1. Introduction

For a little over two decades, computer corpora have been used by linguists to carry out research on grammar or the lexicon. They frequently contain structural information such as part-of-speech (POS) tagging or syntactic dependencies at sentence level. The Web is mostly used for information-retrieval tasks, but it may also be viewed as a very large corpus that is constantly being updated, although at the present time it does not — unlike corpora — come with any pre-encoded linguistic information.

In addition, the makeup of the Web is, by its very nature, constantly evolving and cannot be controlled; languages are still very unevenly represented, with English having the lion’s share. However, the percentage of English content is constantly decreasing, as developing countries acquire the material and necessary skills to post information online.

The question of the representativeness of the sample created by all the texts available on the Web in a given language is a key issue in modern linguistics. From a purely sociological point of view, the type of language used on the Web can be regarded as representative of the way in which the people who have the wherewithal and necessary technical skills to post documents online use the language. The problem of the relative anonymity of the sources and the absence of control mechanisms, such as peer review panels, also raises the question of the acceptability of the forms listed in a corpus that is compiled at random by merely typing key words in the search window of one of the Web’s search engines. Among other problems, it is a well-known fact that a sizeable proportion of the contents that are available in English on the Web is posted by Internet users who are not native speakers. This accounts for the large number of nonstandard sentences that are col-
lected, if one does not use any of the filters available through the “advanced” settings of the various search engines.

2. Searching for rare terms or syntactic patterns

The Web makes it possible to retrieve some rare terms or syntactic patterns for which general language corpora, such as the British National Corpus, are not large enough. This is particularly evident in the field of science. For example, the compound adjective dose-dependent, which is used 7 times in the British National Corpus, has more than 2 million occurrences on the Web (this figure is the average of the counts provided by the Google and Yahoo search engines). b

Mastering the advanced syntax of search engines makes it possible to access lexical information that may be concealed by the order of relevance that is used by most search engines. This is because the algorithm that determines this order of relevance is solely based on the adequacy between the request and the supposed contents of the Web pages, and not on the linguistic interest of the forms that are used. Thus, typing the adjective dose-dependent on a search engine would yield a high percentage of results with a context in which the noun qualified by the adjective is the word effect or its plural, effects. However, the use of the “minus” filter (dose-dependent -effect -effects) makes it possible to locate many other collocates of the adjective (inhibition, response, stimulation…).c

3. Evaluation of the frequency of use of a specialized expression and search for translation equivalents

For translators and writers working in a specialized field, the Web may occasionally serve as a tool for checking the frequency of the terms or expressions that they use. Statistics concerning language distribution on the Web still give figures of over 20% for English, although the percentage has been steadily decreasing in the past few years. It is estimated that 84% of the Web was made up of pages in English in 1997, 64% in 1999, 60% in 2001, and 35% in 2004. In the medical field, one finds an average multiplying factor of 5 for the number of results returned by search engines when entering French terms and their English translation.

Bearing this multiplying factor in mind can help the translator search for translation equivalents that are different from the carbon copy of English structures that is often generated by the growing tendency to translate English sources literally. Thus, the Google search engine returns more than 73,000 English pages for the collocation control the infection, whereas the total obtained for the literal French translation, contrôler l’infection, is less than 730 pages. In the presence of a multiplying factor higher than 100 for two translation equivalents, the translator should assume that there probably are other expressions of higher currency that are used to translate the same concept. In the case of English-to-French translation, the combined use of the quotation marks and the wild card “*” — which stands for the presence of one or several intervening words within an expression — makes it possible to enter such strings as « pour * l’infection » or « en * l’infection » and to quickly detect expressions for which the number of results is higher than those obtained for the literal translation of the collocation control the infection (combattre / éliminer l’infection).

The use of the wild card for solving translation problems may also apply when the exact equivalent of a complex term is unknown. Let us consider the case of the term gluten-sensitiv enteropathy, yielding approximately 80,000 results on the Web, in which the translator might have difficulty finding the accurate translation for the compound adjective gluten-sensitive. Since the result returned by search engines for the literal translation, entéropathie sensible au gluten, is less than 50, one may reasonably assume that there are other translation equivalents. The use of the request « entéropathie * gluten » yields more than 350 results and allows for the identification of the most frequently used equivalents (entéropathie au gluten, entéropathie d’intolérance au gluten, entéropathie induite par le gluten, entéropathie de sensibilité au gluten).

