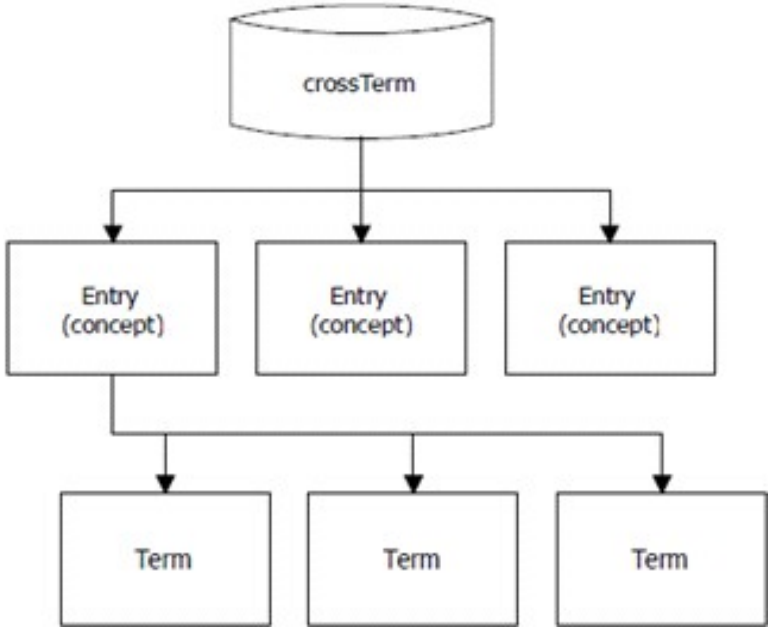


Figure 2: Concept-oriented structure of Minitab termbase

Minitab's corpus comprises 26,128 files containing nearly four million tokens (3,973,265). The files are in three different formats: DITA, XML and HTML.

5.1.2 SAS

SAS is the leader in business analytics software and services, and the largest vendor in the business intelligence market. Based in Cary, North Carolina, the company has more than 13,000 employees located in 400 offices world-wide. The industry focus is on information management, analytics, and business intelligence. SAS software is used at more than 65,000 sites in over 135 countries, including 90 of the top 100 companies on the 2012 Fortune Global 500® list. In 2012 the company reported a revenue of 2.87 billion USD.

SAS began to proactively manage its terminology in 2003. Terminological resources for authoring and for translation purposes are separately developed and managed; the former in North Carolina and the latter in Denmark, in both cases by a full-time terminographer. This research used the authoring-oriented termbase from North Carolina.

Although the terms in the English termbase are ultimately translated since they occur in products that are translated, translators are not using the English termbase itself directly, and therefore, the terms and associated data that we are studying are designed to serve SL employees such as product developers and technical writers and may also appear in English product glossaries. This is the only termbase of the four studied that was developed for authoring processes to the exclusion of translation.

SAS has recently implemented automated controlled authoring by deploying Acrolinx as an authoring aid for some technical writers. However, the lexical data used for controlled authoring is maintained separately from the termbase, in spreadsheets, because the structure of the termbase as designed does not support the granularity of descriptors required. This raises to three the number of separate systems used to manage lexical and terminological resources in the company. Maintaining such inherently overlapping resources in multiple different repositories can lead to significant data redundancy and duplication of effort and resources. The terminologists responsible for managing these three repositories recognise this and are currently collaborating with an ultimate goal of increased consolidation.

The translation of SAS products is carried out in a CAT environment that leverages both terminological resources and TM. Ensuring the consistency of translations of software strings between the software user interface (UI) and citations of those strings in the documentation and online help is paramount. To achieve this consistency, the software strings are translated first, based on relevant TMs. A special terminology dictionary called *Software References* is then created from the source and translated strings. This dictionary is used in the CAT environment alongside the conventional termbase to translate the peripheral materials such as online help. This approach ensures that subsegment-level translations will be consistent, while preserving the distinction between TM-like lexical resources and terminology. SAS uses quite a sophisticated set of proprietary NLP tools to maximise reuse and proper matching of translations and to manage this diverse range of linguistic resources (TMs, TM-based dictionaries, and terminology).

The English termbase is maintained in TermWeb. It was provided in the form of a TBX export file. The following are some statistics about this termbase:

- 4,135 entries
- 4,710 terms
- 575 synonym sets
- 4,135 definitions
- 165 short form, 234 initialism, 11 truncated form, 49 acronym, 22 abbreviation, 3 variant

Each entry contains a definition, which is not typical for company termbases. Coupled with the high proportion of synsets, this reflects the application of the data for authoring and for use in customer-facing glossaries.

The corpus comprises 57,080 HTML files (DDH*) containing 22,136,564 tokens.

5.1.3 Symantec

Symantec Corporation is a global company that produces computer security software. It is head-quartered in Mountain View, California. Symantec develops software for information

security, storage, and systems management. It is one of the world's largest software companies with more than 18,500 employees located in more than 50 countries. It reported revenue of 6.9 billion USD in 2013.

Company terminology is managed by a terminographer based in Ireland. There are two termbases, one for translation/localisation needs and another for controlled authoring. (We studied only the former.) Some of the content in the two termbases overlaps. Symantec uses SDL WorldServer for automated translation project management and computer-assisted translation with TM. Acrolinx is used as the controlled authoring tool.

The localisation termbase was created about ten years ago, and is maintained in the WorldServer TMS, and can be accessed in read-only mode by translators. Writers do not have access since they do not use WorldServer, but use the controlled authoring termbase. One linguistic reviewer per TL acts as a TL terminologist, and has write access to the termbase.

Prior to a translation project, the Acrolinx term harvesting functionality, which is a form of automatic term extraction (ATE), is invoked by an in-house localisation tool and extracts terms from the SL files. The raw output is cleaned and validated by a linguist and the terms are checked against the localisation termbase to identify translations. The TL terminologist translates the terms that lack a translation and checks the translations of the remaining terms to ensure their suitability for the translation project. All the project-specific terms are thus prepared in the WorldServer termbase, which the translators access during translation.

Consolidation of the two termbases was considered but rejected for several reasons. Some Symantec products and other content are not translated and terms from these areas are therefore not required by translators. Some of the terms in the author's termbase are not needed by translators for various reasons, for instance terms that occur very infrequently or are members of the general lexicon. Finally, the data model and data categories in the two termbases need to be different to serve different purposes and target users. It was felt that terms from the authoring termbase and their associated metadata would make the localisation termbase unnecessarily large and more difficult to maintain alongside terms required

for translation. The disadvantages seemed to outweigh the advantages. Nevertheless, the terminologist compares and harmonises the source content between the two termbases on a quarterly basis, and adds terms to one or the other termbase where they would add value.

The localisation termbase has a practical, descriptive focus rather than a normalisation role. Source language terms are added to the termbase if they occur in content to be translated and are deemed to be useful for translators; the term's correctness or perceived quality does not factor into the selection process. There are only two term inclusion criteria: (a) the term is specific to Symantec, in other words, third-party terms are generally excluded, and (b) software strings (error messages, short phrases, etc.) are excluded. Exceptions are permitted. Software strings are excluded from the localisation termbase because a separate database and translation system is used for software localisation. The corpus is checked on an irregular basis to verify the frequency of a term candidate.

The entry structure is intentionally simple to address straightforward needs of term lookup in CAT tools; the termbase is also experimentally used in machine translation. Designed based on requirements of the localisation process, certain limitations of the TMS needed to be accommodated. The termbase is maintained by one full-time terminographer although, as stated earlier, a number of TL linguists contribute entries.

The localisation termbase was exported from WorldServer by the Symantec terminologist and provided in TBX format. It comprises:

- 6,651 entries
- 26 languages
- 3,538 definitions
- 172,907 terms[54] (all languages)
- 6,651 English terms

The fact that the number of English terms equals the number of entries that contain English terms means that there are no English synonym sets in this termbase. This is not surprising given that the termbase is only used for translation purposes. Even though synonym sets for

54 Over half are “TBD,” all of which are place-holders for translations.

the SL fill a definite need in the translation process, by showing for example that two SL terms have the same meaning and can therefore be translated by one and the same TL term, these types of termbases rarely include them[55].

The Symantec corpus comprises 18 TMX[56] files containing 19,808,928 tokens.

5.1.4 Hewlett Packard

The Hewlett-Packard Company (HP) is a multinational information technology corporation head-quartered in Palo Alto, California. HP is the world's leading PC manufacturer. It specialises in developing and manufacturing computing, data storage, and networking hardware, and it also develops software and provides services. Product lines include personal computing devices, enterprise and industry standard servers, related storage devices, networking products, printers and imaging products. Government, health and education are major industry sectors. With over 330,000 employees worldwide and a revenue of over 120 billion USD in 2012, HP is the largest of the companies in this study.

HP has fifteen termbases and various possible selections of corpora. To comply with our research requirement whereby the termbase and the corpus must have a high level of anticipated correspondence, we consulted with the HP terminologist to select an appropriate termbase and corpus for our research.

HP is a very large distributed company. According to the terminologist, the same terms can assume slightly different meanings across the various business units. The company has its own internal translation services, which use SDL Trados for CAT purposes. The terminology is therefore widely used alongside the TM in this environment. Authors and other employees also use the terminology, but only by way of ad-hoc lookup on a Web site. A pilot project is currently under way to determine the feasibility of implementing an automated controlled authoring environment through SDL Author Assistant.

55 This is because localisation termbases are usually produced by translators, who are less informed about SL synonyms than the producers of the SL content.

56 Translation Memory Exchange

In recent years, the focus of HP's globalisation efforts has been on automating translation processes, such as work flows and project management. The development of terminological resources has received less attention and support. Of the 15 termbases, one is a master and 14 are associated with various business units and product lines. The latter are managed on an ad-hoc basis by employees in the business unit, each of whom has other primary roles, such as localisation project manager or translator. The master database contains terms that can be reused across the company, such as global marketing tag lines. The other termbases contain terms, usages of terms, and other expressions that are used in the respective business unit. All the termbases are maintained in SDL MultiTerm.

One full-time terminologist in Grenoble, France, coordinates the effort across the company. Due to the scope of the terminological resources, the diversity of usages, the wide range of subject domains, and a desire to encourage buy-in and ownership among the termbase managers in the business units, it was decided not to consolidate the 15 termbases into one, even though it is acknowledged that keeping them separate results in a significant amount of duplication, in terms of both the data and the maintenance effort. Instead, the strategy is one of a federated system of termbases that ideally, through standardisation of data models, will evolve so that they are complementary and can be utilised together as building blocks. The terminologist calls this “decentralisation with a central gatekeeper.” In terms of structure and design, the master database is the model for the others.

The autolookup terminology functions in SDL Trados are leveraged to ensure translation consistency for any sub-segment level text. This means that the termbases are a repository for virtually any sub-segment level text for which consistency is an objective.

Indeed, as will be shown later, many so-called terms in the termbase are not in fact terms, but rather are longer fragments of text such as partial sentences or even full sentences. After enquiring with the terminologist, we learned how these strings ended up in the termbase. After a translation project is complete, all the software strings (text from the software user interface) and their translations are available in the form of a TM. This TM is imported into the MultiTerm termbase for the business unit and henceforth used in the autolookup

function for future translation projects. Why isn't the TM simply used as a TM? The reason is pragmatic. Software strings tend to be short succinct pieces of text, such as an error message, the label on a button, or the text of a link. They appear on the user interface, often in multiple locations, and are also most likely mentioned in the online help and other user documentation multiple times. The same string is often used in multiple products. Therefore, translating these text strings consistently is very important. However, in many of these contexts the software string is embedded in a larger sentence. The *sentence* is what is matched by the TM function. If the translation of the string alone is present in the TM, it will not be shown to the translator when the string occurs in a larger sentence. Even if a *fuzzy* lookup option is used for searching the TM, the problem still occurs; in HP, a TM is only shown if it corresponds to the sentence being translated by at least 75 percent.

An example might demonstrate this limitation. The expression *Fit to Width* is found in the HP termbase, encoded as a term. This expression does not qualify as a term according to the classical definitions, as it corresponds to more than one concept (action: fit, property: width) and syntactically, it is a phrase (verb+prep+noun). However, it is short: only three words. This string appears in a print dialogue, as a print option. One would expect to see this option described in the online help and user documentation. A typical occurrence might be in a sentence like "To resize the text to fit the selected paper size prior to printing, choose Fit to Width." Thus, this string appears as a sub-segment level text in other sentences. If only in the TM, the translation of *Fit to Width* will not be shown to translators when they are translating sentences like the one above. Adding such strings to the termbase compensates for what is essentially a technical limitation in TM technology. Although this practise is not recommended (Kelly and DePalma 2009: 18), it is frequently required.

Thus, due to constraints of CAT technology, HP considers any frequently-occurring sub-segment level text as a candidate for inclusion in the termbase, regardless of whether this text qualifies as a term in the classical sense.

The termbase selected for our research, IPG DCSL (Imaging and Printing Group), is maintained in MultiTerm, and was provided in MTX format (a proprietary XML format

used in MultiTerm.) The termbase contains:

- 4,221 entries
- 34 languages
- 87,786 terms (all languages)
- 4,403 English terms
- 2,564 term type values[57] (2,512 full form, 20 abbreviation, 29 acronym, 1 noun[58], 2 phrase)
- 182 synsets

HP provided a corpus in the form of two TMX files comprising 400,777 tokens. The corpus represents the DCS (IPG All In One) line of products, which is the corpus most representative of the supplied termbase, according to the HP terminologist.

5.1.5 Summary

All the companies use CAT tools. Symantec and HP use the terminological data primarily for the purpose of computer-assisted translation. Minitab uses its termbase both for computer-assisted translation and for controlled authoring. At SAS, the termbase is used exclusively by writers as a reference and as a source of published English glossaries. All the companies share an interest in using lexical data for controlled authoring, and in fact, all except HP are already doing so to some degree. Curiously, though, only Minitab keeps the lexical data required for controlled authoring in its termbase. SAS keeps this type of data in spreadsheets, and Symantec in a separate database.

Clearly, translating software strings consistently between the UI and peripheral materials such as online help and documentation is a major concern and a challenge, as all four companies have taken specific measures to address this. Minitab references the translations of software UI strings directly into peripheral materials, SAS leverages a combination of TMs and lexical resources in a sequential process, Symantec maintains a separate localisation environment and database for this purpose, and HP has imported software strings and their translations into the termbase. A common objective in all these approaches is to consistently

57 Term type values only occur on the English terms.
58 Obviously a data entry error.

translate *sub-segment level text units*, which are not optimally retrieved in TM.

The approach adopted for managing terminology ranges from highly centralised to decentralised, with Minitab, Symantec, SAS and HP roughly in that order.

5.2 Preparation of the data

Each company's corpus and termbase needed to be prepared for analysis. Preparing the data was very time-consuming. Technical information and support for the concordancing tool, WordSmith, was limited. Error messages were often unhelpful. Techniques to overcome issues had to be developed on the fly using an ad-hoc testing approach, each processing step often took several hours (sometimes many), and some steps had to be repeated several times, each attempt fine-tuned with lessons learned from the previous one.

5.2.1 Preparing the corpora

The corpora needed to be prepared so that they could be analysed by the WordSmith concordancing software.

5.2.1.1 Problems and issues

We experienced numerous problems and issues that overloaded the software and caused processing failures, including the following:

- The number of files was too large
- The files themselves were too large
- The files contained complex markup
- The files contained HTML and XML entities
- Each company provided different file types and used unique markup conventions
- Files contained a mixture of different character encodings
- Some file types were not supported by WordSmith
- The files contained structures that increased the file size, such as redundant spaces and carriage returns

Although WordSmith software supports files with markup, we encountered some limitations with this support. Occasionally we had to make batch changes to the files using UltraEdit and search/replace with regular expressions.

The following samples show some of the different and challenging markup styles in the source files. For ease of readability, in each sample, the actual text is shown in bold italics.

Example 1:

```
<div class="step"><a name="p0evqi20wl8j89n1f3riob0mg14d"></a>
   <div class="paragraph">
   <a name="p0o5ieqbejqya9n1ijrpzj793fg4"></a>Select
   <span class="selectionPath">
   <span class="selection">Model</span>
   <span class="selectionArrow">
   <img src="./images/arrow.gif" alt="then select" width="11"
   height="10"></span>
   <span class="selection">Assignments</span>
   <span class="selectionArrow">
   <img src="./images/arrow.gif" alt="then select" width="11"
   height="10"></span>
   <span class="selection">Show Single Pane</span>
   </span>.
   </div>
</div>
```

Example 2:

```
<tu srclang="EN-US" tuid="24">
   <tuv xml:lang="EN-US">
      <seg>HP is committed to helping customers reduce their
      environmental footprint.</seg>
   </tuv>
</tu>
```

Example 3:

```
<tu creationid="System" changeid="Weronika Semmerling"
tuid="24690790-3" srclang="en" creationdate=
"20100303T112443Z" changedate="20100303T112443Z">
<prop type="x-idiom-source-ipath">
/TRANSPORT_HOME/ILS/Source-English/Nss273_GFX_trans.xls</prop>
<prop type="x-idiom-target-ipath">
/TRANSPORT_HOME/ILS/Target-Romanian/Nss273_GFX_trans.xls</prop>
<prop type="x-idiom-prev-hashcode">1771070271</prop>
<prop type="x-idiom-next-hashcode">-1004421483</prop>
<tuv xml:lang="en">
```

```
<seg>There are malicious files on your computer that may expose
you to identity theft, or cause loss of data</seg>
</tuv>
</tu>
```

HTML and XML entities in the files needed to be replaced by readable text in order to allow a match with corresponding terms in the termbase. The following are a few examples of these entities:

```
Note:   The map object...
If you&#8217;re a Symantec enterprise customer...
Tools &gt; Manage Profiles &gt; Manage
Each party (the &#8220;Indemnitor&#8221;) shall,...
&#8226; 500 MB of available hard disk space
Disclaimers &amp; Legalese
Don&apos;t Know...
Terms &amp;amp; Conditions
IMPORTANT &#8211; The zipped file must be opened...
```

Strings which are not text were sometimes found as tag content, such as in the following example. WordSmith treats the content of tags as text, therefore, these strings would appear in our concordance searches, which would produce some false results. We had to remove these types of elements.

```
<script>function dblclick() function stopError()
{ return true; } window.onerror = stopError;
document.ondblclick=dblclick;</script>
```

There were different character encodings encountered: ascii, UTF-8, and UTF-16. All files had to be converted to UTF-8.

WordSmith cannot process a large number of files at once, or very large files. Therefore, large files had to be split into smaller ones, and large numbers of files merged into a smaller number. Redundant spaces and carriage returns were also removed to reduce the file sizes.

5.2.1.2 Minitab

Minitab's corpus originally comprised 10,245 DITA files totalling 2,646,868 tokens (measured by WordSmith). However, it was later determined that some relevant material

was missing, such as the Web site content, and a large number of HTML and XML files were provided in a second shipment. Out of the total set, some files were found that contained no parsable content, and they were therefore removed:

- 4 graphics
- 661 SCC files
- 850 ditamap files
- 20 dita cache files
- 11 pdf files
- 13 empty txt files

After removal of the above files, as well as several additional small files that were problematic because they contained non-standard markup, the corpus comprises 26,126 files. To facilitate further processing, these files were merged into a more manageable set of 100 files, totalling 3,973,265 tokens.

Initially we used the files in their original format (XML, HTML, DITA) to obtain a word list and an initial batch concordance of the termbase terms in WordSmith. WordSmith has a function to ignore the markup, so we felt that we could preserve the original file formats for the analysis. However, we encountered problems in both performance and output. WordSmith would frequently freeze or crash after several hours of processing. In terms of the output, some of the results indicated that the markup had not been totally ignored. For instance, we found words like *padding* and *float* ranking high in the word list. Upon deeper investigation, it was found that these words only occur in markup (which should have been ignored) as in the following sample:

```
<td colspan=1
     rowspan=1
     style="width: 51px;
          padding-top: 1px;
          padding-right: 1px;
          padding-bottom: 1px;
          border-right-style: None;
          border-bottom-style: Solid;
          border-bottom-width: 1px;
          border-bottom-color: #000000;
          padding-left: 7px;"
     width=51px>
</td>
```

It became clear that the Ignore Markup function was not able to handle markup that spans several lines of text. Removing the markup would have two benefits: the files would decrease significantly in size which would improve processing performance, and the output would not be contaminated with markup strings. Markup was therefore removed using both WordSmith functions and search/replace with regular expressions in UltraEdit.

5.2.1.3 SAS

The SAS corpus is much larger than the Minitab corpus, comprising 57,080 files (DDH*) totalling over 22 million tokens (22,747,120). The file format is HTML. Originally, 71,452 files containing over 30 million tokens were provided, but it was later determined that some of those files correspond to content that was not used by the terminologist when populating the termbase. This content corresponds to the Advanced Analytics Division (AAD), which includes risk assessment and high performance analytics solutions. The documentation for these products is subject to a higher level of security and authorisation. These more sensitive files whose terminology was largely unaccounted for in the termbase were therefore removed from the corpus, after consultation with company representatives.

WordSmith could not complete batch concordances on the SAS corpus, it is thought, due to the size of the corpus. Therefore, the files had to be manipulated in order to make them parsable by WordSmith. The manipulations were carried out by using the WordSmith File Utilities and Text Converter functions:

1. Remove the HTML and XML markup.
2. Merge the 57,080 files into a set of 511 files, averaging 300 Kb in size each.
3. Convert the files to Unicode (8).

5.2.1.4 Symantec

The Symantec corpus was initially provided in the form of 18 TMX[59] files. Initial comparisons of the termbase to the corpus indicated that the gap was very large. We enquired with

59 TMX = Translation Memory Exchange. See: http://www.gala-global.org/oscarStandards/tmx/tmx14b.html

the terminologist and it was realised that some files were missing in the corpus. A new shipment was provided and the initial stages of corpus preparation had to be repeated. This later shipment contained 20 TMX files, ranging in size from small (159 Kb) to very large (454 Mb). Each file contains pairs of segments of text, one is English, and the other is French. The following example shows how these strings are represented in TMX:

```
<tu>
  <tuv xml:lang="en">
    <seg>Increase value.</seg>
  </tuv>
  <tuv xml:lang="fr">
    <seg>Augmenter la valeur ajoutée.</seg>
  </tuv>
</tu>
```

The element <tu> represents a *translation unit*, <tuv> a *translation unit variant*, and <seg> a *segment*.

