COMMUNICATION &LANGUAGE at work
Issue no. 1 | Summer 2012
ICT Tools and Professional Language

Electronic Corpora as Translation Tools:

A Solution in Practice

Vigdis Jensen, Birthe Mousten & Anne Lise Laursen

(pp. 21 - 33)

http://ojs.statsbiblioteket.dk/index.php/claw/article/view/7236

Electronic Corpora as Translation Tools:

A Solution in Practice

Vigdis Jensen
Master's degree in Spanish
Translator and Proofreader at LocaLsoft S.L., Málaga in Spain.
Birthe Mousten
PhD, M.Eng
Department of Language and Business Communication, Aarhus University, Denmark.
Anne Lise Laursen
Associate Professor, Spanish LSP
Department of Language and Business Communication, Aarhus University, Denmark

Abstract

Small bilingual text corpora from a source and target language can be important sources of specialized language tracking for translators. A corpus platform can supplement or replace traditional reference works such as dictionaries and encyclopedia, which are rarely sufficient for the professional translator who has to get a cross-linguistic overview of a new area or a new line of business. Relevant internet texts can be compiled 'on the fly', but internet data needs to be sorted and analyzed for rational use. Today, such sorting and analysis can be made by a low-tech, analytical software tool. This article demonstrates how strategic steps of compiling and retrieving linguistic data by means of specific search strategies can be used to make electronic corpora an efficient tool in translators' daily work with fields that involve new terminology, but where the skills requested to work correspond to being able to perform an advanced Google search. We show the different steps in setting up and working with an ad-hoc corpus, illustrated by means of the software AntConc applied on the SEO area.

Introduction

The work behind this article took its starting point in a Master's thesis by Vigdis Jensen from the Department of Business Communication, Aarhus University (Jensen 2011). The purpose of the article is to operationalize the findings of the thesis and to specify the strategies in order to show how translators can use electronic corpora actively in their translation of specialized texts. In the following, we define electronic corpora briefly. Then we describe the different steps of corpus work: The first step is the compilation of relevant, specialized, electronic comparable corpora. This work is relevant whenever a specialized or emerging area has to be dealt with. A second step, which is a one-time-investment, is the compilation of a reference corpus to sort away function words, such as articles, prepositions, pronouns, etc. An optional step may be to find the threshold of new incoming words in the compilation process, which is now possible thanks to the launching of a software tool which measures the number of new words (types) as a ratio of already existing words (tokens) in the corpora (Corpas Pastor and Seghiri 2007). This is a help if the translator wants to test when there is satisfactory linguistic variation within the specialized area. The last step is the analysis, where the key element is an electronic framework to harvest linguistic data by means of relevant search strategies.

The case study forms the central part of our article. We follow the steps delineated above by using the software program AntConc (Anthony 2006) for the Search Engine Optimization (SEO) area.

Electronic corpora

Over the past couple of decades, large electronic text corpora used as raw data for software processing have gained ground in specific areas such as linguistic research or lexicography, where they can replace past-time beliefs or linguistic intuition as well as past-time activities of manual scrutiny of texts for linguistic evidence. For translators, the Internet itself is a valuable source of linguistic and subject-matter data and so is the huge collection of EU parallel texts. Among translation scholars, however, focus is now turning towards smaller corpora of specialized text, as these are more likely to document the traits of a specific genre or domain. Of specific interest to the translator are the bilingual/multilingual comparable corpora, which can be described as original, non-translated texts from two or more languages that are thematically compiled according to the same set of design criteria (Tognini Bonelli 2010: 21). Li et al. emphasize that corpora must consist "of documents in different languages covering overlapping information" (Li et al. 2011: 473), which ensures matches in the same fields of work. Use of non-translated texts thus means avoiding interference from the source language, which characterizes some parallel or translated texts (Teubert 1996: 247). Therefore, it is an efficient instrument, not only in the translation training stage (Laursen and Arinas Pellón 2012), but also in the subsequent daily practice of professional translation. The latest approach in terms of utilizing text corpora for

translation purposes are corpora compiled 'on the fly' for specific tasks, also referred to as ad-hoc corpora (Sánchez Gijón 2009).

