Evaluation of an automatic process for specialized web corpora collection and term extraction for Basque

Antton Gurrutxaga, Igor Leturia, Eli Pociello, Xabier Saralegi, Iñaki San Vicente[1]
Elhuyar Foundation

Abstract

In this paper we describe the processes for collecting Basque specialized corpora in different domains from the Internet and subsequently extracting terminology out of them, using automatic tools in both cases. We evaluate the results of corpus compiling and term extraction by making use of a specialized dictionary recently updated by experts. We also compare the results of the automatically collected web corpus with those of a traditionally collected corpus, in order to analyze the usefulness of the Internet as a reliable source of information for terminology tasks.

Keywords: web as corpus, automatic term extraction, Basque

1. Motivation

The traditional process for building corpora –out of printed texts, following some selection criteria, linguistically tagged and indexed, etc.– is a very laborious and costly one, so corpora built this way are not as large or abundant as we would like them to be, and even less so in specialized domains. So in recent years the web has been used increasingly for linguistic research, both via tools like WebCorp (Kehoe and Renouf 2002) or CorpEus (Leturia *et al.* 2007a) that query search engines directly and show concordances, or via tools that use the Internet as a source of texts for building corpora to be used the classic way, after linguistic tagging and indexation (Ferraresi *et al.* 2008).

Although the use of the web as a source for building linguistic corpora has its detractors, this approach offers undeniable advantages (Kilgarriff and Grefenstette 2004):

- The corpora that can be obtained are much larger.
- The cost of the automatic building processes is much smaller.

[1] R&D department, Elhuyar Foundation, {a.gurrutxaga, i.leturia, e.pociello, x.saralegi, i.sanvicente}@elhuyar.com

- The web is constantly up to date.

On the other hand, the development of terminological resources is essential for any language that aims to be a communication tool in education, industry, etc. The automation of the term extraction process is a condition for this task to be carried out at a reasonable cost taking large samples of real texts as a data source (Ahmad and Rogers 2001).

If all this is true for any language, it is even more so in the case of a less-resourced language like Basque, so the automation of corpus compilation and terminology extraction processes is very attractive indeed.

2. Corpus collection

The compilation of specialized corpora from the Internet is performed by using an automatic tool (Leturia *et al.* 2008) that gathers the documents via the standard method of search engine queries (Baroni and Bernardini 2004).

The system is fed with a sample mini-corpus of documents that covers as many sub-areas of the domain as possible – 10 to 20 small documents can be enough, depending on the domain. A list of seed terms is automatically extracted from it, which can be manually edited and improved if necessary. Then combinations of these seed words are sent to a search engine, using morphological query expansion and language-filtering words to obtain better results for Basque (Leturia *et al.* 2007b), and the pages returned are downloaded.

Boilerplate is stripped off the downloaded pages (Saralegi and Leturia 2007) which are then passed through various filters:

- Size filtering (Fletcher 2004)
- Paragraph-level language filtering
- Near-duplicate filtering (Broder 2000)
- Containment filtering (Broder 1997)

A final topic-filtering stage is also added, using the initial sample mini-corpus as a reference and using document similarity techniques (Saralegi and Alegria 2007) based on keyword frequencies (Sebastiani 2002). A manual evaluation of this tool showed that it could obtain a topic precision of over 90%.

3. Terminology extraction

Term extraction is carried out using Erauzterm, an automatic terminology extraction tool for Basque (Alegria *et al.* 2004a), which combines both linguistic and statistical methods.

First, a lemmatizer and POS tagger for Basque (Aduriz *et al.* 1996) is applied to the corpus. Then the most usual Noun Phrase structures for Basque terms are detected (Alegria *et al.* 2004b) to obtain a list of term candidates. Term variants are linked to each other by applying some rules at syntagmatic and paradigmatic level. After this normalization step, statistical measures are applied in order to rank the candidates. Multiword terms are ranked according to their degree of association or unithood using Log Likelihood Ratio or LLR (Dunning 1994). Single word terms are ranked according to their termhood or divergence with respect to a general domain corpus, also using LLR. Then those candidates that reach a threshold are chosen. A manual evaluation of the tool reported a precision of 65% for multiword terms and 75% for single word terms for the first 2,000 candidates.

The tool also offers a graphical interface which allows the user, if necessary, to explore, edit and export the extracted terminology.

4. Experiments and evaluation

4.1. Experiments

We used the tools and systems described above to collect three specialized corpora and to obtain term lists from them, and we evaluated the results.