As we are only interested in the English content, it was necessary to remove the French text. We achieved this by running a Python script against the TMX files[60]. The largest files proved problematic for WordSmith. Therefore, we removed additional irrelevant markup by using UltraEdit's search/replace functions with regular expressions. For instance, the following sample shows one English segment accompanied by a lot of redundant markup. We are only interested in the content of the <seg> element and everything else can be removed:

```
<tu creationid="System" changeid="System" tuid="19308014-0"
      srclang="en" creationdate="20090714T124713Z"
      changedate="20090714T124713Z">
      <prop type="x-idiom-source-ipath">
      /VASONT/en_US/VASONT-ffffff9b000122768a702200f7bb2e00-1-
      GMS/tidVASONT-GMS_cmsidv8657968_de_de
      _25697576__VSNT__xml.xml</prop>
      <prop type="x-idiom-target-ipath">
      /VASONT/de_DE/VASONT-ffffff9b000122768a702200f7bb2e00-1-
      GMS/tidVASONT-GMS_cmsidv8657968_de_de_
      25697576__VSNT__xml.xml</prop>
      <prop type="x-idiom-next-hashcode">190543664</prop>
      <tuv xml:lang="en">
            <seg>Quarantine list</seg>
      </tuv>
</tu>
```

60 Script (tmx-filter.py) graciously provided by Sebastian Fleishner, Department of Chinese, Translation and Linguistics, City University of Hong Kong.

After removing the redundant markup, the largest file was just over 200 megabytes in size. We were then able to use the WordSmith Text Converter function to remove some remaining markup tags, and after we split the largest file into six parts, the largest file was reduced to just over 15 Mb in size. The resultant reduced files were processable by WordSmith. The final corpus contains 19,808,928 tokens.

5.2.1.5 Hewlett Packard

At the suggestion of the HP terminologist, our research focused on a corpus reflecting one product area of the company, the Imaging and Printing Group (IPG). The HP corpus was provided in the form of two TMX files totalling 72 MB in size. As with the Symantec files, the TMX files were bilingual, containing English and French text. We removed the French text by using the same Python script as described for Symantec, after which the total file size was 44 MB. We then used the WordSmith Text Converter function to remove all the XML markup from the files. Finally, as one of the files was too large for further processing (over 6 MB), we split it into five separate files, giving us six files in total. The final corpus contains 400,777 tokens.

5.2.1.6 Summary of changes

The following table shows the effects of the manipulations made to the corpora.

	Minitab	**SAS**	**Symantec**	**HP**
Files received	26,126	57,080	20	2
Format	XML, DITA, HTML	HTML	TMX	TMX
Size	271 MB	985 MB	1,126 MB	44 MB
Number of files after splitting or merging	100	511	25	6
Size after cleaning	24 MB	136 MB	124 MB	2.3 MB
% reduction	91	86	89	95

Table 2: Summary of file changes

5.2.2 Preparing the termbases

As our main interest is to investigate how well each company's termbase reflects the important terms in the corresponding SL corpora, we needed to extract only the English terms, and further, only those that are expected to occur in the corpora. In some cases, this meant applying certain filters to exclude terms that are not expected to occur in the corpora, for instance, because they are marked as "deprecated," meaning that writers are requested not to use them. For convenience purposes, we call the resulting list of terms *corpus-valid*.

Corpus-valid terms are terms that can and in all likelihood should occur in the corpus, according to the information available in the termbase, i.e., they are not marked with any usage indicator stating that their use should be avoided by content creators. Unless terms in the termbase are marked with some negative usage indicator, it is reasonable to assume that they should occur in the corresponding corpus as they are therefore deemed to be part of the ordinary language of the company. As we shall see later, this does not mean that they actually do occur in the corpus. When such terms do not occur in the corpus, the question is raised as to why they are in the termbase at all.

In order to isolate the corpus-valid terms, it was necessary to review all the term-related data categories in each termbase, identify any data categories that characterise terms that should not be expected to be found in the corpus, such as a usage flag of "rejected," and then apply filters to exclude these terms from the corpus-valid set. To assist in these tasks, we imported all the termbases into TermWeb.

5.2.2.1 Minitab

As stated earlier, Minitab uses its termbase for both computer-assisted translation and computer-assisted controlled authoring. To facilitate this repurposing, the terms are categorised more granularly than would otherwise be the case if the termbase were only used for translation. The following table shows the data categories used:

Data category	Type	Usage status	Form	Register
TBX	`<termNote type= "Type">`	`<termNote type="Usage_ Status">`	`<termNote type="Form">`	`<termNote type= "Register">`
Values	Idiom Minitab Command Minitab Control Minitab Dialog Title Minitab Message Minitab Other Minitab Output Minitab Session Command Proprietary Stock Phrase - General Stock Phrase - Minitab Symbol	Allowed Preferred Constrained Rejected	Full Short Surface	Jargon Minitab Neutral Qeystone Technical

Table 3: Data categories used for categorising Minitab terms

Note the usage status value *Rejected* and the register value *Jargon.* The question is whether or not terms marked with these usage indicators should be expected to occur in the corpus at all or in any significant measure.

The *Surface* value of the *Form* data category warrants further discussion. This value is used to identify a particular style of term usage that combines one lexical unit (or terminological unit, as the case may be) with another, for instance, a full form and an acronym, such as:

```
<term>acceptable quality level (AQL)</term>
```

Or a primary term and one or more secondary synonyms, such as:

```
<term>hazard function (also called hazard rate or force of
mortality)</term>
```

Note that in these cases there is actually more than one term enclosed in the <term> element, a practise that violates the principle of term autonomy. However, the Minitab terminologist intentionally adopted this practise to help impose a specific content authoring style rule: Use this form on first occurrence of the term in the product content. In other words, the term *hazard function*, when first introduced in a text, must be accompanied by an explanation of synonymous words, which the reader may be more familiar with. This rule

also applies to acronyms: when using an acronym for the first time, include its full form. In all cases, these so-called *Surface* forms comprise two or more terms and the use of parentheses characters. Such style rules are common in companies, but rarely are they imposed by means of termbases. Minitab's termbase is an example of a termbase that contains a purpose-based arrangement of terms, in this case, terms that serve company-specific content authoring goals.

The previous cases raise the question about whether all the terms in the termbase should occur in significant measure in the corpus. Actually, if the content is being produced according to Minitab's standards, terms with a usage status of *Rejected* or a register value of *Jargon* should not occur in the corpus, since their use is obviously being discouraged or is inappropriate for customer-facing texts. Terms with a Form value of *Surface* will only occur insofar as Minitab's technical writers have consistently adopted this particular corporate style rule, and further, they should occur infrequently since the rule states that this form of the term should only be used when first introducing the concept (therefore, in a set of product information files, it should be expected to occur only once). Minitab's terminologist confirmed these observations and stated that rejected terms, jargon, and surface form terms should not be considered corpus-valid terms for the purposes of our study.

We performed a concordance on these terms to validate our statistical assumptions. A concordance on the S*urface* form terms shows that 77 percent do not occur in the corpus, and only six percent occur more than ten times. On average, the *Surface* form terms that are present at all in the corpus occur only twice.

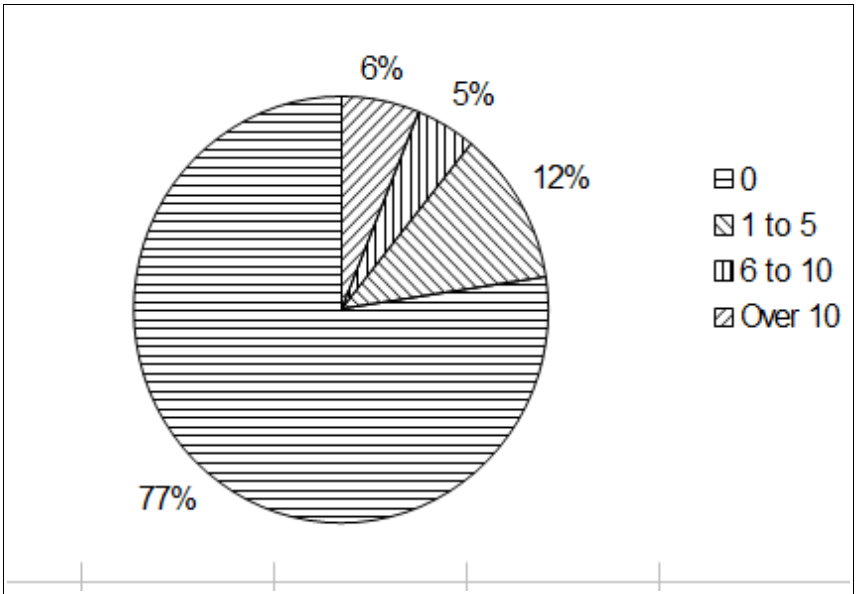


Figure 3: Frequency of Minitab Surface form terms

When these combined terms are split into their individual component terms, it is likely that they will be found more often in the corpus. For instance, take:

```
<term>hazard function (also called hazard rate or force of
mortality)</term>
```

and change it to:

```
<term>hazard function</term>
<term>hazard rate</term>
<term>force of mortality</term>
```

These individually split terms were also checked in the termbase. The number that do not occur dropped from 77 to 20 percent, with most occuring more than ten times.

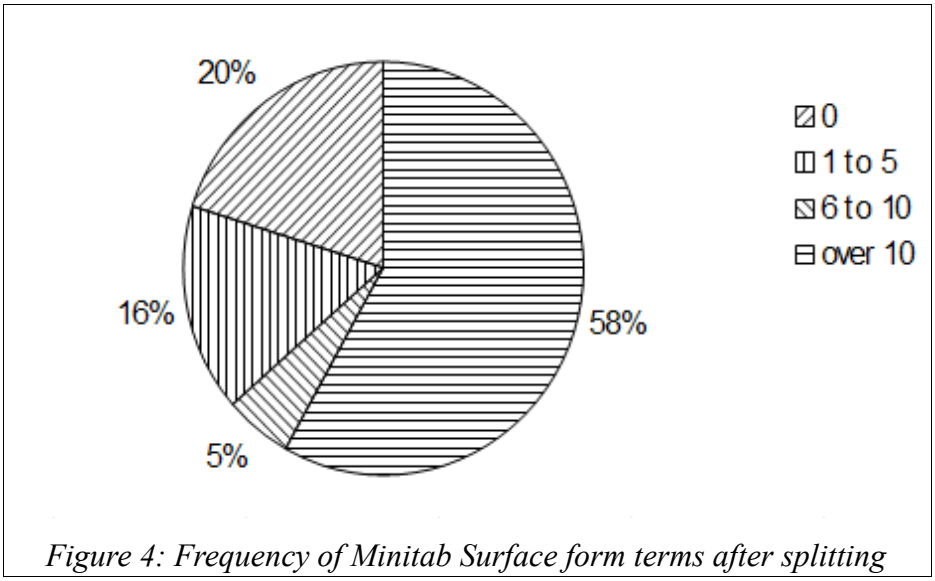


Figure 4: Frequency of Minitab Surface form terms after splitting

The discovery of cases where terms associated with certain data categories in the termbase may not be expected to occur in the corpus led us to realise that these data categories needed to be taken into account in our research. Consequently, we decided to use a TMS, TermWeb, to facilitate searching the data in the termbase according to specific data categories. The use of Xpath to carry out such a filtering procedure on the TBX-encoded format was ruled out, as the data categories of interest are siblings to the term element in the entry structure, and selecting elements based on sibling elements is not possible using Xpath. The following sample excerpt of a TBX entry shows this sibling structure:

```
<tig>
  <term>throughput yield (YTP)</term>
      <termNote type="processStatus">finalised</termNote>
      <termNote type="partOfSpeech">Noun</termNote>
      <termNote type="grammaticalNumber">Singular-Count</termNote>
      <termNote type="Usage_Status">Constrained</termNote>
      <termNote type="Register">Technical</termNote>
      <termNote type="Form">Surface</termNote>
</tig>
```

All the term-related data categories of interest for filtering and research purposes are encapsulated as <termNote> elements, which are siblings to the <term> element in the structure (<termNote> and <term> share the same parent, <tig>). The Xpath query language does not provide a syntax for this type of conditional filtering, i.e. *select <term> elements that have*

a (sibling) <termNote> element with the value "Surface." The Xpath query language only allows selections based on values of parent or child elements.

Using a TMS permits filtering on virtually any data category in the termbase, and even a combination of multiple data categories. The following screen capture from TermWeb shows an excerpt of the results of a filter on the *Surface* value of the Form[61] data category:

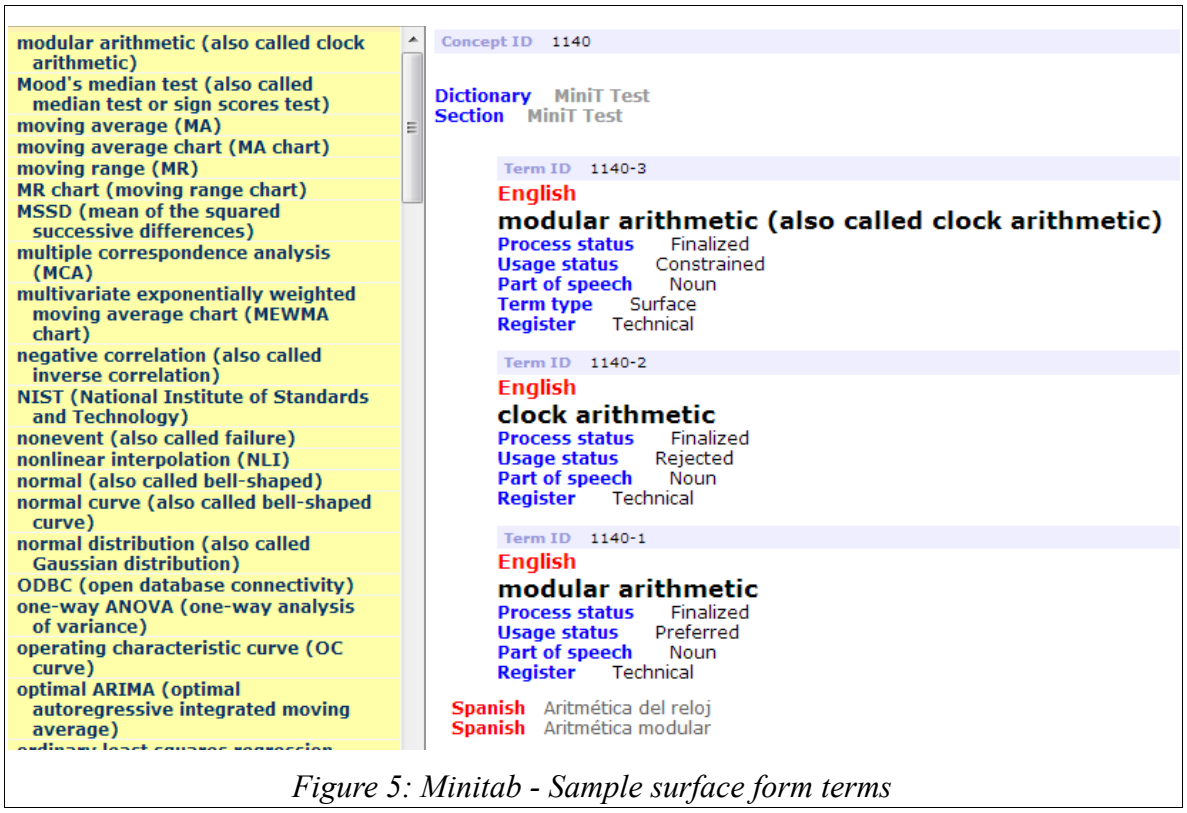


Figure 5: Minitab - Sample surface form terms

Note that the above entry includes the components of the *Surface* term as separate terms themselves (*clock arithmetic*, and *modular arithmetic*). Each of these terms is also adorned with a full set of data categories. This demonstrates that the Minitab terminologist has compensated for the violation of the principle of term autonomy that the *Surface* form term presents, by separately encoding the individual component terms in the concept entry. The following excerpt from the TBX file shows how this is presented in TBX:

61 Since "Form" is not a valid data category name in the TBX standard, it could not be imported into TermWeb under this name. For importing purposes, we mapped it to another available data category, "Term type." This mapping was for convenience purposes only and does not affect the methodology or data analysis in this research.

```
<termEntry>
...
   <tig>
      <term>modular arithmetic</term>
      <termNote type="Usage_Status">Preferred</termNote>
      <termNote type="partOfSpeech">Noun</termNote>
      <termNote type="Register">Technical</termNote>
   </tig>
   <tig>
      <term>clock arithmetic</term>
      <termNote type="Usage_Status">Rejected</termNote>
      <termNote type="partOfSpeech">Noun</termNote>
      <termNote type="Register">Technical</termNote>
   </tig>
   <tig>
      <term>modular arithmetic (also called clock
      arithmetic)</term>
      <termNote type="Usage_Status">Constrained</termNote>
      <termNote type="partOfSpeech">Noun</termNote>
      <termNote type="Register">Technical</termNote>
      <termNote type="Form">Surface</termNote>
   </tig>
      ...
</termEntry>
```

This *Surface* form is a lexical construct recommended by the company style rules for certain contexts. It is comprised of two or more concatenated terms, encoded as a single terminological unit in the TMS. Thus, this construct is treated as a *term*, yet according to the theories of terminology, this construct is not a term. Minitab has departed from the conventions of terminography in order to address a real need in its communications. The *Surface* form is therefore one example where terminography in this commercial environment diverges from conventional theory and practise.

5.2.2.1.1 General lexicon words and expressions

The terms in the Minitab termbase include a large number of words and expressions from the general lexicon. We consider a word or expression to belong to the general lexicon when its meaning is general in nature, that is, not part of a subject field or application specific to the company's content. A few examples of what we consider to be general lexicon expressions, taken from the Minitab termbase, are shown below:

- all
- almost
- also
- can
- did not
- do
- enough
- excellent
- impossible
- minimum
- otherwise
- previous

Many are non-nouns: verbs, adjectives, adverbs, conjunctions, and so forth.

General lexicon words and expressions are not typically included in termbases. They are not terms according to conventional theory, and there is no perceived need to manage them. They do not generally present difficulties for writers and translators, and their meaning is self-evident in their surrounding context. Furthermore, translating a general lexicon word or expression inconsistently usually has little effect on the quality or comprehensibility of the target text. In contrast, inconsistent translations of key terms can have a significant negative effect on the target text (for example, inconsistent names of product features).

We obtained an explanation from the terminologist for the presence of such items in the termbase. Minitab's termbase is used for controlled authoring, which involves controlling not only terms, but also some words and expressions from the general lexicon. A check of several of these items in the termbase shows that each instance involves a terminological entry comprising two words or expressions, one of which has a preferred status and the other a rejected status. For instance, *almost* is preferred, and *nearly* is rejected. In addition, we discovered that the general lexicon expressions had been explicitly marked by the terminologist, in this case by using the Register value of “neutral.” There are a total of 971 words and expressions with this value. (After removal of duplicates due to homographs, this number is reduced to 627.) They are listed in Appendix A.

The following example shows how this metadata is represented in TBX:

```
<termEntry>
 <langSet xml:lang='en'>
  <tig>
     <term>almost</term>
     <termNote type="Usage_Status">Preferred</termNote>
     <termNote type="partOfSpeech">Adverb</termNote>
     <termNote type="Register">Neutral</termNote>
  </tig>
  <tig>
     <term>nearly</term>
     <termNote type="Usage_Status">Rejected</termNote>
     <termNote type="partOfSpeech">Adverb</termNote>
     <termNote type="Register">Neutral</termNote>
     <termNote type="usageNote">Simplified Technical English not
     approved word</termNote>
  </tig>
 </langSet>
</termEntry>
```

Note that the advice to avoid the word *nearly* has come from *Simplified Technical English (STE)*[62], which is a standard for simplified English developed for the aerospace industry that has been adopted in other commercial sectors. It is likely that the recommendation to use *almost* instead of *nearly* also comes from STE, since the two terms are both enclosed in the same entry, although this is not explicit in the TBX markup.

For Minitab, including some general lexicon words and expressions in the termbase is necessary to avoid having to maintain two separate databases: one for general lexicon expressions and another for terms. According to a survey conducted by LISA in 2001, 25 percent of commercial termbases include general lexicon words and expressions (Warburton 2001a: 20). While the survey does not indicate whether these termbases are operated by the same companies that are implementing controlled authoring, we can consider the possibility that they are, given the need to control some general lexicon expressions as part of corporate style. With the increasing adoption of controlled authoring software, this figure is likely to be higher today. Therefore, this is yet another example where terminography in commercial settings diverges from conventional theory and practise.

62 See: http://www.asd-ste100.org/

Prior to running a concordance of the terms from Minitab's termbase against the corpus, it was necessary to remove the items marked with the neutral register value from the list of terms submitted to the concordancer, for two reasons:

1. General lexicon words and expressions would generate an extremely large number of concordances. Not only would this make it impossible to effectively analyse the concordances, but their presence in the total set of concordances would make it difficult to focus on the concordances of domain-specific terms.
2. Our interest is in investigating termbase terms that do not occur, or occur infrequently, in the corpus. General lexicon items are unlikely to fall into this category.
3. None of the other company termbases are used for controlled authoring purposes. The profile of the Minitab termbase therefore diverges from the others. Removing the lexical units that are used for controlled authoring puts the Minitab termbase on equal footing with the others for comparison purposes.

The criteria for determining whether a lexical unit is a member of LSP or of LGP are not clear-cut (see the discussion on semi-technical vocabulary in section 3.2.4). Furthermore, a lexical unit can even traverse from LGP to LSP, a process referred to as terminologisation (Sager 1990: 60; Ahmad and Rogers 2001: 752) and from LSP to LGP, which is referred to as de-terminologisation (Meyer and Mackintosh, 2000). Nevertheless, the 971 words and expressions in Appendix A do, on the whole, appear characteristically general in meaning and usage. However, by identifying these units based on application-specific target usage (controlled authoring), using metadata already in the termbase (neutral register) we adopt functional criteria rather than indeterminate semantic evaluations.

The following figure shows how we used a search filter in TermWeb to obtain the corpus-valid terms.

