Finding domain specific collocations and concordances on the Web

Caroline Barrière
National Research Council of Canada
Gatineau, QC, Canada
caroline.barriere@nrc-cnrc.gc.ca

Abstract

TerminoWeb is a web-based platform designed to find and explore specialized domain knowledge on the Web. An important aspect of this exploration is the discovery of domain-specific collocations on the Web and their presentation in a concordancer to provide contextual information. Such information is valuable to a translator or a language learner presented with a source text containing a specific terminology to be understood. The purpose of this article is to show a proof of concept that TerminoWeb, as an integrated platform, allows the user to extract terms from the source text and then automatically build a related specialized corpus from the Web in which collocations will be discovered to help the user understand the unknown specialized terms.

Keywords

term extraction, collocation extraction, concordancer, Web as corpus, domain-specific corpus

1. Introduction

Collocations and concordances found in corpora provide valuable information for both acquiring the sense and usage of a term or word. Corpora resources are usually complementary to dictionaries, and provide a more contextual understanding of a term. Collocations and concordances are rarely viewed as "static" resources, the way dictionary definitions would, but are rather often considered the disposable result of a tool's process (a concordancer, a collocation extractor) on a particular corpus.

The use and value of corpora for vocabulary acquisition and comprehension is quite known. In language learning mostly [2], its use obviously has advantages and disadvantages compared to dictionaries, and its context of usage might influence its value (self-learning or classroom). Early work on vocabulary acquisition [21] argued that the learning of a new word is frequently related to the incidence of reading or repetitive listening. Even earlier [23], one experiment illustrated that the frequency of a word and the richness of the context facilitates the identification of a word by a novice reader. Even so, computer-assisted techniques for the understanding of unknown words [15] in second language learning are still not widely studied.

In translation studies, the value of corpora has been repeatedly shown [6, 19, 22, 28] and concordancers are the tools of choice for many translators to view a word in its natural surrounding.

Concordances are usually defined clearly as a window of text surrounding a term or expression of interest. Most often, a fixed small window size is established (ex. 50 characters) and the results are called KWIC (KeyWord In Context). Such KWIC views are usually supplemented one-click away by a larger context view (a paragraph), potentially even another click away to access the source text.

Collocations are words which tend to co-occur with higher than random probability. Although conceptually the definition is quite simple, results will largely differ because of two main variables. A first variable is the window size in which co-occurrences are measured. A small window (2-3 words maximum before or after) is usually established for collocations. Longer distances are considered associations, or semantically-related words, which tend to be together in sentences or paragraphs or even documents. A second variable is the actual measure of association used, and there have been multiple measures suggested in the literature, such as Overlap, Mutual Information, Dice Coefficient, etc [10][1].

Even more fundamentally, one key element will largely influence the results of both concordancers and collocation extractors: the source corpus. For the general language, the BNC (British National Corpus) has been widely used by corpus linguists, and recently a Web interface has been developed (BNCweb) to access it [14].

Domain-specific corpora or corpora in other languages than English are not as easily found[2], especially not packaged with a set of corpus analysis tools. The notion of "disposable" or "do-it-yourself" corpora has been suggested as a corpus that translators would build themselves to quickly search for information [7, 26]. Language learners would also often be in need of domain-specific corpora. But the problem resides in the overhead work involved in building such corpus. A process of selection, upload

[1] For detailed explanations and a tutorial on multiple association measures: http://www.collocations.de/AM/

[2] See Serge Sharoff's multi-language Web copus collection (http://corpus.leeds.ac.uk/internet.html).

NLP Methods and Corpora in Translation, Lexicography, and Language Learning 2009 – Borovets, Bulgaria, pages 1–8

(for personal texts) or download (for Web texts) and management (easy storage and retrieval of texts) is involved. Only a few tools exist for such purpose, such as Corpografo [20] and TerminoWeb [5].

This paper presents a new version of the TerminoWeb system[3] which provides the user with the capability of automatically building a "disposable" domain-specific corpus and study some terms by finding collocations and concordances in that corpus. Different steps are necessary for such task. Section 2 presents a working scenario with a single detailed example to explain the algorithms underlying each step. Section 3 links to related work for different aspects of the system, although we do not know of any system which integrates all the modules as TerminoWeb does. Section 4 concludes and hints at some future work.

