Collocations in translation: Personal textbases to the rescue of dictionaries

Uzoma Chukwu

Electronic version
URL: http://journals.openedition.org/asp/2991
DOI: 10.4000/asp.2991
ISSN: 2108-6354

Printed version
Date of publication: 1 December 1997
Number of pages: 105-115
ISSN: 1246-8185

Electronic reference

This text was automatically generated on 2 May 2019.
Collocations in translation: Personal textbases to the rescue of dictionaries

Uzoma Chukwu

Introduction

Collocations have been high on the agenda of linguistic research these last few years, driven by second language teaching and learning research on the one hand and learner’s dictionary lexicography research on the other.¹ Yet, it would be erroneous to think collocations do not pose a problem in translation teaching and learning where students are usually assumed to be already language specialists, especially in the teaching and learning of translation into learners’ B language. The problem here is, for the most part, similar to that posed in ordinary language teaching and learning situations. Usually – and ideally – one of the differences between the translation submitted by a student working into their B language and that submitted by another student working into their A language will be the inability of the former to always come up with the right verbs, adjectives or prepositions that collocate with nouns given certain conditions.

However, if this was just an issue relevant in the context of translating into one’s B language in a training program, then it would have been more expedient to approach it within the framework of second language teaching and learning. The fact is that one does not have to be learning to translate into a foreign language to have first hand experience of how problems of collocation and collocability may constitute the single most irritating linguistic problem in translation.² The question then is: can anything (that is not just ad hoc) be done to remedy the situation? In the rest of this paper, I first try to make a case for the relevance of the issues discussed here for the professional translator. Then I go on to examine the place of dictionaries in the overall scheme of things before introducing the textbase solution proposed here. Finally, I discuss the tools for extracting
collocational information from textbases and considerations for setting up such textbases.

**Collocations in professional translation**

3 There are, at least, two situations in which even the most seasoned translator working into their native language could come against a brick wall in this regard. The first is a situation in which all bilinguals find themselves at one time or the other, viz. when the languages in contact in the bilingual's mind cannot be kept sufficiently apart. This is a common occurrence in translation situations because of the overwhelming presence of the source text. Although there are innumerable benefits accruing from bilingualism, there are also enough reports in the literature of negative influences for bilinguals. The idea has been strong enough to warrant the formulation of what has been called the “balance hypothesis” By this hypothesis, Macnamara (see, e.g., Appel & Muysken 1987) postulated that humans have a “fixed capital” of neural and physiological capacity for learning language and that more proficiency in one language would usually imply fewer skills in others.

4 The problem can also be approached from the point of view of code-switching. The question then would be: why do people switch code? And do findings made possible by research into the conditions in which bilinguals switch code provide support for the assumption that even the most seasoned translators working into their A languages sometimes need corroborating (empirical) evidence for the word combinations they use? Appel and Muysken (1987: 118) argue that code switching can, among other things, serve the referential function since it often involves lack of knowledge of one language or lack of facility in that language on a certain subject and that certain subjects can be discussed with more ease in one language than in the other. It seems reasonable then to argue that if the translator’s original familiarization with the domain underlying the document being translated was made in the B language, finding the right collocations in the A language will not always be as intuitive a process as it should be ordinarily.

5 There is a second and even more compelling argument for the relevance of issues relating to collocation to the top-flight translator working into their A language. The problem here is more complex and arises from the non-homogeneity of language. The notion of “native speaker” has been a largely misused (and even abused) one. How far can the argument of the native speaker’s intuitive knowledge of their language be taken? What is the status of a non-chemist native speaker of English, for example, listening to two chemists discussing the design and results of a chemistry experiment? In fact, the argument has been made that if one discounts the very personal discourse produced within the family and among friends, much of the discourse produced in a language is done within discourse communities that make use of language resources that may not necessarily be available to non-members of the particular discourse communities.

6 Researchers who approach Languages for Specific Purposes (LSPs)\(^3\) from the genre analysis point of view (see, e.g., Swales 1990; Beacco 1992) make the especially interesting distinction between “linguistic communities” and “discourse communities”. A language user will belong to a particular linguistic community by virtue of their having that language as their native tongue whereas belonging to a discourse community is usually a direct outcome of belonging to a socio-professional group. While belonging to the linguistic community in whose language a particular discourse community’s text

\(^3\) ASp, 15-18 | 1997
production is carried out certainly puts the language user at some advantage, it may not be enough to guarantee the same intuitive knowledge the speaker demonstrates in general speech when it comes to handling discourse produced within discourse communities to which they don’t belong. In other words, the competency needed to produce and follow valid discourse within the chemistry discourse community, for instance, is not likely to be part of the intuitive knowledge of English available to every native speaker of the language. Smadja (1993: 146-147) notes, in this regard:

Native speakers are often unaware of the arbitrariness of collocations in nontechnical core English; however, this arbitrariness becomes obvious to the native speaker in specific sublanguages. Linguistically, mastering a domain such as the domain of sailing requires more than a glossary, it requires knowledge of domain-dependent collocations.

