Translating stones: a corpus-based linguistic and lexicographic study in specialized terminology

Viviana Gaballo[1]

Abstract

This study originated from the real-world need to provide a lexicographic reference work for the specialized field of stone processing. Very little is available on this specific niche of the lexicon. This contribution will offer lexicographers and terminologists a first insight into the identification and designation of materials, activities, and processes related to the quarrying and processing of stones. The study was conducted on the data collected to build a pair of comparable corpora, each containing a variety of texts – from brochures to technical specifications – in one of the source languages investigated: English and Italian. The methodology employed derives from the report on a Council of Europe project (see International Journal of Lexicography vol. 9, n. 3, 1996). To advance the inquiry, a number of term candidates were identified – based on the frequency and keyword lists generated from the corpora – and analysed in their contexts of use to eventually formulate hypotheses of equivalence in both languages. This work is the result of the growing convergence of different approaches to meaning, all harnessing corpus evidence.

Keywords: terminology – lexicography – corpus linguistics – contrastive linguistics – translation studies

[1] University of Macerata, Macerata, Italy: viviana.gaballo@unimc.it

Introduction

This article draws from ongoing research work aiming to build an LSP glossary of terms from a sub-domain of construction engineering: stone processing. This very small niche demanded attention as the domino effect triggered by the propagating wave of market globalization compelled many small and medium businesses to review their marketing policies and consider expanding their traditionally regional business scope to encompass the global market. This sudden, epoch-making transformation of their enterprise entailed unexpected, massive use of language services to meet their need for linguistic representation of the field in which they operated, without incurring global-local cultural traps [Hofstede, 2001: 21].

While the language used in some of these sectors (e.g. computer science) had already undergone terminological standardization [Cabré, 1999: 199-203] and language professionals could easily resort to established sources for advice on possible translation solutions, the new vocabulary introduced into some specialized domains (or sub-domains) – more connected to local traditions and cultures – urged for lexicalization in the target language, and sometimes even in the national language (e.g. the food and wine industry).

This article addresses some of the issues raised when analyzing the linguistic material collected with the aim of formulating hypotheses of equivalence for the compilation of a glossary of terms.

1. « Standard » issues

The problem of standardization was first tackled by the Austrian engineer Eugen Wüster who, in his work "Internationale Sprachnormung in der Technik" [1931], formulated the basic principles of terminological modernization and standardization, advocating the creation of special institutions that should be in charge of establishing the principles and methods of terminology. One such institution came later into being as ISO Technical Committee 37 Terminology (Principles and Coordination), and published some twenty standards for terminology and lexicographic work in the 60-year span of its activity. The relevance of this normative production is undeniable, yet the application of a standard, as the 7th principle puts forth [Sager, 1990: 120], is closely linked to the very nature of the standard itself and the dominating industrial, legal, and social conditions that would make the standard applicable. As far as specialized domains are concerned, the need for terminology regulation is often induced by the market, and minor sectors will find it harder to receive the attention and allocation of human/material resources necessary to fill the gap in the terminology planning agenda.

1.1. How "standard" should standards be?

The lack of normalization of the terms used in specific domains forces language professionals to refer to authoritative resources – such as industry associations – that may have issued glossaries to ease professional communication.

One such association – in the stone industry – is the Marble Institute of America. The comprehensive monolingual glossary, and the fully detailed product cards of dimension stones (whose terminology this study will focus on) are just some examples of the valuable linguistic materials that the Institute has made available on its web site. Yet, from a lexicographic perspective, the mismatching designation of industry-related objects (dimension stones) in different regions of the world as reported in some of their product

cards, reveals how extra effort is required from regional and national standardization authorities to break down the culture-bound resistance to full adoption of shared rules (the use of Imperial measures in the UK is a notable case in point).

The mismatching designation mentioned above regards for instance a stone called "Jura Limestone". Part of the specifications of this stone on its product card reads "Internationally this stone is called a marble. For the U.S. and Chinese market this is correct according to ASTM C 119 and GB/T 17670. But in areas of European Standard (EN) jurisdiction, this stone must be defined as limestone because EN 12440 demands a scientific terminology for the classification of natural stones, determined with a petrographic examination according to EN 12407 and 4.2 of prEN 12670."