2. Collocations and concordances

Becoming familiar with the vocabulary in a source text is essential both for reading comprehension (a language learner's task) and text translation (a translator's task).

The understanding process for unknown terms/words could rely on a search for appropriate definitions in a dictionary, as well as a search for collocations and concordances in corpus. We look at this second option and present the following scenario to perform the discovery of collocations and concordances:

1) Source text upload
2) Term extraction on source text
3) Query generation from terms
4) Domain-specific corpus construction
5) Collocations and concordances search

Step 1. Text upload

We take as a starting point a text to translate or a text to understand. TerminoWeb provides an interface for the user to upload (copy-paste) the source text. For illustrating purpose, we arbitrarily take a text on banking fraud issues (http://bankfraudtoday.com/).

Step 2. Term extraction

The term extractor finds single-word and/or multi-word terms in the source document. The number of terms to be found can be set by the user, or estimated automatically based on the document's length and the actual term statistics. The term extraction module implements Smadja's algorithm [25] which is purely statistical and based on frequencies. Such a purely statistical approach has the advantage of being largely language independent, with only a list of stop words necessary for each different language.

TerminoWeb allows term sorting in alphabetical or frequency order, but Figure 1 shows a sample of terms from the uploaded document on bank fraud ordered by specificity. Specificity is approximated by a "hit count" measure which we discuss in the next step of query generation.

Step 3. Query generation

This step is to launch a set of queries on the Web to find documents that are both domain specific (related to the source text) and containing information about the unknown words (words less familiar to the language learner or the translator). The purpose of the query generation (QG) module is to make this task relatively easy for the user. Nevertheless, the following factors which will impact the results must be understood:

a. Unknown terms
b. Domain terms
c. Number of terms per query
d. Number of queries launched
e. Term frequencies

Unknown terms (factor a.), are the ones the user is interested in understanding. In the bank fraud example, they are "closing costs" or "insurance premium" or "predatory lending" (words shown in Figure 1). When the unknown terms are not polysemous (which is more often the case for multi-word terms), domain terms are not necessary to disambiguate them.

But sometimes, unknown terms are common single-word terms taking on a specific meaning in a particular domain, and then domain terms (factor b.) are important for query disambiguation. For example the term "interest" in our present bank fraud domain has a different meaning then in the expression "having an interest in" from the general language. In such case, domain terms "bank" and "fraud" can be specified to be added to all queries.

The following two factors (c. and d.) are number of words per query and number of queries. If for example, 10 terms are unknown, the QG module can generate 10 queries of 1 term each, 4 queries of 3 terms each, or 15 queries of 2 terms each, as the user decides. The QG module will make random combinations of terms to generate the required queries. The number of queries would in theory be better if higher, but this becomes a trade-off between the information gained by more corpus data and a longer waiting period. It will be important in our future work to better measure the gain from more queries versus better chosen or targeted queries.

[3] TerminoWeb 2.0 is available online since June 2009 at http://terminoweb.iit.nrc.ca.

