Terminology mining

by Steve Dyson

This article is primarily for language industry professionals who have logged a good number of hours surfing the Internet for professional information and who are wondering how to access and use such information more efficiently. A secondary aim is to encourage translators, terminologists and others that are not yet regular Internet users to explore this amazing technology further.

The article assumes that you have already discovered something of the depth and breadth of information available on the Internet, that you know how to use search engines to track down the information you need, and that you understand the basics of the Internet and related software.

My aims are

- To convince language industry professionals that have yet to explore the Internet of its immense importance and advantages.
- To suggest products and strategies for making more efficient use of the Internet and the information it offers.
- To raise some questions about terminology management in the Internet age.

But first, a question: Why search for words in context? Lexicographers, terminologists, translators and interpreters, to name but a few types of language professionals, agree on the importance of tracking down terms, usage examples, collocations, definitions and more in authentic language contexts (i.e. in documents written directly by mother-tongue authors without second-language constraints or considerations). The Internet is the most powerful resource ever for doing precisely this. Indeed, it is many orders of magnitude more efficient and cost-effective than, say, ten reserved seats in the world's biggest libraries and ten pairs of eyes scanning all day long in search of solutions to your language problems (if, by any chance, that sounds a little "over the top", be aware that the Internet is expanding so quickly that it is only a matter of weeks before it will be an understatement). On top of this, the Internet offers direct access to a rapidly expanding collection of dictionaries and glossaries in electronic form. Now, how to harness these new resources?

A personal search engine

If you go to the AltaVista homepage (http://www.altavista.digital.com) then move to the bottom of the page, you'll see a box marked: AltaVista Personal Search 97.

If you click on this box and download AltaVista Personal Search, you'll have a search engine of your own capable of indexing every word in most file types stored in all storage devices connected to your computer.

Once you have this set up and operating, you can use AVPS, which automatically operates in conjunction with your Web browser (Netscape Navigator or Microsoft Internet Explorer) to find almost any word or combination of words in any word processor, HTML, Acrobat, Help file, or e-mail (to name but the most common file types) stored anywhere on your computer. To keep track of all new work, file movements, etc., you configure the software to update the index at regular intervals (mine does this every day at lunchtime).

Search and metasearch engines

Depending on what you are looking for, Internet search engines (as opposed to the personal search engine just described) often turn up either too many links or too few. If you get too many, you have to use your preferred engine's Boolean operators and other options to refine your search. If you get too few, you need to consider using a "metasearch engine" to run your search simultaneously on several engines, then consolidate the results. Because each search engine indexes the Internet in its own way, different engines produce different search results. Some focus on documents in just one language or dealing with one subject. Metasearch engines enable you to search many at once in a single operation. WebFerret and WebSeeker are two such engines.

To quote from a recent e-mail from BlueSquirrel, "WebSeeker runs each query through more than 100 Internet search engines simultaneously, delivering the most comprehensive search report available. Quickly refine your results until you find exactly what you're looking for".

Example: A highly technical text on naval warfare contained the terms "lofargram" and "correlogram", neither of which meant anything to me. A metasearch with WebFerret turned up just two or three links, including one to a series of US Navy calls for proposals each featuring a brief summary of the "state of

SteveDyson@compuserve.com

the art". Although written to establish a baseline from which contractors would need to work, these summaries (posted just a few weeks earlier) proved, for the translator, every bit as good as an up-to-the-minute encylopædia article.

Internet magazine 'NetGuide' recently wrote of WebSeeker: "For power searchers who want the ability to find a data needle in the Web haystack, WebSeeker remains an excellent choice".

To sign up for a free Internet research course by e-mail, visit: http://www.bluesquirrel.com/courses/InternetResearch.html

To try out WebSeeker for Windows 95/NT free for a limited time, visit: http://www.bluesquirrel.com/seeker/

A **"wordbot"** - a robot assistant for looking up translations, definitions, synonyms, antonyms, references, etc. of words appearing in a document - is another type of metasearch engine about which we will probably hear a great deal in the near future. For further information, visit:http://www.cs.washington.edu/homes/kgolden/wordbot.html

Off-line browsers

The next thing you need to know is that there are now a number of "off-line browsers" on the market. This class of software allows you to download all or part of any number of Web sites and store them for subsequent . . . you guessed it . . . off-line browsing. The term refers to the ability to jump, or "surf", from page to page or site to site without a direct or on-line connection, in other words without having to wait for a good time of day and without being forced to twiddle your thumbs while you wait for a large image to download.

One example is WebWhacker from BlueSquirrel (at http://www.bluesquirrel.com). WebWhacker is very convenient for many applications. I use it during training courses to show people what is on the Web, or what their competitors have on the Web, without having to set up a modem connection during the course. WebWhacker has, however, one disadvantage: it uses a proprietary file format that AVPS cannot search.

Fortunately, the same company, BlueSquirrel, now offers a tool called Grab-a-Site. This slightly different type of off-line browser downloads Web sites and parts thereof and stores each page as a normal HTML file searchable by AVPS.

If my customer already has a site using validated terminology in one or more languages, I can download everything I need, have AVPS index it all automatically, then use AVPS as my personal terminology assistant.

If my customer does not have such a site, I can download the site of that company's chief competitor in the language of interest and use that as a source of authentic terminology, complete with contexts and source data, not to mention extended contexts including photos, diagrams . . . the works! This is a powerful concept.

