Multiple terminological equivalence of word combinations in scientific discourse: PEPCSAD - pilot study

Katarzyna Kołudzka-Ziętek
University of Silesia
katarzyna.koludzka@us.edu.pl

Abstract

This article presents a pilot study for an ongoing PEPCSAD project (Polish English Parallel Corpus of Specialized and Academic Discourse). It focuses on the issue of terminological equivalence in the translation of word combinations in the specialized discourse. It investigates the actual number of equivalents used by translators and it tries to explain the occurrence of multiple equivalence in specialized discourse.

1. Introduction

The field of specialized terminology is inextricably linked and influenced by specialized translation and undoubtedly, without properly rendered terms and concepts, any translation is doomed to be inadequate and unsuccessful. As a result, specialized terminology in translation attracts the attention of both researchers working in the field of translation and translators themselves, trying to improve the quality of their work. Over the past decades translation studies (also the ones focusing on terminology) have greatly benefited from the use of corpus methodology and as a consequence, in order to enhance the quality of research, PEPCSAD emerged. PEPCSAD (Polish English Parallel Corpus of Specialized and Academic Discourse) is a long-term project with the aim of compiling large, searchable, bidirectional, parallel corpus of English and Polish specialized texts. Our motivation behind its creation is twofold. On one hand, we expect it to be an indispensible source of information about the process of translation, but on the other hand, it will have a more practical application as it will serve both as a terminological data bank, as well as an aid for translator’s training.

The following article presents preliminary results from PEPCSAD pilot study focusing on terminological equivalence in the case of translation of word combinations in domain-specific texts. We wanted to find out the number of equivalents of a given word combination actually used by the translators and its implication for further research.

1.1 Word combinations

Words are definitely not used in isolation. In order to convey desired meaning people choose them from varying semantic sets and they use their combinations to create phrases and roomier clauses. To put it differently, words in any language are characterized by certain level of collocability, that is ability to co-occur in syntagmatic chains and form more or less stable semantic units, which can be termed word combinations. At the same time it should be underlined, that the way words are combined is far from being random and accidental. Such combinations can be perceived as ‘a rich and important source of information both about language and about the world we live in’ (Evert 2005: 15). Undoubtedly, the most interesting (but at the same time the most difficult to identify and define) group of word combinations are collocations. Anyone deciding to go into research on collocations will have to be prepared to face lack of consistency in their description as ‘regrettably, collocation is a term which is used and understood in many different ways’ (Bahns 1993: 57)
In the literature of the subject we may find multitude of very often confusing and sometimes even mutually exclusive definitions of collocations. One of the factors that has led to this proliferation of definitions is the fact that over years the notion of collocation has been widely exploited in various areas of linguistic research (SLA, lexicography, phraseology, discourse analysis, natural language processing etc.). Hence, varying approaches and definitions of collocations have emerged, which as has been claimed ‘depend on the specific aims of the observer’ (Gledhill 2000: 7). As our goals are limited also our definition of collocation (here: collocation will be used interchangeably with the phrase word combination) employed for the purpose of the present pilot study was simplified and restricted only to the occurrence of statistically-significant contiguous nominal phrases.

2. Methodology

2.1 Corpus design

In order to conduct our pilot study we compiled a corpus sample consisting of the 17th edition of The Merck Manual of Diagnosis and Therapy, one of the world’s bestselling and recognizable internal medicine textbook, and its available Polish translation. The investigated English version consists of 1 355 215 words and is written by 286 medical experts, accompanied by 22 consultants and 15 editors. Target text consists of 1 375 949 words and was devised by 39 translators (all of them being medical experts) and 17 editors. In order to ensure text normalization and enhance searchability both texts were converted into txt-format, all tables, figures, images, footnotes and tables of contents were removed. Additionally, corpus was annotated with basic metadata such as: genre, language of the original, title, author or translator; proper annotation was also provided for knowledge domain and subsequently our sample was divided into sub-corpora representing different fields of medicine.

2.2 Collocation extraction

In order to determine the nodes of investigated collocations, frequency-based list of words, showing the occurrences of all words in the investigated sample, was created using WordSmith’s WordList Tool. Next, the list was limited to a hundred most commonly occurring items and after the initial removal of grammatical words only nouns were chosen as nodes. By using Concord Tool option we were able to detect all collocates of the previously selected nodes. Next, frequency-based and statistical tests were employed (see table 1) in order to filter out statistically significant word combinations and consequently only combinations with f>5, MI>3 and t-score>2,576 were chosen for further analysis (chi-square test was used as an additional test only for border line cases). Finally, 128 different collocations (3380 individual instances) were selected and analyzed.

Then, concordance lines (using WordSmith’s Concord Tools) were created for all 3380 instances of investigated collocations. As the preliminary data processing did not involve the use of sentence aligner (which will be used in the latter stages of analysis) consequently, corresponding concordance lines were extracted manually from the Polish version of the corpus. Additionally, concordance lines were supplemented with the translator’s code. One of the last stages of analysis was enlisting all possible translation equivalents of investigated word combinations.