The domains chosen were Computer Science, Biotechnology and Atomic & Particle Physics. The collection of the corpora from the Internet did not have a target size, because the Internet in Basque is not as big as that in other languages, and the number of pages we would want to collect for a particular domain might not exist. So we simply launched the collecting processes and stopped them when the growing speed of the corpora fell to almost zero, thus obtaining corpora that were as large as possible. Then we applied the terminology extraction process to the corpora and obtained three term candidate lists. These lists were automatically validated against a recently compiled specialized dictionary, *Basic Dictionary of Science and Technology* (http://zthiztegia.elhuyar.org), which contains 25,000 terms. The best ranked ones of the remaining candidates were manually evaluated by experts to decide if they were terms or not.

Figure 1 shows the size of the corpora obtained, the number of terms extracted and the number of terms validated manually or by the dictionary, for each of the three domains.

Corpus	**Atomic and Particle Physics**	**Computer Science**	**Biotechnology**
Sample corpus size	32 docs, 26,164 words	33 docs, 34,266 words	55 docs, 41,496 words
Obtained corpus size	320,212	2,514,290	578,866

Extracted term list size	46,972	163,698	34,910
Dictionary validated	6,432	8,137	6,524
Manually evaluated	1,147	905	628
Terms	887	513	432
Not terms	260	392	196

Figure 1. Corpus and term list sizes obtained for each of the three domains

4.2. Evaluation

We evaluated the domain precision of the lists obtained from the Internet, by analyzing the distribution of the terms across the domains, taking the domains of the specialized dictionary as a reference. The results of this evaluation are shown in Figure 2, where we can observe that all three lists show peaks in or around their respective domains, which proves that the corpora are indeed specialized and that the term lists automatically extracted belong mainly to the desired domains.

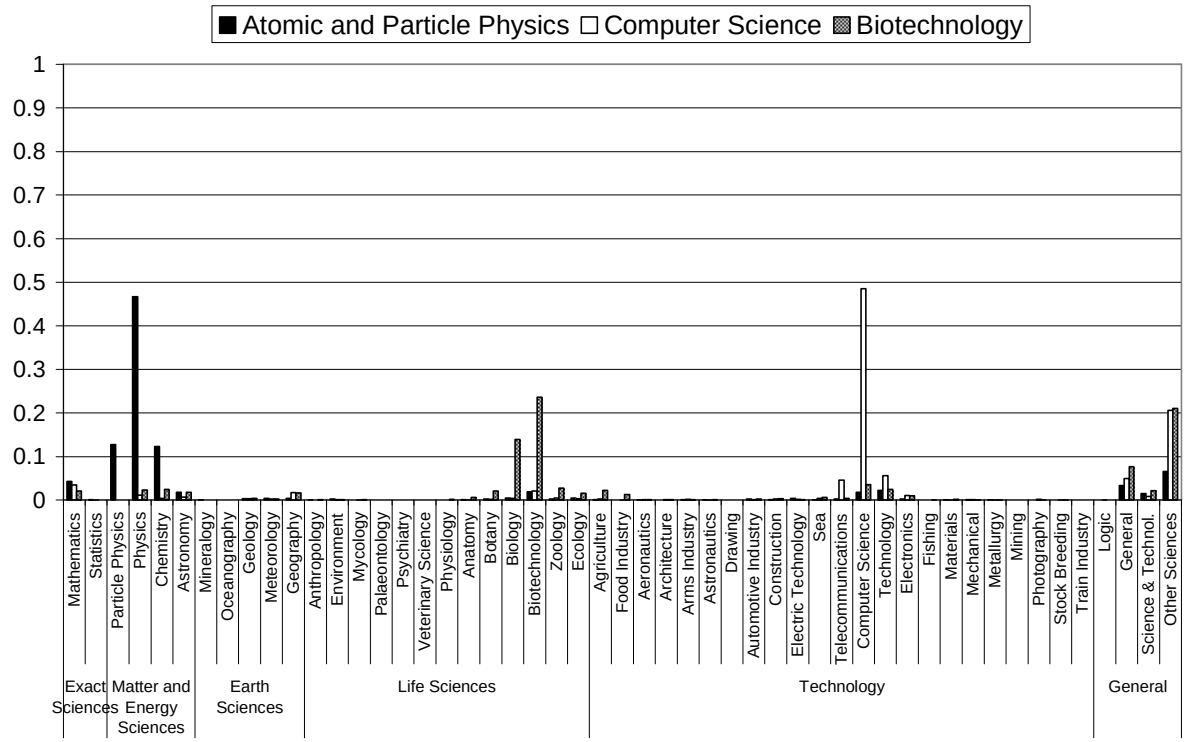

Figure 2. Domain distribution of the extracted term lists

The precision of the extracted term lists, that is, the percentage of the extracted terms that really belonged to the desired domain, was also evaluated. Figure 3 shows the evolution of this precision as the number of candidate terms grows. Here we can observe that the results are different for each of the domains. As a general rule, we can say that pure sciences perform better than technologies, which might indicate that these domains are more "terminologically dense", although we cannot be sure about this, because it could also be due to the different nature –extension, diversity,

production– of the domains. Besides, we believe that the seed document selection might also affect the quality of the resulting corpora and term lists.