If I already have a terminological database, I can use these tools to speed up database enrichment. If I am entering a new field, I may find it more cost-effective to use only these tools and abandon the traditional terminological database entirely. Lateral thinking. Revolution?

Compile your own on-line encyclopædia

If you work in any of the fields that are particularly well represented on the Internet, you can quickly compile the equivalent of a huge on-line encyclopædia then use AVPS to index it. Suppose you work into defence technologies. Visit, say:http://www.nttc.edu/solicitations/awards/dodsbir96/abs1af.html

You will find files of several hundred K each that can be downloaded to yield the equivalent of an up-to-the-minute encylopædia of US military technology (when imported into your word processor, these files will need some cleaning up, but a few hour's work will yield well over 1 Mbyte of ready-to-use documentation).

Surfing around this sort of area, you will then encounter specialist dictionaries such as that on photonics at http://www.laurin.com/DataCenter/Dictionary/CD/wrdlsts4.htm

These examples are drawn from my own special-interest areas, but I can assure you that you'll do just as well in almost any subject area and almost any language. And if you don't find what you are looking for tomorrow, you probably will in a few months' time.

A point of interest on lesser-used languages is that many of these language communities are waking up to the advantages of the Internet faster than some of the larger language communities. This means that people working in, say, Basque, Slovenian or Latvian you now have access to previously undreamt of volumes of material for reference purposes.

Download complete dictionaries and glossaries, too

Many translators-cum-Net-surfers will have discovered a host of dictionaries, glossaries and other terminological resources on the Net. Perhaps you have even thought of downloading one or two, converting them from HTML to WP format, and pasting the

multiple files together to generate something more usable. This is feasible with single-file glossaries - provided you have a good macro to clean up all the "junk" (spaces, carriage-returns, tags, etc.) that HTML leaves in WP files - but it is very tedious for large multi-file glossaries, even with the help of macros.

For macros, shareware and freeware, visit: http://www.onlineworld.com/software.html or ZDNet's software library at: http://www2.isl.net/infopet/software.htm For more information on search engines and Internet indexes, visit: http://www.magnet.gr/magnet/InternetDirectory.html

If you work in communications, you will be pleased to know that the complete US Federal Standard 1037C, or FCC Glossary of Telecommunications Terms, can be downloaded from: http://glossary.its.bldrdoc.gov/fs-1037@b

The Grab-a-Site/AVPS combination sidesteps all these problems at a stroke. Just download the entire site or section and let AVPS index it for you.

And do not be put off by the fact that some dictionaries and glossaries on the Web are unvalidated or of doubtful quality. If AVPS leads you to something you do not like or trust, you do not have to use it. On the other hand, it can be very gratifying to find something - sometimes almost anything - in just ten seconds, when your conventional resources turn up nothing at all.

Hardware requirements

Tools such as these demand real computing power. AVPS is not available for Mac. On a PC, a 120-MHz Pentium with 24 Mbytes of RAM is about the minimum. If you really like the technology, you'll probably find you need a large additional hard disc to store HTML and other files you download and save for indexing. I have a dedicated 4.6-Gigabyte hard disc plus a 1-Gb Iomega Jaz removable hard disc drive for added flexibility.

My whole approach to terminology research has changed since the day I fired up AVPS almost a year ago. I also use terminology on CD-Rom (Termium since I work from French into English), but I suspect I will never buy a paper dictionary again.

Looking further ahead

Suppose you have been asked to do a translation from language B into language C about an organisation that has a Web site in several languages including B and C. Suppose further that, although not perfect, the versions of this site in B and C represent a good starting point for customer-specific terminology and turns of phrase (after all, it is already on their site, isn't it?).

You can use Grab-a-Site to download both language versions. Soon, I suspect (Are you listening Trados?), you will be able to run a memory-based translation and alignment tool to produce aligned files of such Web pages. Next, you might run automatic term identification software to generate a terminological database, while keeping AVPS as a fallback for searches involving terms that the automatic tools get wrong.

Using this approach, you would have a swag of HTML reference files, the aligned text files corresponding to these HTML reference files, the AVPS index, and an automatically generated terminological database before you actually start translating. Not bad for a few minutes' work.

Of course, much of this automatic indexing, terminology identification and so on is vastly inferior to carefully prepared human-input terminological research . . . until you look at the bottom line, or lines, namely cost-effectiveness and speed.

Terminology management or terminology mining?

"Data mining" is a new area of information technology dealing with extracting nuggets of information from vast masses of data. There is plenty of information on this subject on the Internet itself. Just fire up your search engine and off you go. "Terminology mining" may come to mean the extraction of nuggets of terminological information from vast masses of documentation accessible via the Internet. As mentioned above, the problem we now face is how to find data needles in the Web haystack.

The more I think about it, the more it strikes me that terminological databases are essentially tools to help us make optimal use of limited volumes of documentation. The challenge we now face is of a different nature since documentation is now readily available in electronic form in enormous quantities (I imagine that lexicographers face a similar situation with their corpora).

I believe we are looking at a new ball game. Though the shape of the pitch remains unclear and the rules have yet to be finalised, a strategy is beginning to emerge. ■

Steve Dyson, an Australian living and working in France, is a technical translator and communicator, a consultant on translation methods and quality assurance, and an instructor in effective writing.