The use of the wild card also proves useful whenever the translation equivalent of only one term belonging to several frequently associated terms is known to the translator, while the other is not. Let us consider the case of the French term petite circulation, which is not listed as an entry in the Grand Dictionnaire Terminologique, unlike another term that is frequently mentioned in the same context, grande circulation. A quick examination of the first 20 results returned when entering “systemic and * circulation” reveals that most of the results contain the expression systemic and pulmonary circulation. The entry for pulmonary circulation in the Grand Dictionnaire Terminologique actually indicates that the term has a near-synonym, lesser circulation, that can reasonably be assumed to be equivalent to the French term petite circulation.

4. Current and future uses of the Web for terminological purposes: diachronic corpora and grammar check

The Web’s search engines now make it possible to search for translation equivalents and carry out bilingual term extraction by using Web pages selected for a pair of given languages as comparable language data bases or corpora. The concordancer KWICFinder, which will be described below in further detail, is a particularly invaluable tool for such purposes. For tracking the coining of new words (neology), the creation of diachronic corpora is facilitated by Google’s daterange function — in particular, in fields like data processing and medicine, which are heavily represented on the Web. The daterange function makes it possible to restrict the results of the request to the pages that were last updated prior to the date entered by the Internet user. Search engine indexing started in the late ’80s, so it is now possible to compare the current state of the language with usage that prevailed 15 years ago.

In the near future, the Web will probably make it possible to check the grammar of a text written in English by purely statistical means. The machine analysis of the most frequent word sequences, which has been used for many years in natural language processing for automatic part-of-speech (POS) tagging (using the 4-gram method), makes it possible to detect word sequences in a given sentence whose
The blood rich in carbon dioxide returns to the right ventricle.

Table 1: Number of Web pages containing the 4-word sequences of the sentence “the blood rich in carbon dioxide returns to the right ventricle”.

<table>
<thead>
<tr>
<th>Phrase</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>health-care reform</td>
<td>12</td>
</tr>
<tr>
<td>health-care system</td>
<td>9</td>
</tr>
<tr>
<td>health-care costs</td>
<td>7</td>
</tr>
<tr>
<td>health-care professionals</td>
<td>6</td>
</tr>
<tr>
<td>health-care plane</td>
<td>4</td>
</tr>
<tr>
<td>health-care facilities</td>
<td>3</td>
</tr>
<tr>
<td>health-care reforms</td>
<td>3</td>
</tr>
<tr>
<td>health-care systems</td>
<td>3</td>
</tr>
</tbody>
</table>

Table 2: Result of the n-gram search “health-care N” on Phrases in English

6. Using the Web for syntactic disambiguation of complex noun phrases.

6.1. The N1 N2 N3 syntactic pattern

Research in terminology (Frantz et al. 2001) has revealed that complex terms — that is, terms that are at least two words long — are frequently nested within longer terms or collocations. This is a characteristic that is actually taken into account by authors of terminological extraction software. When faced with such structures, the translator has to make a choice between two distinct ways to interpret the syntactic dependencies: the first noun (N1) may modify the rest of the noun phrase (N2-N3), as in the expression placebo control group, or the sequence made up of the first two nouns (N1 N2) may modify the last noun (N3), as in bone marrow transplant. In most cases, the grouping together of N1 and N2 or N2 and N3 is not a problem if the two-word collocation or complex term can be readily identified. Modern computer-assisted translation (CAT) software frequently makes use of specialized dictionaries that store such sequences and their translations, and thus allow the disambiguation of the corresponding expressions. However, as pointed out by Rouleau (2003), some translation problems may also arise when shorter sequences (such as N1 N2) have to be translated into Romance languages, when a choice may have to be made between translating N1 as a noun complement or an adjective. For example, in the term bone marrow, bone is most frequently translated as an adjective in French (moelle osseuse), but the French translation for spine surgery will use a noun complement (chirurgie du rachis).

