Specialized Corpora from the Web and Term Extraction for Simultaneous Interpreters

Claudio Fantinuoli

1 Introduction

There is no doubt that the Web is a mine of language data of unprecedented richness and ease of access (Kilgarriff and Grefenstette 2003). As more people use the Web for more tasks, it provides an increasingly representative machine-readable sample of interests and activity in the world (Henzinger and Lawrence 2004). Despite some drawbacks, the Web is an immense source of disposable corpora (Varantola 2003) that can be used for specific purposes such as translation or interpretation tasks. Many language professionals use the Web as a source of information to study the language and process the specific terminology; in some cases, they also build a corpus to be looked up with a concordancer, but this is done through manual queries and downloading. Obviously this is an extremely time-consuming task. The time investment is perceived as particularly unjustified if the final result is meant to be a single-use corpus. If the aim is that of constructing a corpus big enough to allow terminology extraction, then an automated process to bootstrap corpora from the Web is the best solution to speed up the process.

When preparing themselves for a highly specialized conference, interpreters must acquire linguistic and extra-linguistic information in order to perform a good interpretation task (Gile 1995). As

Kalina (1998) points out, the elaboration of the preparatory documentation can help interpreters to advance the workload and to improve the working conditions in the booth. This preparation is nowadays very traditional, i.e., it is done manually and it includes: collection of parallel texts, reading (acquisition of extra-linguistic information) and elaboration and memorization of glossaries containing the specific terminology (language learning). This task appears to be time consuming and not efficient enough if we take into account the time factor, i.e., the time conditions under which a professional interpreter is used to work. To facilitate this process, we propose an approach to "Corpus Driven Interpreters Preparation". The process of "knowledge acquisition/language learning" needed by interpreters in order to prepare themselves for a conference can be optimized by making it "terminology-driven", or "bottom-up": from the terminology to the conceptual structure of a particular domain. Corpora can be the source of a potentially endless "serendipity process" (Johns 1988), as one word or phrase leads to another, depending on the user's intuition and individual proficiency, interests or needs. In this approach, the interpreter will "explore" the corpus starting from a list of specialized terms. In this way s/he will learn the terms, their meaning and usage in context,granting that amount of flexibility and active interaction typical of the interpreter's preparation. A list of specialized terms, the starting point of this kind of preparation, can be obtained by automatically extracting the specific terminology from a corpus. To speed up the process, the corpora can be automatically created using tools such as BootCaT (see section 2 below) and the Web as a source of specialized texts. The interpreter will then look up the corpus using a concordancer.

In this experiment we compare two procedure of terminological extraction using two different specialized corpora: the first is a manual corpus built by a terminologist in order to manually extract the specialized terms of the domain (childhood acute lym-

phoblastic leukemia), the second is a corpus automatically generated by the BootCaT tool using the Web as a corpus and a series of starting seeds that are expected to be representative of the domain under investigation. This list of seeds closely resembles what many interpreters have at their disposal as preparatory documentation in real life, i.e., the keywords of the abstracts given to the interpreters from the conference organizers. Our first aim is to evaluate BootCaT and in general the use of the Web as a corpus for specialized purposes. In our study we consider professional interpreters to be the target users of this tool. Interpreters represent a special user typology and the terminology needed varies according to the needs of the interpreter. Thus we will propose three different criteria for evaluating the tool. The experiment is conducted with Italian, German and English corpora.

2 The BootCaT procedure

In the last few years several experiments have used the BootCaT toolkit to bootstrap corpora from the Web in order to extract linguistic information such as terms or collocations. See, for example, Baroni and Bernardini (2004), Baroni and Ueyama (2004) and Sharoff (this volume). The multi-word term extraction method we implement has some similarities with the one proposed by Baroni and Bernardini (2004).

The basic BootCaT procedure is very simple.[1] Basically two main tasks are accomplished by the tool: 1) building a corpus of specialized texts from the Web; 2) extracting the relevant terminology from the downloaded corpus.