This dual designation in English, due to differing international product standards, is mirrored by a dual designation in Italian based on the context of use. To relate the example discussed above to the Italian counterpart, we should very briefly say that limestone is classified as a natural stone. More precisely, natural stone is a term used to differentiate non-manufactured stone products such as limestone, travertine, marble and granite from human-made "cultured stone" products. As reported by the Lombardy Marble Association (QUOTE: "Commercialmente si usa il termine pietre verso quei materiali non lucidabili o non abitualmente lucidati perché non hanno il grado di diffrazione della luce che è associato allo splendere dei marmi. Il termine pietra non viene usato in ambito scientifico, la consuetudine settoriale vi include diversi tipi di rocce …"), the commercial denomination of limestone in Italian, oddly enough, is simply *pietra* (stone), while the scientific designation depends on its composition (*roccia ignea*, *sedimentaria* or *metamorfica*), size (*argillite*, *arenaria*, *tufo*, *conglomerato*), or form (*breccia*).

To add up to the already disarrayed linguistic patterns, the Lombardy Marble Association informs us that "Pietra Naturale" is now also a quality mark (as the renowned "Pura Lana Vergine" for example).

All this information is significant for the different purposes of both the lexicographer and the terminologist. But the question as to how "standard" a standard should be remains open.

2. Interlingual lexicography issues

Lexicography, as well as all other branches of applied linguistics, is generally characterized as interdisciplinary by nature.

When discussing the interdisciplinary relations between lexicography and translation studies, Gerhard Hartman [2006:157-158], noted how the overall improvements that had been "diagnosed for the study of monolingual dictionaries have not been fully applied to reference works involving more than one language", and advocated – among other things – more systematic studies on the translation-related complexities of the information categories (words, phrases, meaning, grammar, usage, names, etc.) in various interlingual reference works.

Before listing his set of lexicographic desiderata for dictionary-making and dictionary research, Hartman [2006:158-159] urged for more empirical studies on some specific issues such as intercultural diversity, translation equivalence and directionality. This last issue, along with translation equivalence, will be the main focus in the following paragraph.

2.1. Bi-directionality and consistency in terminology

While directionality in translation studies usually refers to trainees' and professionals' greater competence in translating from a foreign language into their mother tongues

[Lonsdale, 2004: 63-64], in terminology, directionality is realized in either of two ways depending on the context, on its being either monolingual or bi-/multilingual. Sager [1990: 223-224] clarifies that "term banks are directional in the sense that they have a monolingual database with translation equivalents which may then have pointers to another monolingual database which gives the full range of information on a 'homonymous' term in that language", but he also adds that what is really needed to assist translation from and into any language is "term banks with fully reversible entries".

Terminology databases are often designed in such a way as to be open for external contributions, as is the case with IATE (Inter-Active Terminology for Europe), the EU inter-institutional terminology database. This valuable multilingual resource is constantly updated with terms suggested by language and field experts, and monitored for coherence and consistency by EU terminologists.

A case of inconsistency in the IATE terminology bank is documented below for the sake of scholarly speculation, certainly not to diminish the value of the huge amount of work involved in maintaining this priceless collection of specialized terms in the different European languages. In a spirit of collaboration, the inconsistency was promptly notified to and equally promptly acknowledged by the IATE team. This example will be used here to warn about the possible misuse of the interdisciplinary quality of terminology, and to stress the need for bidirectionality.

The case at issue refers to the apparent replacement of the onomasiological process typical of terminology by the semasiological process typical of lexicography [Cabré, 1999: 38]. It is questionable whether a term in a multilingual term bank can be matched with a definition rather than the equivalent term in another language (or a paraphrase) should the term not exist in the TL [Sager, 1990: 42-44; Baker, 1990: 26-42]. Definitions are undeniably part and parcel of both lexicographic and terminographic work; what worries here is certainly not the presence of a definition (which should receive its proper space in a term bank) [Sager, 1990: 44-50], but the absence of the equivalent term. For this reason, the suggested match of the English term "flame texturing" – one of many stone processing techniques – with the Italian description of the technique instead of the equivalent term "fiammatura" is rather questionable (IATE copyright notice does not permit reproduction of any part of the term bank content, therefore the description used cannot be reported in this paper).