3. Results and discussion

Specialized discourse is supposed to be very precise and thus, one of its features mentioned by Gotti is its monoreferentiality (Gotti 2003: 33). Monoreferentiality means that ‘in a given context only one meaning is allowed’ (ibid). However, as Pamela Faber noticed ‘Although

<table>
<thead>
<tr>
<th>collocate</th>
<th>f(x)</th>
<th>node</th>
<th>f(y)</th>
<th>total</th>
<th>t-score</th>
<th>chi-square</th>
<th>MI</th>
</tr>
</thead>
<tbody>
<tr>
<td>liver</td>
<td>1144</td>
<td>disease</td>
<td>4406</td>
<td>181</td>
<td>13,17717</td>
<td>697,954</td>
<td>5,6048111</td>
</tr>
<tr>
<td>blood</td>
<td>496</td>
<td>flow</td>
<td>2102</td>
<td>196</td>
<td>13,945049</td>
<td>32419,629</td>
<td>7,9930569</td>
</tr>
<tr>
<td>abdominal</td>
<td>674</td>
<td>pain</td>
<td>1970</td>
<td>196</td>
<td>13,930017</td>
<td>27298,131</td>
<td>7,6442155</td>
</tr>
</tbody>
</table>

Table 1. Example of statistical tests employed to extract collocations from the corpus.
specialized language initially aspired to the ideal of having one linguistic designation for each concept with a view to imbuing specialized communication with greater precision, reality has turned out to be quite different (Faber 2009: 8). As we may presume from the obtained data a very similar trend is seen in the translation of specialized discourse.

The mean number of equivalents for all investigated collocations was 6.21 (the maximum number of equivalents used was 27 for renal disease). A closer look into data reveals that translators use significantly more (96%) multiple equivalents. (see diagram 1).

The question of course arises why is it so.

One of the reason may obviously be connected with the frequency of occurrence of collocations in the corpus. Diagram 2 clearly shows that the number of equivalents is directly proportional to the number of occurrences of a given word-combination - the more occurrences, the more potential equivalents.

An attempt was also made to determine the relation between the use of a given equivalent and translator’s specific idiolect. In other words, we tried to detect whether the appearance of multiple equivalents may be attributed to translators’ personal preferences as ‘translator’s idiolect may have certain relevance within the translating of a given text’ (Barlow 2001: 156). In order to investigate whether we can observe any link between translator’s preferences and the use of a very specific equivalent we used translator’s codes to prepare statistics for any equivalent with f>2. It turns out that out of all possible equivalents for all investigated word combinations 72% were used by more than one translator (e.g. translation choroba nerek [f=82] for renal disease was used by 13 different translators). Then, we focused on the remaining 28% (that is equivalents being used by only one translator)
and we reversed the statistics trying to find out whether this translator restricts himself only to the use of this specific equivalent for a given word combination. The statistic shows that in 83% of cases translators do not stick blindly to one possibility but juggle with multiple choices and then tailor them to meet their needs. This result may prove Kerremans statement that ‘in specialized texts, authors sometimes deviate from the traditional view of using only one term consistently throughout the text to refer to a particular concept’ (2010: 1-2).

Next stage of our analysis was the comparison of the number of dictionary equivalents with their actual number used in the corpus. Surprisingly, it was impossible to find any dictionary equivalent in the case of 77% of investigated word combinations, one equivalent was provided for 18% and two equivalents for 5% of them. Dictionary is therefore insufficient source of translation equivalents especially for word combinations as ‘in specialized dictionaries it is very often neglected that besides their referential function (i.e. they point to domain-specific concepts), terms fulfill additional roles in specialised discourse’ (Kerremans 2010: 1)

2. Conclusion
In conclusion, we must definitely underline the fact that translation is highly context dependent. When we take a closer look into the patterns of equivalent variants used by translators (all of them being professionals in the field of medicine) we may discover that they are not always synonyms. Very often translators provided more general translations (e.g. choroba układu krążenia <cardiovascular disease> instead of heart disease) or used much more specific terms (e.g. ból nadbrzusza <epigastric pain> instead of abdominal pain) or they even decided to use descriptive variants in order to make their translations more cohesive. Thus, instead of providing one dictionary equivalent consistently, experienced translators may use multi-equivalence as a way to enhance the quality of their translation, to make it more understandable and naturally-sounding for the target audience. The results of this pilot study should be reflected in the way translation courses are structured and consequently more emphasis should be put on terminological variation in translation and its context-dependence.

References


Faber, P. 2009. The pragmatics of specialized communication. Entreculturas, no. 1: 61-84


Słomski, P. 1996 Podręczny słownik medyczny angielsko-polski i polsko-angielski. PZWL: Warszawa

---

1 English-Polish medical dictionary was used as a reference (Słomski, 1996)