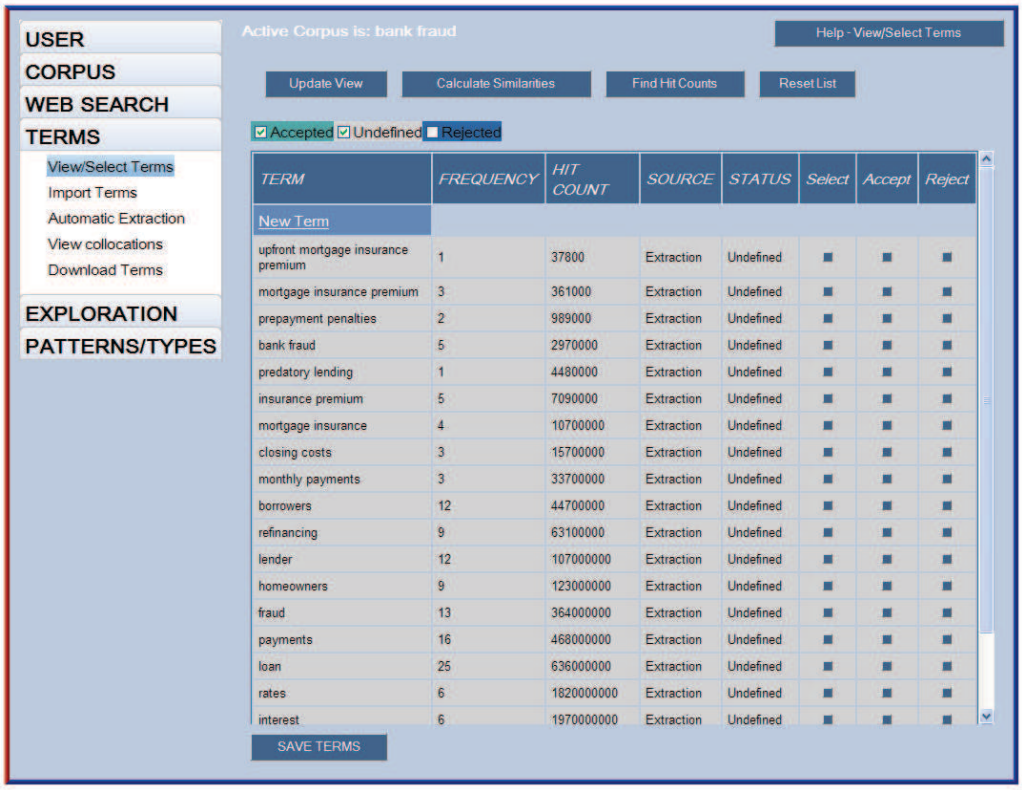


Figure 1 – Extracted Terms with source text frequencies and Web hit counts

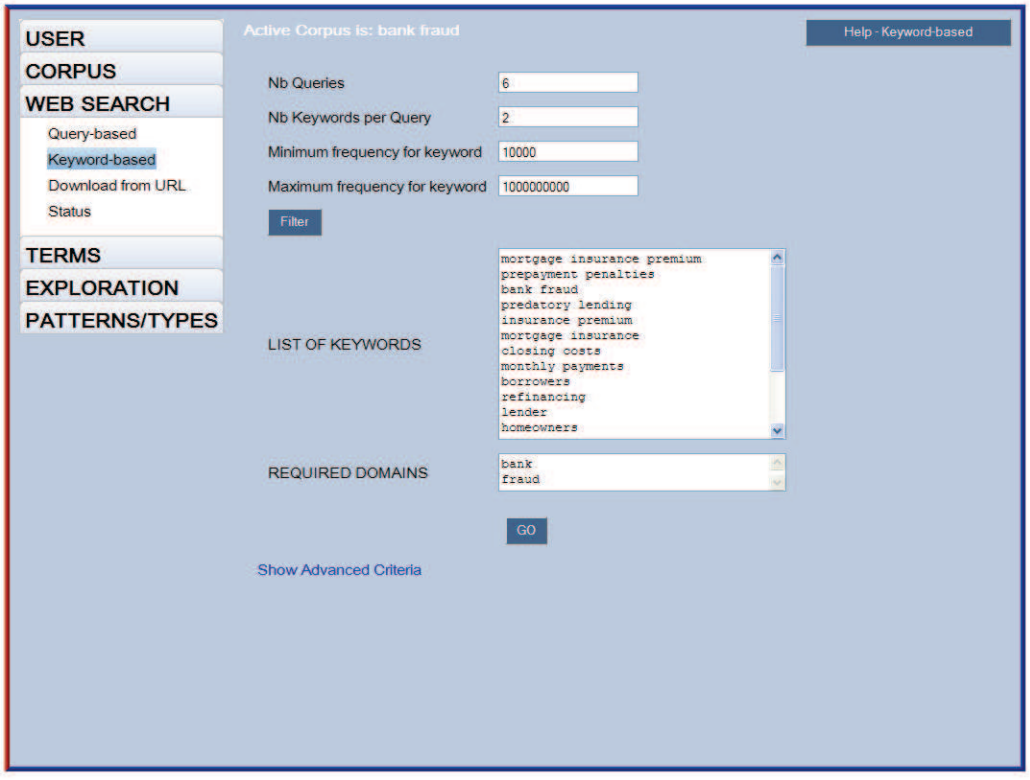


Figure 2 – Query Generator Module Interface

Figure 2 shows the QG Module interface which gives the user much flexibility in specifying domain terms, unknown terms, number of queries and number of terms per query.