BootCaT compares frequencies in specialized and reference corpora to look for terms typical of the former. This is a fairly common idea in terminology extraction and corpus comparison

[1] For a more detailed description of the procedure see Baroni and Bernardini (2004).

work. See, for example, Rayson and Garside (2000) and Kilgarriff (2001). The tool uses an iterative algorithm to bootstrap corpora from the Web and extract unigram terms. It then proceeds to extract multi-word terms on the basis of the downloaded corpus and of the unigram term list extracted in the previous phase. The bootstrapping process, using the Google search engine,[2] starts with a small list of seeds that are expected to be representative of the domain. The seed terms are randomly combined and each combination is used as a Google query string. The top n pages (HTML, PDF and doc files) returned for each query are retrieved and formatted as text. The unigram terms are extracted from the corpus of retrieved pages by comparing the frequency of occurrence of each word in this set with its frequency of occurrence in a reference corpus. Frequencies are compared using the Mutual Information (Church and Hanks 1990) and the Log Likelihood (Dunning 1994) association measures.

To make it to the final candidate lists of simple and multi word terms, the extracted terms must fulfill two criteria: 1) they must correspond to a specific morphosyntactic pattern (section 7); 2) they must contain at least one of the extracted unigrams.

3 Empirical assessment

Evaluating the performance and the differences between the terminological extractions from an automatic downloaded corpus and a manual corpus is not an easy task. In this case, the situation is further complicated because we try to take into account a well defined potential user of the extracted data, the professional interpreter. With this in mind, we base our evaluation on: the quality of terms based on human assessment – i.e., well- or ill-formed – and on their degree of specialization; the level of specialization of

[2] `http://www.google.com/apis`

	Words	Bytes
Italian	108,016	763,455
German	88,895	738,695
English	286,346	2,037,176

Table 1. Manually collected specialized corpora

the extracted terms in light of the needs of interpreters; the comparison of the extracted terms with a reference term list manually created by a professional terminologist.

The reference term lists (RTL) were created from manually constructed corpora (see table 3) collected by a terminologist in a multilingual project on "childhood acute lymphoblastic leukemia" (Bordoni 2001). The Italian RTL contains 136 terms; the German one 158 terms; the English one 155. The collection of the texts, mainly from the Internet (PDF, doc and HTML), but also from printed papers, and the extraction procedure were all done manually, i.e., searching for suitable websites, evaluating the quality of the texts and then extracting from them the specialized terminology.

In order to make the comparison of the manual and the automatic terminology extraction methods more fair, we excluded from the manual lists the terms that were extracted from printed texts by the terminologist and were not found in her corpus. Notice, however, that we base the evaluation on terms that were extracted by the terminologist from her manually compiled corpus. Thus, when we compare the quality of term extraction between the manual and automatically constructed corpus below, we are actually giving an advantage to the manual corpus, given that we use a list of terms that were extracted from it as our golden standard.

4 Evaluation of the candidate terms

4.1 Five-level taxonomy

The candidate terms were divided into five groups according to their level of specialization and well-formedness:

1. specialized terms contained in the reference term list;
2. specialized terms not contained in the reference term list;
3. general medical terms;
4. "general" terms;
5. incomplete or ill-formed terms.

Category 1 contains terms that were manually extracted by the terminologist (and therefore are contained in the RTL), e.g.: *epatosplenomegalia*, *intrathekale Chemotherapie* and *bone marrow aspiration*. In category 2 we find highly specialized terms that were not detected by the professional terminologist, e.g.: *leucemia mieloblastica acuta*, *myeloische Leukämie* and *allogenic peripheral blut*. Category 3 contains non-specialized terms that are commonly used in the field of medicine, e.g.: *apparato urinario*, *antibiotische Therapie* and *bone*. In category 4 we find general terms that are not specific to the medical field, e.g.: *fattore*, *statistische Auswertung* and *Journal*. Category 5 contains ill-formed, incomplete expressions and fragments, e.g.: *sempre alla stessa*, *Kind selten* and *recurrent childhood*.

All extracted terms were evaluated according to this grid. Of course there is always an amount of arbitrariness in this kind of evaluation, even though we aimed for consistency: make the same judgment for the same term independently of the extraction method.

4.2 The target user: The interpreter

As Kurz (1996) points out, interpreters may need both specialized and less specialized terms in order to prepare themselves for a conference. Depending on whether the interpreter is interpreting into or out of the foreign language or whether s/he is used to interpreting in that specific domain or not, we can have two main scenarios:

a the interpreter needs only the highly specialized terms regarding the subject field (in our case leukemia); or

b the interpreter needs the specialized terms plus the more general medical terms.

