Using the Web as a Source of LSP Corpora in the Terminology Classroom

Sara Castagnoli

1 A short introduction to corpus-based terminology

Corpus-based terminology can be described as a working method which consists in exploring a domain-specific corpus in order to investigate terminological issues (Gamper and Stock 1998).

Even though its theoretical grounds are similar to those on which corpus-based lexicography is founded, it has taken longer for corpus-based terminology to become an established procedure; this is probably due to the different nature of the corpora involved, which are large and general – and therefore easily reusable – in the former case, domain-specific and smaller – i.e., difficult to re-use – in the latter. Terminologists and translators usually need to build a new corpus every time they embark on a new task, and the consequent reduced cost-effectiveness has often been adduced as the main argument against the construction of "disposable" (as defined in Varantola 2003) corpora, especially in relation to those domains in which most reference material used to be available only on paper, thus requiring manual checking or scanning. Today, however, the increased availability of texts in electronic format enables to speed up the process of collecting and processing corpora to an extent which was unthinkable until not so long ago.

2 And here comes the Web...

Being an unparalleled, virtually unlimited and ever expanding source of machine-readable texts, encompassing almost every language and knowledge domain (Fletcher 2004), the Web can play a leading role for the use of corpora to become common practice both in translation and terminology. While we do not believe that the Web can be considered a corpus – and certainly not a specialized corpus – in itself, since its contents are not assembled according to any specific criteria, we will argue that it may represent a good source for LSP (language for special purposes) corpora and terminology, for a variety of reasons.

First of all, as mentioned above, it is possible to find on the Internet texts on virtually any specialized subject, written in a variety of genres and communicative settings (expert-expert, expert-initiated/uninitiated[1] and, even if less interesting for the purposes of terminological research, initiated-initiated/uninitiated), which allows terminologists to choose among sources characterized by different levels of specialization, and to study variation and synonymy across different text types.

Secondly, while being a drawback for other types of linguistic research, the fact that new documents appear or are updated on the Web on a daily basis is an asset for terminologists: since terms are continually being invented and evolving, in relation to both their meaning and usage, it can be argued that a Web-based open corpus is more likely to contain up-to-date terms and state-of-the-art concepts than a static corpus.

Lastly, besides the fact that Web access is becoming increas-

[1] Pearson (1998) describes how specialized terms can occur in different communicative settings, arguing that terminological density varies according to the degree of specialization of the participants. "Initiates" are defined as people having some knowledge of a given specialized field, whereas "uninitiated (...) are not necessarily involved, either professionally or through their leisure interests, in a particular subject field".

ingly easier and inexpensive, and that it is constantly available, most translators are already familiar with it and use it in their everyday work,[2] which makes it reasonable to suggest that tools for corpus creation and analysis based on the Web would be easily integrated into their workstations.

3 Teaching corpus-based terminology using the Web

Given that terminological research constitutes a substantial part of the translator's work, and that corpora – both general and specialized – have been suggested to be effective tools in enhancing the quality of translations (Gavioli and Zanettin 1997), the principles and methodology involved in creating corpora and extracting terminology from them have become part of the teaching curriculum at the School for Interpreters and Translators of the University of Bologna, Forlì, Italy.

This paper reports on a classroom experience carried out in Spring 2005 with a group of ten trainee translators taking an optional 48-hour course in Terminology and LSP. The main objective of the course was to teach students why and how to use corpora in two stages of terminology work, namely term extraction and terminography (i.e., the recording and presentation of terminological data, most often by means of databases). The course was mainly organized along these two axes, developed by two different teach-

[2] A questionnaire circulated to professional translators during the period April-June 2005 in the framework of the European project MeLLANGE (*Multilingual e-Learning in LANGuage Engineering*, `http://mellange.eila.jussieu.fr/`) revealed that – over 623 respondents, located mainly in the UK, but also in France, Italy and Germany – 93.4% of translators use Google to research terminology, with more or less refined strategies, 43.3% regularly visit websites belonging to specific companies, 29.6% regularly visit websites acting as domain portals and 21% regularly visit other kinds of websites.

ers;[3] the module on corpus creation – preceded by a few lessons on the Unix operating system – was integrated with introductory notes on corpus annotation, XML, POS tagging and collocation extraction, whereas the final lessons were dedicated to illustrate the use of termbases within some CAT tools. Our aim was to consider terminological work both as an autonomous discipline and as a component of the translation process. Corpora were thus created and analyzed in two different teaching situations, i.e., during the terminology course proper and for the end-of-course project.