When multiple very specific terms are combined, the resulting set of documents is likely to be empty (no documents found). When few general terms are used (one at the limit) the resulting set is likely to be extremely large and inappropriate (imagine a query with "credit" or "stolen"). Empirically, we have found that queries of more than 3 terms often lead to empty sets, although the size of the result set is not solely dependent on the number of terms in the query but rather very closely related to the actual frequency of those terms.

A quick estimate of how specific or general a word or expression is can be provided by a "hit count" measure using a search engine. In our experiment, we use Yahoo Search Engine[4]. Figure 1 shows the term list sorted on hit counts. The sample of terms shown is to provide the reader a sense of the large range from specificity to generality. The term "mortgage insurance premium" is more specific (hit counts: 36100) than "monthly payments" (hit counts: 33700000) which is more specific than "rates" (hit counts: 1820000000).

The QG interface, shown in Figure 2, allows the user to filter query terms based on lower-bound (too specific) and upper-bound (too general) hit counts (factor e.).

Figure 3 shows the queries status. It shows combinations of 2 unknown terms combined with two mandatory domain terms. In TerminoWeb, queries can be "in progress" still looking for documents, "finished" as they have retrieved the requested number of documents (here 10) or "aborted" if something went wrong during the search.

Step 4. Corpus construction

The resulting documents from all the queries are put together to form a large corpus. The maximum number of documents would be equal to the Number of Queries * Number of documents per query, but that is an upper bound since queries could return a smaller set than what is desired (if too specific), some queries could "abort" and also, there will be document overlaps in the returned sets[5].

When a query leads to many documents, then a larger set is analyzed and scored to only keep the 10 most *informative* ones as part of the corpus. Although not the purpose of the present article, we briefly mention that TerminoWeb's focuses on the discovery of informative texts on the Web. Much research efforts have been devoted to TerminoWeb's capability to attribute an "informative score" to each text based on a few criteria such as domain specificity, definitional indicators, text size, sentence length, etc. Much effort has been spent on the exploration of definitional indicators, in the form of knowledge patterns representing different semantic relations. For example, "is a kind of" indicates hyperonymy and "is also known as" indicates synonymy. The presence of such knowledge patterns in a document will increase its informative score. TerminoWeb can show the documents in order of their informative score.

The corpus management module allows the user to inspect each document by providing a link to the original web page. The user can then decide to accept or reject some pages, limiting the documents in the corpus. This step is optional in the present process and mostly useful for thematic searches in which terminologists would like to inspect each source text from which they will select terms and contexts. If this step is not performed, the user will simply perform the next step (explore documents) on a larger set of documents.

Step 5. Collocations and concordances search

The user can now select a term to study and see (1) concordances for this term, (2) collocations generated from the term and (3) concordances for the collocations.

Figure 4 shows concordances for the term "refinancing", illustrating TerminoWeb's capability at providing KWIC views, larger context views, and links to source Web pages.

Figure 5 shows collocations with the word "refinancing". Only two collocations would have been found in the original source text, and many more domain-specific collocations are found in the extended corpus. Calculation of collocations is performed the same way as terms were found. Smadja's algorithm [25] allows the search for non-contiguous collocations. We indicate them with a % for a missing word. The maximum number of missing words was set to one, but could be larger if needed.

Figure 6 shows the concordancer used to highlight concordances around the found collocations. These are ordered alphabetically[6].

Another interesting feature of TerminoWeb is to allow users to find hit counts for collocations to approximate their specificity/generality, the same way as we presented earlier for terms. Figure 5 shows the hit counts for the different collocations.

[4] Yahoo! provides a java API which can be used for research purposes.

[5] As a research prototype, TerminoWeb can only process html and text documents, and it also filters out "almost-empty documents" containing only links or a few lines.

[6] Figures 4 and 6 contain a "no relation" column, meaning the contexts shown do not contain predefined knowledge patterns for different semantic relations.