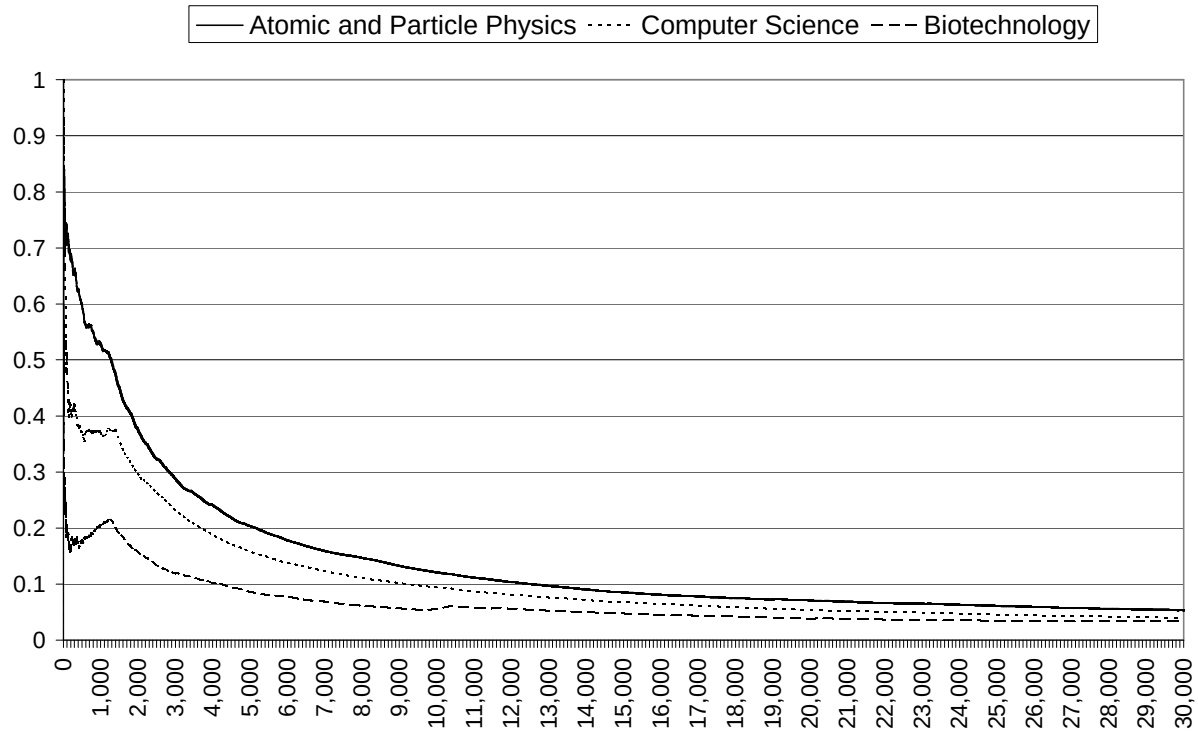

Figure 3. Domain precision of the extracted term lists

Also, the size of the collected corpora does not seem so important as far as the term extraction task is concerned: the Atomic and Particle Physics corpus achieves better results than the Biotechnology one, the former being almost half the size of the latter (Figure 1). As we have already pointed out, the nature of the domain is more important.

We also compared the extracted term lists with the lists on the domains of a specialized dictionary compiled and recently updated by experts, and look at the recall, that is, the percentage of the dictionary achieved, and the number of new terms extracted that were not in the dictionary. These two pieces of data are shown in Figures 4 and 5. By looking at the recall, we could draw the conclusion that the corpus building process is not good enough for compiling a quality dictionary, but we will see later that a traditional corpus does not do better. The use of corpora lacking representativeness could be put forward as a reason for that flaw. But another possible explanation for this fact could lie in the current situation of Basque terminology and text production. Although Basque began to be used in Science and Technology thirty years ago, it cannot be denied that there is a given amount of highly specialized terminology that is published *ex novo* in dictionaries, with little document support if any. That could be the reason why several terms chosen by experts and published in the dictionary do not occur in either of the two corpora. However, we can see in Figure 5 that many new terms appear, so the process proposed is definitely interesting for enriching or updating already existing specialized dictionaries.