4.3 Second-level taxonomy

To account for the needs of interpreters (section 4.2), the 5 categories of the original taxonomy (section 4.1) were merged in what we call T2a e T2b. To evaluate the precision of the system to extract only the highly specialized terminology of the domain, we use the taxonomy T2a:

$$T2a = \{A1,B1\} \text{ where } A1 = \{1,2\} \text{ and } B1 = \{3,4,5\}$$

A1 are the acceptable highly specialized terms, i.e., the sum of the terms belonging to category 1 (extracted terms that were also manually detected) and to category 2 (highly specialized terms that were not manually detected).

We evaluate the quality of the system in extracting terms from the medical domain – highly specialized terms and otherwise – with the taxonomy T2b:

$$T2b = \{A2,B2\} \text{ where } A2 = \{1,2,3\} \text{ and } B2 = \{4,5\}$$

A2 are the acceptable medical terms, i.e., the sum of the terms belonging to group 1 (extracted terms that were also manually detected), to group 2 (terms specific to the domain that were not manually detected) and to group 3 (generic medical terms).

4.4 Recall

We evaluate terminological extraction from the two different corpora in terms of precision and recall, using the two taxonomies just described. In our study we define *Recall* (for category 1) as follows:

$$Recall = \frac{AUTOTERMS}{MANTERMS} \times 100$$

$AUTOTERMS$ is the number of category 1 terms that were automatically extracted, and $MANTERMS$ the number of terms manually identified by the terminologist.

We consider the manually extracted terms as being the only terms contained in the corpora and compute the recall value on the number of terms retrieved manually. The recall gives us an idea of the amount of terms contained in the reference terminology list (the one compiled by the terminologist) that are retrieved by the semi-automatic system. In this way we compare the results of the manual and the automatic term extraction procedures (given that recall is based on terms that were extracted from the manual corpus, we would expect, in principle, higher recall when automated extraction is performed on the manual corpus).

5 Corpus construction

We started the bootstrapping process with a series of 9 seeds for each language (table 2). As far as interpreters are concerned, we can suppose that the initial terms can be obtained from the

Italian	German	English
leucemia	Leukämie	Leukemia
“midollo osseo”	Knochenmark	“bone marrow”
LLA	ALL	ALL
chemoterapia	Chemoterapie	chemotherapy
trapianto	Transplantation	transplantation
“leucemia acuta linfoblastica”	“akute lymphatische Leukämie”	“acute lymphoblastic leukemia”
linfocita	Lymphozyt	Lymphocyte
“puntura lombare”	Liquorpunktion	“lumbar puncture”
leucociti	Leukozyten	Leukocytes

Table 2. Initial seeds used to create the corpora

	Italian	German	English
URLs	308	128	304
Bytes	12,519,130	7,555,510	3,086,908

Table 3. Number of URLs and size of the corpora

conference abstracts delivered to the interpreter. Note that in order to grant similar initial conditions for all languages, we used the translation of the same seeds in every extraction.

As BootCaT allows the user to control several important parameters, such as the number of queries issued for each iteration, the number of seeds used in a single query, the number of pages to be retrieved, etc., we downloaded files using the following parameters: 3 seeds for each query; 20 tuples, each used for a query; a maximum of 20 pages to be downloaded for each query. The number of URLs, without counting duplicates, obtained with this method is shown in table 3. Then we proceeded by automatically downloading and converting the detected URLs into text files; the size of the corresponding corpora is also reported in table 3.

The size of the downloaded corpora varies considerably among the languages and this even though the initial conditions were

	Italian	German	English
Reference corpus	3,288,496	3,109,525	3,388,390
Specialized corpus (Web)	1,512,766	813,817	422,037
Specialized corpus (manual)	105,890	85,074	274,215

Table 4. Size of the corpora in tokens

virtually the same for all extractions (seeds and BootCaT parameters). Interesting enough, the language with the least amount of text is English, the language of international scientific communication.

6 Extraction of unigram terms

We first tokenized the specialized and the reference corpora with command-line scripts (table 4).