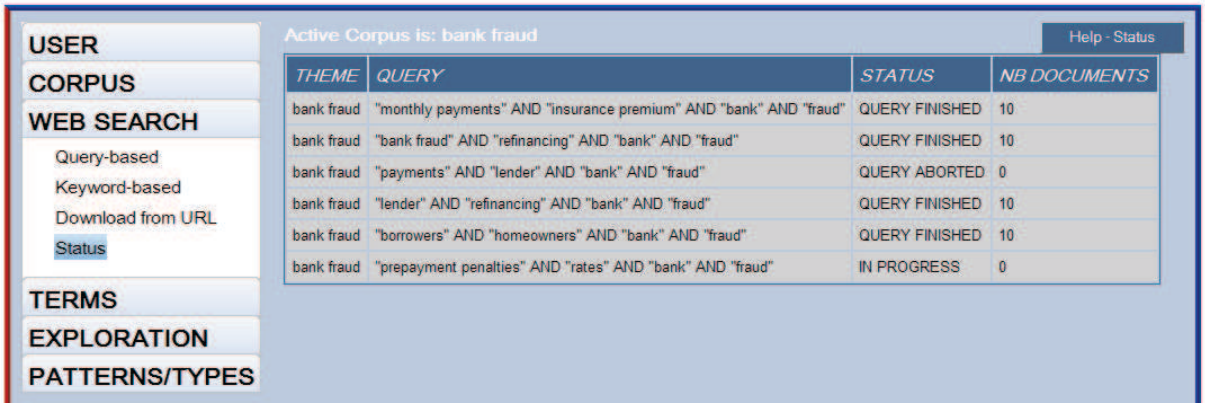


Figure 3. Status of queries launched

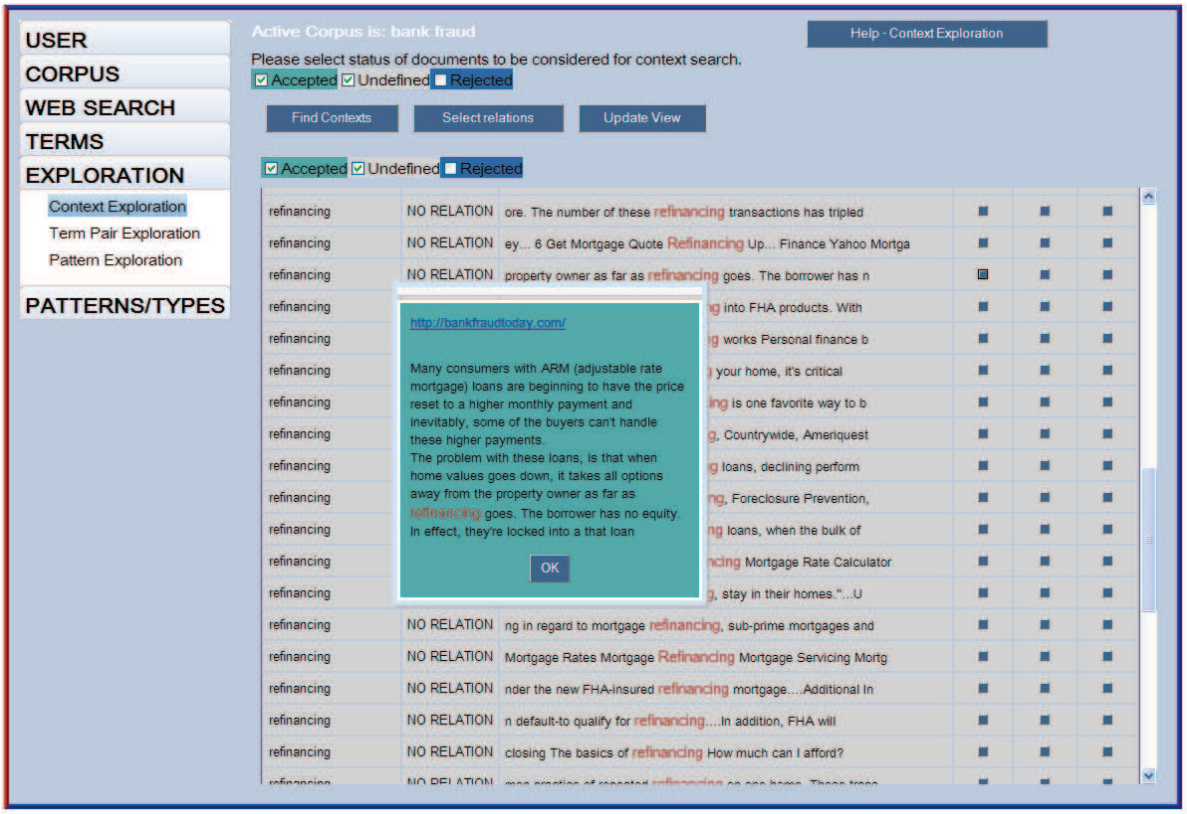