Compilation of specialized corpora

The composition of a corpus or subsets of bilingual corpora depends on the specific domain or genre of the task to be translated and the specific line of business in question. This could be for instance annual reports for companies or technical descriptions for the automation industry. No matter which genre or domain, a common search strategy is needed to compile relevant texts. The compilation can take place via keyword search on Google, which, in fact, does not differ from the normal procedure in connection with a translation, except that in this approach texts are stored and subsequently processed by the software.

Firstly, it is possible to specify a whole class of sites by using Boolean search. The colon, for instance, permits you to specify the origin of the documents: if you insert 'site:es' in the search field together with a relevant keyword, Google will return results from only Spain. The +sign and –sign will exclude sites or combine search queries. By using a –sign as in e.g. 'quarterly report –annual' only quarterly reports will appear and by using the +sign as in 'annual report +quarterly', sites with both quarterly and annual reports will be returned. Secondly, a search can be enhanced by using keywords to move into more specific areas or domains. Should you be preparing a corpus for an automation-industry task, the combination in Google's advanced function of the keywords automation and robotics in the 'all words' field will generate sites with specialized vocabulary that might give inspiration for further search options in the 'any of these words' field; for an automation corpus, keywords like two-axis, multi-axis, CNC and footprint could be relevant to find more specific sites.

The process can be synthetized like this: 1) make a few initial search queries for texts related to the basic concepts of the texts to be translated, 2) iterate the process one or more times with the new keywords found in the first round(s) and 3) store the texts as flat texts (.txt) with a file name for the source and a date for the collection of the text (Jensen 2011: 25).

A search within a URL address is adequate and effective when a translator wants to track corporate language and terminology within a specific company, institution or organization. A specific search may also be made in portals or web directories, which are organized link collections of companies operating in a specific field (an example is dmoz). This may at first glance seem more direct. However, the search for relevant URL-addresses may be difficult if the translator is not familiar with the line of business, which is typically the problem at the outset of this process; another problem may be that the translator has to navigate further within a company site to find relevant pages.

Compilation of a reference corpus

By means of a reference corpus, the most common function words, viz. articles, prepositions, conjunctions, determiners, common adverbials, etc. can be excluded from the list of frequent words in the specialized corpora. Including a reference corpus for translation purposes is not an obligatory step, but is an efficient element to get an overview of keywords in any specialized corpus. These keywords can be extracted by the software tool through a comparison of the two corpora. Reference texts are common-language texts of any kind, e.g. private emails, short stories, simple newspaper articles, etc. that can be collected into one long document just containing running, general-purpose text. The compilation of the reference texts can be considered a one-time investment since this corpus is reusable for extraction of specialized terms in other fields, too. The size of a reference corpus should be proportional to the size of the corresponding specialized corpus.

Software tools and methods

The size of specialized, comparable corpora can be approached pragmatically, which most corpus users have done until now. However, now software has been developed where measurements can be made by looking at the added output of extra texts at specific points of the compilation. The measurements by this software show the saturation point, which is the point where additional electronic texts do not add any significant amounts of new words. In other words, the addition of more texts will not add significantly to the expansion and clarification of the vocabulary in the field, and the measurements function as a stop signal to prevent random expansions of corpora. In fact, our case study below suggests that a subset of corpora can be quite limited and still enjoy terminological representativeness.

The software used to retrieve data from the corpora is a key element in the process. While a test of the saturation point is optional, the retrieval of linguistic data by an electronic tool is essential to track the span of linguistic variation in the corpora. The retrieval can be optimized by means of a number of specific search strategies, as it will appear from the following case study.

Corpora in practice—our case study

Our comparable corpora show how a translator can use ad-hoc corpora in a practical assignment. The translator can benefit from the different language corpora by harvesting the specialized vocabulary and in this way make a high-quality translation.