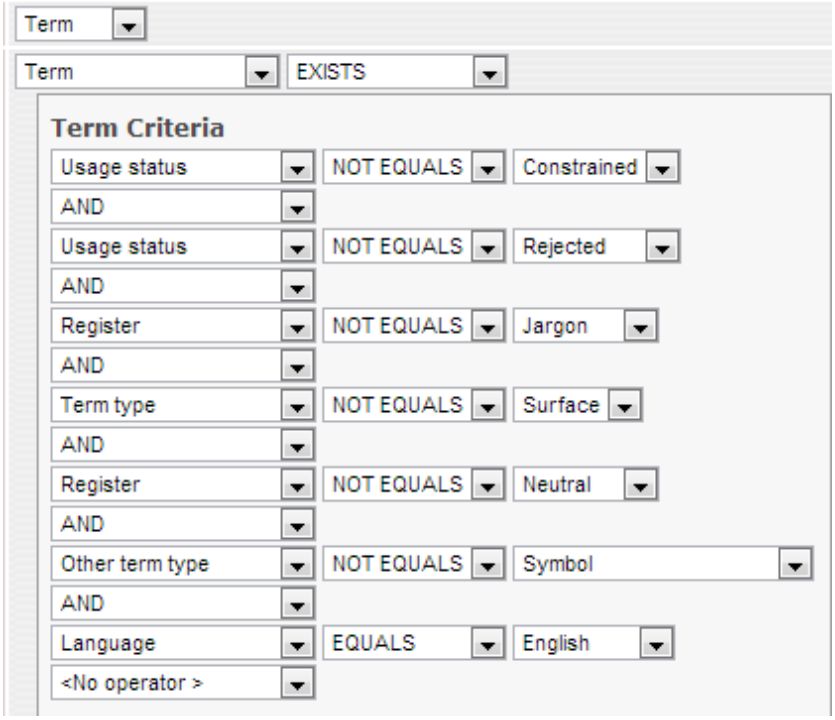


Figure 6: Minitab - Filter for corpus-valid terms

The number of occurrences of each data category value in the termbase is shown below.

Data category	Number of occurrences
Usage status: Constrained	751
Usage status: Rejected	1,063
Register: Jargon	7
Register: Neutral	971
Term type: Surface	219
Other term type: Symbol	27
Total[63]	3,038

Table 4: Minitab data categories marking non-corpus-valid terms

The final set of corpus-valid terms comprises 1,565 entries and 1,785 terms. After removing duplicate lemmas due to homographs and polysemes, as well as three terms that contained a trademark symbol, the set comprises 1,777 terms.

63 The number 3,038 is not the total number of terms, it is the number of data category values used. Some terms are marked with more than one value, for instance, both a Usage status of "rejected" and a Register value of "neutral."

5.2.2.2 SAS

The SAS termbase was provided as a TBX file which had been exported from TermWeb. The termbase contains only English terms.

The termbase does not contain any general lexicon words or expressions, a surprising finding given that SAS uses Acrolinx for controlled authoring, and given the large number of such expressions found for Minitab. As we stated earlier, further investigation with SAS representatives revealed that SAS maintains most of its controlled authoring vocabularies in separate spreadsheet files.

The original termbase provided by SAS contained 4,510 terminological entries and 5,195 terms (after removal of duplicates due to homographs, 4,555 unique terms). An initial evaluation revealed a lower than expected correspondence between the termbase and the corpus, with 17 percent of the termbase terms not found in the corpus at all.

We met with the company terminologist to seek an explanation. Apparently, a subset of the terms in the termbase should not have been included as these terms are used internally only and are therefore not expected to occur in company documentation. A new shipment of the termbase was provided, containing 4,135 entries. This termbase contained two sections, one called *Draft* (292 entries) and the other *Product terms* (3,843 entries)[64]. We were advised by the terminologist to only use the section called *Product terms*.

The Product terms section contains 4,384 terms. All terms have a usage status value: *allowed* (498 terms), *preferred* (3,840 terms), *do not use* (43 terms) and *unspecified* (3 terms). Terms with a *do not use* value are not expected to occur in the corpus to any significant degree, and therefore, they were eliminated from the corpus-valid terms. The final set of corpus-valid terms contains 3,841 entries and 4,341 terms, which were obtained by applying the following filter in TermWeb:

64 There were two other sections as well, but they were empty.

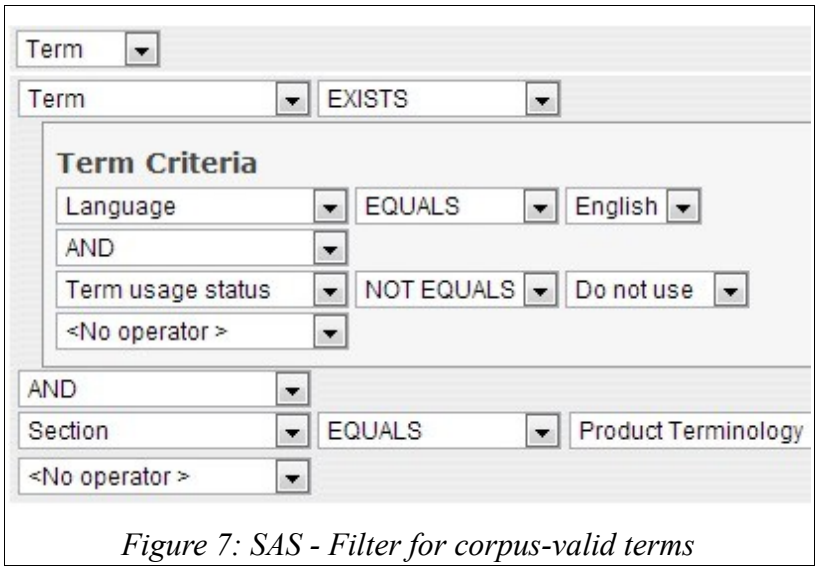


Figure 7: SAS - Filter for corpus-valid terms

After removal of duplicates due to homonyms and polysemes, the final corpus-valid set contains 4,195 terms.

5.2.2.3 Symantec

The Symantec termbase was provided in the form of a TBX file that had been exported from WorldServer, totalling over 150 megabytes in size. It was multilingual, with 26 languages represented. Prior to importing the file into TermWeb, we deleted redundant data categories such as dates and user names, which reduced the file size from 140 to 30 Mb.

Upon examining the terms, we found 31 general lexicon words which we needed to remove for the same reasons that were given for Minitab, as well as three non-English terms which were likely there by mistake. We added a data category in TermWeb (genlex="yes") and assigned it to these entries. We were then able to apply the following filter to obtain the set of corpus-valid terms. The removed terms are shown in Appendix B.

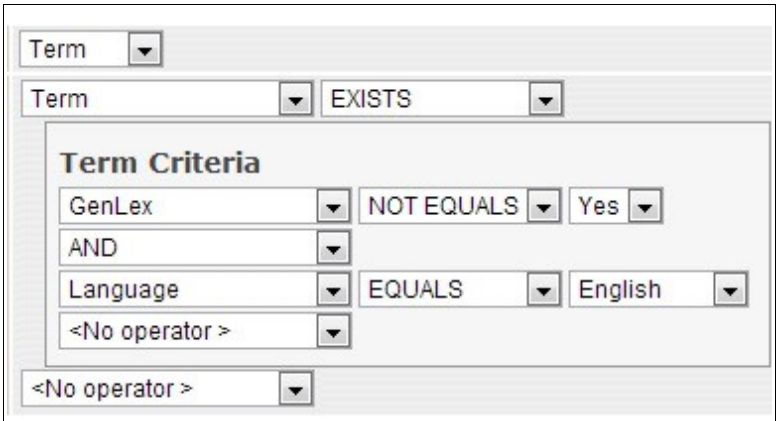


Figure 8: Symantec - Filter for corpus-valid terms

The final set of corpus valid terms contains 6,617 entries and 6,617 terms. This matching number indicates that there are no synsets in the termbase. After removal of the duplicate terms due to homonyms, polysemes, and redundant duplication, there are 6,441 unique terms that comprise the corpus-valid set.

5.2.2.4 Hewlett Packard

The IPG DCSL termbase was provided in MultiTerm XML format (MTX). Using search and replace with regular expressions in UltraEdit, we converted the MTX markup into standard TBX markup, and then imported it into TermWeb. Since none of the entries contain usage information, all are valid based on the information provided, and therefore, no export filter was applied. The termbase contains 4,221 entries and 4,403 terms. After removal of the duplicates due to polysemes and homonyms, there are 4,385 terms.

5.2.2.5 Corpus-valid terms

Once the termbases were prepared and the English terms extracted as described in the previous sections, the following number of unique termbase terms remained. These are the terms that we consider to be corpus valid for the purposes of comparisons with the corpora.

	Total termbase terms (includes duplicates due to homographs)	**Unique corpus-valid terms**
Minitab	4,121	1,777
SAS	4,710	4,195
Symantec	6,651	6,441
HP	4,403	4,385

Table 5: Corpus-valid terms

By *unique*, we mean that multiple instances of terms having the same surface form, i.e. homographs, were reduced to one instance. This is what was meant when we stated, for each company, that duplicates caused by homographs and polysemes were removed. If concept orientation was respected, a homonym or a polyseme would result in two occurrences of the same surface form of a term in the list of corpus-valid terms. For example, the term *plot* may exist in a termbase in two separate entries, once as a noun and once as a verb. (There are also cases of unjustified duplicates in the termbases, such as two entries for *plot* having the same meaning.) However, in our research we are running concordances of terms based only on their surface form, that is, for the most part, the part of speech value of a term is not taken into account when we consider its frequency in the corpus. This constraint is due to the fact that our corpora are not part-of-speech tagged[65]. But also, we found that the part of speech metadata in the termbases was not reliable for two of the four companies studied. This will be further explained later.

65 It is out of the scope of this research to tag the corpora. This condition is acknowledged in the concluding remarks about limitations and future work.

CHAPTER 6 ANALYSIS OF THE DATA

In this chapter, we analyse the termbases, and study the terms they contain in relation to the corpora. In particular, we identify terms in the termbases that occur infrequently in the corpora, and attempt to explain why they ended up in the termbase. We also examine features of terms in the corpora, such as word class and variants.

6.1 Analysing the termbases

6.1.1 Review of key standards

Prior to analysing the termbases, in this section we wish to briefly describe several key ISO standards that motivate our assumptions regarding sensible termbase modelling.

ISO 16642 (2004), *Terminological Markup Framework (TMF)*, provides a meta-model for termbase design. According to this standard, a terminological entry describes one and only one concept and consists of three hierarchical sections: terminological entry (TE), language section (LS), and term section (TS)[66], as shown in the following figure.

66 An optional fourth level, term component section (TCS), can be included to record information about the single-word components of multi-word terms. However, this is rare in commercial termbases and is typically associated with lexical resources developed for natural-language processing applications.

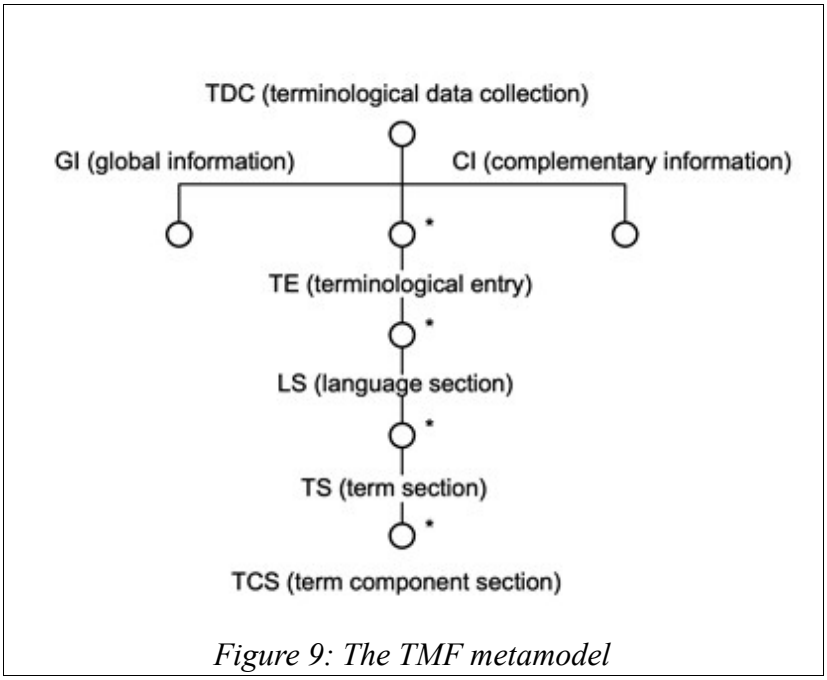


Figure 9: The TMF metamodel

The terminological entry section usually contains descriptive information pertinent to a concept, such as a definition and subject field, and administrative information about the entry. For this reason, it is often referred to as the concept section or level. The language section is a container for all the term sections for a given language, as well as information pertaining to the concept in that language. For example, it may contain a definition in the given language. The term section contains exactly one term, and information about the term, such as the part of speech, term type, and a context, and is repeatable to allow for lexical synonyms and variants of the main entry term.

The TMF standard did not define this structure as a novel concept but was merely reflecting existing widely-recognised best practises. The leading termbases of the time already reflected this structure[67].

ISO 30042 (2008), *TermBase eXchange (TBX)*, defines a framework for specifying XML-based terminology markup languages (TMLs), and includes one such TML, called TBX-Default. TBX complies with TMF. The aim of the standard is to facilitate the development of terminological resources that are interoperable. TBX complements TMF by specifying

67 For instance, Termium, Eurodicautom, and so forth.

over 100 data categories that could be included in a termbase, and where in the TMF metamodel they can occur. For instance, it allows definitions to occur in any of the three sections, but restricts subject fields to the terminological entry section, and the part of speech to the term section. Again, TBX reflects established industry best practises.

TBX Basic is an industry-recognised TML that is a simpler version of TBX-Default, comprising only 24 data categories instead of over a hundred. It was developed by the Terminology Special Interest Group of the Localization Industry Standards Association in 2009 and updated by its successor, TerminOrgs[68] in 2014. TBX Basic adds several further restrictions to TBX, for instance, definitions are not allowed in the term section, and the part of speech is restricted to six pre-defined values (noun, verb, adjective, adverb, proper noun, other), whereas in TBX-Default the set of possible values is entirely open. TBX Basic is considered a good general guideline for developing termbases.

The ISO TC37 *Data Category Registry* (DCR)[69] is an online repository of descriptions of data categories that can be used in termbases and other language resources such as lexical works. It complies with ISO 12620 (2009), *Specification of data categories and management of a Data Category Registry for language resources*. The DCR was intended to standardise data categories, thereby increasing interoperability. While this objective has yet to be reached, it remains a useful resource of information about data categories.

6.1.2 Entry model and data categories

The four termbases were compared with respect to the data entry hierarchy (data model) and the selection of data categories. For this comparison, the exported file was used; the HP database is in MultiTerm XML format, and the remaining three are in TBX format. In all cases, the entries are structured in three levels: concept, language and term. The following table shows the various data categories found in the termbases as well as the level in the entry model where each data category occurs. There are no data categories at the language level, except the language identifier itself.

68 www.terminorgs.net
69 www.isocat.org

Level	Data category	Values	HP	Minitab	SAS	Symantec
Concept	Concept ID		x	x	x	
	Section				x	
	Category		x			
	Subject field			x		
	Domain			x		
	Product group		x			x
	Project					x
	Platform					x
	Created by		x	x		x
	Creation date		x	x		x
	Modified by		x	x		x
	Modification date		x	x		x
	Definition		x	x	x	x
	Source of definition			x	x	
	Definition review status				x	
		finalised			x	
		unprocessed			x	
		provisionally processed			x	
	Definition comments				x	
	Context					x (Note 3)
	Thesaurus descriptor			x		
	Comment					x
	(target language) comments					x
	Status (Acrocheck validation)					x
	Part of speech					x
		noun				x
		proper noun				x
		adjective				x
		adverb				x
Language	Lang ID		x	x	x	x
Term	Term		x	x	x	x
	Term ID			x	x	
	Status (process)		x	x	x	x
		approved	x			
		finalised		x	x	
		in review	x			

Level	Data category	Values	HP	Minitab	SAS	Symantec
		provisionally processed		x	x	
		unprocessed			x	
		processed				x
		proposed				x
	Status (usage)			x	x	x (Note 1)
		preferred		x	x	
		rejected		x		
		allowed		x	x	
		constrained		x		
		admitted				x
		do not use			x	
		unspecified			x	
	Register			x		
		Neutral		x		
		technical		x		
		Minitab		x		
		Qeystone		x		
	Subject field			x		
	Program/project		x			
	Part of speech (See note)		x	x	x	
		noun	x	x	x	
		Proper noun		x		
		Noun phrase			x	
		verb	x	x	x	
		adjective	x	x		
		adverb	x			
		other	x	x		
	Gender			x		
		Neutral		x		
	Number			x		
		Singular-count		x		
	Term type		x	x	x	
		Full form	x	x	x	
		Short	x	x	x	
		Abbreviation	x		x	
		Initialism/acronym	x		x	
		Surface		x		
		Phrase	x			
	Usage note			x		
	Context		x	x		
	Source of context			x		
	Comment				x	

Level	Data category	Values	HP	Minitab	SAS	Symantec
	Created by		x		x	x
	Creation date		x		x	x
	Modified by		x		x	x
	Modification date		x		x	x
	Type					
		Do not translate		x		
		Idiom		x		
		Proprietary		x		
		Stock phrase – General		x		
		Stock phrase – Minitab		x		
		Symbol		x		
		Transcription		x		
		Transliteration		x		

Table 6: Data categories in the termbases

Notes:

1. Although the Usage Status data category is present in the Symantec termbase, all entries have the same value: admitted. Also, there are no synonym sets in the termbase. These two facts render this data category redundant.
2. The HP termbase contains 87,770 terms (4,220 entries, multilingual version), but only 620 instances of the part of speech data category. Thus, less than one percent of the terms are marked with a part of speech value.
3. In the Symantec termbase, the Context and Definition often contain the same text. This suggests that an automatic extraction process was used, which confirms the information provided by the terminologist (see section 5.1.3).

Data categories used by at least three of the four companies are shown with a grey background. Aside from the administrative data categories (dates, identifiers, etc.), which are inserted automatically, the notable common data categories are:

1. Definition
2. Part of speech
3. Process status
4. Usage status
5. Term type

The part-of-speech is present in all four termbases. Symantec encodes it at the concept

level, and the remaining at the term level. The term level is the correct location according to the TBX standard. Encoding it at the concept level has the consequence that all terms in the concept entry must have the same part-of-speech. Although this is ordinarily the case, exceptions can occur, such as in the area of software localisation, where a label on a software user interface can be expressed by a verb in one language and by a noun in another.

It should be noted that, when stating that a data category is used, we mean that it is present in the termbase entry model and structure. This does not mean that the data category always contains any content or values. For instance, as just noted for HP, the part-of-speech is often omitted. While all termbases have a field for definitions, only SAS includes definitions for each entry, and Minitab provides a definition for about half its entries. The other companies rarely include definitions. Further, the Usage status occurs in three termbases, but in reality only two, since all terms in the Symantec termbase have the same value.

All termbases feature sub-setting at the concept level, with different data category names for the subsets (Section, Category, Subject field, Domain, Product group, Project, Platform). The subject field data category is present only in the Minitab termbase, and it is repeated at the term level (this occurred as a result of the structure of the imported translation glossaries). Furthermore, all the values are “general” with the exception of about 125 which are “unknown.” Minitab subsequently added a replacement field, called Domain. This field will eventually be useful for categorising the termbase according to semantic criteria, but at this early stage of the termbase's development, the use is limited: 797 occurrences (covering about 30 percent of the termbase), nearly all of which are “statistics” (716), 35 are “Minitab,” 21 are “general,” and 25 are “computing.”

6.1.3 Size of the corpus in relation to the termbase

It is important for demonstrating the validity of our findings that we first establish that the ratio of the size of each company's corpus to the size of its termbase is similar. Major differences in this ratio could signal a problem of representativity. Let us first consider the views of scholars about corpus size in general. First, we recall the size of our four corpora.

	Corpus size in tokens
HP	400,777
Symantec	19,808,928
Minitab	3,973,265
SAS	22,136,564

Table 7: Size of the corpora

Domain-specific corpora used for terminography can apparently be smaller than corpora used for lexicography (Rogers and Ahmad 1994: 849; Engwall 1994: 50); Rogers and Ahmad suggest tens or hundreds of thousands instead of millions of words[70]. We suggest that this assessment is suitable for the classical motivations for terminology work, that is, building terminologies as conceptual networks. Indeed, Meyer and Mackintosh heavily emphasise conceptual richness (1996: 267) more than size as a key criterion for corpora used for terminology research purposes[71]. Chung suggests that a corpus for specialised purposes should contain at least 100,000 words in order to provide reliable statistics about the nature of lexical items (2003: 225). There is general agreement that a corpus that exceeds 250,000 words cannot be considered *small* (Flowerdew 2004: 19). However, in commercial settings, where termhood is, as we maintain, significantly motivated by frequency considerations, the larger the corpus the greater the correspondence will be between term selection and needs for production-oriented terminology. Nonetheless, these views seem to confirm that the size of the corpora used in our research is more than adequate.

The following table also includes the number of termbase terms. The size of the corpus in relation to the termbase is then calculated by dividing the number of corpus tokens by the number of termbase terms. Thus, for instance, HP's corpus is 91 times larger than the termbase (400,777 divided by 4,385).

70 In her study of terminological variation in the field of genetic engineering, Rogers used a corpus of 34,000 words (1997: 222)

71 Much research has subsequently been carried out in Canada on semi-automatic construction of knowledge-rich corpora and extraction of knowledge-rich contexts, as resources for terminographers. See for example works by Ingrid Meyer, Caroline Barriere and Lynne Bowker.

	Corpus size in tokens	Corpus-valid terms from termbase	Size of corpus in relation to termbase
HP	400,777	4,385	91
Symantec	19,808,928	6,441	3,075
Minitab	3,973,265	1,777	2,236
SAS	22,136,564	4,195	3,074

Table 8: Size of the corpora in relation to the termbases

The HP corpus is disproportionately small, being only 91 times larger than the termbase, whereas for all the other companies the corpus is several thousand times larger than the termbase. This could be a warning sign that the data analysis for HP may not be reliable, as many termbase terms might not be expected to occur in a corpus that is insufficiently large. On the other hand, the HP corpus is not too small according to the scholars cited above.

6.1.4 Observations

All the termbases exhibit some problems. The way Minitab handles the subject field data category constitutes data redundancy, and the values used are not meaningful. The replacement field, domain, is not yet mature with sufficiently diverse values. In the Symantec termbase, having language-specific comments, the part of speech, and the context at the concept level violates all the standards mentioned in section 6.1.1. As it turns out, this was done because of a limitation in the SDL WorldServer TMS: it does not support exporting and importing of data in comma-separated-value (CSV) format at the term level. The HP termbase violates the best practises of term autonomy and concept orientation by placing terms in fields other than the Term field, such as in the Definition field. Conversely, comments can be found instead of terms in the Term field. The SAS termbase is well designed and properly used, with the exception that the language section is repeated for each English term in an entry, a practise that also violates the aforementioned standards. However, this could be due to the export routine. Also, abbreviations are repeated in the definition field, which is a form of redundancy.