6.2. The ADJ N1 N2 syntactic pattern

The ADJ N1 N2 syntactic pattern is frequently used in medical English and poses problems that are quite similar to the 3-noun sequence described above. Again, there is a choice to be made between two syntactic dependency patterns, depending on whether the adjective modifies the noun that immediately follows it (N1) or the N1 N2 expression. Thus, the sequence coronary artery can be readily identified.
in the noun phrase [coronary artery] disease, whereas the sequence coronary heart does not exist outside of the expression coronary [heart disease], a syntactic dependency that can be logically inferred by anyone who knows the term heart disease. Previous research (Maniez, 2001) has shown that such ambiguous structures can be disambiguated by using a comparison of the frequencies found for the ADJ N1 and ADJ N2 sequences in specialized corpora or on the Web.

6.3. Coordination in complex noun phrases
The problem posed by the assignment of modifiers to the various nouns that follow them within syntactically ambiguous structures is occasionally compounded by coordination. The following example (1) shows the potential for ambiguity when premodification and coordination are jointly used in scientific English:

(1) The ability of PET to detect cancer is based on the altered substrate requirements of malignant cells, which result from increased nucleic acid and protein synthesis and glycolysis.

Translating sentence (1) from English into any other language requires answering a number of questions:

- Does the adjective nucleic modify the noun acid, the sequence acid and protein, or the noun synthesis?
- Does protein modify the noun synthesis alone or both of the coordinated nouns that follow (synthesis and glycolysis)?
- Does the adjective increased modify acid, synthesis, or the sequence synthesis and glycolysis?

The binary choices that are available actually create a dozen different solutions for the various syntactic combinations. Obviously, both linguistic and scientific skills come into play to rule out such dependencies as nucleic protein or protein glycolysis, but looking for such expressions within a specialized corpus using a concordancer (or “googling” such sequences) may help dispel doubts if such reassurance is needed. If we symbolize syntactic dependencies with brackets, we obtain the following structure: increased [ [ nucleic acid] and [protein] ] synthesis and glycolysis. A possible French translation for the noun phrase would be « augmentation de la glycolyse et de la synthèse des protéines et de l’acide nucléique », a wording that is actually less ambiguous than the original in that the second term of the second coordination (glycolyse) is clearly out of the scope of the modification by the translation of the segment nucleic acid and protein.

6.4. Two examples of complex noun phrase patterns: <ADJ1 ADJ2 N1 (AND/OR) N2 N3> and <ADJ N1 (AND/OR) N2 OF N3>
The following patterns were isolated using a corpus compiled from the CD-ROM Annals of Internal Medicine for the year 1993. The corpus has 4.5 million word forms (tokens) and contains the text of all the articles that were published in the year 1993 in the following publications: New England Journal of Medicine, Journal of the American Medical Association, Annals of Internal Medicine, Lancet, and British Medical Journal. In order to study the syntactic patterns of interest, we used an automatic part-of-speech tagger, Winbrill.

6.4.1. The <ADJ N1 (AND/OR) N2 OF N3> pattern
After studying the 589 sequences that fit the <ADJ N1 (AND/OR) N2 OF N3> pattern, we identified four distinct combinations of syntactic dependencies:

a) [ADJ [N1 AND N2] ] OF N3
The adjective modifies both N1 and N2, and the complement OF N3 applies to both of the coordinated nouns. This combination was the most frequently encountered (55%).

- effective [identification and care] of patients
- optimal [dosage and duration] of therapy

b) [ADJ N1] AND [N2 OF N3]
The adjective modifies only the first noun, and the complement OF N3 applies to N2. This combination was seen in 26% of the selected patterns.

- [maternal age] and [length of gestation]
- [ethnic group] and [place of birth]

c) ADJ [N1 AND [N2 OF N3] ]
The adjective modifies both N1 and N2, but the complement OF N3 only applies to N2. This combination was seen in 11% of the selected patterns.

- substantial [overlap and [loss of discrimination] ]
- invaluable [ context and [source of information] ]

The adjective modifies only the first noun, and the complement OF N3 applies to both of the coordinated nouns. This combination was seen in 8% of the selected patterns.