The original case study focused on Danish and Spanish corpora only (Jensen 2011). In our case, we have supplemented the work with an English corpus to test how the combination of corpora and software can shed light on the terminological trajectory from English into other languages—as used in the SEO business (Search Engine Optimization). We have tested the corpus platform on a set of corpora of promotional webpages offering services for Search Engine Optimization (SEO). The SEO

technology is relatively new, and the corresponding terminology includes items such as conversions, doorways, and crawlers; terms that at first sight might seem well-known, but here stand for quite new concepts. These results popped up in our English corpus and requested yet more detailed research to find equivalents in the other corpora, which will appear from our analysis below.

Specialized field of SEO

We began our compilation of the corpora by drawing mainly on Google search queries for relevant seeds (= keywords). Our point of departure was the term SEO itself, and then we followed the procedure described previously, when we repeated the process on the basis of apparently new terms found in the first round, e.g. SE Optimization, link building, optimization of webpages, etc. Alternative sources, such as portals or web directories, proved not to be so efficient, because they often link to the companies' main webpage on the website, which involves further internal search queries for relevant SEO-related pages.

Size of corpora

As to the size of the corpora, the saturation point was 186,000 words for the Spanish corpus and 161,000 for the Danish corpus, respectively. However, the English SEO corpus did not show any significant volume of new words at the point of 55,000 words, which might be due to an established vocabulary in this field. It should be added that it is not necessary to implement this control measure as a daily routine. We highlight this point in order to establish a numerical benchmark for ad-hoc corpora as such.

Software program: AntConc

As to the actual analysis of the SEO field, the software chosen is AntConc, which is a user-friendly freeware concordance program[1]. The interface shown in fig. 1 illustrates how a simple query—in this case for the term search—generates concordance lines listed in KWIC (keyword-in-context) format in the centre of the screen shot. Leftmost, Corpus files shows the list of texts included, and the File list shows in which files the examples appear. Below the result window, all the search parameters are shown. In the example, the Kwic Sort function has been used to sort the context words to the right and left in alphabetical order. This is one of the additional features of the software program.

[1] Available online 25 April, 2012 at http://www.antlab.sci.waseda.ac.jp/software.html