Minitab's termbase exhibits the greatest variety of data categories. This may be due to the fact that the termbase is newer (therefore more recently designed), the terminologist has received some formal training, and the termbase is used for both controlled authoring and translation. Terminologists in the early stages of developing a termbase may include more data categories than they actually need, preferring to remove data categories later discovered redundant or unnecessary for some reason, rather than add new ones mid-stream.

The SAS termbase is the only one that contains a definition for each entry. Definitions are important for this termbase because it is used to produce English product glossaries. This also explains why SAS has a review status for definitions, and a data category for comments about the definition. This data category can be used as an export filter to ensure that only finalised definitions are exported for published glossaries.

Symantec is the only company that does not include a term type data category, to represent acronyms and other variants. Further investigation shows that this information is sometimes included in the language-specific comment field at the concept level, for instance:

```
<descrip type='German_Comments'>acronym: TCO</descrip>
```

There are several problems here. First, the content of the field does not match the purpose of the field. This field is for comments, and “acronym: TCO” is not a comment. Second, *TCO* is a term. By failing to put it into the term field, this term will not be retrieved through a search (either by a human user or an automated application), and will not appear in the alphabetical navigation list. Furthermore, this practise violates the principle of term autonomy. Third, this field contains two different types of information, a term, and a type value for this term. Combining multiple types of information in one data field violates the principle of data elementarity. Finally, marking the term type in this manner makes it impossible to apply a filter based on term type, for instance, to find all acronyms in the termbase. Other problems include field misuse. For instance, the definition field in the Symantec termbase includes usage notes as well as definitions.

SAS and Minitab use their termbase for controlled authoring. This explains the wide range of data categories available to track usage of terms. But the companies have adopted different approaches. At Minitab, all information required for Acrolinx is maintained in the termbase, whereas SAS maintains most of the required data in a separate repository (spreadsheets), and then imports subsets of that data into the termbase so that it is also visible through the Web interface. This may explain why SAS does not have a dedicated field for usage notes; the Comment field is sometimes used for this purpose.

Given that the notion of subject field is so pivotal to LSPs and to terminology, it is puzzling that none of the termbases have adopted this data category (it is present in the Minitab termbase, but not used effectively, as previously noted). We wonder if this has any bearing on the criteria for termhood that have been used; if there are no subject fields, and rarely any definitions, are any semantic criteria at all used to establish termhood?

6.2 Analysing the termbase terms

In the following sections we examine various properties of the terms found in the termbases, such as case, length, variants, and word class. Our aim is to eventually establish a co-relation between certain properties and low corpus frequency. But first, we establish the frequency of the termbase terms in the corpus as our baseline statistic.

6.2.1 Frequency

In Section 2.5.1, we noted that frequency of occurrence determined through corpus analysis is a key criterion used by lexicographers to decide which words to include in a dictionary. We maintain that it should also be a guiding criterion for commercial terminographers due to the production-oriented needs of their working environment. Indeed, Daille and Kageura observed that core frequency is a good indicator of terminologisation (Jacquemin 2001: 38-39), as did Cabré (1999-b: 137). The first and most important data analysis task is therefore to determine, for each company, how frequently the termbase terms occur in the corresponding corpus. This allows us to measure the size of the gap between the termbases and

the corpora. This measurement is also our baseline against which we aim to identify and evaluate techniques to reduce the gap. To establish this measurement, we ran a concordance of the corpus-valid terms against the corpus twice, the first being case-sensitive, and the second case insensitive.

Because the number of termbase terms, and the respective corpora, are of different sizes for each company, it was necessary to first establish a common statistical basis for comparison.

6.2.1.1 Normalising the frequency counts

Since our four corpora are quite different in size, the frequency of individual terms in the four corpora need to be normalised to a standard corpus size, in order to be directly comparable to each other. Biber et al note that, "if the texts in a corpus are not all the same length, then frequency counts from those texts are not directly comparable" (1998: 263). This principle is also relevant when comparing different corpora (as opposed to different texts). Through normalisation, the raw frequency counts from different-sized corpora are adjusted so that they can be compared accurately. McEnery and Hardie (2012, 49) provide the following formula:

$$NF = \frac{F \times B}{C}$$

where NF is normalised frequency count, F is frequency of a word in a corpus, B is the base of normalisation, and C is the corpus size in tokens. The formula reads as follows:

"The normalised frequency count of a word equals the frequency of the word in a corpus multiplied by the agreed base of normalisation and divided by the total word count (tokens) of the corpus." The base of normalisation is a corpus size chosen as the standard for comparison purposes, such as: one million.

We considered normalising the frequency counts using a base of normalisation of one million, four million and ten million tokens. The figures required to normalise the frequency counts for each base of normalisation, known as the normalisation factors, are

shown below. The normalisation factor is calculated by dividing the base of normalisation by the actual corpus size for each company (B divided by C in the formula). For instance, the normalisation factor for SAS for a corpus size of four million tokens is calculated as follows:

$$\frac{4{,}000{,}000}{22{,}136{,}564} = 0.181$$

This means that the raw frequency counts of SAS terms must be multiplied by 0.181 in order to be transformed into a normalised frequency count.

	Minitab	**SAS**	**Symantec**	**HP**
Size of corpus (number of tokens)	3,973,265	22,136,564	19,808,928	400,777
Normalisation factor to 1M	0.2517	0.0452	0.0505	2.4952
Normalisation factor to 4 M	1.006	0.181	0.202	9.98
Normalisation factor to 10 M	2.517	0.452	0.505	24.952

Table 9: Normalisation factors

To avoid distorting the results through inflation by excessive factorisation, we chose to normalise the frequency counts to a corpus size of four million tokens.

6.2.1.2 Average frequency of termbase terms

Using the normalisation factors to a corpus size of 4 million calculated in the previous section, we can now compare the average frequency of the termbase terms in the corpora.

	Minitab	**SAS**	**Symantec**	**HP**
Total concordances	355,072	2,283,004	2,368,703	84,328
Number of termbase terms	1,777	4,195	6,441	4,385
Average frequency of terms	199.82	544.22	367.75	19.23
Normalisation factor	1.006	0.181	0.202	9.98
Average frequency, normalised	201.01	98.50	74.29	191.93

Table 10: Average frequency of termbase terms

The results are shown in the following figure.

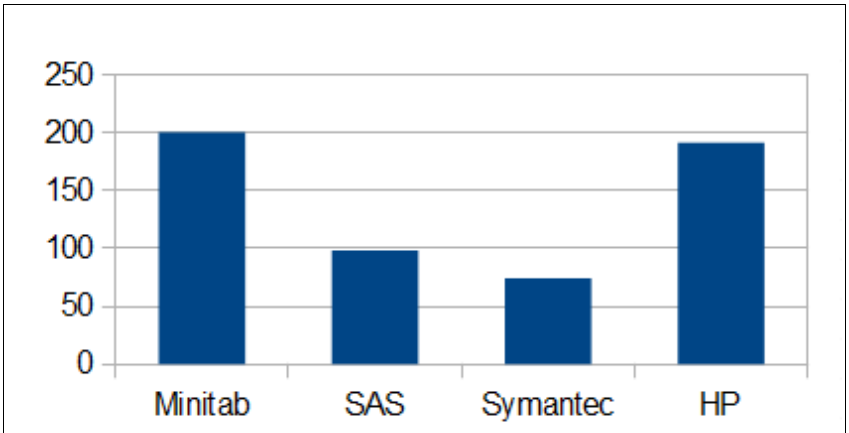


Figure 10: Average normalised frequency of the termbase terms

Here we can see that Minitab performs the best, followed by HP. However, the figure for HP is likely distorted by over-factorisation resulting from a normalisation factor that is nearly ten fold that of Minitab. A term such as *printer*, which occurs over 5,000 times in the HP corpus, may not be expected to occur 50,000 times in a corpus that is ten times larger, especially since such a corpus is more likely to cover a wider diversity of product areas. Conversely, the figures for SAS and Symantec might be distorted by under-factorisation. A term like *hard drive*, which occurs nearly 4,000 times in the Symantec corpus, might not be reduced to only 800 occurrences in a corpus one fifth of that size. Nevertheless, regardless of the potential imprecision of these figures, what we can state with near certainty is that the average frequency of termbase terms is higher for Minitab than for the other companies.

6.2.1.3 Establishing comparable frequency ranges

For certain observations, we are interested in measuring the number of terms in a frequency range, rather than individual frequency values of terms. Since Minitab's corpus of nearly four million tokens was used as the basis for our term frequency normalisation factor in the previous section, we used it also as the factor against which to set frequency ranges for the occurrence of termbase terms in the corpus. We chose ranges of 0, 1 to 10, 11 to 50, and over 50 as frequency ranges for Minitab, and set frequency ranges for the other companies

accordingly. These frequency ranges were set rather arbitrarily, simply to establish how many terms do not occur in the corpus or occur infrequently. In a corpus of four million tokens, we assume that less than 100 occurrences of a term is fairly infrequent, and by extension, certainly less than 50 or less than 10 is even more so. In this case, we are considering the frequency ranges in each corpus separately, as opposed to normalising the frequency to one corpus size. Thus, for SAS, Symantec, and HP, we calculate frequency ranges that are comparable to those defined for Minitab. For example, what is equivalent to a range of 1 to 10 in a corpus of 3,973,265 tokens when the corpus size is 22,136,564 tokens? We produce the required factor by dividing 22,135,564 by 3,973,265, which is 5.57. In other words, SAS's corpus is 5.57 times larger than Minitab's. We then multiply the numbers in the range by that factor, so 1 to 10 in Minitab equals 1 to 58 in SAS.

	Minitab	**SAS**	**Symantec**	**HP**
Size of corpus in tokens	3,973,265	22,136,564	19,808,928	400,777
Factor	1	5.57	4.98	0.1
Frequency: 0	0	0	0	0
Frequency range A	1 to 10	1 to 58	1 to 52	1
Frequency range B	11 to 50	59 to 278	53 to 250	2 to 5
Frequency range C	over 50	over 278	over 250	over 5

Table 11: Comparable frequency ranges

6.2.1.4 Number of termbase terms that occur at frequency ranges

In the following table, we show the number of corpus-valid termbase terms that occur at the comparable frequency ranges:

	Minitab	**SAS**	**Symantec**	**HP**
Frequency: 0	203	530	2,240	3,150
Range A	422	2,103	2,467	187
Range B	500	841	891	319
Range C	652	721	843	729

Table 12: Number of corpus-valid termbase terms that occur at frequency ranges.

These figures are only directly comparable if we represent them as a percentage of the corpus-valid terms from the termbase:

	Minitab	SAS	Symantec	HP
Frequency: 0	11.42	12.63	34.77	71.84
Range A	23.75	50.12	38.30	4.26
Range B	28.14	20.04	13.85	7.27
Range C	36.69	17.21	13.09	16.62

Table 13: Percentage of corpus-valid termbase terms that occur at frequency ranges.

The above table is represented in the following graph:

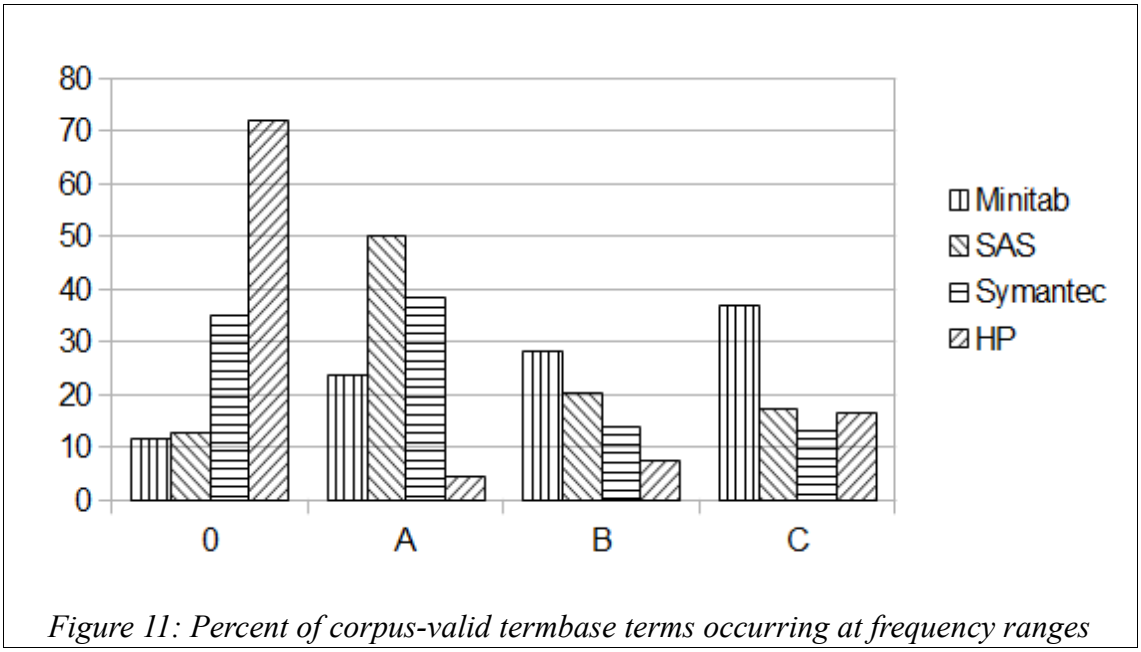


Figure 11: Percent of corpus-valid termbase terms occurring at frequency ranges

Here, we can easily see that HP stands out among the four companies as having the most termbase terms that do not occur in the corpus: a staggering 72 percent. Of the remaining companies, Symantec has the next largest gap with 35 percent of termbase terms not present in the corpus. This finding appears to confirm our earlier concern that the HP corpus is too small to present reliable evidence about the validity of terms in the termbase in a study such as ours that relies heavily on frequency factors.

It is also worth noting that only Minitab's results resemble a presumed ideal, that is, the number of terms increasing in prevalence from 0 occurrences to the highest range (C). Overall, this suggests that Minitab's termbase has the smallest gap in relation to its corpus.

6.2.2 Case

Because we will be investigating how the case of termbase terms affects their match rate in the corpus, we need to know the relative proportion of upper and lower case (corpus-valid) terms in the termbases.

	% lower case	% upper case
Minitab	77	23
SAS	76	24
Symantec	46	54
HP	10	90

Table 14: Proportion of upper case and lower case terms in the termbases

Here we note a marked contrast between the first two and the latter two. In the case of Minitab and SAS, only about one quarter of the termbase terms are in upper case, whereas for Symantec and HP the percentage is much higher.

6.2.3 Length

Given that multi-word terms (MWTs) are known to be very important in terminology (see Section 2.3.5), we decided to quantify the termbase terms with respect to their length (in number of tokens). We anticipated that this data might help to explain why the gap between the termbase and the corpus is greater for HP and Symantec than for the other two companies. The data is provided in the following table.

Term length	Minitab	SAS	Symantec	HP
1 token	464	1,041	1,030	1,066
2 tokens	880	2,314	2,924	1,348

Term length	Minitab	SAS	Symantec	HP
3 tokens	324	655	1,454	1,056
4 tokens	77	153	615	518
5+ tokens	31	33	419	391

Table 15: Number of termbase terms by term length

These figures are shown as a percentage of the total termbase terms in the following table and chart.

Term length	Minitab	SAS	Symantec	HP
1 token	26.11	24.82	15.99	24.33
2 tokens	49.58	55.14	45.38	30.79
3 tokens	18.23	15.61	22.57	24.10
4 tokens	4.33	3.65	9.55	11.84
5+ tokens	1.74	0.79	6.51	8.94

Table 16: Distribution of termbase terms, by length, as a percentage of termbase terms

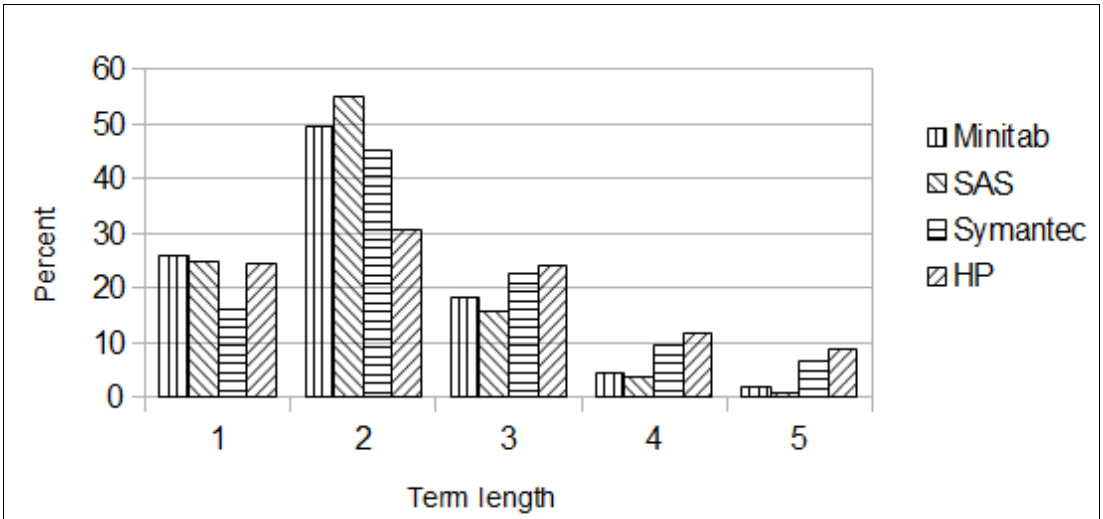


Figure 12: Distribution of termbase terms by length, as % of total termbase

Two-token terms are the most frequent in all termbases. This confirms the general perception and research (Daille et al 1996: 204) that bigrams are be more terminologically relevant than unigrams due to their greater semantic specificity, whereas as the term length exceeds beyond bigrams to trigrams and more, repurposing potential diminishes with decreasing frequency of occurrence.

Interestingly, our data shows that trigrams are almost as common as unigrams in the termbases, and for Symantec they are more frequent. MWTs are clearly perceived to be very important by commercial terminologists. However, note also that Symantec and HP have significantly more terms with a length of four tokens and more (17 and 21 percent respectively) than the other two (5.5 percent). Perhaps not coincidentally, as noted earlier these two companies have a larger gap between the termbase and the corpus than the others. It is logical to presume that longer terms, having more lexical qualifiers, are often more specific in meaning than shorter counterparts, and are therefore likely to occur less often. Furthermore, Rogers observes that long terms are often truncated in running text, a process she calls "stripping," which again reduces the occurrence of the longer form (1997: 220). This finding suggests that long terms contribute to the gap between termbases and their corresponding corpora, and that perhaps the relative importance of terms of two and three tokens should be considered when selecting terms for a termbase. However, as a best practise, the multi-word terms that are needed by writers or translators for various reasons, such as to ensure syntactic consistency or to provide the full form of an abbreviated term, should be included in a termbase regardless of their length or their frequency in the corpus.

If we examine terms comprising six tokens or more, we can see by their nature that they are unlikely to occur frequently in the corpus. In the case of HP, many of these terms are actually full phrases; as such, they belong in a TM, not a termbase:

- Discounts from trusted partners for valued HP customers
- Reset location and privacy settings for shopping online
- Create a separate file for each scanned page
- Click Print to print a test page
- The printer is out of paper

In the case of Symantec, many are product names:

- Symantec Web Security for Microsoft ISA Server 2004
- Altiris Inventory for Network Devices from Symantec
- Brightmail Spam Folder Agent for Exchange
- Norton Smartphone Security Premier Edition User's Guide
- Veritas Backup Exec Remote Agent for Windows Servers

We will validate this assumption in Section 6.3.4.

6.2.4 Word class

We concur with the generally-accepted notion that most terms are nouns. However, earlier we challenged assertions arising out of the GTT that terms are almost exclusively nouns due to their bound relationship with concepts and the objects they represent. We are therefore interested in examining the part of speech, or word class, of the terms in the termbase. In this count, we consider all terms in the termbase regardless of their usage indicator (preferred, deprecated, and so forth). But we exclude the *Surface* form entries for Minitab and the general lexicon units, since they are not terms per se, and out of a concern to maintain comparability of the termbases as described in section 5.2.2.1.[72]

	Minitab	**SAS**	**Symantec (see note)**	**HP (see note)**
Noun	2,339	4,278	4,312	491
Verb	73	60	85	60
Adjective	122	11	72	43
Adverb	4	0	3	1
Other	2	0	0	27

Table 17: Part of speech of the termbase terms

Notes:

1. In the Symantec termbase, the word class (part-of-speech) occurs at the entry (concept) level, as opposed to the term level where it is required by standards. Furthermore, only about two thirds of the entries are marked with a part of speech value (4,472 out of 6,626).
2. Only 622 of the terms in the HP termbase have a part-of-speech value.
3. For SAS and Minitab, nearly all the terms in the termbase have a part-of-speech value (SAS: 4,349 out of 4,384; Minitab: 2,540 out of 2,896).

We can see that most of the terms that are marked with a part-of-speech value are nouns.

72 As such, this selection of terms is slightly different than the set of corpus-valid terms, which explains why these numbers do not match those provided in Section 5.2.2.5

The proportion becomes more clear when shown as a percentage of the termbase terms that have a part-of-speech indicated in the termbase, as in the following table and graph.

	Minitab	SAS	Symantec	HP
Noun	92.09	98.37	96.42	78.94
Verb	2.87	1.38	1.90	9.65
Adjective	4.80	0.25	1.61	6.91
Adverb	0.16	0	0.07	0.16
Other	0.08	0	0	4.34

Table 18: Part of speech of termbase terms, as % of pos-marked terms

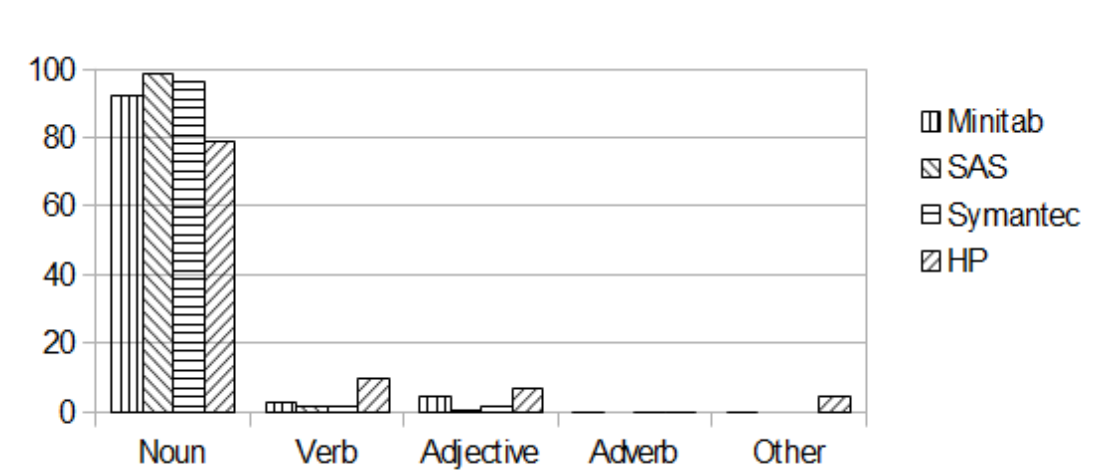


Figure 13: Part of speech of termbase terms, as % of pos-marked terms

In all cases except HP, over 90 percent of the terms are nouns, and we are reminded that the data for HP may be less reliable due to the larger number of terms that do not have a part of speech value.