Equally questionable would be the level 3 reliability assigned to the match, which seems to imply and justify the intentional use of a definition in the place of the equivalent term.

The entry also provides us with the opportunity to stress the need for bi-directionality as demonstrated by the missing match in the reverse search, i.e. from Italian into English. None of the hits returned for the search term "trattamento superficiale" contains the long definition that was used to translate "flame-texturing".

Moreover, it appears that the inconsistency mentioned before occurs more than once in the IATA database since, even in the case of terms commonly used in the metal and plastics industry, e.g. *sfridi* (IT) and *swarfs* (EN), their definitions are provided instead, in both English and Italian. But, going back to the sub-field discussed in this article, i.e. stone processing, and in this paragraph, i.e. directionality, even the search for the Italian term "fiammatura" did not return any occurrences of the equivalent term "flame-texturing".

Here, in addition, another issue arises: is a shift in rank [Catford, 1965] plausible in a terminology bank, i.e. can a verb be used to translate a noun (if the latter is available in the TL) and vice versa – reference is made here to the terminological units *dar la fiammatura* and *flash effect*. As a matter of fact, when full bi-directionality fails, the term bank user (e.g. a translation trainee) should be alerted to a potentially poor match, which would call for more investigation. To this purpose, the IATA data bank has been provided with a tool, the reliability star ranking, that informs the user about how safe the match is. Yet, the use of this

tool should be more consistent, considering that three stars – meaning "reliable" – have been assigned to the shift in rank being discussed here.

Full bi-directionality should be aimed for, as Sager [1990: 223] noted, especially "in a bilingual or multilingual context, and where tools are required to assist translation from any one into any of the other languages."

3. Corpus issues

While the usefulness of investigating corpora for language description has long been acknowledged [Sinclair, 1991; Teubert, 2007], it took a while until corpus-based terminology became an established procedure [Maia, 2002], probably because of the different nature of the corpora involved: in the former case, large corpora are used to validate hypotheses of a higher generalization power, in the latter small corpora are used as a valuable aid for the treatment of specific research questions, particularly in translation and contrastive linguistics.

Although corpora need not contain large amounts of text to be relevant, they do have to contain the most appropriate types of texts based on the aim of the study: in our case, the aim was the compilation of a bilingual glossary of stone industry related terms with a focus on a subset of natural stone – limestone. To this purpose, a bilingual comparable corpus was built: unlike a parallel corpus, a comparable corpus contains original, non-translated texts in two languages, sharing some features such as topic, text type, time frame, degree of technicality, etc., and is also particularly useful in uncovering culturally relevant information.

3.1. An "unbalanced" corpus

As stone processing companies introduce their products on brochures, leaflets, catalogues and web sites, these text types were naturally elected for inclusion in our corpus. However, the limited availability of Italian texts on the specific domain resulted in an extremely unbalanced comparable corpus: more than 120,000 words in the English (UK) sub-corpus vs. fewer than 10,000 words in the Italian sub-corpus. The imbalance could have been worrying, had not domain specialists, including the commissioner of the glossary, put their expert knowledge at our disposal. Their competence in the field made up for any missing or dubious information in the sub-corpus.

Actually, the absence of some domain-specific terms such as *stonemason* and *tagliapietra*, was noticed in the relevant sub-corpora. The terms would require additional investigation which is beyond the scope of this article, but would certainly be challenging for any further analysis or research. Possible lines of development could be: 1) some diachronic investigation to check whether the terms have become obsolete, e.g. due to the introduction of machine processing tools; 2) a sociolinguistic analysis to understand if and why the human activity (i.e. cutting) is represented but its actors are not, considering that all other synonyms of the word (ledgeman, stone cutter, etc., in their singular and plural forms) are missing as well.