Since corpus creation was not the main subject of the course, and since designing and constructing "well-made" corpora would have required much more time and effort than available in the classroom, students were asked to work on corpora assembled automatically using the BootCaT toolkit, a suite of Perl programs designed to bootstrap specialized corpora from the Web (Baroni and Bernardini 2004). Other reasons behind this choice included the desire to introduce students to a tool which they might find helpful for their future activity as translators, as well as to provide them with new IT competences. Advantages and disadvantages of automatic corpus compilation were then discussed with students on the basis of their analysis of the usefulness of their corpora, which proved to be a very instructive activity. Some of the conclusions that were reached are reported in the following sections.

3.1 During the course: Practicing term extraction

After introducing students to the basic principles of terminology (languages for special purposes vs. general language, terms vs. words, terms vs. concepts, etc.), and having illustrated the advantages of corpora over traditional dictionaries, students were asked to choose domains they were familiar or had already worked

[3] Alessandra Matteucci was in charge of the part of the course about terminography.

with in other translation courses, and to provide a list of terms they presumed to be typical of such domains, to be used as seeds for the Web mining procedure. The aim of the exercise was to collect a corpus on which to practice term extraction through a variety of techniques, such as the production of word or cluster (bi-grams, tri-grams, etc.) lists, statistical measures (frequency, mutual information and log-likelihood), and morphosyntactic analysis (based on POS tagging, i.e., retrieving all occurrences of given combinations of POS tags which are hypothesized to be typical patterns for terms, such as ADJ+NOUN or NOUN+NOUN in English). The reason we asked students to work on domains they were already acquainted with is that we wanted them to be able to judge the results of the above methods, in order to start a discussion on which term extraction techniques they considered to be more profitable.

The three groups decided to work on medicine (nervous system disorders), law (Italian company law) and technology (cell phones), the first two subjects having been dealt with during a translation course and the third being chosen on the spot, as a domain known to all the members of the group. Table 1 shows the terms chosen as domain key-words for the automatic downloading of Web pages.

Students were then allowed to decide the size and number of tuples to be formed as well as the maximum number of URLs they wanted to retrieve for each tuple, while keeping numbers low enough for the retrieval process not to be too long. Table 2 shows that, although students made more or less the same decisions, the final result – i.e., the size and/or quality of their corpora – differed remarkably. In the following paragraphs we will try to identify possible reasons for this phenomenon by analyzing some data taken from the medical corpus and the cell phone corpus.[4]

[4]Interim data about the Company law corpus are not available because students were not required to document and save data about each single stage

medicine	**company law**	**cell phones**
neurotrasmissione	diritto societario	cellulare
sistema nervoso centrale	decimi	scheda SIM
noradrenalina	pubblicità	PIN
dopamina	azioni	PUK
catecolamina	creditore	GSM
sistema nervoso autonomo	riforma	WAP
sostanza nera	registro delle imprese	UMTS
neurone	spa	GPRS
cellula nervosa	amministratore	SMS
sistema dopaminergico	srl	T9
	pignoramento	MMS
	conferimento	videofonino
	regolamento	bluetooth
	partecipazioni	caricabatteria
	società unipersonale	auricolare
	società pluripersonale	batteria al litio
	consiglio di amministrazione	infrarossi
		videochiamata
		vivavoce
		scrittura intuitiva
		schermo a cristalli liquidi

Table 1. Seeds for the Web mining procedure

As shown in the first two rows of table 2, different choices were made in relation to the number of tuples. Having chosen a limited number of highly specialized terms, the group working on the medical domain decided to form twenty 2-term tuples, in order to avoid specifying search criteria so narrow that they would probably have resulted in a very small corpus. Instances of such tuples include ["sistema nervoso autonomo" "sistema nervoso centrale"] (*autonomic nervous system, central nervous system*), "sistema nervoso centrale" "sostanza nera"] (*central nervous system, substantia nigra*), ["cellula nervosa" "sostanza nera"] (*nerve cell, substantia nigra*), [noradrenalina neurone] (*noradrenaline neuron*). On the other hand, the cell phone group decided to create fifteen 3-term tuples, such as [cellulare videochiamata GPRS] (*mobile phone,*

of the corpus creation process.