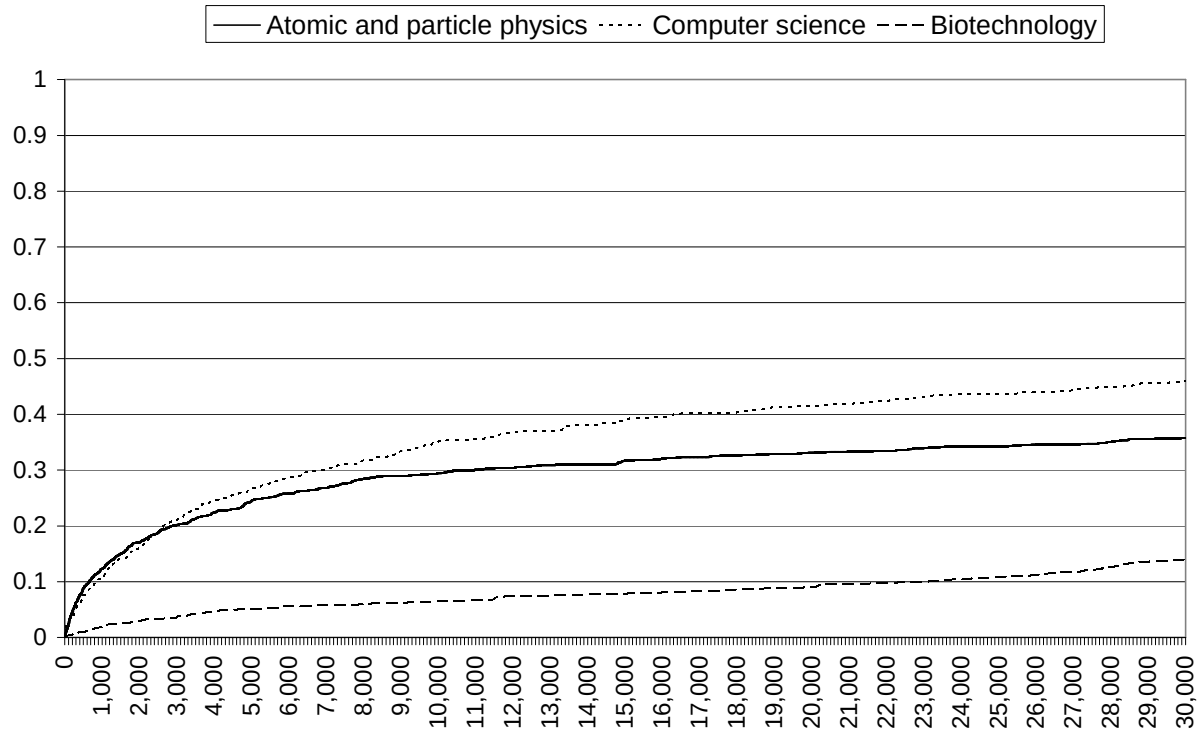

Figure 4. Recall of the extracted term lists compared with the dictionary

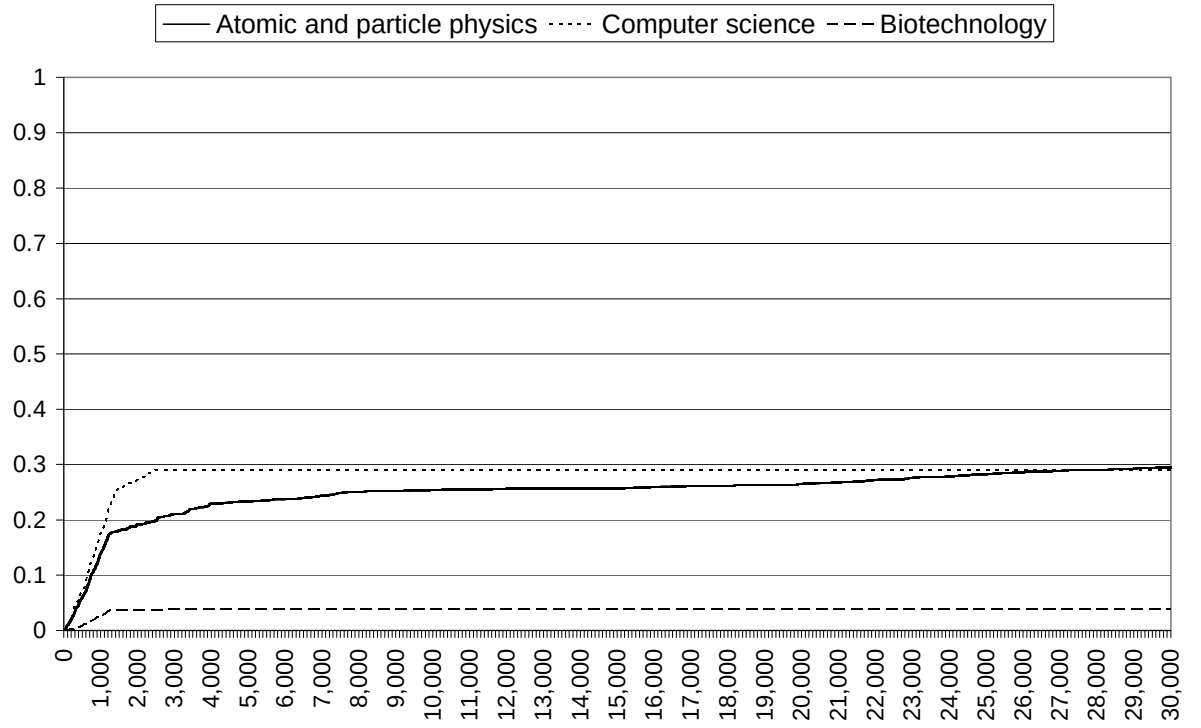

Figure 5. New terms in the extracted term lists that were not in the dictionary

Finally, we compared the term list extracted from a corpus automatically collected from the web with the term list extracted from a classical corpus. So a sub-corpus of the Computer Science domain was extracted from a traditional corpus, the ZT Corpus (Areta *et al.* 2007; http://www.ztcorpusa.net), and terminology was extracted with the

same method used with the Computer Science web corpus. Then both lists were compared. Figure 6 shows data on these two corpora and their respective term lists.

Corpus	Computer Science	Computer Science - ZT Corpus
Sample corpus size	33 docs, 34,266 words	-
Obtained corpus size	2,514,290	332,745
Extracted term list size	163,698	24,283
Dictionary validated	8,137	3,389
Manually evaluated	905	1,022
Positive	513	479
Negative	392	543

Figure 6. Corpus and term list sizes obtained for the web and traditional corpora

Figures 7, 8, 9 and 10 show, respectively, the domain distribution, domain precision, recall compared with the dictionary and new terms that were not in the dictionary of the two extracted term lists. They prove that we can obtain similar or, in some aspects, even better results with the automatic corpus collection process. As the cost is much lower, we believe that the process proposed in the paper is valid and very interesting for terminological tasks.