3.2. The "driving" force

The methods used to extract terminology from documentation have undergone increasing automatization, from total human control (manual underlining or highlighting of term candidates) to total computer control (fully automatized term extraction). A middle measure between the time-consuming activity of the former and the relatively unreliable results of the latter can be found in the use of corpora and corpus analysis tools. A method inspired by the

corpus-driven approach suggested by Tognini-Bonelli [2002: 84-100] was used to retrieve data and speed up term selection as it allows for the disambiguation of meanings that would otherwise be very difficult to distinguish without resorting to the relevant context.

In applying the approach to our study, we noticed however that while word lists are exceptionally helpful in exposing the lexical structure of a corpus hinting at possible underlying interpretations, they do not perform as well when it comes to detecting specialized vocabulary. This task is carried out particularly well by keyword lists.

3.3. A "key" issue

A keyword is generally intended as a means to access digitally stored information. Yet, the meaning it acquires when used with corpus analysis tools is substantially different: these identify keywords by comparing patterns of frequency. A word can be identified as a keyword if it occurs with unusual frequency in a corpus as compared with its frequency in a larger corpus that is used as a reference corpus. Thus, by comparing the wordlist from an LSP corpus – like the STONES Corpus designed for this study – with the wordlist from a LGP corpus – like the BNC, our reference corpus – we are able to harvest a number of keywords that may already be considered candidate terms. A quick look at the results obtained after running the Keyword tool (one of the tools in the Wordsmith Suite) on the STONES Corpus seems to confirm this (see figure 1).

stones.kws (keyness)

N	WORD	FREQ.	S~1.TXT %	FREQ.	R~2.LST %	KEYNESS
1	LIMESTONE	1.574	1,36	799		17.973,0
2	STONE	2.097	1,81	8.204		17.601,1
3	SLATE	1.159	1,00	551		13.312,9
4	TRAVERTINE	966	0,83	15		12.729,7
5	TILES	1.036	0,89	1.057		10.920,5
6	MARBLE	1.063	0,92	1.400		10.814,9
7	CAITHNESS	739	0,64	153		9.038,4
8	FLAGSTONE	689	0,59	28		8.952,0
9	HONED	706	0,61	117		8.742,6
10	SQ	764	0,66	1.000		7.777,9
11	VAT	838	0,72	2.361		7.503,1
12	NATURAL	1.171	1,01	14.044	0,02	7.401,4
13	STONES	798	0,69	3.227		6.642,2
14	UNDERFLOOR	504	0,44	27		6.507,1
15	GRANITE	579	0,50	583		6.111,6
16	EX	790	0,68	4.735		6.013,7
17	PAVING	554	0,48	459		5.993,2
18	BASKET	624	0,54	1.286		5.911,0
19	£	506	0,44	394		5.514,5
20	SANDSTONE	484	0,42	367		5.291,1

Figure 1. The first 20 keywords from the STONES Corpus (Software used: Wordsmith Tools)

The concept of keyness is a relative concept, highly dependent on the corpus used for reference; in our case, the use of the BNC corpus as our reference corpus strengthened our expectations that the results would be reliable. Close observation of the first 20 keywords from the STONES Corpus provides us with the certainty that the corpus is representative of the terminological niche we are investigating. Our objective was to analyse terms belonging to a sub-subset of the stone industry – *limestone* – and the fact that this word ranks first in the keyword list is reassuring about the reliability of the corpus.

The keyword list also shows some interesting differences and analogies between the two English corpora (LSP/LGP): although the adjective *natural* belongs to general language, it has nonetheless been detected as one of the top-20 keywords in the STONES Corpus, which clearly underlines the importance of the concept that is to be related to the macro-category of the stones at issue, i.e. *natural stones*. Once again, this confirms the reliability of the corpus. Yet, an odd word made its way into the keyword list that only to the non-expert eye could appear as a candidate term, i.e. *Caithness*: this word only shows as a whole set of abstract words indicating physical, generally measurable properties of objects, but totally lacks the referent in such a category, as it simply denotes a geographical area in Northern Scotland. Interesting information from the Caithness.org website completes the picture:

> **The Quarry**
>
> The stone was obtained from the quarry which ran from below the car park and extended up through the Castletown village for almost one mile. The facility of Caithness sandstone to split into thin slices or flags (from the Norse word 'Flaga' – a flag) occurred because 370 million years ago a Great Lake covered the whole of the north of Scotland and out into the North Sea. The bed of this lake was formed by successive layers of sediment. Between the layers the lake dried out, leaving dead fish (fossils) and the subsequent planes along which the rock now splits. The flags were raised by hand using levers. Very little blasting was employed to break up the rock. At the height of production in 1902, 35363 tonnes of flag was produced, valued at £23,239 – a huge sum in those days.

The word *Caithness* only apparently disturbs the "purity" of the keyword list; on the contrary, since in most of its occurrences (535) it co-occurs with *flagstone*, it is extremely important to the language expert as it confirms the use of a geographical modifier to qualify a type of stone, which in turn justifies the use of the modifier *Lecce* to qualify the typical stone quarried in the area around the "Florence of the South", as they lovingly nicknamed Lecce in Italy. A web search on UK or US sites will confirm specialized use of the functional unit "Lecce stone".

From a lexicographic viewpoint, the first 20 keywords from the STONES Corpus contain the whole subset of Natural Stone – in order of frequency: limestone, travertine, marble and granite, plus a member of the limestone class: sandstone. A keyword list could not be more complete in its initial screen. The general term *stone* is quite obviously recurrent in both its forms, singular and plural. Products (semi-finished and finished products) and processes are represented as well: *slate*, *tiles*, *flagstone*, *honed*, *paving*.

4. "Method" issues

The research methodology applied in this study is a combination of the corpus-driven approach (Tognini-Bonelli, 2002: 84-100), by which we let the corpus inspire our investigation and suggest strategic solutions in advance of our queries, and the reverse formulation of the procedural steps in the identification of translation equivalence suggested by Tognini-Bonelli (2202: 135), by which we intend to counter-balance the more contrastive linguistics prone model with one that is better suited to translation studies and terminology.

TL Comparable Corpus	**SL Comparable Corpus**	**SL/TL Comparable Corpus**
Step 1	*Step 2*	*Step 3*
from TL Formal Patterning to TL Function(s)	identification of possible SL candidate terms/ translation equivalents for each function	verification of correspondence between SL translation equivalents and TL functions

Table 1. Translation Equivalence Model – A Functional Perspective – Stage 1

It is argued that, in order to achieve functional equivalence in a specialized language, our enquiry should begin by analyzing the target language corpus first. Perusal of word lists and keyword lists from the TL corpus – as well as skimming of concordance lines – will provide the appropriate linguistic pre-conditioning in line with the domesticating technique (Venuti, 1995: 61) which should be given priority whenever we want a translated text not to sound foreign, as in the case with technical and scientific texts.

This methodology for investigating translation equivalence represents a departure from the methodological stages suggested by Tognini-Bonelli (2001: 135) but at the same time it re-affirms the need for a truly and fully corpus-driven approach, not only in linguistics but also in translation studies and terminology.

Stage 1 of the TE Model may seem a time-consuming activity but it requires no more time than the time spent by language professionals in collecting information before starting their job; in addition, it provides them with the linguistic outfit needed to appreciate the analogies or differences between industry-related cultures, while providing some final solutions for the SL terminology/text, which have not been biased by the source language. Further parsing in reversed order will complete the search.

SL Comparable Corpus	**Translation Corpus/Translator's Experience (SL/TL)**	**TL/SL Comparable Corpus**
Step 1	*Step 2*	*Step 3*
from SL Formal Patterning to SL Function(s)	identification of possible TL candidate terms/ translation equivalents for each function	verification of correspondence between TL translation equivalents and SL functions

Table 2. Translation Equivalence Model – A Functional Perspective – Stage 2

By going through Stage 1, in a sort of top-down approach, the terminologist/translator will assess what is likely to be found in the specific domain of the target language and what is (im)possible to associate to the source language while, by going through Stage 2, in a sort of bottom-up approach, s(he) will assess what is likely to be missing in the TL corpus that should be searched in other sources.