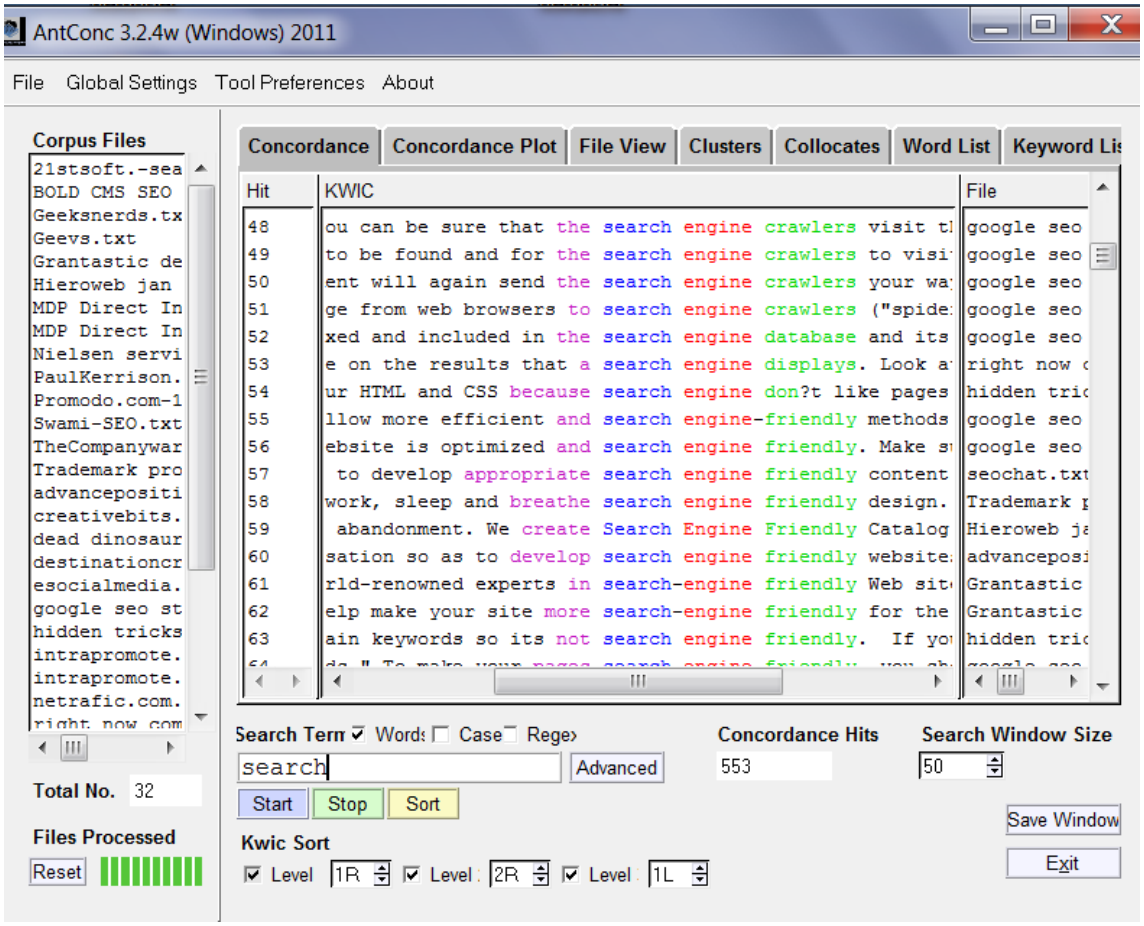

Figure 1: KWIC format listing for the term search

Search strategy for keywords

For systematic extraction of specific keywords, the software compares the specialized corpus with the reference corpus. In the SEO case, the top 10 keywords generated in this way appear from table 1.

English corpus	Spanish corpus	Danish corpus
search, site, SEO, engine, page, website, keyword, content, Google, tag	web, buscadores, SEO, posicionamiento, sitio, página, resultados, búsqueda, Google, clave	SEO, Google, søgemaskineoptimering, søgeord, hjemmeside, optimering, links, søgemaskiner, website, sider

Table 1: Top 10 keywords in English, Spanish and Danish

A further retrieval of linguistic data from the corpus involves the use of different *search strategies* depending on the specific feature or linguistic item you are looking for.

Search strategy for multiterms

Multiterms can be found by using frequent keywords as 'seeds' and by sorting the context words to the right and/or the left of the keyword. Departing from the top 10 keywords in the English corpus, the software generated alphabetical groupings of concordance lines, where typical SEO-related clusters emerged, such as *search engine crawlers* (ref. figure 1), *search engine ranking*, *search engine results*, *search engine robots*, *landing page*, *destination page*, *doorway page*, *gateway page*, *keyword density*, *keyword phrases*, *keyword prominence*, *site content and on-site conversion rate*, *(web)site traffic and (web)site visitors*.

Search strategy for synonyms

Synonyms can be found or supposed synonyms can be verified by means of a simple search looking for synonymy indicators in the concordance lines. Jensen mentions a number of such indicators, viz. *parenthesis, slash, or* and *also referred to as* (Jensen 2011: 75). A direct search for a parenthesis or slash can be performed by typing "(" or "/" in AntConc's *Search Term* box. In addition, the *Case* box must be ticked, as illustrated in figure 2, to generate the concordance lines.