- [absolute number] and percentage] of lymphocytes
- [natural history] and epizootiology] of Lyme disease

Theoretically, the translator is faced with two choices that combine to create the four solutions above. The choices are about whether to validate the distant syntactic dependencies, with ADJ N2 on the one hand and N1 OF N3 on the other. Unlike translating software, the translator has access to lexical meaning and can check the validity of those distant syntactic dependencies using his or her knowledge of both medicine
adverse events and yields of efficacy for sotalol and the other drugs were compared by the log-rank test.

6.4.2. The <ADJ1 ADJ2 N1 (AND/OR) N2 N3> pattern

Using the same corpus, a search for the pattern <ADJ1 ADJ2 N1 (AND/OR) N2 N3> resulted in the identification of 338 forms, which correspond to the realization of 5 distinct combinations of syntactic dependencies:

a) [ADJ1 ADJ2 N1] and [N2 N3] [central nervous system] and [bone marrow].

b) ADJ1 ( [ADJ2 N1] and [N2 N3]) good ( [hygienic practices] and [hospital policies] )

c) ( [ADJ1 ADJ2 N1] and [N2] ) N3 [serious adverse event] or [withdrawal] ) form

d) ADJ1 ( ADJ2 [N1 and N2] ) N3 persistent ( central [excitability and pain] ) behaviours

e) ADJ1 ( ADJ2 [N1 and N2 N3] ) rheumatoid ( synovial [fluids and tissue cultures] )

Once again, checking the validity of distant dependencies on the Web is of assistance in the deduction of the correct syntactic structure. In sentence (4), which presents the structure listed above under (d), the high frequencies of such collocations as persistent pain and central pain will help confirm the ADJ1 N2 and ADJ2 N2 dependencies. Likewise, in sentence (5), the relatively low counts obtained for serious withdrawal and adverse withdrawal rule out those dependencies, while the (ADJ2 N1) N3 dependency (adverse event form) is confirmed by over 10,000 results.

(4) By contrast, the same treatments are less effective when administered only minutes later, once the persistent central excitability and pain behaviours have been established.

(5) Each serious adverse event or withdrawal form was forwarded promptly to the medical department, where its receipt was documented but the code not broken.

Even if no such use of the Internet is either available or wanted by the teacher, attracting students’ attention to such complex noun phrases can only contribute to sharpening their syntactic awareness and consequently improving their translating skills.

7. Conclusion

The use of electronic resources is certainly no replacement for translators’ intimate knowledge of their field of interest and of both of the languages that are at work. However, the use of specialized corpora and the Web’s search engines as tools in translators’ quest for lexical data will probably increase in the coming years, as search modules that were originally meant to provide information about factual content are gradually upgraded with increasingly sophisticated linguistic features. Obviously, a great deal of work remains to be done in teaching members of the medical profession how to write clear scientific prose, but the principle of economy that is inherent in language and the concision that is required of scientists impose the use of premodification in English, with its share of complex and potentially ambiguous noun phrases. It is therefore a worthwhile investment for any teacher of medical translation to devote a fair amount of instruction time to the structures we have discussed in this article, and to point out the many reading comprehension pitfalls that the use of such structures often creates.

Notes

a) The British National Corpus is a text archive of over 100 000 words that was compiled in the early 1990s and whose aim was to provide linguists with a representative sample of the English language.

b) These figures, however, only measure the “surface Web,” i.e. the part of the Web which is indexed by search engines, and not the part known as the “deep Web,” which is not accessible to search engines because of the nature of its contents. According to certain estimates, the size of the deep Web could be approximately 500 times larger than that of the surface Web.

c) For the Google search engine, the online guide available at <www.googleguide.com> provides a very complete description of the ways in which such searches can be narrowed down for linguistic purposes. The API (applications programming interface) guide,
Many compound adjectives have literal translations in the field of data-processing (case-sensitive = sensible à la casse, date-sensitive = sensible à la date), but the French translation of English technical terms formed after the adjective sensitive often uses greco-latin suffixes (heat-sensitive/temperature-sensitive = thermosensible, light-sensitive = photosensible, moisture-sensitive = hygrométrique, speed sensitive switch = contacteur tachymétrique). These examples are taken from the Grand Dictionnaire Terminologique, which is available online at the following address: <www.grand-dictionnaire.com>.

It is, however, used in the definition of the term left to right shunt (shunt gauche-droite in French).

Carbon-dioxide laden is a less frequently used alternative.

Bibliography