6.2.5 Variants

We demonstrated in the literature review that terminological variation is recognised by some scholars as a natural form of communication even in LSPs. We are therefore interested to see how prevalent it is in our case studies.

Since in terminography a variant is a term in its own right and also has the same meaning as another term, it is meant to be encoded in the same terminological entry as the term of which it is a variant. For instance, *access control list* and *ACL* should be encoded as two terms in the same entry. The terminologist may decide to use a term type data category to indicate the type of variant, such as *acronym*.

A terminological entry that contains more than one term in a given language can be said to represent a set of synonyms, or *synset*. Synonyms include variants, but also other semantically-equivalent terms which have no similarity at the level of the surface form (for example, in the Minitab termbase, *separator* and *delimiter*, or in the SAS termbase, *rich client* and *thick client*), which as we stated earlier we call *lexical synonyms* to distinguish them from variants.

In a TBX file, synsets are encoded as follows (some elements are not shown for simplicity purposes):

```
<termEntry>
      <langSet>
            <tig>
                  <term>first term</term>
            </tig>
            <tig>
                  <term>second term</term>
            </tig>
      </langSet>
</termEntry>
```

Thus, if we search for the following string in a TBX file, we can find and count the synsets:

```
</tig>
<tig>
```

This pattern will identify all synsets, of which variants would be a subset.

For this exercise, we are interested in establishing the proportion of termbase terms that are variants, regardless of whether or not they are corpus-valid, that is, regardless of whether or not they should be expected to occur in the corpus. Terms that are not corpus-valid in our identification procedures are terms that have a style value in the termbase stating that they

should not be used. However, this does not mean that they do not exist or never existed as variants in the corporate language. Terms with such a usage value should not be *expected* to occur in the corpora to any significant degree, because we assume that writers adhere to such guidelines. If we cannot expect them to occur, then we cannot judge the termbase negatively if they do not occur; we cannot therefore include these terms when we quantify the gap between the termbase and the corpora. That is why we normally exclude such terms from our corpus searches.

However, variants in termbases are presumed to have been added to the termbase at some point in time because they were found to be in use at that time. At some point later in time, some of these variants were given a usage value so that their use could be effectively reduced. We therefore feel that to quantify the prevalence of variants in the termbase all variants should be considered, regardless of any usage indicator.

6.2.5.1 Minitab

Since the Minitab termbase contains many general lexicon entries, and at this stage we are exclusively interested in the occurrence of terminological variants, we need to apply a filter to exclude the general lexicon entries. We also want to exclude the entries with the *Surface* form value, since they are a concatenation of several terms intended to impose stylistic rules, and are not variants. We used the following filter in TermWeb:

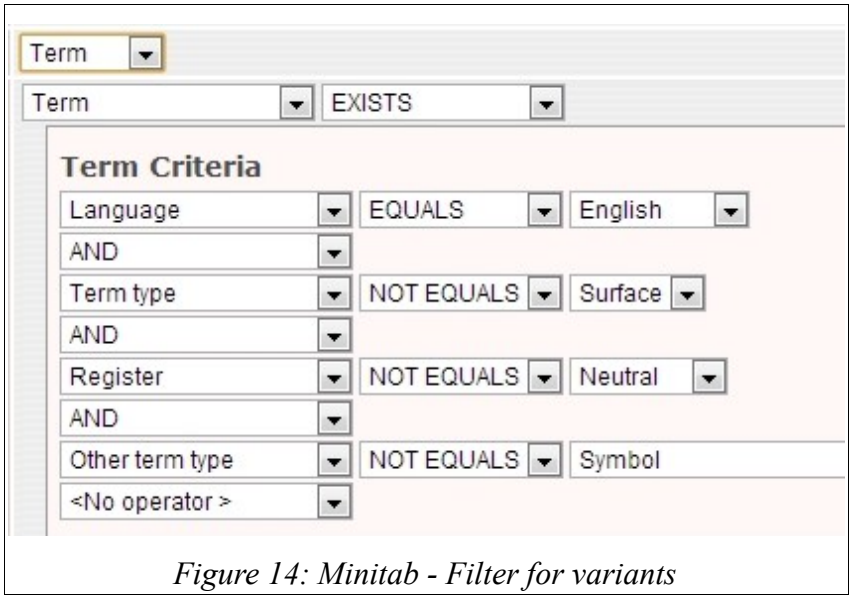


Figure 14: Minitab - Filter for variants

The exported file contains 1,783 entries and 2,892 terms. The following table shows the results of our counting procedure which identifies synsets, a synset containing potentially a variant or a lexical synonym.

	Entries	**Terms**	**Variants or synonyms**
Termbase	1,783	2,896	
Entries with 1 English term	1,039	1,039	0
Entries with 2 English terms	529	1,058	529
Entries with 3 English terms	128	384	256
Entries with 4 English terms	47	188	141
Entries with 5 English terms	26	130	104
Entries with 6 English terms	8	48	40
Entries with 7 English terms	1	7	6
Entries with 8 English terms	4	32	24
Entries with 10 English terms	1	10	9
		Total	1109
Percent of terms that are lexical synonyms or variants			38.35 %

Table 19: Minitab - Synsets

Thus, the termbase contains 1,109 terms that are either a variant or a lexical synonym of another English term. This accounts for nearly 40 percent of all the terms. A cursory

examination of these 1,109 terms indicates that lexical synonyms are less frequent than variants. The following screen capture shows a randomly-selected subset of the output in a spreadsheet. Lexical synonyms are shown in boldface with a grey background to easily distinguish them from variants. These 24 synsets contain a total of 48 synonyms (in the second and third rows), 12 of which are lexical synonyms and 36 are variants.

MANOVA	multivariate ANOVA	multivariate analysis of variance
multi vari	multi-vari	multivariate
model fitting method	model-fitting method	model-fitting procedure
Minitab Support	**Minitab Technical Support**	**Minitab Customer Support**
multivariate exponentially weighted moving average chart	multivariate EWMA chart	MEWMA chart
MSA	measurement system analysis	measurement systems analysis
device	measuring device	measurement device
mean square	MS	mean squares
McQuitty's linkage method	**weighted average linkage method**	McQuitty linkage method
casement display	**draftsman plot**	**matrix plot**
LCL	Lower CL	lower confidence limit
lower boundary	LB	lower bound
actual capability	**overall capability**	**long-term capability**
logistic likelihood	log likelihood	log-likelihood
LRT	likelihood-ratio test	likelihood ratio test
right-skewed	right-skew	**left-tailed**
left skewed	**negative-skewed**	left-skewed
Kolmogorov-Smirnov test for normality	Kolmogorov-Smirnov normality test	Kolmogorov-Smirnov test
KCC	Kendall's coefficient of concordance	Kendalls coefficient of concordance
concordance	**inter-rater agreement**	**inter-rater reliability**
IQRange Box	IQR box	interquartile range box
in-control	**under control**	in control
individuals-moving range chart	Individuals/MR Chart	I-MR chart
IO	input/output	I/O

Figure 15: Minitab - Sample synsets

We decided to further explore the metadata to determine if we could produce a more precise figure than 40 percent. Indeed, the "short" value of the Term Type data category is used to mark acronyms, abbreviations, and truncated forms of long terms. This group of terms accounts for most, but not all, variants. Spelling variants and hyphenated forms, for instance, do not have this value, as shown in the following example.

```
<tig>
   <term>off-target</term>
</tig>
<tig>
   <term>off target</term>
</tig>
```

However, not even all short forms or acronyms are properly marked, as shown in this example:

```
<tig>
   <term>PC</term>
</tig>
<tig>
   <term>personal computer</term>
</tig>
```

The Term Type data category has therefore not been consistently applied. Nevertheless, applying this value to the search filter identifies 478 such terms. There are therefore more than 478 variants in the termbase, which is 17 percent of all terms. We can conclude that probably at least 20 percent of the termbase terms are variants (given that the combined percentage of variants and lexical synonyms is 40).

6.2.5.2 SAS

For SAS, to assess the prevalence of variants in the termbase, we also do not want to exclude any terms based on a usage indicator. For this purpose, there are 3,843 entries and 4,384 terms in the termbase.

	Entries	Terms	Variants or synonyms
Termbase	3,843	4,384	
Entries with 1 English term	3,335	3,335	0
Entries with 2 English terms	481	962	481
Entries with 3 English terms	23	69	46
Entries with 4 English terms	3	12	9
Entries with 5 English terms	0	0	0
Entries with 6 English terms	1	6	5
		Total	541
Percent of terms that are lexical synonyms or variants			12.34 %

Table 20: SAS - Synsets

The SAS termbase includes 541 terms that are either a lexical synonym or a variant, or 12 percent of all terms. We examined these terms and we catalogued 65 lexical synonyms and

476 variants. Indeed, 466 variants were marked with a term type value: 163 short form, 232 initialism, 4 truncated form, 47 acronym, 19 abbreviation and 1 variant. Therefore, about 11 percent of the terms in the termbase are variants.

6.2.5.3 Symantec

The Symantec termbase does not contain synsets or any term type values indicating the presence of variants. This does not mean that the Symantec termbase does not contain variants. Some acronyms and other types of variants are present, but they are encoded as single terms in an entry, without any marker nor with any indication of the term that they are a variant of. For instance, the acronym *AMS* is entered without any indication of its full form, and the two terms *auto-fill* and *autofill* are found in separate entries although there is evidence that they have the same meaning (they have the same translations). Thus, it appears that the principle of concept orientation has not been adopted. There is yet another problem. In Section 6.1.4, we provided sample markup that showed variants encoded in a comment field. Given these multiple instances of lack of compliance with standard markup procedures for variants, it is not possible to quantify variants in the Symantec termbase.

6.2.5.4 Hewlett-Packard

We obtained the following statistics about potential variants from the HP termbase.

	Entries	**Terms**	**Variants or synonyms**
Termbase	4,221	4,403	
Entries with 1 English term	4,044	4,044	
Entries with 2 English terms	172	344	172
Entries with 3 English terms	5	15	10
		Total	182
Percent of terms that are lexical synonyms or variants			4.1 %

Table 21: HP - Synsets

The percentage of 4.1 is significantly lower than for SAS and Minitab. We checked the 182 instances of a </tig> followed by a <tig> and we found that nearly all of them correspond to case differences, and a few to singular/plural variants, as in the following two examples:

```
<tig>
  <term>Black and White</term>
</tig>
<tig>
  <term>black and white</term>
</tig>

<tig>
  <term>Occasions</term>
</tig>
<tig>
  <term>occasion</term>
</tig>
```

These are not true variants, but merely different surface forms that are attributed to the requirements of the running text, such as using title case for the label of a user interface object. These can therefore not be qualified as variants at all.

We then checked the metadata for markers of variants. In the HP termbase, 51 terms are marked with a term type value: 20 abbreviation, 29 acronym, 2 phrase. One would assume that these terms are variants of a corresponding main term, however, upon closer inspection we find that they exist alone in the entry, such as the following two examples, each of which contains only one <tig>...</tig> section and therefore only one term.

```
<termEntry>
    <admin type="termbaseSection">HP IPG DCSL</admin>
    <descrip type="definition">Domain Name Service. When
     you use the web or send an e-mail message, you use
     a domain name to do it.</descrip>
    <langSet xml:lang="en">
     <tig>
      <term>DNS</term>
      <termNote type="partOfSpeech">noun</termNote>
      <termNote type="termType">acronym</termNote>
     </tig>
    </langSet>
</termEntry>
```

```
<termEntry>
    <admin type="termbaseSection">HP IPG DCSL</admin>
    <descrip type="definition">Hexadecimal. The base 16
     numbering system, which uses the digits -9 plus
     the letters A-F.</descrip>
    <langSet xml:lang="en">
     <tig>
      <term>HEX</term>
      <termNote type="partOfSpeech">noun</termNote>
      <termNote type="termType">abbreviation</termNote>
     </tig>
    </langSet>
</termEntry>
```

In these two entries, there is actually a corresponding main term (*Domain Name Service* and *Hexadecimal* respectively), but it has been inserted into the definition field. These terms therefore would be irretrievable in a term search or in an autolookup function. We checked the 51 instances of the term type data category and all of them have the same structure as the two examples given above.

Thus the handling of variants in the HP termbase is problematic. The corresponding main form terms of variants are not entered as terms in the termbase, but are in the definition field. This problem is similar to that found for Symantec, where variants were recorded in a comment field. The <tig> element, intended to encode lexical synonyms and variants in the concept-oriented structure, has been used to store morphological variations that need not be encoded in a termbase at all. This approach was likely adopted to compensate for the limitations of TM matching during auto-lookup of terms in the CAT tool.

Given these issues, we only consider the 51 terms that are accompanied by the term type data category to be true variants, representing, only one percent of the termbase.

6.2.6 Observations

The most interesting and important finding in our analysis of the termbase terms for this research is their frequency in the corpus. This frequency constitutes the baseline figure against which we will measure the results of our proposals on ways to close the gap

between termbases and corpora. To recap, the combined percentage of corpus-valid termbase terms that do not occur or occur infrequently in the corpus is as follows:

- Minitab : 35%
- SAS : 63%
- Symantec : 73%
- Hewlett-Packard : 76%

In the case of HP, 72 percent of the termbase terms do not occur in the corpus. This suggests that the cause of this gap is not totally linguistically motivated.

The two companies with the largest gap between the termbase and the corpus also have a larger proportion of long terms (in number of tokens) and upper case terms in the termbase. These two companies also have a problematic approach to recording variants in the termbase, and in HP's case only one percent of the termbase terms are variants.

In contrast, eleven percent of the SAS termbase terms are variants. While we were not able to precisely calculate the percentage of terms that are variants in the Minitab termbase, we have confidently estimated that they are in the range of 20 percent. The difference between the two is likely due to the fact that Minitab keeps the terminology required for controlled authoring in the termbase, whereas SAS does not. Thus, it appears that the termbases presenting a lower gap with the corpus also have recorded a significant number of variants.

Finally, the two companies with the largest gap failed to properly document the part-of-speech of the terms. Although this omission may have no direct relation to the lower corpus correspondence, it may reflect an approach to term identification that has less rigour, or relies more on automated processes. Indeed, we have already noted in the case of HP that some of the termbase terms are in fact TM segments (Section 5.1.4). A TM segment need only occur once in order to end up in the termbase, and the text in which it appeared could be modified or eliminated from the company materials at any time. In the case of Symantec, an automatic term harvesting technology was used to identify many terms (Section 5.1.3). This function is activated on a project basis, using only a small corpus.

Furthermore, unless the manual clean-up of the extracted term candidates leveraged corpus statistics, one can expect a high degree of boundary-setting problems with such a process.

These findings demonstrate that adopting certain practises with respect to term length, case, variants, and word class can help to increase the correspondence between termbases and corpora. We will learn more about the impacts of these features by studying the termbase terms that occur in the corpora in various frequency ranges, as well as other terms discovered directly from corpus investigations.

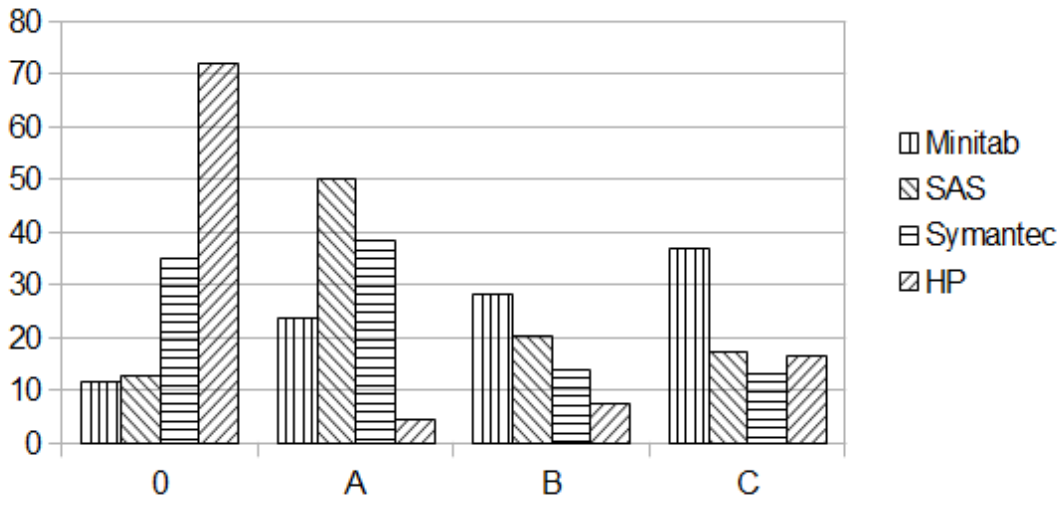
6.3 Termbase terms that do not occur in the corpus

Terms in a company's termbase may not occur in the corpus for various reasons. First, it is difficult, perhaps even impossible, to produce a termbase that is perfectly aligned with any identifiable corpus. For one reason, a company's textual content changes on a daily basis. Secondly, the content is often physically distributed, making it difficult to consolidate in the form of a single corpus. Furthermore, some content may be locked in binary file formats that are unparsable as text, such as FrameMaker files and graphics files, and therefore cannot be included in a corpus that is to be analysed by concordancing software. New information is constantly being produced, for new products or other customer-facing materials. On the other hand, textual content associated with products or services that are no longer offered by the company may be deleted or archived in the company, and yet terms associated with that content may have been added to the termbase some time earlier and remain there. For example, we found the term *Netscape Communicator* in the HP termbase; this is the name of a Web browser that ceased to exist in 2002. Termbases also tend to lag slightly behind their corresponding corpus because of the time required to find terms and add them. And in the case of some Minitab terms, the corpus is slightly behind the termbase; some termbase terms were selected from industry standards and had not yet been adopted by writers when the corpus was provided. Therefore, there will always be some gap between commercial termbases and corpora. In the case of HP, it is suspected that these types of pragmatic factors are contributing to the large gap.

Nevertheless, linguistic properties of some of the terms in the termbase do contribute to the gap. Identifying some of those linguistic properties is a goal of this research. For this purpose, we examined the terms that do not occur in the corpus in an effort to identify features that may be contributing to this problem. In particular, we examined case, number (plural vs singular forms), length in tokens and the presence of pre-modifiers and post-modifiers as well as modifiers that are proper nouns. For convenience purposes, we use the adjective *nonextant* to refer to the termbase terms that do not occur in the corpus.

6.3.1 Distribution

We reproduce here the graph shown in Figure 11. Range 0 corresponds to termbase terms that do not occur in the corpus (*nonextant terms*). The actual percentages are 11, 13, 35 and 72 for Minitab, SAS, Symantec and HP respectively.



Percentage of corpus-valid termbase terms occurring at frequency ranges

6.3.2 Differences in case

According to best practise, terms should be recorded in a termbase in their proper case (Wright 1997: 17). Proper case is the case in which the term normally appears in a text.

Our search of terms in the corpus that produced the numbers in the first row of the following table was case sensitive, meaning that if the term exists in the corpus in a different case than in the termbase, it will not be counted as a match.

	Minitab	SAS	Symantec	HP
Nonextant terms – case sensitive matching	204	530	2,240	3,150
Nonextant terms – case insensitive matching	165	468	1,849	2,792
Case differences between termbase and corpus	39	62	391	358
Percent of nonextant terms that may be due to case differences	19.11	11.69	17.46	11.37

Table 22: Nonextant terms comparing case sensitive and case insensitive results

A case sensitive search is wise if we are to assume that the best practise is followed. With a case insensitive search, a proper name such as *Windows* and its common noun counterpart *windows* will be matched as equal, when they are actually entirely different terms and concepts. Likewise, an acronym such as *IT* for *information technology* would be considered a match with the pronoun *it*, and many proper nouns in product names would be considered matches with their common noun counterparts, such as *Database Manager* and *database manager*. Thus a case insensitive search will produce a certain amount of false positives.

Nevertheless, valid case differences do occur between the termbase and the corpus, so we decided to quantify terms that are absent from the corpus in any case. The results are shown in the second row of Table 22. We then sought to determine if the difference was justified.

Setting aside those terms that still occur very infrequently in the case insensitive search, we highlight some interesting examples from Minitab, SAS, and Symantec where the case of a term in the termbase differs from the case in the corpus. For this exercise, we chose not to consider the HP data given the unusually large number of nonextant and upper-case terms.

Nonextant term	Term found in the corpus	Freq.	Valid match?
acceptance sampling by variables	Acceptance Sampling by Variables	147	Yes.

Nonextant term	Term found in the corpus	Freq.	Valid match?
SMART	smart, Smart	16	No. The termbase term is an acronym for *Specific, Measurable, Achievable, Realistic, Timebound*. The term in the corpus has the meaning of *intelligent*.
ALT	Alt	54	No. The termbase term is an acronym for *Accelerated Life Test*. The corpus term is the keyboard key, *Alt*.
Fill Out Mode	Fill Out mode, fill out mode	72	Yes.
Kendall's tau-a	Kendall's Tau-a	25	Yes. But the termbase term is justified (new standard).
certificate application	Certificate Application	93	Yes.
Symantec corp.	Symantec Corp.	1,331	Yes.
ghost boot partition	Ghost boot partition, Ghost Boot Partition	168	Yes.
managed delivery	Managed Delivery	110	Probably. The term is capitalised in the corpus because it is a product name. The general concept of managed delivery does not exist in the corpus.
Answering Machine	answering machine	158	Yes.
Error Message	error message	42	Yes.
type II error	Type II error	100	Yes. But the termbase term is justified (new standard).
information server	Information Server	54	Yes.

Table 23: Frequency of some nonextant terms with case adjusted

Some of these nonextant terms appear to have been entered in the wrong case in the term-base, such as *Error Message* and *Symantec corp*. Yet others might be caused by improper use of case by writers, such as for *Fill Out mode* and *fill out mode*. Each potential issue needs to be examined by the terminologist to determine where the problem lies, or if indeed there is a problem at all. The termbase terms *Kendall's tau-a* and *type II error*, for example, are correct even though they do not occur in the corpus. These terms have been added to the termbase to correct a current error in usage, and the corpus will soon change to reflect this.