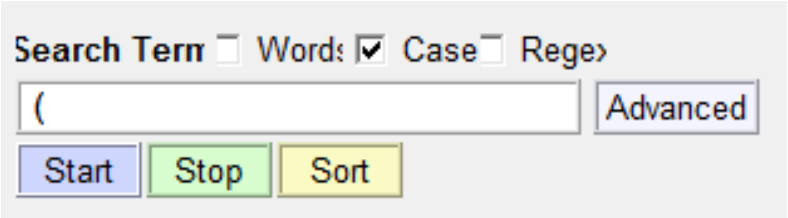

Figure 2: Example of a synonym indicator in the search phase

When we applied this strategy, the concordance lines yielded a number of co-occurrences of acronyms vs. full terms from the English corpus, e.g. *SEO (search engine optimization)*, *Pay Per Click (PPC)* and *search engine page results (SERPs)*. Other examples found by using the parenthesis were *search engine crawlers ("spiders")* and *Keyword weight (density)*. A simple search for *or* resulted in the concordance lines *"natural" or "organic" search* as well as *search engine spiders or crawlers or bots*. By tracking synonyms in this way, the translator can see which synonyms have a high frequency or low frequency within the specific language field, as well as which synonyms are at the core of the cross-company use of words and which words are specific for maybe even just one company. In some cases when nothing in the context proves synonymous relations, but when a semantic familiarity seems to exist, external sources have to be consulted. This was the case with *gateway pages* and *doorway pages* where it was not immediately visible and discernible in the corpus contexts whether the *or* signified synonyms or alternative options.

The alternative way of finding synonyms is by means of *context words (collocates)*. Inspired by the phrase *top 10 rankings*, we use for example *top* as seed, and our results

then indicated that *position(s)*, *(search engine) ranking* and *placement* might be used as synonyms. The choices at hand appear from figure 3.

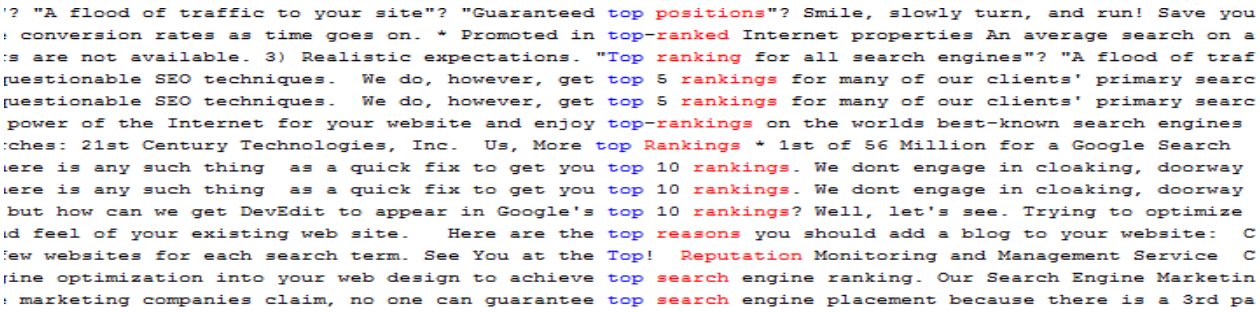

Figure 3: Finding synonyms by means of context words (collocates)

Search strategy for equivalents

A quite successful strategy for the identification of *equivalents* is the *trial-and-error strategy*. In the SEO case, we supposed that Anglicisms were to a certain degree integrated in the Spanish and Danish vocabulary so we chose to look for this as a first try. A query for e.g. *landing page* in the Spanish corpus confirmed the Anglicism, but at the same time the concordance lines delivered the domestic equivalents *página de aterrizaje* and *página de destino*, as appears from figure 4.

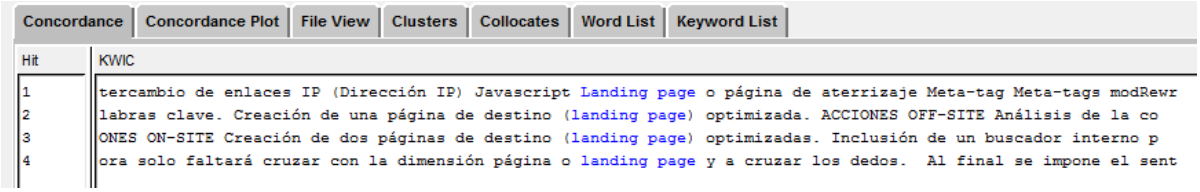

Figure 4: Trial-and-error strategy as a starting point

A corresponding query in the Danish corpus showed a frequent use of *landing page*–and indications of domestic synonyms, viz. *landingsside* and *informationsside.* A trial-and-error search for *destinationsside (*i.e. a literal translation from the term found in the English corpus*)* confirmed its existence in the Danish terminology, as well as its synonymous relation to the term *landing page*, as it appears from the following sequence from the corpus:

"destinationsside, det er denne side som brugerne skal **lande på** når de søger i Google"

[destination page; this is the page on which the users should land when searching in Google]

The above strategy of trial-and-error in terms of queries for 1) Anglicisms and 2) literal translations proved to be successful in most cases, because the queries were instantly capable of confirming or refuting the occurrence of English loan words and/or domestic competitors. Apart from that, synonyms indicators like parenthesis or slash—as referred

to above—pointed to further term variants. In general, the tests yielded various synonyms in the target languages, as shown in table 2, where "—" means no loan word, but with local terms and synonyms used instead shown underneath.

English corpus	Spanish corpus	Danish corpus
keyword weight keyword density	— — densidad de palabras clave densidad de keywords	— keyword density søgeordstæthed
crawler spider	crawler spider rastreador araña	crawler spider edderkop
conversion rate	— ratio de conversión tasa de conversión índice de conversión	— konverteringsgrad konverteringsratio

Table 2: Term extraction from target language corpora (trial-and-error strategy)

The terminological move from source language to target language in this way is quick. The simple trial-and-error query, looking for 'loans' or literal translations (calques), can confirm or refute a supposed target term in a second.

For more complex search queries, the word-in-context method and a bit more creativity are needed. As an example of this, a search query for equivalents of the term *placements*, or rather *top (10) placements*, in the Spanish corpus proved to function by inserting *entre* (among) in the search box. The outcome of this search was a number of different target solutions for a translation into Spanish as shown in figure 5.

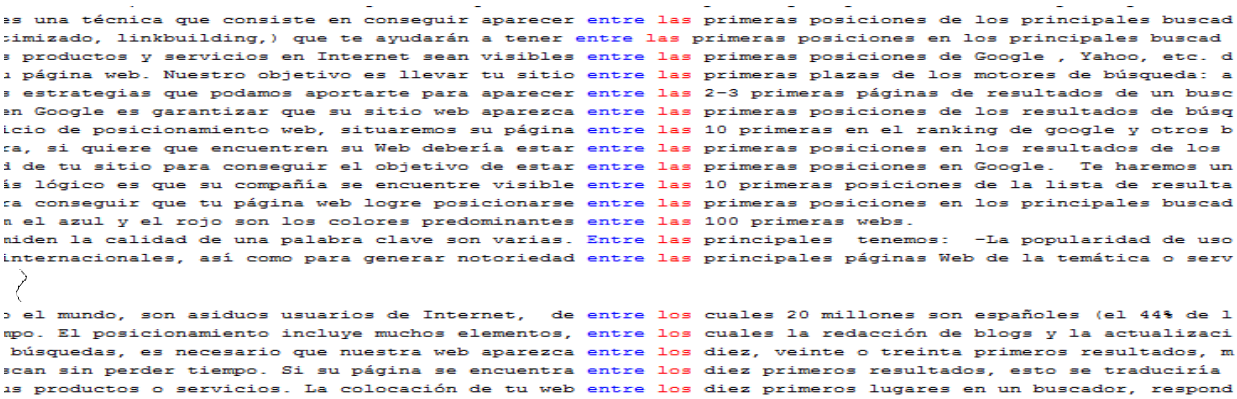

Figure 5: Entre used as the creative entrance to finding alternative solutions

The continuation of the search with the verb *aparecer* (appear), which proved to be a frequent collocate to the left in the first query, rendered a number of alternative equivalents, e.g. *aparecer en el llamado top 10* (to appear in the so-called top 10),

aparecer en el Top10 de resultados (appear in the Top10 results), or the Spanish domestic version *tope* in *aparecer en el tope de resultados*.