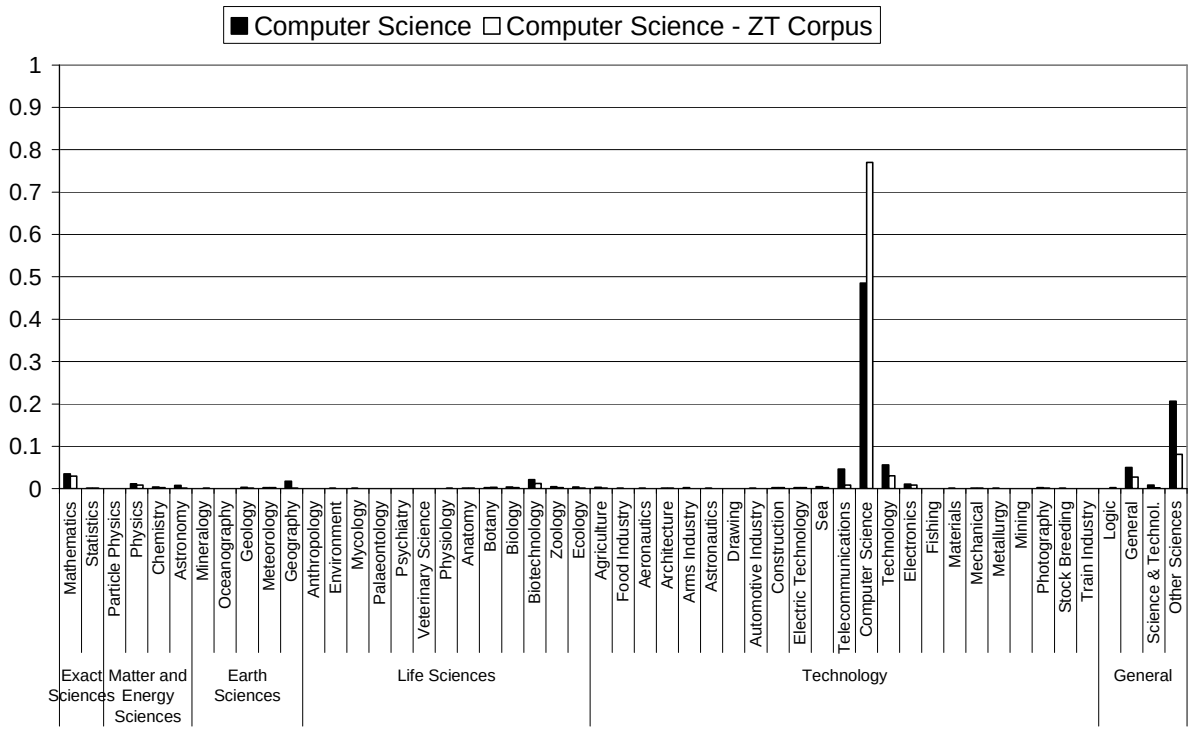

Figure 7. Domain distribution of the extracted term lists

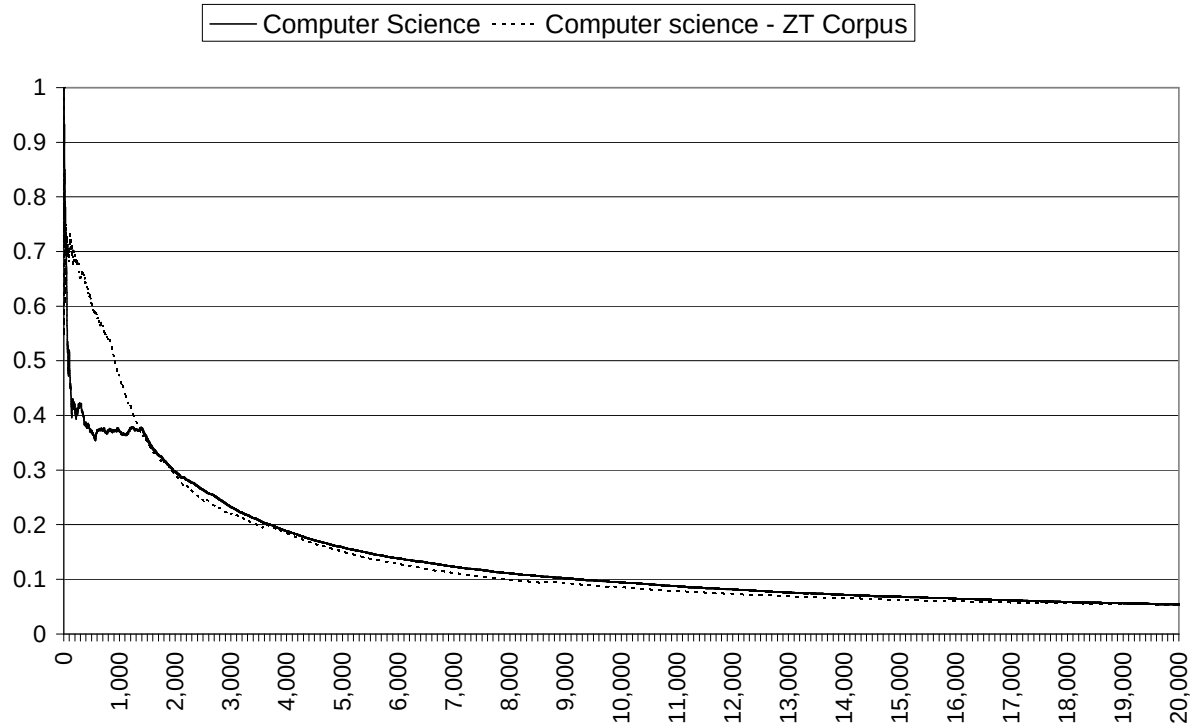

Figure 8. Domain precision of the extracted term lists

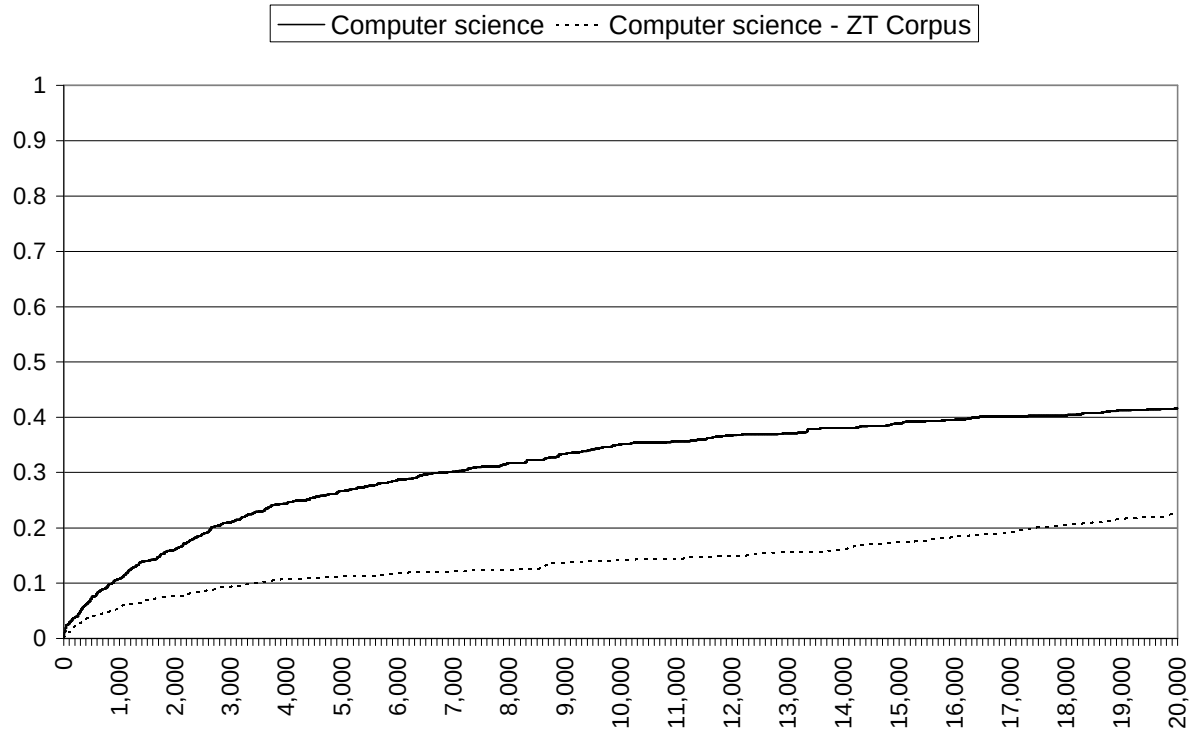

Figure 9. Recall of the extracted term lists compared with the dictionary