The relevance of this method lies in the further reduction of the existing gap between translating from and into one's mother tongue.

5. Concluding remarks

This article has addressed some of the issues raised when analyzing the linguistic material collected with the aim of formulating hypotheses of equivalence for the compilation of a glossary of terms.

The brief overview of the linguistic domain of natural stone processing allowed us to raise some points regarding standardization, bi-directionality in term bases, and methodology in interlingual terminology and translation.

Terminologists' ultimate goal is terminological standardization in specialized domains so that terminographic products such as dictionaries may be created. Standardization, however, has been achieved in a few sectors only, e.g. computer science. While glossaries have been issued by industry organizations to ease professional communication in the domain of stone processing, which lacks a standard, the language professional is left with nagging doubts about the designations of dimension stones in different regions of the world. Extra effort is required from regional and national standardization authorities to break down the culture-bound resistance to full adoption of shared rules.

A case of inconsistency in the IATE terminology bank has been documented as an example of replacement of the onomasiological process typical of terminology by the semasiological process typical of lexicography – equivalent terms have been replaced by their definitions – and used to support bi-directionality in terminology banks.

Finally, the criterion of keyness has been used to elicit data and speed up term selection from our STONES bilingual comparable corpus. This procedure for term extraction has proved to be effective, and has been incorporated into a new Translation Equivalent Model derived from Tognini-Bonelli's methodological stages (2001: 135).

This multidisciplinary contribution is deemed to be of some interest to all language professionals involved in lexicographic and terminological studies.

Bibliography

BAKER Mona, *In Other Words*, London, Routledge, 1992.

CABRÉ M.Teresa and **SAGER** Juan C. (ed.), *Terminology : Theory, Methods and Applications*, Amsterdam-Philadelphia, John Benjamins, 1999.

CATFORD J.C., *A linguistic Theory of Translation*, Oxford, Oxford University Press, 1965.

HALLIDAY M.A.K., **TEUBERT** Wolfgang, **YALLOP** Colin and **ČERMÁKOVÁ** Ann, *Lexicology and Corpus Linguistics*, London-Ney York, Continuum, 2004.

HARTMAN Reinhard R. K., "Desiderata in Lexicography: Looking Back at Some Problems, and Forward to Solutions", Sica Giandomenico (ed.), *Open Problems in Linguistics and Lexicography*, Monza, Polimetrica International Scientific Publishers, 2006.

HATIM Basil and **MUNDAY** Jeremy, *Translation: An Advanced Resource Book*, New York, Routledge, 2004

HOFSTEDE Geert, *Culture's Consequences: Comparing Values, Behaviours, Institutions, and Organizations across Nations*, Thousand Oaks, Sage, 2001 [1980].

LONSDALE Allison B., "Direction of Translation: Directionality", Baker Mona (ed.), *Routledge Encyclopedia of Translation Studies*, London, Routledge, 2004 [1998].

MAIA Belinda, "Corpora for terminology extraction: The differing perspectives and objectives of researchers, teachers and language service providers", *Language Resources for Translation Work and Research*, LREC 2002 Workshop proceedings, 25-28.

SAGER Juan C., *A Practical Course in Terminology Processing*, Amsterdam-Philadelphia, John Benjamins, 1990.

SINCLAIR John McH., *Corpus, Concordance, Collocation*, Oxford, Oxford University Press, 1991.

TEUBERT Wolfgang (ed.), *Text Corpora and Multilingual Lexicography*, Amsterdam-Philadelphia, John Benjamins, 2007.

TOGNINI-BONELLI Elena, *Corpus at Work*, Amsterdam-Philadelphia, Benjamins, 2001.

VENUTI Lawrence, *The Translator's Invisibility: A History of Translation*, London, Routledge, 1995.

WÜSTER Eugen, « Die Allgemeine Terminologielehre – ein Grenzgebiet zwischen Sprachwissenschaft, Logik, Ontologie, Informatik und den Sachwissenschaften», *Linguistics*, Vol. 119, 1974: 61-106.