**How dictionaries fit into the picture**

7 Dictionaries and lexicography (at least, European lexicography) have come a very long way since Samuel Johnson (see, e.g., COBUILD, Longman Language Activator, The Cambridge Word Route bilingual series). And the changes that come to mind here are not just cosmetic. Every one of the dictionaries mentioned above represents a bold attempt to redefine the mission of dictionaries and dictionary-making and to break with the monolithic view of the problems users bring to dictionaries. Contrary to the pervasive assumption that has driven dictionary making until fairly recently, users do not always flick through the pages of dictionaries in search of the meaning of difficult words. Sometimes, they are just in search of words to express ideas in their heads, or the right verbs, adjectives or prepositions to combine with particular nouns, etc. From the point of view of collocations, COBUILD is perhaps the most relevant of the three cited above. Although, it innovates in other quite significant ways (e.g., how it characterizes meaning), its enduring value derives from the tons of collocational information it provides the user and how they have been presented to facilitate retrieval. Mention must also be made of the BBI Combinatorial Dictionary of English (more commonly referred to as BBI, the initials of its compilers). Strangely enough, the emergence of such new dictionaries has also served to draw attention to the scope of the problem and how even the most laudable efforts in this regard barely succeed in defining the problem. Dictionaries (individually or collectively) may never be able to completely satisfy the language user’s collocation needs because of the immensity of the territory to cover.

8 The second reason – closely linked to the first – is that the dictionary maker’s sense of what might be a source of problem may not always be representative of that of the spectrum of potential users. While theoretical knowledge on what constitutes a collocation may be fairly precise (see, e.g., the introduction of BBI; Smadja 1993), the actual decision of the dictionary maker to record a particular collocation will nearly always also be influenced by practical considerations. After all, who wants to clutter their dictionary – dictionaries are already cluttered enough with the current levels of information – just to mention word sequences which, although technically qualify as collocations, can be easily computed? But assuming a dictionary could damn the consequences, there still remains the question of how such abundance of information will affect retrieval time. Until fairly recently, dictionary makers offered only one way to retrieve information from dictionaries, irrespective of whether they have been implemented electronically or they are print dictionaries: call up the headword and then
plod through the tiny print of the entry in search, for example, of the collocation one is interested in. Although this is a procedure that is guaranteed to find the collocation if it had been recorded, it has serious overhead implications for the translator’s time. While physical time is not an unlimited resource in the translation business and must be taken into consideration, pressure on psychological time is even more difficult to manage and may have more serious consequences for the quality of the translation.

The textbase solution

It is clear from the foregoing that the provision of collocational information has become one of the primary tasks of dictionaries. Yet, to insist that dictionaries provide “all the collocations available in a language” would also be to miss the point that collocations do not form a closed set category like the lexicon. In fact, according to Leech (1991), even the syntax and the lexicon of a language are more open-ended than we may have been led to believe. The idea of “providing all the collocations in a language” is therefore an impractical one. And even if the dictionary maker could find a way around the theoretical and practical problems involved here, to compile a dictionary that provided all the collocations available in the language, such a dictionary would be for all practical purposes unusable.

The only other feasible alternative that comes to mind is for individual (professional) users to pick up the thread from where the dictionary left off by setting up personal textbases. This is an idea that should not be too hard to embrace since most translators already maintain textbases, although under the name of parallel texts and even if, for the most part, the documents comprising such textbases have been hardcopy documents, managed manually. Aside from the essentially linguistic use to which the textbases can be put (and which interests us most here), they should also provide the translator with invaluable domain knowledge within a few strokes of the keyboard.

Once the principle of the textbase solution has been agreed to, the two issues of how to put the textbases together and how to manage them must be addressed. There is perhaps no special reason why any of these two issues must be taken before the other; however, for purely organizational reasons I will consider the tools first before considering sourcing.