Domain	Medicine	Company law	Cell phones
tuple size	2	n.a.	3
tuples	20	n.a.	15
URLs	183	n.a.	138
URLs/tuples	9.15	n.a.	9.2
lines	37,073	34,821	40,749
words	281,015	281,736	160,298
characters	2,010,356	1,931,760	1,120,754
words/URLs	1,535.60	n.a.	1,161.58
ch.s/URLs	10,985.55	n.a.	8,121.41

Table 2. Corpora statistics

video call, GPRS), [SMS videochiamata UMTS] (*text message, video call, UMTS*), [caricabatteria SMS GSM] (*battery charger, text message, GSM*), [WAP bluetooth "scrittura intuitiva"] (*WAP, bluetooth, predictive text*). Both groups decided to retrieve a maximum of 10 URLs for each tuple, with similar URLs/tuples ratios.

Table 2 shows that there is a remarkable difference in size between the medical corpus and the cell phone corpus, which can be only partly explained by the lower number of tuples searched.

Analysis of average words/URLs and characters/URLs ratios actually allow us to state that webpages related to cell phones are much shorter (by 347 words and 1,864 characters, respectively) than those belonging to the medical domain. Inspection of retrieved URLs and further analysis of the cell phone corpus through word lists (e.g., table 3) and concordances suggest that this is due to the kind of webpages that were downloaded, i.e., pages belonging predominantly to commercial sites or to Web portals offering different kinds of cell phone services (downloading of ringtones and wallpapers, comparison of technical specifications, etc.). Normally such websites are not rich in descriptive or informative pages, but rather conceived with a persuasive purpose and therefore stylistically characterized by eye-catching images and lists; this idea

is corroborated by the evident disparity in the number of lines between the two corpora (see table 2).

On the other hand, observation of the most frequent nouns in the two other corpora suggests that these are largely characterized by highly specialized and formal texts.

medicine		company law		cell phones	
995	sistema	1,864	articolo	871	suonerie
661	cellule	1,770	comma	606	Foto
483	parte	1,689	società	474	telefono
471	cervello	1,179	Art	442	colori
374	cellula	1,147	soci	375	Prezzo
372	neuroni	830	capitale	259	Siemens
371	attività	816	decreto	256	dati
369	membrana	774	azioni	254	Band
354	effetti	753	numero	230	acquisto
351	malattia	746	caso	226	credito
333	azione	705	socio	218	tecnologia
327	corpo	680	amministratori	218	band
307	neurone	668	atto	216	Provenienza
307	farmaci	652	societa'	213	Spese
305	recettori	637	diritto	211	servizi

Table 3. 15 most frequent nouns in the three corpora

One of the first conclusions that can be reached is, therefore, that the automatic creation of corpora from the Web for terminological research is more effective and productive for domains which are highly specialized, whereas it is difficult to retrieve specialized texts concerning more popular domains (e.g., cell phones), in relation to which there is an overflow of information on the Web. Specialized terms belonging to such domains (e.g., "*LCD*", "*lithium battery*") have become so common in everyday language (it might be argued that they have gone through a process of "determinologization", i.e., they have lost their specificity to become part of general language), that it seems impossible to use them to auto-

matically identify specialized text to be used as reference material for a terminological task.[5]

Students were also encouraged to think about other possible problems concerning corpora which are automatically assembled from the Web. The quality and reliability of the texts (and of the terms employed in them) cannot be taken for granted; questions of register and style should be taken into account, as well as their relevance to the task.

However, the quality of a corpus ultimately depends on the quality of information the translator/terminologist is able to extract from it (Varantola 2003). Besides being used for term extraction, DIY specialized corpora can be rich sources of other information to be recorded in a terminological sheet, such as definitions, contexts, semantic relations etc.