The reference corpora are part of the EuroParl corpus, a large collection of texts from the European Union.[3] They cover a large variety of topics and this makes them suitable to be used as a benchmark for corpus comparison. Using the UCS tools,[4] we compared frequencies between the reference and the specialized corpora. We computed both the Mutual Information (MI) and the Log-Likelihood (LL) association measures in order to account for terms with low and high frequencies (Evert and Krenn 2001). In our experiment MI and LL are not used to compute the proximity factor of two words in a given text (the probability that a word occurs with another word – collocation), but to compare the occurrences of a given word in two different corpora, as illustrated by Sharoff (this volume). We extracted the final unigram term lists considering only the first 200 words obtained with every

[3] `http://people.csail.mit.edu/people/koehn/publications/europarl/`

[4] `http://www.collocations.de`

	Italian	German	English
corpus (Web)	390	355	298
corpus (manual)	409	399	468

Table 5. Size of unigram lists

Italian	German	English
N+ADJ+ADJ	ADJ+ADJ+N	ADJ+ADJ+N
N+ADJ	ADJ+N	ADJ+N
N	N	N
N+N		N+N
N+PRE+N		N+N+N

Table 6. Morphosyntactic patterns

association measure. In addition we extracted acronyms simply by searching for capital letter words longer than 1 and shorter than 4 characters. We merged the three lists obtaining the numbers of candidate unigram terms reported in table 5 (some examples: for Italian, *anemia, induzione, EFS, leucociti, citogenetica*; for German, *B-ALL, Blasten, Blutbild, Chemoterapie, Erbrechen*; for English, *cyclophosphamide, cyclosporine, cytarabine, leukemia, MRD*).

7 Extraction of multi-word terms

The unigram lists and the corpora were used to extract multi-word terms. We first tagged the specialized corpora using the TreeTagger[5] and then built bigrams and trigrams.

We extracted multi-words terms that satisfied the POS patterns shown in table 6 and that contained at least one unigram from the lists previously extracted (section 6).

[5] `http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTree/Tagger.html`

	Italian	German	English
corpus (Web)	353	333	317
corpus (manual)	290	324	335

Table 7. Candidate terms

Taxonomy	Extracted terms	%	Recall
1	13	3.68	9.56%
2	85	24.08	
3	201	56.94	
4	30	8.5	
5	24	6.8	
Tot terms	353	100	

Table 8. Results (Web): Italian

The lists of single and multi word terms were then merged (table 7).

8 Evaluation

8.1 General

We assigned a value to every candidate term according to our taxonomy (section 4.1). As pointed out above, we focused primarily on consistency. We manually assigned each term to a category of our grid (the terms were evaluated in random order and without knowing their source). For the five categories we obtained the results reported in tables from 8 to 13 (while reading these tables, please keep in mind our categorization from section 4.1 – 1: Specialized terms contained in the reference term list; 2: Specialized terms not contained in the reference term list; 3: General medical terms; 4: General terms; 5: incomplete or ill-formed terms).

As far as the 5 categories are concerned, we can easily see that there are similarities among the languages. The obvious difference

Taxonomy	Extracted terms	%	Recall
1	50	15.01	32.64%
2	145	43.54	
3	78	23.42	
4	35	10.51	
5	25	7.5	
Tot terms	333	99.98	

Table 9. Results (Web): German

Taxonomy	Extracted terms	%	Recall
1	48	15.14	30.97%
2	139	43.85	
3	87	27.44	
4	30	9.46	
5	13	4.1	
Tot terms	317	99.99	

Table 10. Results (Web): English

Taxonomy	Extracted terms	%	Recall
1	59	20.34	43.38%
2	91	31.38	
3	77	26.55	
4	57	19.65	
5	6	2.07	
Tot terms	290	99.99	

Table 11. Results (manual corpus): Italian

Taxonomy	Extracted terms	%	Recall
1	53	16.36	33.54%
2	139	42.9	
3	33	10.18	
4	57	19.65	
5	21	6.48	
Tot terms	324	99.99	

Table 12. Results (manual corpus): German

Taxonomy	Extracted terms	%	Recall
1	38	11.34	24.51%
2	152	45.37	
3	91	27.16	
4	38	11.34	
5	16	4.78	
Tot terms	335	99.99	

Table 13. Results (manual corpus): English

concerns the value obtained for the Italian automatically downloaded corpus. If we pay attention to the distribution of terms within Italian, we see that most terms are in the third category, i.e., general medical terms. This means that the downloaded Italian corpus is less specialized then the German and the English ones, even though the initial seeds were the same. Again, this is an interesting starting point to further investigate differences in Web document availability in different languages.

If we consider the recall values, we see that the automatic extraction of highly specialized terms from the downloaded corpora leaves out many terms that were considered important by the terminologist. While this may cast some shadows upon the effectiveness of the automatic method of terminology extraction used (from the terminologist's prospective), it does highlight the fact that both corpora – manual and automatic – are of comparable quality (from the extraction's perspective). This is especially interesting since the manual set used for recall assessment was extracted from the manual corpora, thus we know that the manual set terms are present in the latter, that, in principle, should thus provide higher recall than the automatically constructed corpora.