Tools for transforming textbases into linguistic information

Tools for transforming raw text data (corpora or textbases) into information abound today and can be deemed to come in two flavors: concordancing and full text retrieval (FTR) tools. Viewed from a distance, tools in the two categories may appear to be largely the same; however, such an impression proves to be erroneous upon closer examination. Tools in the first category incorporate functionalities to make texts provide all sorts of sophisticated information that can be used not just to make various types of judgement about language per se and the functioning of texts but also to perform content analysis. Although such tools may be used by anyone, one would think they are essentially scholarly and lexicographic research tools. Some of the best known ones include Micro OCP, TACT, WordCruncher and Wordsmith.
The second category comprises tools that are almost mono-function to the extent that they offer only (or essentially) FTR functions. These are, properly speaking, tools designed with the business environment in mind, the kind used by the typical help desk to ferret out information for a customer on the other end of the phone line. Like concordancing tools, they are also usually very sophisticated although sophistication here has been mostly channelled into the number and complexity of search algorithms implemented, indexing method, and retrieval speed. Typical examples of packages in this category include: DTSearch, InfoFerret, Isys and ZyIndex.

To determine the suitability of tools of these two categories for our purposes, let us measure them against a number of standard criteria, using one standard package from each category (DTSearch and TACT) as reference.

Table 1. Summary of features of the two categories of tools for manipulating textbases

<table>
<thead>
<tr>
<th></th>
<th>Full Text Retrieval (FTR) Tools</th>
<th>Concordancing Tools</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Indexing</strong></td>
<td>classic indexing</td>
<td>classic indexing only</td>
</tr>
<tr>
<td></td>
<td>advanced indexing (takes letter case into consideration)</td>
<td>relative high speed of indexing although not as high as that of FTR tools</td>
</tr>
<tr>
<td></td>
<td>index updating</td>
<td></td>
</tr>
<tr>
<td></td>
<td>very high speed indexing</td>
<td></td>
</tr>
<tr>
<td><strong>Query</strong></td>
<td>Boolean, proximity</td>
<td>Boolean, proximity</td>
</tr>
<tr>
<td></td>
<td>phonetic, fuzzy, viewers</td>
<td>query via regular expressions</td>
</tr>
<tr>
<td></td>
<td>transparent and user-friendly query procedure</td>
<td>procedure not transparent and not particularly user-friendly</td>
</tr>
<tr>
<td><strong>Preprocessing</strong></td>
<td>no preprocessing modules</td>
<td>possibility of preprocessing (tagging, lemmatization, etc.)</td>
</tr>
<tr>
<td><strong>Information provided</strong></td>
<td>full text retrieval</td>
<td>full text retrieval</td>
</tr>
<tr>
<td></td>
<td>keyword-in-context (KWIC)</td>
<td>keyword-in-context (KWIC)</td>
</tr>
<tr>
<td></td>
<td>no concordances</td>
<td>concordance generation</td>
</tr>
<tr>
<td></td>
<td>no text statistics</td>
<td>text statistics</td>
</tr>
</tbody>
</table>

A few comments are in order here. First, from the point of view of indexing, the possibility offered by FTR tools to update existing indexes is a major advantage of this category of tools in the context of maintaining textbases which are dynamic, as opposed to corpora that are static (see below). The second comment concerns the querying of textbases. Although concordancing tools are quite competitive in terms of query algorithms that have been implemented, the user is required to come up in each case with the exact query statement for retrieving the information sought. This is perhaps explained by the fact that concordancing tools are typically research tools and they need to come with the power and flexibility to permit the researcher things that are not run-of-the-mill. The third comment concerns the type of information one can expect to obtain from textbases using these tools. Although some FTR packages do have KWIC displays (in addition to the traditional full text display) just like concordancing packages,
the definition of what constitutes a KWIC display varies a great deal. While Figures 1 and 2 below are KWIC displays for TACT and DTSearch respectively, in the first case we have a veritable concordance display whereas in the second (and despite all the configuring done) it was impossible to obtain the same kind of display. Interestingly, from the point of view of the user interested in collocations, only the concordance display is truly useful. Only a concordance display can allow the user to make the kinds of observation necessary for making collocability decisions.

Figure 1. Concordances for the word “card” in an electronic commerce textbase generated using TACT

The display was configured to not only prefix the word immediately to the left of the keyword with ‘>’ to facilitate study of collocation pattern but also to sort concordances according to the word immediately to the left.

Figure 2. KWIC display for the word “card” in an electronic commerce textbase generated using DTSearch

No amount of tweaking could give the kind of display possible with TACT (see Figure 1).

From the foregoing, it is clear that concordancing tools are better adapted to our purposes than FTR tools. Naturally, it will need to be added that for concordancing tools
to be entirely satisfactory, they will need to embrace the dynamic archival vision of corpora that has motivated the choice of the term “textbases” in this paper (see below) and implement index updating. It is not an entirely cheering prospect to have to completely re-index the textbase each time a new document is added to the collection.