From this point of view, all the corpora collected by the different groups turned out to be relevant to the task. Students were encouraged to look for definitions, contexts, synonyms and variants of terms with the aid of a concordancer.[6] They were, for instance, advised to search for defining expressions and linguistic signals such as "*is a kind of*", "*consists in*", "*known as*", "*also called*" etc. Some explicit definitions were present in each corpus, but it was interesting to notice that – where the need arose to infer definitions from the text – the less formal texts often proved to be more useful than the more specialized ones, possibly because of the need to explicate concepts for the less expert audience involved.

Discussion was therefore triggered about the pros and cons of

[5]Because of the time constraints of doing such activity in the classroom, we did not reiterate the bootstrapping procedure using unigrams and multi-word terms extracted from the first downloaded corpus, as suggested by Baroni and Bernardini (2004). This might have helped to retrieve more specialized texts, but it might equally have degraded the output.

[6]In this case, the IMS Corpus WorkBench (Christ 1994) was used to encode and index the students' corpora, and the associated Corpus Query Processor (CQP) was used for concordancing.

automatically building corpora from the Web: despite the drawbacks pointed out above (mainly, the lack of control over text sources, but also the incompleteness of the material – i.e., it was not possible to find definitions and useful contexts for all terms), most students stated that they were favorably impressed by the possibility of collecting such large amounts of reference materials they could use for any translation task, with such little effort and in such a short time. The group working on cell phones also realized that manually creating a corpus from the Web for their domain (i.e., "hunting" via Web queries through search engines; cf. Fletcher 2004) would be equally difficult and more time-consuming, as there is too much information online whose relevance needs to be evaluated before finding the "right" texts for a corpus like this one.

3.2 Applying experience to the end-of-course project

The final test for the course consisted in a composite project, based on the English-to-Italian translation of a text on the domain of asparagus cultivation; the source text was chosen by the teachers mainly on the basis of its degree of specialization, i.e., rich of domain-specific terms but not too technical. Students were given the source text to be translated, and were asked to collect reference corpora in both source and target language as well as to produce a given number of terminological sheets with information extracted from the corpora. Following our classroom discussions, they were let free to decide whether to build the corpus automatically or manually, and they were asked to provide feedback on the reasons underlying their choice.

As expected, all the students who have taken the exam at the time of writing decided to try and work on automatically assembled corpora. We will first analyze the procedure followed to create corpora for the source language, then moving on to target language corpora and corpus use.

Concerning the choice of the seeds on which to base the boot-

strapping process for the source language corpus, most of the students identified specialized terms within the source text and added a few more general terms, such as "*cultivation*", which were not present in the text but which were perceived to be relevant. Most of them also demonstrated an understanding of the Web mining procedure by increasing – compared to what was done in class – the number of tuples to be searched as well as the number of webpages to be retrieved for each tuple, in order to retrieve larger corpora.

As far as the target language corpus was concerned, some students reported that they had chosen the seeds by guessing – and verifying with dictionaries – potential equivalents of source terms. Two students, on the other hand, decided to use a search engine to identify some relevant and (presumably) authoritative webpages in the target language and to extract candidate seeds for the bootstrapping procedure from such pages. In one case, this proved to be a good intuition, which allowed the student to reduce the risk of "circularity" (Varantola 2003), i.e., the risk of choosing wrong (translations of) keywords and to build corpora on such unsuitable terms. In the other case, however, the suitability of the extracted terms was not evaluated carefully, and the student (a non-native speaker of Italian) ended up choosing an extremely rare word, i.e., *brattea* ("bract"), which probably spoilt the results of some automatic searches. It is important to always keep in mind the need for careful evaluation of seeds and the limitations of automatic corpus creation from the Web.

After examining their target language corpora in view of the compilation of the termbase and of the translation, however, most of the students found that their material was not sufficient to retrieve all the information needed, i.e., suitable definitions and domain-relevant contexts, and some of them decided to build another corpus semi-automatically, with the aid of a program (Text-

Stat)[7] which allows users to assemble corpora by specifying the URLs of the webpages to be downloaded, which might be either previously known or discovered through a search engine. According to their reports, this process of focusing on and downloading predetermined reliable websites, which we might call "grazing" (Fletcher 2004), proved to be very effective: not only could they evaluate the relevance and quality of texts before including them in the corpus, they could also build corpora rich in useful information while keeping them to an easily manageable size. Moreover, as many authors have already pointed out (see, e.g., Zanettin 2002, Maia 2002), the fact of having to find and read candidate reference texts prior to the translation task proper helped students to familiarize themselves with the specialized subject, thus enhancing their understanding of the domain and, possibly, of the source text; some students actually reported that visiting several websites allowed them to find pictures and images which helped them to better understand the structure of the asparagus plant.