Many of the terms that occur in the corpus in a different case than in the termbase are written with initial upper-case characters in the termbase. As noted in Section 6.2.2, the proportion of termbase terms that are in upper case is 23, 24, 54, and 90 percent for Minitab, SAS, Symantec and HP respectively. Minitab and SAS have been emerging as better-performing termbases with respect to corpus correspondence. The fact that they also have the fewest upper case terms may not be coincidental. In the case of HP, having 90 percent of the terms in upper case is very unusual, and is yet another possible factor explaining why the gap between the termbase and the corpus for HP is so high.

Determining what proportion of the nonextant termbase terms are in upper case, and comparing that figure to the proportion of upper case terms in the termbase as a whole, would show whether upper case terms are contributing disproportionately to the incidence of nonextant terms. These figures are 32, 25, 63 and 96 percent respectively. With the exception of SAS, there is, indeed, a marked increase.

6.3.3 Differences in number

Best practise dictates that terms should be entered in their canonical form in termbases (Wright 1997: 17). For nouns, this means the singular form, unless the term is normally encountered in the plural form. We are therefore interested in determining whether the termbases contain terms in the plural form which would be more productive if they were in the singular form. In other words, to what extent is the number property of termbase terms (singular vs plural) contributing to the gap between termbases and corpora? More precisely, how many plural termbase terms should have been entered in their singular form? The following table provides a picture of the situation.

Plural terms	Minitab	SAS	Symantec	HP
Number in termbase	91	107	150	513
Number that have concordances in plural form	81	90	94	107
Number that do not have concordances in plural form	10	17	56	406

Plural terms	Minitab	SAS	Symantec	HP
Number that have concordances when singularised	64	54	56	67
Number that do not have concordances when singularised	27	53	94	446

Table 24: Corpus frequency of plural termbase terms when singularised

In all cases, singularising the plural terms does not improve the situation overall (the number of concordances is greater for the plural form). The figures demonstrate that some terms are exclusively encountered in their plural form. We checked the terms and identified some that occur exclusively in the plural form and a good number that occur more frequently in plural. The following table provides a few examples:

	Singular	Plural
authentication credentials	5	94
Data Transfer Services	0	95
credentials	232	1,648
user rights	21	43
confidence bands	2	32
settings	625	2,426
test results	6	346
Remote Library Services	0	87

Table 25: Terms occurring more frequently in plural form

We are interested in determining if any of the nonextant plural terms are found in singular form. For this purpose, we ran a concordance on those specific terms after singularising them (10 for Minitab, 17 for SAS, 56 for Symantec and 406 for HP). Of Minitab's ten nonextant plural terms, only one was found after being singularised (*capability index*, 71 occurrences), and this term is already in the termbase. The results for the other three companies are shown in the following tables.

Nonextant plural term	Singular term	Frequency of singular term
stack traces	stack trace	4

Table 26: SAS - Nonextant plural terms that are found in singular form

Nonextant plural term	Singular term	Frequency of singular term
AutoPay Renewal Services	AutoPay Renewal Service	7
Instant Messaging Security Services	Instant Messaging Security Service	5
Key Management Services	Key Management Service	42
PGP Keys	PGP Key	10

Table 27: Symantec - Nonextant plural terms that are found in singular form

Nonextant plural term	Singular term	Frequency of singular term
Add Contacts	Add Contact	1
Add Photos	Add Photo	2
Auto Adjust Images	Auto Adjust Image	1
Check for Updates	Check for Update	3
Color Photos	Color Photo	1
Groups	Group	24
HP Photosmart Essentials	HP Photosmart Essential	2
HP Premium Photo Papers	HP Premium Photo Paper	5
HP Products	HP Product	4
Incompatible Cartridges	Incompatible Cartridge	1
Individuals	Individual	3
Network Diagnostics	Network Diagnostic	21
Occasions	Occasion	4
Open Projects	Open Project	1
Photosmart Essentials	Photosmart Essential	2
Share Photos	Share Photo	1
Software Updates	Software Update	44
Themes	Theme	2

Nonextant plural term	Singular term	Frequency of singular term
Toolbars	Toolbar	1
alternatives	alternative	1

Table 28: HP - Nonextant plural terms that are found in singular form

These figures show that few of the plural termbase terms are unjustified in the plural form (0, 1, 3 and 4 percent respectively, rounded, of all the plural termbase terms). Another data finding that confirms this observation is the total number of concordances of plural terms and their corresponding singular forms:

	Minitab	SAS	Symantec	HP
Plural terms	21,188	23,570	20,907	4,443
Singular form of plural terms	16,607	8,635	4,464	9,850

Table 29: Number of concordances of plural terms and their singular form

For three of the companies, the total number of concordances for the singular forms is much less than for the plural forms. The figures for HP do not follow this pattern due to the interference of several unigrams that are highly productive in singular form, such as *devices* (173) versus *device* (972), and *Photos* (83) versus *Photo* (459).

These findings suggest that the incidence of singular terms being mistakenly entered in plural form in the termbase is minimal, and is not contributing significantly to the gap between termbases and corpora. This is not to say that this type of error does not occur, but apparently it is not widespread.

A contrasting manifestation of this type of problem is when a plural concept is mistakenly entered in singular form in the termbase. Here, we are referring to terms that exclusively, or almost exclusively, occur in plural form. Such terms are quite uncommon, a few examples are *Windows* and *Norton Utilities*. Other terms are found more frequently in the plural form than in singular, such as *settings* and *properties*. We did not find any example of plural concepts mistakenly entered in singular form in the termbase.

In summary, the plural as a non-canonical form is not contributing significantly to the gap between the termbases and the corpora. Nevertheless, terminologists should avoid entering a term in the termbase in plural form if the singular form is used in any significant measure.

6.3.4 Term length

Generally speaking, the more words a MWT comprises, the less frequently it will occur (Rogers 2000: 17). Writers tend to shorten long terms for purposes of economy (Cabré 1999-b: 227; Freixa 2006: 61). For instance, when writing about access controls for a software program, a writer may start with the precise full form of the term *access control list* and use it several times in this manner, but then shorten it to *control list* and even to *list*. Provided that the context is clear, readers know that the *list* is the *access control list* previously mentioned. Terminologists need to be careful to recognise these patterns and thus to avoid including only truncated forms of MWTs in the termbase. Nevertheless, we are interested in establishing, through empirical data, the relation between a term's length and its absence from a corpus. First, we looked at the distribution of nonextant terms by length:

	Minitab	**SAS**	**Symantec**	**HP**
Number of nonextant terms	203	530	2,239	3,144
% 1 token	17.73	11.70	5.45	12.37
% 2 tokens	45.81	53.96	41.49	31.17
% 3 tokens	24.14	25.09	29.16	29.36
% 4 tokens	6.90	8.11	14.25	15.11
% 5 tokens or more	5.42	1.13	9.65	11.99

Table 30: Distribution of nonextant terms by term length

This is also shown in the following graph.

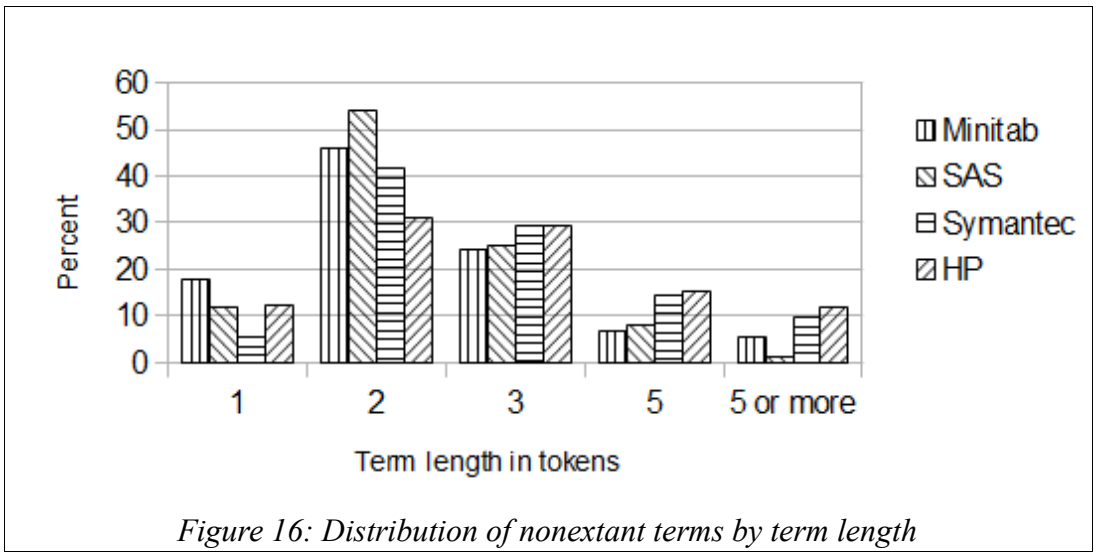


Figure 16: Distribution of nonextant terms by term length

This graph is not particularly informative, since it resembles the distribution of terms by length in the termbase as a whole. In other words, there are more bigrams in the termbase than trigrams and so forth. The graph suggests that bigrams are mostly responsible for non-extant terms, which is not necessarily the case. It would be more informative to determine the percentage of each set of n-gram terms that are nonextant. This would give us a better understanding of the types of terms, by length, that tend to be missing from the corpus.

	Minitab	**SAS**	**Symantec**	**HP**
% of 1 token terms that are nonextant	7.76	5.96	11.84	36.49
% of 2 token terms that are nonextant	10.57	12.36	31.77	72.70
% of 3 token terms that are nonextant	15.12	20.31	44.91	87.41
% of 4 token terms that are nonextant	18.18	28.10	51.87	91.70
% of terms 5 tokens or longer that are nonextant	35.48	18.18	51.55	96.42

Table 31: Percent of n-gram terms that are nonextant

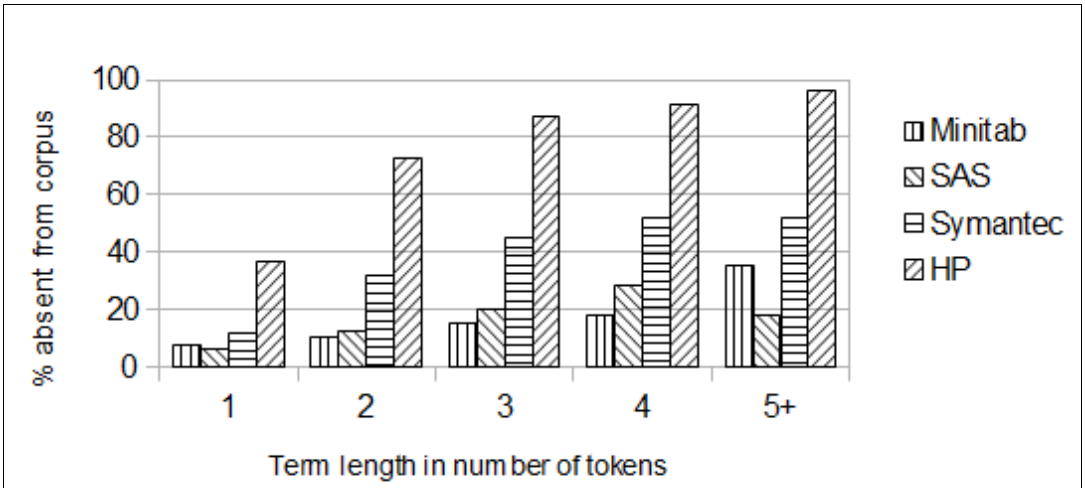


Figure 17: Percent of n-gram terms that are nonextant

This graph shows that the longer the term, the more likely it will be absent from the corpus (except for a slight dip from four to five tokens for SAS). The overall percentages are much higher for HP and to a lesser degree Symantec because these companies have proportionally more nonextant terms than the others.

If length in n-grams is a factor in the frequency of a termbase term in the corpus, perhaps we should look at the possibility to reduce the length of certain under-productive terms by eliminating non-essential parts, such as certain types of pre-modifiers and post-modifiers.

6.3.4.1 Resetting the boundaries of MWTs

We have just shown that the likelihood of a term being nonextant increases as the length in tokens increases. We are therefore interested in discovering whether removing certain components of some nonextant MWTs would render them more productive as elements of the corpus. Components of MWTs that are not domain-specific or do not present any particular properties that are important for translation purposes (such as high frequency or visibility), may not be essential in a terminology strategy that is motivated by production-oriented factors (as opposed to semantic description). Furthermore, in English, many noun phrases are compositional, i.e. the meaning of the noun phrase is equal to the sum of the meaning of its component words. It may not be necessary to include certain compositional

noun phrases in a termbase. For example, the term *dynamic link library function* may not be necessary in a termbase alongside *dynamic link library.* In all likelihood, it would occur in a corpus less frequently than the latter. Yet, the termbase may in fact include the former and not the latter. This is where repurposability comes into play, strategically, it would be better to include the latter in a termbase and not the former. We believe that many termbases have not optimised families of related MWTs in this manner.

Cabré describes the difference between "phrasal terms[73]" and "free structures": "phrasal terms are lexical structures with a terminological value, whereas free structures are just phrases that occur in discourse" (1999-b: 91). She acknowledges that there is a continuum between these two structures; certain free structures that occur frequently can therefore assume terminological value. We recognise that various phrasal structures that involve changes in word order, the introduction of prepositions, and so forth, can also constitute valid variants that could be productive alternates to our nonextant terms. However, because we are now considering how to address the issue of term length, we will focus the next analysis on truncated forms of MWTs.

When setting term boundaries, one has to balance the various critera for term selection that are suitable in a commercial setting, such as desires for repurposability and productivity (which emphasise frequency), semantic relevance (domain specificity), contextual environment (degree of visibility), the degree of lexicalisation versus free form, and the needs of end-users (translation difficulty, risk of inconsistency). If the so-called non-essential word has a domain-specific meaning or connotation, it may be necessary to retain it in the MWT entered into the termbase even if the resultant MWT is not a frequent one. For translators, this term may require a unique translation, and writers may need to access the definition in order to understand the meaning and proper usage. In this case, the terminologist may decide to enter both versions of the MWT in the termbase: the one with the non-essential part and the one without it, particularly if the latter is productive in forming other MWTs.

73 Cabré uses the terms "phrasal term" and "terminological phrase" for what we call multi-word terms.

Components of MWTs that denote occasional properties rather than intrinsic features are usually of little terminological interest since these features are not essential to the underlying concept. At the same time, MWTs containing such non-essential or words that contribute to the MWTs meaning in a strictly compositional manner are likely to occur less frequently in the corpus than their shorter counterparts, thereby reducing the repurposing potential of the termbase as a whole.

In the next two sections, we examine the effect of removing certain words in the first position and in the final position from nonextant MWTs.

6.3.4.1.1 Words in the first position

Many of the words found in the first position of a MWT are pre-modifiers of a core term. The following tables show the frequency of occurrence of these terms when the boundary is reset by removing the first word or words. We begin with Minitab.

Nonextant term	Adjusted term	Frequency of adjusted term
absolute correlation coefficient	correlation coefficient	330
individual fitted values	fitted value	270
initial communality estimate	communality estimate	0
joint probability density function	probability density function	111
left arrow key	arrow key	47
long-term benchmark Z statistic	benchmark Z statistic	0
short-term benchmark Z statistic	benchmark Z statistic	0
partial least squares coefficient plot	least squares coefficient plot	0
partial least squares loading plot	least squares loading plot	0
partial least squares response plot	least squares response plot	0
partial least squares score plot	least squares score plot	0
sequential mean squares	mean squares	129
up arrow key	arrow key	47

Table 32: Nonextant Minitab terms with front-end boundary adjustment

Here, we have reduced the number of nonextant terms from 13 to seven. Examining the adjusted terms that are still nonextant, we can adjust the boundaries even further:

Nonextant term	Adjusted term	Frequency of adjusted term
least squares coefficient plot	coefficient plot	21
least squares loading plot	loading plot	26
least squares response plot	response plot	11
least squares score plot	score plot	26
benchmark Z statistic	Z statistic	15

Table 33: Nonextant Minitab terms with front-end boundary adjustment

Since the modifier *least squares* is quite unique, its absence as a collocate with *plot* raises questions about whether this term occurs with other primary terms. Indeed, it occurs 610 times in the corpus, by itself and frequently with *regression*, *estimate*, *means*, *estimation*, *method*, *model*, and other headwords. Some of these terms are in the termbase. By adjusting term boundaries, we have reduced the number of nonextant terms from 13 to two.

Nonextant term	Adjusted term	Frequency of adjusted term
absolute frequency variable[74]	frequency variable	301
active data source	data source	7,201
average backorder wait time	backorder wait time	0
average inventory wait time	inventory wait time	0
average revenue per user	revenue per user	0
generalised least-squares method	least-squares method	2
minimum processing time	processing time	100

Table 34: Nonextant SAS terms with front-end boundary adjustment

Here, we have reduced the number of nonextant terms from seven to three. However, because *backorder wait time* and *inventory wait time* still produce frequencies of zero, we checked *wait time*, and it occurs 44 times, which now reduces the nonextant terms to one. However, unlike *average* or *minimum*, *backorder* and *inventory* are not quantifiers which could be considered non-essential to the core term. Nevertheless, the investigation enables

74 The term *absolute frequency* is also nonextant.

the terminologist to replace nonextant terms with terms that actually exist in the corpus. Further examination of the concordances of *wait time* should be carried out to confirm the concept and to identify any additional complex terms that are based on this core term.

Nonextant term	Adjusted term	Frequency of adjusted term
active risk scan	risk scan	2
automatic incremental backup	incremental backup	521
bad cluster	cluster	8,490
online backup agent	backup agent	75
original span split	span split	8
private certificate authority	certificate authority	271
proactive performance alert	performance alert	18
remote Notification Server	Notification Server	1,001
replicated network traffic	network traffic	672
secure digital signatures	digital signature	522
special processing needs	processing needs	5
third-party security removal tool	security removal tool	0

Table 35: Nonextant Symantec terms with front-end boundary adjustment

Of these 12 terms, only one was not found in the corpus after the bounary adjustment. Minor changes to 11 terms results in 11,592 occurrences in the corpus. Highly frequent terms, such as *cluster*, can be examined for other productive collocates, such as *cluster server* (3,069 occurrences), *storage foundation cluster* (160 occurrences) and *global cluster manager* (153 occurrences).

Nonextant term	Adjusted term	Frequency of adjusted term
automatic photo feeder	photo feeder	0
available disk space required disk space total required disk space	disk space	49
basic device drivers	device driver	7
Found Network Printers	network printer	4
important uninstall information	uninstall information	0
more fax settings	fax settings	51

Nonextant term	Adjusted term	Frequency of adjusted term
newer operating system unknown operating system	operating system	67
rear paper slot	paper slot	0

Table 36: Nonextant HP terms with front-end boundary adjustment

Clearly, a term like *newer operating system* is inappropriate. The terms *photo feeder* and *paper slot* could refer to obsolete concepts, or, these terms are possibly further evidence that the corpus is incomplete.

When present in a MWT, proper names typically occur as pre-modifiers. Proper names are subject to frequent changes as new products are developed, replacing existing ones. We felt that proper names warranted special examination among the set of nonextant terms.

For HP and Symantec, terms that begin with the company name, such as *HP Image Editor* and *Symantec Workflow Server* figure prominently in the list of non-extant terms. For Symantec, 292 such terms begin with *Symantec*, 13 with *Altiris*, 109 with *Norton*, 32 with *Veritas*, and dozens if not hundreds of terms begin with other trademarks (*Dell, Windows, HP*, etc.). Counting only these mentioned figures, the total (446) accounts for 25 percent of the nonextant terms. The situation for HP is similar. Its nonextant terms include 221 that begin with or contain *HP* or *Hewlett-Packard*, 14 with *Adobe*, 12 with *Corel*, 34 with *Microsoft*, 22 with *Windows*, and an array of others. The total of the five mentioned here is 303 terms, or about 10 percent of the nonextant terms.

In contrast, the termbases for Minitab and SAS contain relatively few such terms in their nonextant set. For SAS, there are 11 terms with *SAS*, three with *Siebel*, one with *Windows*, and a scattering of other trademarks. For Minitab, we find three with *Minitab* and no other identifiable trademarks[75]. This represents only about three percent and two percent of the nonextant terms respectively.

75 Note that it is not possible for us to identify absolutely all trademarks.

6.3.4.1.2 Words in the final position

Words in the final position of a MWT can result in a mismatch between the termbase term and the corpus. The following tables provide some examples from each company.

Nonextant term	Adjusted term	Frequency of adjusted term
actuarial survival function	actuarial survival	11
conditional maximum likelihood estimate	conditional maximum likelihood	10
half normal plot of the standardized effects	half normal plot	34

Table 37: Nonextant Minitab terms with back-end boundary adjustment

The last nonextant term above is a new standard and is therefore justified in the termbase.

Nonextant term	Adjusted term	Frequency of adjusted term
comma-separated values format	comma-separated values	83
control chart for attributes control chart for variables	control chart	243
critical success factor component	critical success factor	540
Extensible Stylesheet Language expression	Extensible Stylesheet Language	12
organization chart analysis	organization chart[76]	42
transition matrix report	transition matrix	145

Table 38: Nonextant SAS terms with back-end boundary adjustment

Nonextant term	Adjusted term	Frequency of adjusted term
Patch Management Import page	Patch Management Import	64
Policy-Based Encryption service	policy-based encryption	9

Table 39: Nonextant Symantec terms with back-end boundary adjustment

76 Note that *organization chart* is not the best term; *organizational chart* occurred 315 times.

Nonextant term	Adjusted term	Frequency of adjusted term
red eye removal	red eye	41
system requirements details	system requirements	64
wireless network name	wireless network	173
printhead failure	printhead	275
proof sheet error	proof sheet	221

Table 40: Nonextant HP terms with back-end boundary adjustment

6.3.5 Observations

The gap between the termbase and the corpus is larger for HP and Symantec than for the other two companies. We have shown that their termbases contain a significantly larger proportion of upper case terms and proper names, both of which have contributed to their larger set of nonextant terms. We do not claim that these types of terms do not belong in termbases, but they should be chosen judiciously and backed by corpus evidence, and their proportion compared with other types of terms should be balanced.

We have demonstrated the value of using corpus evidence when making decisions about the optimal boundaries of MWTs. We investigated a random selection of nonextant terms that contained a word of potentially minor significance or a word that introduced an additional layer of semantic precision, and found that in almost every case a greater matching could be achieved by adjusting the term boundary accordingly. We need to acknowledge, however, that particularly in controlled authoring environments, such as for Minitab and SAS, new terms are sometimes added to termbases before their adoption by writers, such as terms for new standards. In this case, a nonextant term will occur and is justified.

So far, we have only considered nonextant terms in this section. Termbase terms that occur very infrequently in the corpus (such as once or twice) likely present similar boundary-setting problems. Setting term boundaries that are optimised to achieve the repurposing demands of production-oriented applications is a challenge for terminologists.