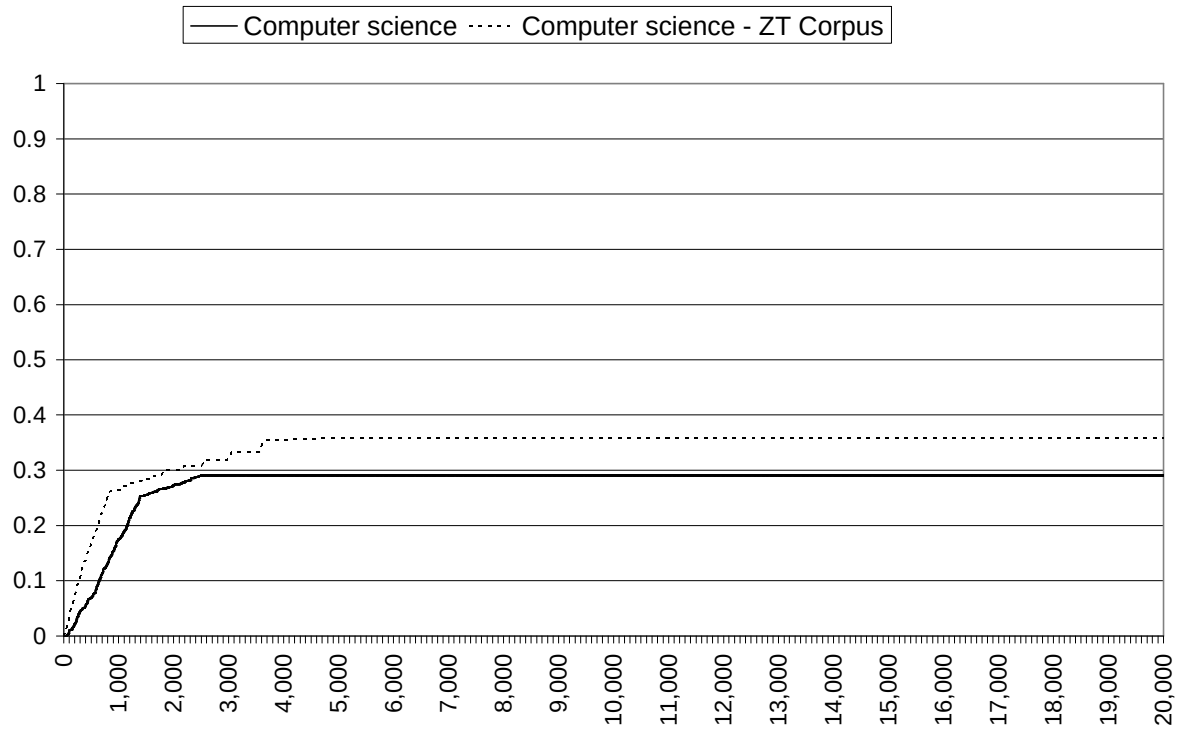

Figure 10. New terms in the extracted term lists that were not in the dictionary

5. Conclusions

The distribution and domain precision graphs prove that the corpora and term lists obtained are indeed specialized on the desired domains and lead us to believe that the automatic corpus collection and term extraction process can be valid for terminology tasks.

The evaluation also shows that results almost as good as from a traditional corpus can be obtained regarding precision or new terms, and even better in the case of recall.