Apart from the extraction of specific terms, the corpora offer direct access to word associations or collocations in abundance—and on a much deeper level than can be found in any dictionary and with an instant result that beats the alternative Google search queries. In addition to *aparecer,* figure 5 above shows other collocates, e.g. *ser visible*, *posicionarse* and *encontrarse*.

Simple queries revealed quite a few collocating adjectives and verbs, e.g. *search engine friendly* (cf. figure 1), *quality traffic, drive more traffic for a website, convert into sales, index pages, follow links, exchange links, etc.*

In the translation process, the trial-and-error search can again be of great help. Looking for equivalents of *search engine friendly* in the Spanish corpus, we gave it a try with the direct translation of *friendly* (amigable) with a confirmed outcome, e.g. *amigables a los buscadores, amigables con los buscadores, amigable a los motores de búsqueda*, as appears from the screen shot in figure 6.

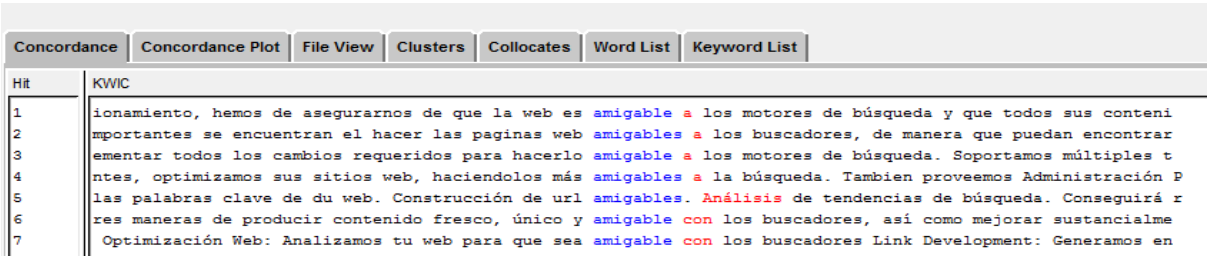

Figure 6: Results from a direct translation strategy: friendly/amigable

Again, a word-in-context strategy can be applied in the search for translation equivalents. For the English phrase *drive more traffic for a website*, the target solution in the Spanish case can be traced by means of the seed *más tráfico* (more traffic). Here a number of alternative solutions can be found in the concordance lines, viz. *enviar más tráfico a su sitio web* (send more traffic to your web site), or—inversely—*conseguir más tráfico, recibir más tráfico* and *lograr más tráfico de visitas a tu sitio* (to get more … traffic to your site). The observant reader notices that the English word *drive* with a push direction can be expressed in Spanish with a push direction, too, through the word *enviar*, but also with a pull direction through the words *consequir*, *recibir* and *lograr*. Such small subtleties will seldom be found by using standard dictionaries.

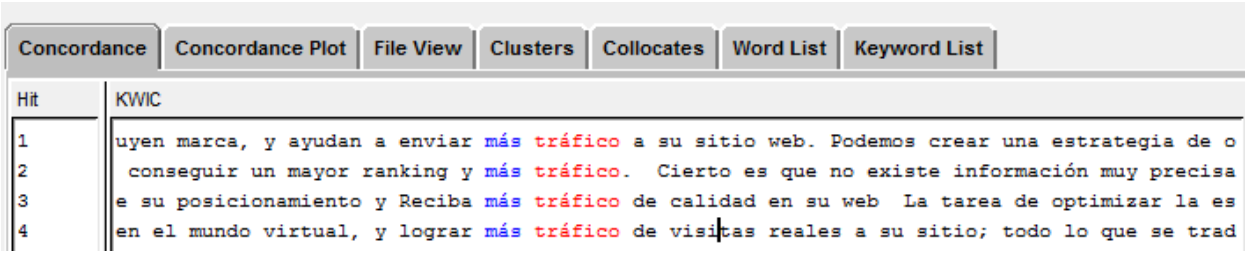

Figure: 7 Word-in-context strategy yielding a verbal structure

The examples and strategies shown and exemplified above are only some of the simple search strategies that can be made in the corpus by means of AntConc. More complex search strategies can be made to show more details, by means of the settings (Advanced, Word, Case and Regex, cf. figures 1 and 2). However, our results can be used straightaway by translators who are used to work in Google. As the translator ventures deeper into AntConc, the functions of the program may gradually become well-known tools.