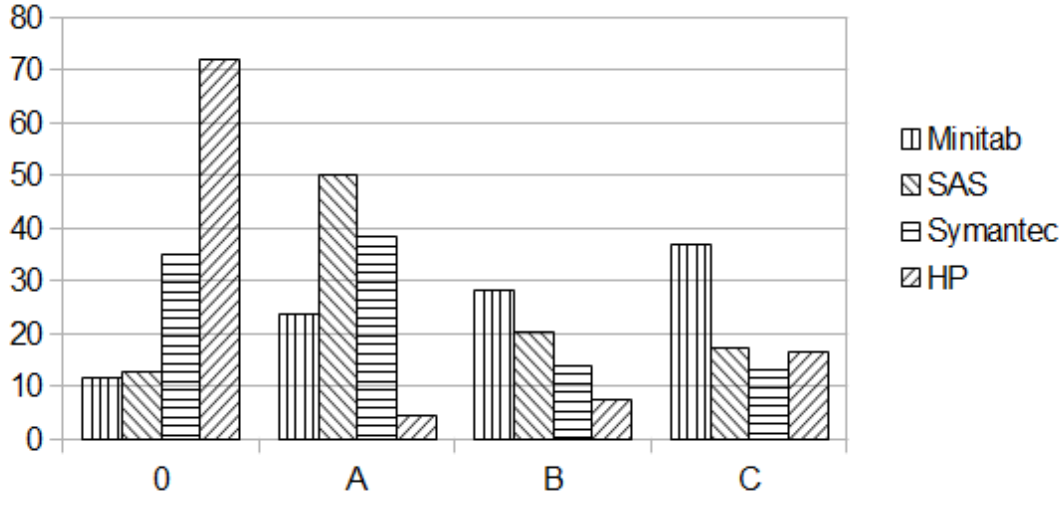
In this section, a somewhat ad-hoc approach was adopted, whereby long MWTs that do not occur in the corpus were detected in the termbase, parts of these terms that are potentially interfering with a corpus match were identified and removed, and the shorter terms searched anew in the corpus to verify a match. Later, we will test a corpus-based approach whereby the keywords in the corpus are first identified, and then their frequent collocates. Rather than eliminating redundant terms from a termbase after-the-fact, this approach avoids documenting them altogether. We believe that the keyword method will produce higher-quality, more repurposable terminology for a termbase.

6.4 Termbase terms that occur infrequently in the corpus

Terms that occur infrequently in the corpus are likely to share many of the same properties and phenomena already demonstrated for terms that do not occur in the corpus at all, with respect to properties such as length, case, and so forth.

6.4.1 Distribution in the termbase

We reproduce here the graph shown in Figure 11.



Percent of corpus-valid termbase terms occurring at frequency ranges

Range 0 are terms that do not occur (nonextant terms), and range A are terms that occur infrequently in the corpus. Thus, using Minitab as an example, about ten percent of the terms do not occur, and about 24 percent occur infrequently. Minitab has the fewest infrequent terms, followed by Symantec and SAS. (HP can hardly be considered for range A since over 70 percent of its termbase terms do not occur in the corpus.)

The following table provides the details of this graph for range A:

	Minitab	**SAS**	**Symantec**	**HP**
Frequency range A	1 to 10	1 to 58	1 to 52	1
Number of terms in range A	422	2103	2467	187
% of termbase terms in range A	23.74	50.12	38.26	4.3

Table 41: Infrequent termbase terms

To better understand the nature and prevalence of termbase terms that occur infrequently in the corpus, we quantified them in more granular low-frequency categories. We chose four categories in the range 0 to 100 occurrences in a corpus of 4 million tokens. Therefore, the raw frequencies were normalised using our standard calculation (see Section 6.2.1.1.)

- Minitab: raw frequency x 1.006
- SAS: raw frequency x 0.181
- Symantec: raw frequency x 0.202
- Hewlett-Packard: raw frequency x 9.98

The following figure shows the distribution of termbase terms occurring at incremental ranges of low frequency, as a percentage of the termbase terms occurring 100 times or less (normalised).

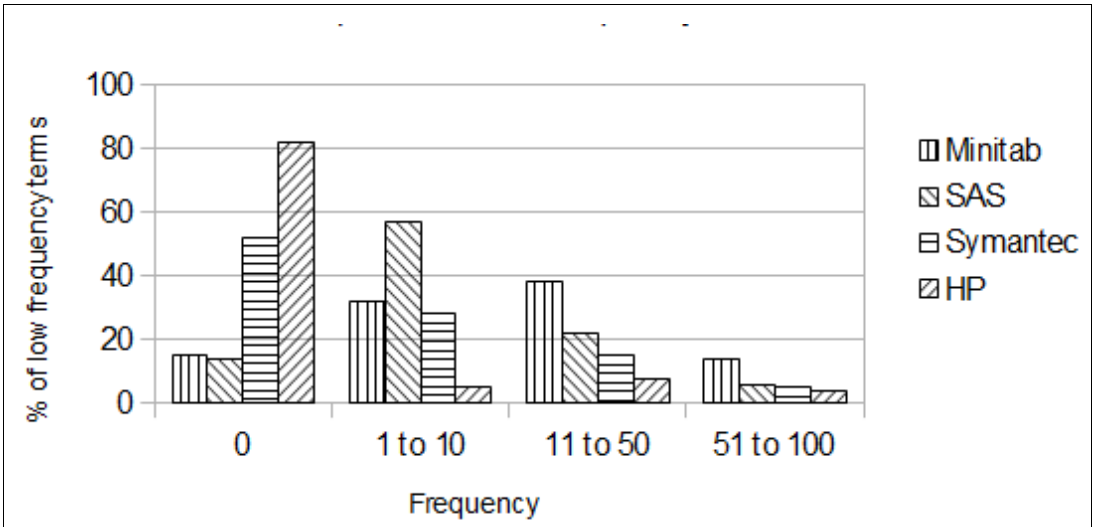


Figure 18: Percent of infrequent termbase terms occurring at low-frequency ranges

HP terms are the most problematic; over 80 percent of the infrequent terms do not occur in the corpus. For Symantec, over half the infrequent terms do not occur and a further 30 percent occur one to ten times. Minitab and SAS have a wider distribution of low-frequency terms, with Minitab performing the best, having the greatest proportion of terms in the higher frequency areas of the low frequency range.

These graphs demonstrate that the problem of low frequency terms in the termbase is least pronounced for Minitab, followed by SAS, Symantec and finally HP.

6.4.2 Term length

As we were for the nonextant termbase terms, we are interested in the length of terms that occur infrequently. For this purpose, we could use frequency range A (see Section 6.2.1.3).

However, the percentage of terms that fall in this range is high (with the exception of HP, but this is because over 70 percent of its terms do not occur). Evaluating the length of terms in this range is therefore likely to produce results very similar to the length of termbase terms as a whole. Since we are interested in discovering any possible salient differences in term length for infrequent terms, we suggest examining the terms that occur at even lower

frequency than above, as shown below. (In this case, we cannot lower the frequency for HP any lower than 1, so its value remains the same.) We call this range A-s[77].

	Minitab	SAS	Symantec	HP
Frequency range A-s	1 to 2	1 to 11	1 to 10	1
Number of terms in range A-s	127	1,086	1,358	186
% of termbase terms in range A-s	7.14	25.88	21.08	4.24

Table 42: Very infrequent termbase terms

The following table and graph show the relevant data for range A-s.

	Minitab	SAS	Symantec	HP
% of range A-s are 1-token terms	15.75	9.48	8.98	28.49
% of range A-s are 2-token terms	45.67	58.47	51.47	42.47
% of range A-s are 3-token terms	26.77	23.29	24.37	19.35
% of range A-s are 4-token terms	9.45	7.00	8.10	8.06
% of range A-s are 5+token terms	2.36	1.84	7.07	1.61

Table 43: Very infrequent termbase terms by term length

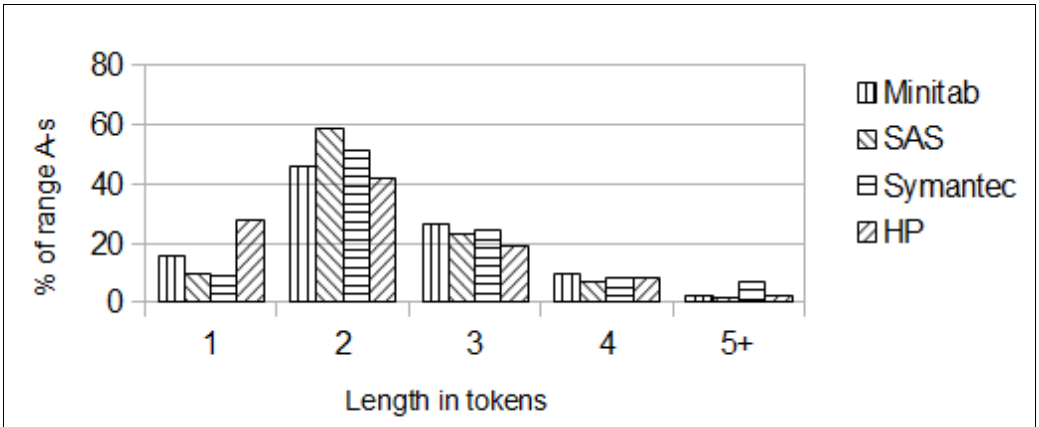


Figure 19: Distribution of very infrequent termbase terms by length

Once again, this graph is not particularly informative as it resembles the distribution of terms by term length as a whole. We may get a better sense of how term length affects frequency by measuring the proportion of infrequent terms in each set of n-grams.

77 "s" for "smaller than A."

	Minitab	SAS	Symantec	HP
% 1-token terms are in range A-s	4.31	9.89	11.84	4.97
% 2-token terms are in range A-s	6.59	27.44	23.91	5.86
% 3-token terms are in range A-s	10.49	38.47	22.76	3.41
% 4-token terms are in range A-s	15.58	49.67	17.89	2.90
% 5+token terms are in range A-s	9.68	60.61	22.91	0.77

Table 44: Percent of n-token terms that are very infrequent

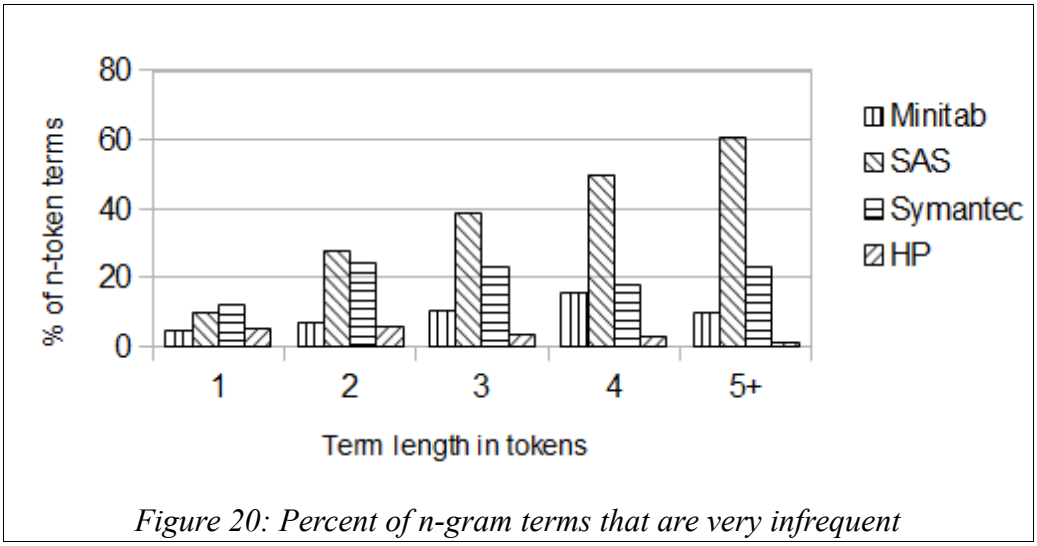


Figure 20: Percent of n-gram terms that are very infrequent

The figures for HP are unreliable due to the high incidence of terms that do not occur in the corpus and the fact that range A-s is the same as range A. For SAS, there is a progression that supports the assumption: the longer the term the more likely that it occurs infrequently. The same can be said for Minitab, except for a slight dip beginning at five tokens. For Symantec, the likelihood of infrequency is about the same for all MWTs, hovering around 20 percent.

6.4.3 Other properties

It is safe to assume that the properties of terms that occur infrequently in the corpus will be similar to those of terms that were not found at all, as described in Section 6.3. Nevertheless, we would like to validate this assumption with some data. Our analysis, however, will

not be as extensive as was carried out for the nonextant terms. For this purpose we will examine the terms in range A-s.

The proportion of the terms in range A-s that are in upper case is 24, 24, 56 and 92 percent for Minitab, SAS, Symantec and HP respectively. Among these upper case terms, proper names figure most prominently in the set for Symantec; many of the terms appear to be the names of products, with 40 percent of the upper case terms beginning with familiar company names such as *Symantec*, *Veritas*, and *Norton*. While HP has the highest percentage of upper case terms, compared to Symantec, fewer appear to be proper names proportionally speaking. Instead, many are the result of TM segments being added to the termbase, which replicate sentence case or title case, such as *Clean Printer* and *Preparing to print*.

These figures do not differ significantly from the percentage of upper case terms in the termbase as a whole (see Section 6.2.2). Interestingly, this finding actually contrasts with that for nonextant terms, which have a significantly higher percentage of upper-case members compared to the upper-case terms in the termbase, at least, in three of the four termbases. Therefore, case differences may not be a significant contributing factor towards the prevalence of termbase terms that are *infrequent* in corpora. This difference in case contributions between nonextant and infrequent terms is not easily explained without further empirical evidence.

To determine the effect that front-end and back-end adjustments in the boundaries of infrequent MWTs would have on the termbase-corpora correspondence, we need to examine the terms manually and perform some comparable concordances. Due to the manual effort involved, we focus on the terms from Minitab, given that this is the smallest set and also is emerging as the most reliable, but we also validate those findings with a random check of the terms from SAS. We start with Minitab.

Twenty-two terms (17 percent of the infrequent terms) have a potentially non-essential word in the first position, the absence of which may increase matches with the corpus. A few examples are shown in the following table.

Infrequent termbase term	Frequency	Adjusted term	Frequency
right arrow key	2	arrow key	47
two-sided CUSUM chart	1	CUSUM chart	62
one-color ramp	3	ramp	27
average linkage method	2	linkage method	34

Table 45: Infrequent terms with front-end boundary adjustment

Twelve terms (10 percent) have a proper name or a possessive construct in the first position. A few examples are provided in the following table.

Infrequent termbase term	Frequency	Adjusted term	Frequency
Anderson-Darling goodness-of-fit statistic	2	goodness-of-fit statistic goodness-of-fit	15 594
Student's t-distribution	5	t-distribution	81
Kendall's rank-order correlation coefficient	1	rank-order correlation coefficient	3

Table 46: Infrequent terms with front-end boundary adjustment

Fifteen terms (12 percent) contain low-performing words in the final position. A few examples are provided in the following table.

Infrequent termbase term	Frequency	Adjusted term	Frequency
covariance structure of the data	2	covariance structure data covariance structure	17 13
global worksheet variable	2	global worksheet worksheet	70 5,824
centroid linkage method	2	centroid linkage	8
randomized block design experiment	1	randomized block design	17
regression fit line	1	regression fit	77

Table 47: Infrequent terms with back-end boundary adjustment

We also found MWTs whose components combine more frequently with other collocates, and as such, should be entered individually in the termbase. We confirmed this for 12 terms (10 percent), but there are likely more. The following are a few examples:

Infrequent termbase term	Frequency	Adjusted term	Frequency
hyperbolic arcsine	2	hyperbolic arcsine	24 34
exponential growth trend model	2	exponential growth trend model	46 87

Table 48: Infrequent terms whose components combine frequently with other collocates

Thus, nearly 50 percent of the Minitab termbase terms that occur infrequently in the corpus can lead to the discovery of much more frequent terms when the boundary is adjusted. This mirrors the situation with nonextant terms.

It is impossible to precisely determine the scope of the potential to improve term identification by resetting term boundaries for the other companies, as the set of terms to manually examine is too large. A scan of the 1,086 SAS termbase terms in this group suggests the following scope of the types of problems we found in Minitab:

- about 40 terms (four percent) with potentially non-essential words in the first position
- about 60 terms (six percent) with potentially non-essential words in the final position
- about one hundred terms with other boundary setting issues

We also found 20 occurrences of phrase structures with articles, such as *attach a database* and *update the path*. Thus, certainly, at least 20 percent of the SAS terms in this range are under-optimised with respect to term boundaries.

6.4.4 Observations

It is not possible to quantify the gains that would be achieved by examining the boundaries of infrequent termbase terms to identify frequently-occurring alternates for all four companies, as this would necessitate a separate concordance be carried out, and the results carefully analysed, on over 5,000 terms (see Table 41). However, tests with Minitab and

SAS suggest that for at least 20 percent and possibly as much as 50 percent of the infrequent termbase terms, other terms can be derived that are much more frequent in the corpus, by resetting term boundaries according to corpus evidence.

Nevertheless, there are several valid reasons why infrequently-used terms would be present in a termbase. For controlled authoring applications, effecting changes to the existing use of terminology or style often requires terms to be added to the termbase that have been used only infrequently before, if at all. This was indeed the case for a number of the infrequent Minitab terms noted in the previous tables. Definitions may need to be provided for infrequent terms that convey a specialised meaning, so that they are used and translated correctly. We do not, therefore, suggest that infrequent terms are always unnecessary and should be removed. Each term needs to be considered individually. However, at the least, frequently-occurring terms that we can identify by adjusting the boundaries of infrequent terms need also to be included in a termbase so that the economies of scale offered by production-oriented systems such as CAT tools can be realised.

It should also be noted that the two types of terms discussed in this section, i.e. long MWTs and their shorter counterparts, do not necessarily share the same meaning, and therefore, would not necessarily be entered in the same entry of a termbase. Our analysis is confined to syntactic adjustments. The meaning behind each term needs to be determined by the terminologist and the entries constructed accordingly.

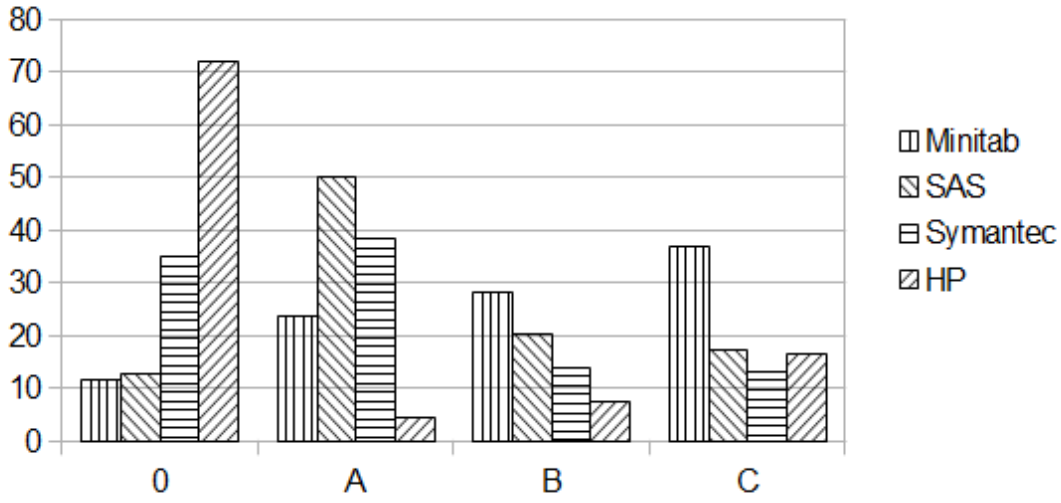
6.5 Termbase terms that occur frequently in the corpus

By examining the properties of termbase terms that occur frequently in the corpus, we may identify some effective term selection criteria.

6.5.1 Distribution in the termbase

We reproduce here the graph from Figure 11. Range C corresponds to the frequent terms. The actual percentages of frequent termbase terms are 37, 17, 13 and 17 for Minitab, SAS,

Symantec and HP respectively. Once again, our analysis is based on the corpus-valid terms.



Percentage of termbase terms occurring at frequency ranges

The following table provides the details for range C.

	Minitab	**SAS**	**Symantec**	**HP**
Frequency range C	Over 50	Over 278	Over 250	Over 5
Number of terms	653	721	843	730
% of termbase in range C	36.75	17.19	13.09	16.65

Table 49: Frequent termbase terms

We now examine the frequent terms to determine any salient properties.

6.5.2 Term length

Shorter terms are likely to occur more frequently, for reasons already stated. Nevertheless, we are interested in the statistical evidence. For this purpose, we analysed the length of the termbase terms that occur at the highest frequency range (C).

	Minitab	SAS	Symantec	HP
Frequency range C	Over 50	Over 278	Over 250	Over 5
% of termbase in range C	37	17	13	17
Number of terms	653	721	843	730
% 1 token	42.57	63.94	50.89	69.18
% 2 tokens	45.33	31.48	34.88	23.84
% 3 tokens	9.80	4.44	10.20	5.07
% 4 tokens	1.53	0.14	2.97	1.51
% 5 or more tokens	0.77	0	1.07	0.41

Table 50: Distribution of frequent termbase terms by length

The percentages are more easily compared graphically:

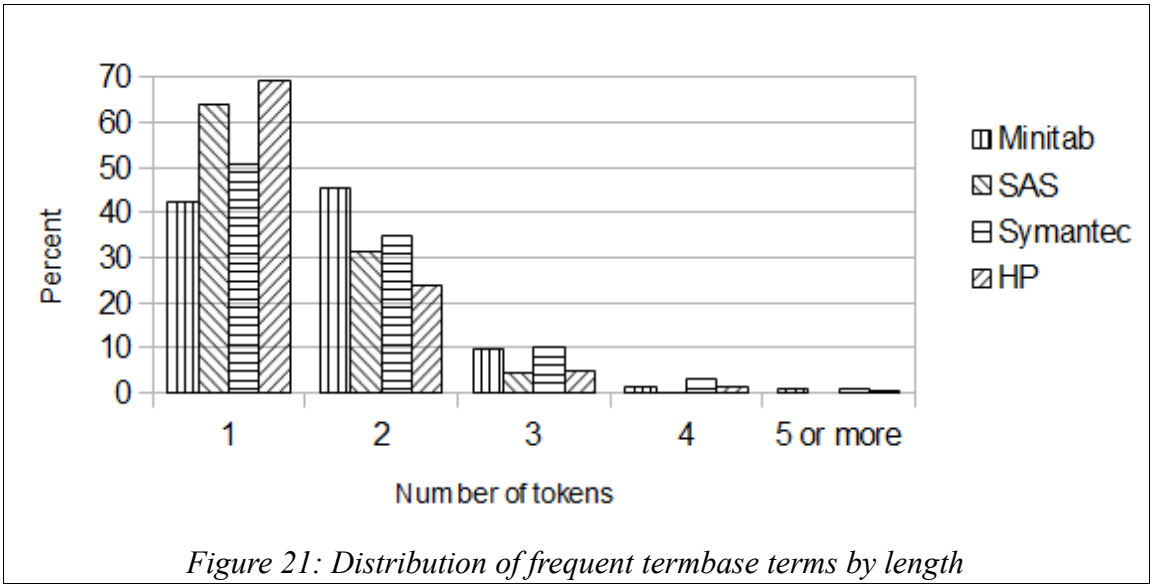


Figure 21: Distribution of frequent termbase terms by length

Here, we can see that monograms and bigrams make up the vast majority of frequent terms, which is to be expected. It is interesting to observe that for Minitab, the distribution of frequent terms is about equal for unigrams and bigrams before it drops dramatically, whereas for the other companies we see a marked decline with each increase by one token in term length. This suggests that Minitab is paying more attention to MWTs.