Problems of setting up textbases

It is important at this point to say a few words on a major terminological choice made in this paper, i.e., the use of “textbases” (instead of corpus). A corpus is generally a finalized and static collection or body of texts, a closed set: as at the time of exploitation of a corpus, the collation process is deemed to have been completed. Considering the uses to which corpora have been generally put (text analysis, lexicographic research, etc.), this is quite understandable. However, the sort of text collection in mind in this paper will of necessity be an evolving one, an archive. Not only is it unreasonable to expect the translator to postpone using the textbase until collation has been completed or the textbase has become representative enough (to employ a criterion used in evaluating how accurately corpora mirror the larger population of texts of which they are supposed to be only a sample), it is difficult to see how such considerations can be translated in concrete terms. The concept of textbase has therefore been introduced to emphasize the dynamic and the largely opportunistic (see below) nature of the collection.

Another major issue concerns how the textbases will be sourced. Even though much of the publishing done today is in computer-enabled environments, resulting in the availability of nearly all published texts in computer-readable format, the problem of corpora sourcing has still not been solved. Texts are intellectual property and those who hold the copyright quite understandably do not give them away without good reason. Big translation agencies may have the resources to negotiate permission to use copyrighted material but the majority of individual translators are certainly not in any position to do that and may have to do what has been the practice in the corpora business, that is collect texts opportunistically, “according to what sources of data can be made available and what chances for collection arise” (Leech 1991:10). The Internet is today the most readily available source for such an opportunistic collection of texts. Naturally, because much of what is available over the Internet has not gone through the editorial process, they usually carry no warranties of reliability and the professional user will need to come up with criteria for weeding out the worst. Such criteria may include reading part of or the entire document, consideration of the source of the document, etc. before including it in the textbase.

A second major issue that needs to be addressed here concerns the choice between maintaining just one textbase or several textbases. Although the first option may seem to offer the advantage of reaching critical mass quite fast, it most certainly is not the ideal choice. There are certainly innumerable collocations in the general component of a language but one would think that professional language users like translators will have the most problems with collocations that are restricted to speech within different discourse communities. In other words, if the translator wants the collocational information obtained by querying their textbase to be as reliable as possible, then they will need to maintain separate textbases for the different discourse communities identified. Another important reason why this multiple textbase solution is preferable derives from the need to make the textbases double as knowledge bases.
Concluding Remarks

As it was pointed out in the body of this paper, the textbase idea suggested and discussed here as a way of addressing the collocation problem in translation does not require any real “paradigm shift” to embrace. Most translators are already into it although, like it was also pointed out, such bases have been mostly made up of hardcopy texts, managed manually, and have been used more as (non-linguistic) knowledge bases. However, this is not to say it is an idea that can co-exist harmoniously with some of the most foundational assumptions about language that underlie current practices in translation. Here, I think specifically of the assumption of the native speaker’s intuitive knowledge of their language and how it is affected by the non-homogeneity principle of natural languages also introduced in the body of this paper. There is a sense in which to ask the translator (native speaker) to set up and maintain textbases for the sole purpose of getting their collocations right may be tantamount to asking them to abandon (at least, partially) a major marketing argument. Yet, if the non-homogeneity principle is correct, it is hard to see how this can be completely avoided. But the translator needs not simply seek to pursue some fine linguistic principle with such textbases; there are enormous benefits to be reaped at the end of the process since such textbases can also double as domain knowledge bases.

The problem of embracing the idea is therefore one of sensitization. This is why this discussion opened with a reference to the translation teaching and learning context. Sensitization to the reasons why even a top-flight translator working into their A language needs some readily-available source for verifying word combinations and how only personal textbases can adequately address the problem is more easily achieved in a training program. If trainee translators are made to see why they cannot rely solely on their intuitive knowledge of their native languages to get the translation of that medical (or legal, or economic, etc.) translation right, they are more likely to evolve into translators that consider a concordancing software package as a standard component of the translator’s workbench. The widespread availability of these tools for the personal computer environment has made projects of this magnitude to cease to be feasible only in the context of well-funded research laboratory or international cooperation efforts.

BIBLIOGRAPHY


**Dictionaries**


**NOTES**

1. Natural Language Processing (NLP) is another important area that has helped galvanize collocations to the top of linguistic research agenda. The realization that the difference between the native and the non-native speaker is the relative ease with which they come up with those non-computable word combinations that are called collocations has led researchers to seek ways of characterizing such combinations for NLP systems (see, e.g., Smadja 1993).