Overall, the evaluation results are encouraging and indicate that acceptable results can be obtained with much less work than by means of a completely manual process.

References

ADURIZ, I., ALDEZABAL, I., ALEGRIA, I., ARTOLA, X., EZEIZA, N. and URIZAR, R. (1996). EUSLEM: A Lemmatiser / Tagger for Basque. In *Proceedings of EURALEX'96*. Göteborg: EURALEX: 17-26.

AHMAD, K. and ROGERS, M. (2001). Corpus Linguistics and Terminology Extraction. In S. E. Wright and G. Budin (eds.) *Handbook of Terminology Management Volume 2*. Amsterdam: John Benjamins: 725-760.

ALEGRIA, I., GURRUTXAGA, A., LIZASO, P., SARALEGI, X., UGARTETXEA, S. and URIZAR, R. (2004a). An Xml-Based Term Extraction Tool for Basque. In *Proceedings of the 4th International Conference on Language Resources and Evaluations (LREC)*. Lisbon: ELRA: 1733-1736.

ALEGRIA, I., GURRUTXAGA, A., LIZASO, P., SARALEGI, X., UGARTETXEA, S. and URIZAR, R. (2004b). Linguistic and Statistical Approaches to Basque Term Extraction. In *Proceedings of GLAT 2004: The production of specialized texts*. Barcelona: ENST Bretagne: 235-246.

ARETA N., GURRUTXAGA A., LETURIA I., ALEGRIA I., ARTOLA X., DÍAZ DE ILARRAZA A., EZEIZA N. and SOLOGAISTOA A. (2007). ZT Corpus: Annotation and tools for Basque corpora. In *Proceedings of Corpus Linguistics*. Birmingham: University of Birmingham.

BARONI, M. and BERNARDINI, S. (2004). BootCaT: Bootstrapping corpora and terms from the web. In *Proceedings of the 4th International Conference on Language Resources and Evaluations (LREC)*. Lisbon: ELRA: 1313-1316.

BRODER, A.Z. (1997). On the resemblance and containment of documents. In *Proceedings of Compression and Complexity of Sequences 1997*. Los Alamitos: IEEE Computer Society: 21-29.

BRODER, A.Z. (2000). Identifying and filtering near-duplicate documents. In *Proceedings of Combinatorial Pattern Matching: 11th Annual Symposium*. Montreal: Springer: 1-10.

DUNNING, T. (1994). Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics*, 19(1): 61-74.

FERRARESI, A., ZANCHETTA, E., BARONI, M., and BERNARDINI, S. (2008). Introducing and evaluating ukWaC, a very large web-derived corpus of English. In *Proceedings of the 4th Web as Corpus workshop*. Marrakech: ELRA: 47-54.

FLETCHER, W.H. (2004). Making the web more useful as a source for linguistic corpora. In U. Connor and T. Upton (eds.) *Corpus Linguistics in North America 2002*. Amsterdam: Rodopi: 191-205.

KEHOE, A. and RENOUF A. (2002). WebCorp: Applying the Web to Linguistics and Linguistics to the Web. In *Proceedings of the WWW2002 Conference*. Honolulu: W3C.

KILGARRIFF, A. and GREFENSTETTE, G. (2004). Introduction to the special issue on the Web as corpus. *Computational Linguistics*, 29: 333-348.

LETURIA, I., GURRUTXAGA, A., ALEGRIA I. and EZEIZA A. (2007a). CorpEus, a web as corpus tool designed for the agglutinative nature of Basque. In *Proceedings of the 3rd Web as Corpus workshop*. Louvain-la-Neuve: Presses Universitaires de Louvain: 69-81.

LETURIA, I., GURRUTXAGA, A., ARETA, A., ALEGRIA, I. and EZEIZA, A. (2007b). EusBila, a search service designed for the agglutinative nature of Basque. In *Proceedings of Improving non-English web searching (iNEWS'07) workshop*. Amsterdam: SIGIR: 47-54.

LETURIA, I., SAN VICENTE, I,. SARALEGI, X. and LOPEZ DE LACALLE., M. (2008). Basque specialized corpora from the web: language-specific performance tweaks and improving topic precision. In *Proceedings of the 4th Web as Corpus workshop*. Marrakech: ELRA: 40-46.

SARALEGI, X. and ALEGRIA, I. (2007). Similitud entre documentos multilingües de carácter científico-técnico en un entorno web. *Procesamiento del Lenguaje Natural*, 39: 71-78.

SARALEGI, X. and LETURIA, I. (2007). Kimatu, a tool for cleaning non-content text parts from HTML docs. In *Proceedings of the 3rd Web as Corpus workshop*. Louvain-la-Neuve: Presses Universitaires de Louvain: 163-167.

SEBASTIANI, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1): 1-47.