4 Concluding remarks

The course in Terminology and LSP was designed, among other objectives, to sensitize students to the great possibilities offered by a more conscious and profitable use of a tool – i.e., the Web – with which they are already acquainted, by showing them how easy it can be nowadays to build corpora which could be used, along with traditional online dictionaries or glossaries, as performance-enhancing tools within some specific translation or terminological task.

While preparing their end-of-course projects, students realized that the advantages of automatically assembling corpora from the Web were counterbalanced by the need to carefully assess the qual-

[7] Freely downloadable from `http://www.niederlandistik.fu-berlin.de/textstat/software-en.html`

ity of the results, but that it was simple for them to use the Web itself to adjust their corpora by adding other relevant material with more or less automated methods. Nonetheless, it is only after having acquired some competence on a specific domain that it is possible to see the need for and to carry out such "corrections".

Our conclusion is therefore that the degree of usefulness of LSP corpora automatically assembled from the Web depends first and foremost on the user's familiarity with the specialized domain in question. Studying the terminology belonging to a domain which is totally – or mostly – unknown to the user through corpora created automatically can be quite risky, as the user would not have the necessary knowledge to judge the appropriateness of the output. As far as terminology is concerned, however, such output would mainly depend on the content of webpages, and less on the quality of the Web mining tool; in this respect, it might be argued that even search engine results can be difficult to interpret for the non-expert eye, the Web being rich in unreliable, non-authoritative materials. On the other hand, when the user has – or has acquired – sufficient domain-specific knowledge to be able to critically evaluate texts/terms retrieved with no – or limited – human supervision, the possibility to collect large quantities of data in such a short time cannot but prove of great value for terminologists and translators alike.

References

Baroni, M. and Bernardini, S. (2004). BootCaT: Bootstrapping corpora and terms from the Web. In *Proceedings of LREC 2004*, 1313-1316.

Christ, O. (1994). A modular and flexible architecture for an integrated corpus query system. *Proceedings of COMPLEX'94: 3rd Conference on Computational Lexicography and Text Research.*

Fletcher, W. (2004). Facilitating the compilation and dissemination of ad-hoc Web corpora. In Aston, G., Bernardini, S. and Stewart, D. (eds.) *Corpora and language learners*, Amsterdam: Benjamins, 273-300.

Gamper, J. and Stock, O. (1998). Corpus-based terminology. *Terminology* 5(2), 147-159.

Gavioli, L. and Zanettin, F. (1997). Comparable corpora and translation: A pedagogic perspective. *First international conference on Corpus Use and Learning to Translate.*

Maia, B. (2002). Corpora for terminology extraction: The differing perspectives and objectives of researchers, teachers and language service providers. *Language Resources for Translation Work and Research, LREC 2002 Workshop Proceedings*, 25-28.

Pearson, J. (1998). *Terms in context*, Amsterdam: Benjamins.

Pearson, J. (2000). Teaching terminology using electronic resources. In Botley, S. , McEnery A. and Wilson, A. *Multilingual corpora in teaching and research*, Amsterdam: Rodopi, 92-105.

Sager, J. (2001). Terminology compilation: Consequences and aspects of automation. In Wright, S. and Budin, G. (eds.) *Handbook of terminology management: Volume 2, application-oriented terminology management*, Amsterdam: Benjamins, 761-771.

Varantola, K. (2003). Translators and disposable corpora. In Zanettin, F., Bernardini, S., and Stewart, D. (eds.) *Corpora in translator education*, Manchester: St. Jerome, 55-70.

Zanettin, F. (2002). DIY Corpora: The WWW and the translator. In Maia, B., Haller, J. and Ulrych, M. (eds.) *Training the language services provider for the new millennium*, Porto: Universidade do Porto, 239-248.