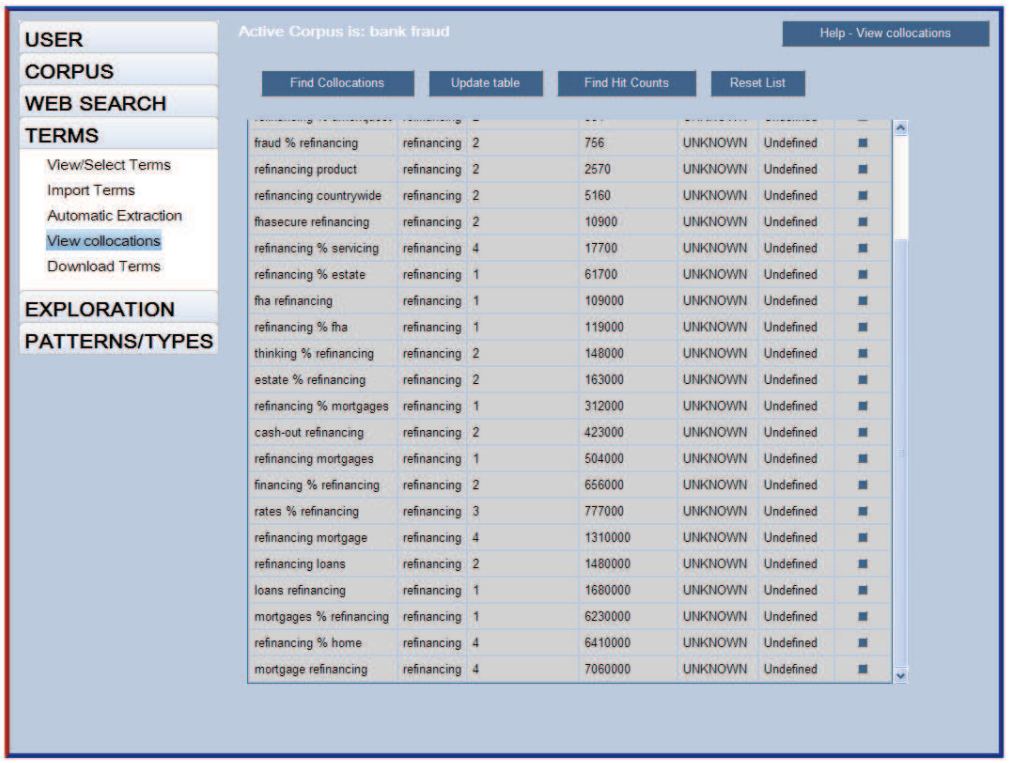
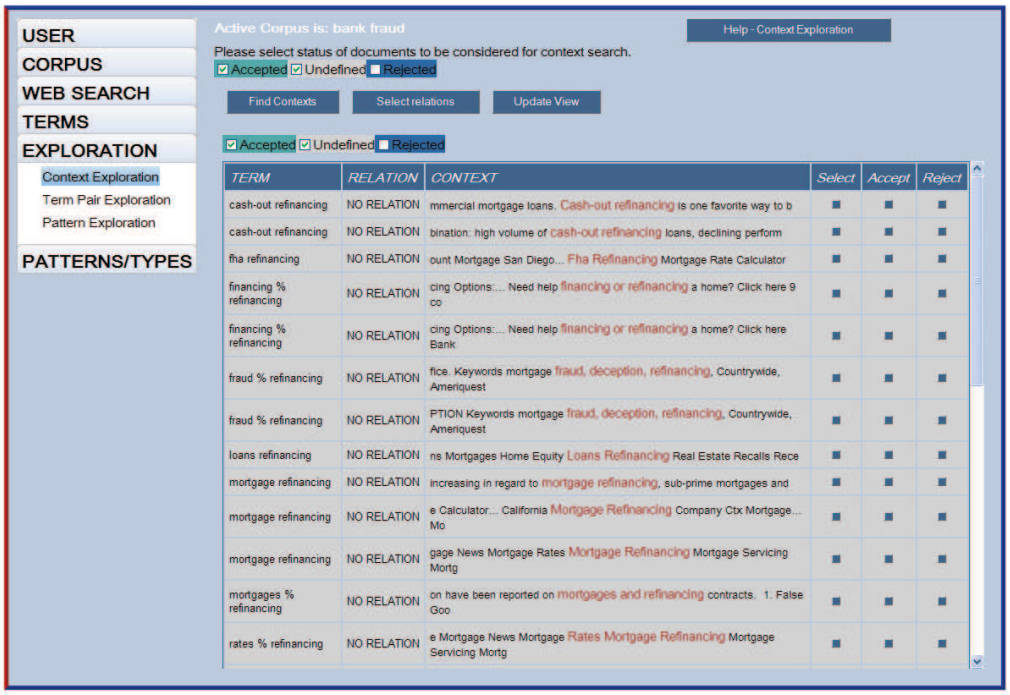