Conclusion

Applying the ad-hoc corpora strategy to the terminology of the SEO is in fact a kind of litmus test for the practical application of this tool in a professional translation context. By compiling specialized comparable corpora, we find that it is possible to avoid the noise which typically occurs in Google results. The corpus texts can be stored and analyzed and, by means of specific strategies, focused queries can be made, which will yield a range of options. The software facilitates a road mapping of the special terminology in the field as well as an identification of possible synonyms while the target corpora can give you a range of options that facilitate the choice for an adequate solution in a translation.

Authors

Vigdis Jensen

LocaLsoft S.L., Málaga, Spain

Vigdis Jensen, Master's degree in Spanish. Currently employed as a translator and proofreader at LocaLsoft S.L., Málaga in Spain.

Contact:

vigdis.jensen@hotmail.com

Birthe Mousten
Department of Language and Business Communication,
Aarhus University, Denmark

Birthe Mousten, PhD, M.Eng., external lecturer at BCom, Aarhus University. Teaches technical translation and general business courses at the university. Offers courses in technical writing and translation as well as legal writing and translation for trade and industry professionals.

Contact:
bmo@asb.dk

Anne-Lise Laursen
Department of Language and Business Communication,
Aarhus University, Denmark

Anne Lise Laursen, Associate professor, Spanish LSP. Coordinator of Spanish studies at BCom, Aarhus University. Teaches translation of financial texts. Research areas lexicography, contrastive linguistics and corpus linguistics.

Contact:
all@asb.dk

References

Anthony, L. (2006): "Developing a freeware, multiplatform corpus analysis toolkit for the technical writing classroom." In *IEEE Transactions on Professional Communication*, 49(3), 275-286.

Corpas Pastor, G. and Seghiri, M. (2007). "Specialized Corpora for Translators: A Quantitative Method to Determine Representativeness." In *Translation Journal*, 11 (3). [Online, 29 February, 2012 at http://translationjournal.net/journal/41corpus.htm].

Jensen, V. (2011). "*Undersøgelse af et sammenligneligt, bilingvalt ad hoc-korpus som fagsprogligt og fagspecifikt hjælpemiddel ved oversættelser inden for et nyt fagområde.*" [Research on a comparable, bilingual, ad-hoc corpus as an LSP specific tool for translation of specialized texts in a new field]. Master's Thesis, Aarhus University.

Laursen, A. L. and Arinas Pellón, I. (2012). "Text Corpora in Translator Training: A Case Study of the Use of Comparable Corpora in Classroom Teaching." In *The Interpreter and Translator Trainer* 6(1), (in press). Manchester: St. Jerome.

Li, B.; Gaussier, E. and Aizawa, A. (2011). "Clustering Comparable Corpora For Bilingual Lexicon Extraction." In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: shortpapers*, Portland Oregon, June 19-24. Association for Computational Linguistics. [Online, February 29, 2012 at http://www.aclweb.org/anthology/P/P11/P11-2083.pdf], 473-478.

Sánchez Gijón, P. (2009) "Developing Documentation Skills to Build Do-It-Yourself Corpora in the Specialised Translation Course". In Beeby, A., Inés, P.R. and Sánchez-Gijón, P. (eds) (2009). "*Corpus Use and Translating."* Amsterdam: John Benjamins, 109-127.

Teubert, W. (1996). "Comparable or Parallel Corpora?" In *International Journal of Lexicography* 9(3). Oxford: Oxford University Press, 238-264.

Tognini Bonelli, E. (2010). "Theoretical overview of the evolution of corpus linguistics." In O'Keefe, A. and McCarthy, M. (eds) (2010). "*The Routledge Handbook of Corpus Linguistics*." Abingdon: Routledge, 14-27.