	Extraction from Web corpus	Extraction from manual corpus
	Italian	
A1	27.76	51.03
A2	84.70	78.27
	German	
A1	58.55	59.26
A2	81.97	83.33
	English	
A1	58.99	56.71
A2	86.43	83.87

Table 14. Comparison between A1 and A2 (in percentage)

8.2 Interpreter-targeted evaluation

As we pointed out before, the ultimate criteria to evaluate the tool are the needs of professional interpreters. This is why we evaluate it according to the taxonomies T2a and T2b, i.e., according to the capacity to extract highly specialized terms (A1) or specialized terms plus the more general medical terms (A2).

The results (table 14) are similar across the different languages, besides the expected exception of A1 with the Web corpus in Italian. For the category A1 – specialized medical terms – the best result was obtained with the manual corpus for the German language (59.26%). But the results obtained with the Web corpus are very close to this value: German 58.55% and English 58.99%, the latter being the best result obtained with this language. For the category A2 – specialized and generic medical terms – the best result was obtained with a Web-derived corpus (English, 86.43%).

Again, these results have to be interpreted by keeping in mind that a portion of the terms in A1 and A2 (namely the terms in the manual set) have been extracted from the manual corpus, which is, thus, advantaged in terms of the evaluation procedure.

9 Conclusion

We showed that term extraction from manually compiled and automated Web-derived corpora leads, in general, to comparable results (further research is needed on the reasons for poor performance of the Web-based procedure in Italian).

Given how time-consuming it is to build a corpus by hand, automated Web-based corpus construction is a very promising way to reach good results with limited efforts.

Using the BootCaT procedure, interpreters preparing for a conference can obtain a list of relevant terms and texts within minutes, even when targeted preparatory materials have not been made available by the conference organizers (as is often the case in professional settings). While the current version of the BootCaT toolkit requires computational skills beyond what is reasonable to expect from interpreters, the graphical interface currently being tested (Baroni et al. 2006) has the potential to make BootCaT a very popular tool for our target community.

References

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

Baroni, M., Kilgarriff, A., Pomikálek, J., Rychlý, P. (2006). WebBootCaT: Instant domain-specific corpora to support human translators. *Proceedings of EAMT 2006*, 247-252.

Baroni, M. and Ueyama, M. (2004). Retrieving Japanese specialized terms and corpora from the World Wide Web. *Proceedings of KONVENS 2004*.

Bordoni, F. (2001). *Leucemia linfoblastica acuta in età pediatrica:*

Proposta di glossario terminologico trilingue (italiano - tedesco - inglese). Unpublished dissertation, SSLMIT, Bologna.

Church, K. and Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics* 16(1), 22-29.

Dunning, T. (1994). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics* 19(1), 61-74.

Evert, S. and Krenn, B. (2001). Methods for the quantitative evaluation of lexical association measures. *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, 188-195.

Gile, D. (1995). *Basic concepts and models for translator and interpreter training*, Amsterdam: Benjamins.

Henzinger, M. and Lawrence, S. (2004). Extracting knowledge from the World Wide Web. *Proceedings of the National Academy of Sciences of the United States of America* 101, 5186-5191.

Kalina, S. (1998). Kognitive Verarbeitungsprozesse. In Snell-Hornby, M., Hönig, H., KuSSmaul, P. and Schmitt, P. (eds.) *Handbuch Translation*, Tübingen: Stauffenburg, 330-335.

Kilgarriff, A. (2001). Comparing corpora. *International Journal of Corpus Linguistics* 6, 1-37.

Kilgarriff, A. and Grefenstette, G. (2003). Introduction to the special issue on the Web as corpus. *Computational Linguistics* 29(3), 333-347.

Kurz, I. (1996). *Simultandolmetschen als Gegenstand der interdisziplinären Forschung*, Wien: WUV-Univ. Verlag.

Johns, T. (1988). Whence and whither classroom concordancing? In Bongaerts, T., de Haan, P., Lobbe, S., and Wekker H. (eds.) *Computer applications in language learning*, Dordrecht: Foris, 9-27.

Rayson, P. and Garside, R. (2000). Comparing corpora using frequency profiling. *Proceedings of Workshop on Comparing Corpora at ACL 2000*, 1-6.

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.