Figure 4. Term “refinancing” concordances



fraud % refinancing	refinancing	2	756	UNKNOWN	Undefined
refinancing product	refinancing	2	2570	UNKNOWN	Undefined
refinancing countrywide	refinancing	2	5160	UNKNOWN	Undefined
fhasecure refinancing	refinancing	2	10900	UNKNOWN	Undefined
refinancing % servicing	refinancing	4	17700	UNKNOWN	Undefined
refinancing % estate	refinancing	1	61700	UNKNOWN	Undefined
fha refinancing	refinancing	1	109000	UNKNOWN	Undefined
refinancing % fha	refinancing	1	119000	UNKNOWN	Undefined
thinking % refinancing	refinancing	2	148000	UNKNOWN	Undefined
estate % refinancing	refinancing	2	163000	UNKNOWN	Undefined
refinancing % mortgages	refinancing	1	312000	UNKNOWN	Undefined
cash-out refinancing	refinancing	2	423000	UNKNOWN	Undefined
refinancing mortgages	refinancing	1	504000	UNKNOWN	Undefined
financing % refinancing	refinancing	2	656000	UNKNOWN	Undefined
rates % refinancing	refinancing	3	777000	UNKNOWN	Undefined
refinancing mortgage	refinancing	4	1310000	UNKNOWN	Undefined
refinancing loans	refinancing	2	1480000	UNKNOWN	Undefined
loans refinancing	refinancing	1	1680000	UNKNOWN	Undefined
mortgages % refinancing	refinancing	1	6230000	UNKNOWN	Undefined
refinancing % home	refinancing	4	6410000	UNKNOWN	Undefined
mortgage refinancing	refinancing	4	7060000	UNKNOWN	Undefined

Figure 5 – Collocations found with “refinancing” in the domain specific corpus.



TERM	RELATION	CONTEXT	Select	Accept	Reject
cash-out refinancing	NO RELATION	mmercial mortgage loans. Cash-out refinancing is one favorite way to b	■	■	■
cash-out refinancing	NO RELATION	bination: high volume of cash-out refinancing loans, declining perform	■	■	■
fha refinancing	NO RELATION	ount Mortgage San Diego... Fha Refinancing Mortgage Rate Calculator	■	■	■
financing % refinancing	NO RELATION	cing Options:... Need help financing or refinancing a home? Click here 9 co	■	■	■
financing % refinancing	NO RELATION	cing Options:... Need help financing or refinancing a home? Click here Bank	■	■	■
fraud % refinancing	NO RELATION	fice. Keywords mortgage fraud, deception, refinancing, Countrywide, Ameriquest	■	■	■
fraud % refinancing	NO RELATION	PTION Keywords mortgage fraud, deception, refinancing, Countrywide, Ameriquest	■	■	■
loans refinancing	NO RELATION	ns Mortgages Home Equity Loans Refinancing Real Estate Recalls Rece	■	■	■
mortgage refinancing	NO RELATION	increasing in regard to mortgage refinancing, sub-prime mortgages and	■	■	■
mortgage refinancing	NO RELATION	e Calculator... California Mortgage Refinancing Company Ctx Mortgage... Mo	■	■	■
mortgage refinancing	NO RELATION	gage News Mortgage Rates Mortgage Refinancing Mortgage Servicing Mortg	■	■	■
mortgages % refinancing	NO RELATION	on have been reported on mortgages and refinancing contracts. 1. False Goo	■	■	■
rates % refinancing	NO RELATION	e Mortgage News Mortgage Rates Mortgage Refinancing Mortgage Servicing Mortg	■	■	■