What is possibly more interesting is the percentage of each set of n-grams that is in the high-frequency range, shown in the following table.

	Minitab	SAS	Symantec	HP
% of 1 token terms in range C	59.91	44.28	41.65	47.37
% of 2 token terms in range C	33.64	9.81	10.05	12.91
% of 3 token terms in range C	19.75	4.89	5.91	3.50
% of 4 token terms in range C	12.99	0.65	4.07	2.12
% of terms 5 tokens or longer in range C	16.13	0	2.15	0.77

Table 51: Percent of n-token termbase terms that are frequent

These percentages are shown graphically in the following figure:

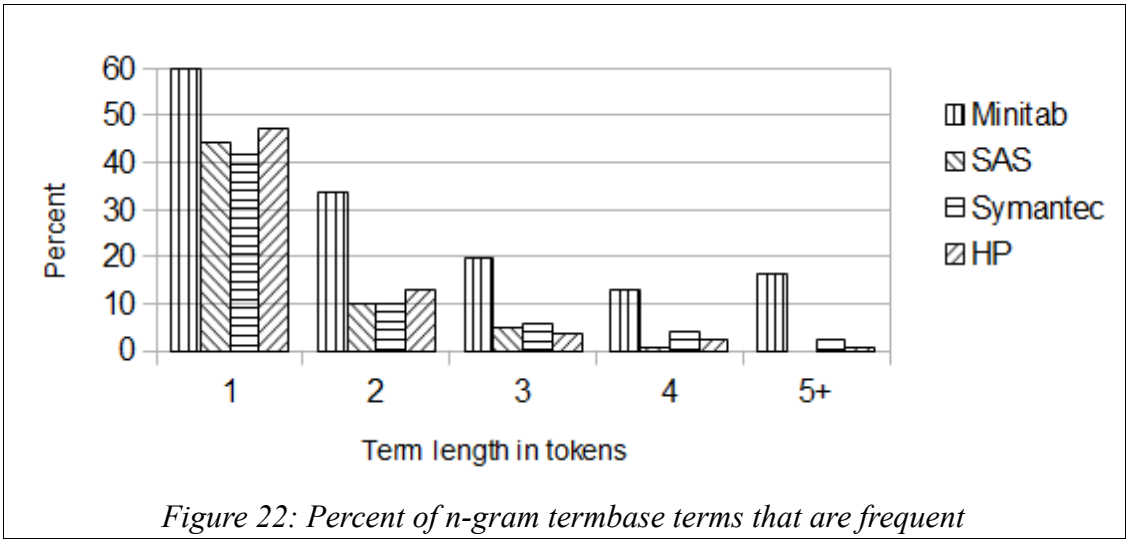


Figure 22: Percent of n-gram termbase terms that are frequent

Minitab has performed the best in terms of selecting frequently-occurring terms in each set of n-grams. For instance, over 30 percent of its bigram terms occur frequently in the term-base, compared to only about ten percent for the other three companies. Overall, the longer the term, the less likely that it will occur frequently, which again is to be expected. What is interesting is the more gradual decline for Minitab, whereas for all three other companies the decline is much more abrupt, with only unigrams comprising a significant amount of frequent terms. Again this suggests that perhaps with the exception of Minitab (although it could be argued that here as well the frequency distribution could be improved), not enough attention is being paid to frequently-occurring MWTs.

6.5.3 Other properties

We would like to examine the nature of frequent terms, beyond their length, to discover any salient morpho-syntactic properties. We already know, through our examination of properties of terms that do not occur or occur infrequently, that avoiding some of the problems identified in those terms will increase the frequency matching between the termbases and the corpora. Therefore, logically-speaking, the more frequently occurring termbase terms in our sample termbases are likely to present a low incidence of these characteristics. For instance, the incidence of upper-case terms, proper names, and modifiers will be relatively low. A cursory examination of the terms in Range C shows this to be the case.

Rather than repeating the analysis of such properties on frequent terms, validating yet again what we have demonstrated with infrequent terms, we would like to manually examine the frequent terms to see if there are any other salient features. For this purpose, Range C offers too many terms for manual examination. To narrow this range even further to very frequent terms, we chose a normalised frequency of 500 or more in a corpus of 4 million. As before, the raw frequencies were multiplied by the normalisation factor to produce normalised frequencies (see Section 6.2.1.1). The relevant data is provided in the following table.

	Number of terms occurring 500+ times (normalised to 4M)	**Percent of corpus-valid termbase terms**
Minitab	139	7.82
SAS	175	4.17
Symantec	192	2.98
HP	248	5.66

Table 52: Very frequent terms

As stated in Section 6.2.1.2, the figures for HP are likely inflated due to over-factorisation.

In the following pages, we attempt to categorise the terms into various groups. The categories are not mutually exclusive and therefore some items can belong to more than one

category, such as a word being both an adjective and a member of the general lexicon. Furthermore, for all companies, many terms could not be categorised beyond stating that they denote domain-specific concepts, such as *e-mail* and *disk*. Our aim is to identify some salient properties of very frequent terms and not to provide a rigorous statistical analysis.

For Minitab, 73 percent of the terms in this very high-frequency range are unigrams, and the longest term has three tokens.

Type of term	Number found	Percent of total	Example
Acronym	8	5.76	SS PPM
Function word	0	0	
General lexicon word	1	0	instead
Proper name	7	5.04	Anderson-Darling
Verbs	3	2.16	click
Adjectives	11	7.91	average[78] normal random binary
User interface concepts	4	2.88	worksheet toolbar dialog box Help

Table 53: Properties of Minitab's frequent terms

Here we note that the majority of terms (75 percent) correspond to domain-specific concepts that defy further categorisation, such as *percentile*, *probability*, and *formula*.

For SAS, 90 percent of the terms are unigrams, and the longest term is three tokens. We find the following types of terms:

78 Adjectives such as *average* and *normal* are domain-specific in the field of statistics whereas they could be considered general lexicon units in other domains.

Type of term	Number found	Percent of total	Example
Acronym	10	5.71	ODS SCL
Function word	0	0	
General lexicon word	6	3.43	level line observation
Proper name	1	0.57	Java
Verbs	6	3.43	include request export
Adjectives	2	1.14	mean constant
User interface concepts	7	4.0	pop-up menu button wizard

Table 54: Properties of SAS's frequent terms

Here, the proportion of terms denoting domain-specific concepts that defy further categorisation is about 83 percent.

For Symantec, 78 percent of the terms are unigrams, and the longest term is four tokens.

Type of term	Number found	Percent of total	Example
Acronym	1	0.52	IT
Function word	0	0	
General lexicon word	1	0.52	multiple
Proper name	31	16.16	Symantec Security Response Norton Utilities
Verbs	17	8.85	back up click deploy recover
Adjectives	6	3.12	trusted powerful virtual

Type of term	Number found	Percent of total	Example
User interface concepts	4	2.08	console dialog box Help icon

Table 55: Properties of Symantec's frequent terms

For Symantec, the proportion of domain-specific terms that could not be further categorised drops slightly to 71 percent. This is possibly due to the higher occurrence of proper names.

For HP, 83 percent of the terms are unigrams, and the longest term is three tokens. We find the following types of terms:

Type of term	Number found	Percent of total	Example
Acronym	15	6	ADSL
Function word	9	4	To Yes
General lexicon word	24	10	language stop number black
Proper name	17	7	Windows Mac
Verbs	42	17	Save Add Start
Adjectives	20	8	original default normal
User interface concepts	4	2	Viewer Menu Control panel Button

Table 56: Properties of HP's frequent terms

For HP, the proportion of domain-specific terms that could not be further categorised drops significantly to 46 percent. This is due to the higher number of general lexicon words, verbs, and function words.

The number of verbs and adjectives in this set is worth noting; they represent 25 percent of this group, which is higher than the figure generally cited as a benchmark for termbases (less than 10 percent). While some of these terms may also be nouns, such as *start* and *copy*, one might also consider that their high frequency may be partially explained by the fact that these lexical items are commonly found on user interfaces, such as in Menus, where they usually express a verbal concept. We can therefore expect that a large portion of the occurrences of these terms are verbs. We verify this assumption in the next section.

6.5.3.1 Validation of verbs

To verify that the verb usage of homographs is more frequent than any other usage, we examined the concordances directly. For instance, let us consider *change* from HP as an example. This verb might be considered a general lexicon word out of context, but the concordances and keyness value provided by WordSmith suggest otherwise (keyness will be described in detail later, suffice it to state here that it is a measure of domain-specificity). We find contexts such as "change the printer cartridge" and "change settings," etc., some of which would require a TL equivalent akin to a more specialised synonym such as *replace* and *modify*. In addition, verbs like *copy* and *save* assume a special domain-specific status due to their prevalence on software user-interfaces and related documentation.

In any case, we do find that verb usages are in the top-ranking collocates, such as *to change* (255), *change the* (389) and *can change* (63), while in contrast, noun usages such as *the change* (7) and *a change* (2) are infrequent, as shown in the following figure:

N	Word	Texts	Total	L3	L2	L1	Centre	R1	R2	R3
1	CHANGE	6	820	3	1		768		1	3
2	THE	6	719	35	16	7		389	13	29
3	TO	6	391	9	13	255		11	6	26
4	SETTINGS	5	276	13	10	9		43	76	75
5	YOU	6	235	68	67	32			6	7
6	AND	6	138	8	8	25		9	15	15
7	OR	6	110	5	7	27		2	6	20
8	CAN	6	106		21	63				2
9	PRINT	5	106	3		1		14	50	15
10	OF	6	103	6	3				15	54
11	A	6	98	15	4	2		5	13	3
12	FOR	6	94	4	1			3	10	22
13	IN	6	67	1				7	4	11
14	CLICK	4	65	2	5	20			6	5
15	SETTING	5	62	2	2				15	10
16	PRINTER	5	54	5	5	2		5	20	
17	WANT	4	54	7	43					
18	ON	6	49	7	1			1	1	11
19	DEFAULT	6	48	1	2			8	20	7
20	THEN	4	43	2	10	6		2	4	6
21	FROM	5	41	2	1			11	1	4
22	IMAGE	6	39	4	1	1			18	1

Figure 23: Collocates of the word change from HP

We manually examined all the usages of the homographic verbs in the previous tables and observed that the verb usage dominates. The results shown in the following table.

Homograph	Verb predominates	Other POS predominates	Balanced distribution
Minitab			
click	x		
fit	x		
run	x		
SAS			
copy			x
export			x
forecast		x	

Homograph	Verb predominates	Other POS predominates	Balanced distribution
request	x		
Symantec			
alert		x	
boot		x	
chat		x	
check	x		
click	x		
display	x		
download	x		
filter			x
fix	x		
load	x		
monitor	x		
risk		x	
rule		x	
run	x		
scan		x	
update			x
upgrade			x
HP			
change	x		
copy		x	
edit	x		
load	x		
print	x		
save	x		
scan			x
share	x		
slide	x		
stop	x		
update		x	
view	x		

Table 57: Validation of verb homographs

6.5.4 Observations

Unigrams and bigrams make up the vast majority of termbase terms that occur frequently in the corpus. Very frequent terms are short: 73 to 90 percent are unigrams and rarely do they exceed trigrams.

In the very high frequency range, we note a prevalence of terms that present the situation of homography, i.e. terms the surface form of which can be both a noun and a verb (such as *copy* or *request*), or a noun and an adjective (such as *mean* and *constant*). In the preceding section, we have attempted to only include in the counts for particular word classes those words that are likely to occur more frequently in the word class indicated. However, the property of homography is contributing to the high-frequency status of these terms in the corpus. Indeed, previous research has shown that homographs are particularly common in the computing domain (Lam Kam-mei 2001). Lam Kam-mei studied the phenomenon of anthimeria (functional shift, or conversion), in which one part of speech is substituted for another, as in the following sentence where *interrupt* functions as an adjective:

It may be necessary to turn off the interrupt feature of the printer.

and concludes that "anthimeria is a powerful developmental force in English, particularly in the realm of computing" (2001: xi).

Homographs are therefore important for commercial terminologists to record in termbases, not only because of their frequency, but also because translations of the individual homographs will frequently differ (the French translation of *print* is *imprimer* for the verb and *impression* for the noun.) Thus, a separate entry is required for each instance associated with a given word class. For homographs, the part-of-speech value is important distinguishing metadata and should be included in the entry, a practise which, as we noted in section 6.2.4, has not always been adopted in these four companies.

In all cases except SAS, the non-noun termbase terms (verbs and adjectives) account for a larger percentage of the frequent termbase terms than their weighting in the termbase as a

whole (12 percent for Minitab and Symantec, 25 percent for HP, whereas the percentage of documented non-nouns in the termbases is 7, 4 and 20 respectively, see Section 6.2.4). This suggests that non-nouns are under-represented in the termbases. The lower proportion of non-nouns in the SAS termbase can be explained by the fact that this is a monolingual termbase primarily used by writers and as a source of published glossaries. These two functions require fewer non-nouns than cross-lingual applications such as translation[79].

The termbases contain low amounts of frequently-occurring terms; only three to eight percent of the termbase terms occur very frequently, and for three of the companies only 13 to 17 percent of the termbase terms are in the frequent range C. Only Minitab has managed to produce a termbase of which a moderate proportion of the terms we examined are frequent in the corpus (37 percent of termbase terms).

6.6 Verbs in the corpus

In Section 6.2.4, we demonstrated that over 90 percent of terms in commercial termbases can be expected to be nouns. We wonder if sufficient attention has been paid to other open word classes, such as verbs, adjectives and adverbs. We decided to focus on the case of verbs. We used the keywords function of WordSmith to identify and study a limited number of verbs that rank highly as keywords in the corpus (for this, we chose a keyness measure of 15,000 or above). (Keywords are further explained in Chapter 7.)

The following eight verbs meet the above keyness criteria in the Minitab corpus:

- choose
- plot
- click
- enter
- use
- design
- test
- display

79 Note that controlled authoring also requires more non-nouns than we observed in the SAS termbase, but SAS does not use its termbase for controlled authoring.

Two of these verbs are missing from the termbase (*enter* and *design*). Four of the six that are present in the termbase were marked by the terminologist with the neutral register value, indicating that she considers them to be part of the general lexicon (*choose*, *use*, *test*, and *display*. See section 5.2.2.1.1 and Appendix A.). The verb *plot* is marked with a rejected usage value, yet one can find many concordances of its use as a verb in the corpus (there are 484 occurrences of three randomly selected verb usage patterns: *to plot*, *can plot*, and *you plot*). Perhaps this is a case where controlled authoring efforts have not yet fully taken effect in the corpus.

Some of these verbs are very frequent, even without searching inflected forms, for example, *choose* (16,081 occurrences) and *click* (14,491 occurrences). In comparison, the average frequency of the 47 verbs currently in the termbase (38 from the corpus-valid set, and 9 from the general lexicon group) is only 1,489 occurrences. Since the termbase verbs already include six of the verbs listed above, this suggests that too few of the remaining 41 verbs are frequent in the corpus. Indeed, four of the termbase verbs do not occur, and another five are very rare.

For SAS, 10 verbs have a keyness measure of 15,000 or above.

- specify
- select
- create
- click
- contain
- plot
- display
- access
- use
- run

These 10 verbs occur over 500,000 times; they are highly productive domain-specific verbs. None are documented in the termbase. The 59 verbs currently in the termbase occur 83,000 times overall, however, several homographs could be artificially inflating this number by including noun occurrences, particularly *log*, which occurs over 21,000 times.

For Symantec, nine verbs have a keyness measure of 15,000 or above.

- protect
- provide
- use
- download
- access
- select
- install
- restore
- click

Only three of these verbs are documented in the termbase: *select*, *restore*, and *click*. The six which are not found in the termbase occur 195,189 times in the corpus.

It is interesting to note the difference in the number of verbs in the termbase for each company: Minitab six, Symantec three, and SAS none at all[80]. This can be understood by considering the purpose and use of the termbases. Most of the verbs recorded by Minitab are used in controlled authoring. The SAS termbase is primarily used as a technical reference for English writers and as a source of published glossaries; these users and end uses do not require these types of verbs. However, should this termbase ever be extended to other purposes, the lack of verbs could become a problematic lacuna. The Symantec termbase is used for translation. We suspect that, in this case, more verbs would be useful in the termbase, particularly industry-specific verbs such as *protect* and *restore*.

In summary, we have shown that certain verbs are domain-specific due to their elevated frequency when compared to the reference corpus. These verbs tend to be under-documented in company termbases, at least, in the IT domain we are studying. Mining more verbs through keyword analysis with an expanded keyness threshold would help to optimise the set of verbs in each termbase.

80 Proportionally, Minitab records even more than these figures suggest compared to Symantec, given the smaller size of its termbase.

6.7 Variants in the corpus

In this section, we consider the prevalence of variants in our corpora. For the purposes of our research, a variant is a term with a surface form that is similar to or derived from the surface form of another term with which it shares the same meaning (see section 2.3.3). Terminological variants include acronyms and other abbreviated forms, and differences due to spelling, hyphenation, case, and the use of spaces (for example, *checkbox* and *check box*). However, differences in case or morphology that are present due to grammatical or stylistic requirements of running text, such as the initial capital of a sentence or capitalisation styles of headings, or a term used in the plural form, are not considered variants.

In Section 6.2.5, based on two companies that provided observable data, we estimated that between 10 and 20 percent of termbase terms are variants. We are now interested in determining how frequent variants are in the corpora, as this will help to demonstrate the prevalence of variants in the language used in companies, more specifically, companies in the IT domain in our case. For this purpose, we will determine the frequency of some variants in relation to the frequency of the so-called main term.

It is not technically within our means to search the corpora for all types of variants, as that would require a term extraction tool with variant detection functions, which is not readily available. We can, however, search the corpora for the variants that are encoded in the termbase. Since only Minitab and SAS have encoded variants in such a way that they can be identified, we will search only the Minitab and SAS corpora.

Among the 1,565 entries containing corpus-valid terms for Minitab, 46 contain both a full form and an acronym both of which are corpus-valid (they do not have a *rejected* or *constrained* usage value). The acronyms occurred 10,951 times and their full forms 7,395 times in the corpus. In nearly all cases, the acronym occurs more frequently than the full form.

Sixty-two entries contain both a full form and an abbreviated or short form, also corpus-valid, for example:

- fitted distribution line / fitted line
- Fiducial CI / fiducial confidence interval
- Control Panel / Windows Control Panel

The 62 longer forms occurred 1,658 times in the corpus, with 17 not occurring at all. The 66 shorter forms occurred 4,263 times, with only six not occurring at all. Again, the shorter variants are more prevalent than their corresponding full forms.

For SAS, we found 252 corpus-valid acronyms in the termbase, occurring 143,394 times, with 29 not occurring at all, and 22 occurring fewer than 10 times. There are 255 corresponding full forms found 21,784 times, with 42 not occurring at all, and 77 occurring fewer than 10 times. Thus acronyms are used six times more often than the full forms.

There are about 460 other types of variants, but most are truncated MWTs, for instance, *process flow diagram* and *diagram*, or *client authentication* and *authentication*. As it would be impossible to ensure that all the occurrences of the truncated form correspond in meaning to the full form, a frequency count in this case would be unreliable. Instead, to test our assumption that variants do occur significantly in the corpus, we chose several with more unique morphological or syntactic structures that would reduce these risks of polysemy. In all cases, the shorter variant occurs more frequently than its longer counterpart.

Term	Frequency	Variant	Frequency
graphics output device	40	graphics device	147
Web Archive file	0	WAR file	54
extended server memory	6	extended memory	15
classification variable	631	class variable	1,165
classification effect	52	class effect	61

Table 58: Examples of MWT and variants, showing frequencies

We checked about 370 SAS variants and their corresponding main terms (typically a longer form) in the corpora and we found that the former is much more common than the latter.

These two findings demonstrate that variants are common in commercial texts and must be accounted for in commercial termbases.

6.8 Observations

In all four companies, the gap between the termbase and corpus is too large, but the size of the gap differs significantly, and this difference has allowed us to make some observations and offer some potential explanations. In general, terms with the following properties tend to have a low correspondence with the corpus:

- Multi-word terms containing a word that is potentially non-essential or introduces over-specificity
- Terms longer than three tokens in length
- Upper case terms (proper nouns)

Variants occur frequently in commercial texts and therefore, documenting variants in termbases is important. Also, domain-specific verbs and other closed word classes should be documented, especially those that share the same surface form as another word class.

Setting term boundaries optimally is quite challenging and requires a corpus-based approach to term identification.

Among the four sets of data that we are studying from four different companies, the data from HP diverges considerably from the other three, in the following fundamental areas:

1. In relation to the size of its termbase, its corpus is very small compared to the other companies. (See Section 6.1.3)
2. Over 70 percent of the termbase terms do not occur in the corpus. This proportion is much higher than that of the other companies. (See Section 6.2.1.4)
3. It has the highest proportion of long MWTs (terms with more than three tokens),

indeed, four times that of two of the other three companies. (See Section 6.2.3)

4. The termbase contains an unusually high proportion of upper case terms and proper nouns. (See Sections 6.2.2 and 6.3.5)
5. Many of the terms in the termbase are not actually terms, but are full or partial sentences. This is because the termbase was populated with strings from a TM, including, for example, full sentences and function words. (See Section 5.1.4).

These findings demonstrate that there is insufficient correspondence between the corpus and the termbase for our research. For this reason, HP has been excluded from many of the remaining stages of our data analysis.

In the next chapter, we explore the potential of keywords to raise the correspondence between termbases and corpora.