2. There are a number of possible reasons why the problem may not have received the attention it deserves in the translation community. First, the translation community will like to believe, rightly or wrongly, that the young men and women in translation training programs have complete mastery of all their working languages. Mission statements such as: Translation teaching is about the problems of how to transfer meaning that has been expressed in one language into a second language as opposed to perfecting one’s skills in any of the languages implied are indicative of the assumptions underlying the curriculum of translation schools. The second explanation has to do with the politics of the translation business. By over-exalting the native speaker, the politics of the translation business has led many translators to develop an unhealthy confidence in their intuitive knowledge of their native languages (see, e.g., Massardier-Kenney 1997).

3. The increasing interest in LSP research and teaching is a direct indication of the recognition by linguists of this non-homogeneity of natural languages.

4. Specialized dictionaries have not been entirely left behind in the revolution sweeping through lexicography. Most of them (see, e.g., Routledge bilingual dictionary series) are providing more
and more collocational information although they too are still nowhere near satisfying all the collocational needs of the sophisticated professional user.

5. It is encouraging in this regard that more recent electronic implementations of dictionaries are including full-text search, a functionality that enables the querying of an entry to find any particular piece of information that would otherwise require reading significant portions of the entry to ferret out. For example, the current implementation of the French language dictionary, *Petit Robert*, comes with functionalities for full-text search, for the retrieval of cited material, and for the dictionary plan view permitting the visualization of only the principal senses in which the headword is used.

6. There are number of projects with such (or similar) ambitions. A good example is the *Dictionnaire explicatif et combinatoire du français contemporain* (DECF), which is a direct application of the meaning-text theory of Igor Mel’cuk. Incidentally, the DECF is intended to be a theoretical dictionary and is therefore only useful to the researcher. *The Dictionnaire contextuel de français pour la géologie* (DCFG) is another bold attempt in this regard although limited to the narrow and specialized area of geology. The final product of DCFG is a two-volume document weighing close to eight kilograms and that requires the patience of a researcher to ferret out any collocational information.

7. It is true that, in some cases, there is some tweaking to be done on the texts (tagging, lemmatization, etc.) at the pre-processing stage for the program to be able to provide the sort of information the researcher is interested in.

8. Although *WordCruncher* is a very powerful tool, it is with some hesitation that I have included it in the category of concordancing tools. It was impossible, using the WordCruncher 5.2 viewer, to obtain the one-line KWIC display that qualifies as concordance.

9. The KWIC display in DTSearch is termed “Variable Context Display” in TACT which reserves the name KWIC to the one-line display of the keyword in its context of occurrence (see Figure 1). Wordsmith’s use of the name “concordance” for this display is perhaps more helpful as it removes all ambiguities.

10. Typically, one would expect that a document emanating from a reputable research organization or government agency will be more reliable than one emanating from an Internet enthusiast who is no way staking their reputation.

ABSTRACTS

Les problèmes de collocation constituent facilement les problèmes linguistiques les plus intraitables en traduction, compte tenu surtout de la non-homogénéité des langues naturelles. Si les dictionnaires modernes sont riches en collocations, ceux-ci ne sauront peut-être jamais répondre entièrement aux besoins d’une catégorie d’usagers professionnels aussi avisés que les traducteurs pour la simple raison que les collocations ne constituent pas une classe fermée (comme pourrait l’être le lexique). Le présent article propose une étude de la solution qui passe par la constitution par chaque traducteur de base de données textuelles à partir desquelles pourront facilement être générées, selon le besoin, des concordances pour un mot ou un terme quelconque.

Problems of collocation are easily the most daunting linguistic problems encountered in translation, in particular due to the non-homogeneity of natural languages. While most modern dictionaries are providing more and more collocational information, they are still no where near
completely satisfying the needs of sophisticated professional users like translators, especially considering that collocations do not constitute a closed set – like the lexicon. This paper discusses how translators can rise to the challenge by setting up and maintaining personal textbases from which concordances for any word or term can easily be generated as the need arises.

INDEX

Mots-clés: base de données, collocation, concordance, corpus, traduction
Keywords: collocation, concordance, corpus, textbase, translation

AUTHOR

UZOMA CHUKWU

Uzoma Chukwu trained as translator/terminologist and holds a PhD in Anglophone Studies from Université Lumière Lyon 2 (France). He is currently visiting professor at the Graduate School of Translation and Interpretation, Monterey Institute of International Studies in California (USA), where he teaches translation and terminology management. His research interests include terminology theory, translation pedagogy and LSPs. uchukwu@gowebway.com