Figure 6 – Concordances around collocations for “refinancing”

.

3. Related Work

Our research covers a wide range of topics, uses diverse natural language processing strategies, and includes the development of multiple algorithms for all steps, from term extraction to query generation to collocation search. As our purpose in this article is to present a proof of concept of an integrated system, we do not present any quantitative comparisons with other algorithms or systems, but rather highlight some research related to corpus building and analysis.

Our research relies mainly on the principle of "Web as corpus"[7] [17] and exploiting the Web for language learners and translators. In the book Corpus Linguistics and the Web [16], a distinction is made between "accessing the web as corpus" and "compiling corpora from the internet". Our system relates to both views. The hit count specificity/generality approximations relate to the former view. The corpus building modules gathering results from the query generation module relates to the latter view.

Search for Web documents is usually associated to the field of information retrieval. A large body of research exists within that area and we borrow from it. Searching for a particular document to answer a particular question (an information retrieval task) is different than searching for domain-specific documents to "augment" a user's knowledge. The former has a specific goal, finding an answer to a question, and the latter has a discovery purpose.

Nevertheless our query generation module faces the same problems as those of query expansion in information retrieval [12, 27]. Query expansion is a delicate task, as using general terms which tend to be polysemous can lead to off-topic documents, and using very specific terms will not help as they will not return any documents. Our approach was to allow the inclusion of domain-words for restriction and then do a random selection of terms for expansion.

Our query generation module was inspired by the work of Baroni [3, 4] who suggested query combinations of common words to build a corpus of general knowledge or specialized language. Earlier work by Ghani et al. [11] presented a similar idea for minority languages. TerminoWeb includes a unique re-ranking of documents based on an "informative score" as defined in [1]. It then builds informative sub-corpora from the Web.

Although, systems such as WebCorp [24] and KWiCFinder [13] do not build sub-corpora, they use the Web as a large corpus to find collocations and concordances, providing user with easy-to-use real-time systems.

For corpus analysis per se, TerminoWeb combines different modules performing term extraction, collocation extraction and concordance findings. A large pool of research exists in computational terminology around the problem of term extraction. Although a simple frequency based approach is implemented in TerminoWeb, there are more sophisticated algorithms being developed in the community (see [8] for a review of earlier systems and [9] for a new trend of term extraction based on comparing corpora). For collocations, we refer the reader to Smadja [25] for the algorithm we implemented, and to [10] for a review of different measures. Finding concordances does not require any statistical corpus linguistic knowledge, and is simply a window of text capture.

The Sketch Engine [18] system provides a good comparison point to position TerminoWeb. Overall, TerminoWeb's corpus analysis capabilities are simpler than the ones in Sketch Engine. The purpose is quite different, as TerminoWeb's main objective is to provide an integrated platform for understanding terms related to a domain or a source text. For doing so, the emphasis is on easy real-time construction and simple analysis of disposable corpora. No text-preprocessing is necessary, but then, no part-of-speech analysis is available either. We want the user to be able to quickly search for specialized information on the Web to understand important concepts via an integrated system for term extraction and term collocation and concordances finding. This is different from studying language patterns and understanding the uses of words or phrases as can be done in a better way in Sketch Engine [18].

4. Conclusions

Overall, although the value of "disposable corpora" for translators [7, 26] and for language learners [2] is acknowledged, the difficulty of performing text selection based on some principles implemented by natural language processing algorithms, and then the difficulty of doing efficient corpus management certainly prevents most users from building their own corpus. They are in need of tools, such as TerminoWeb, which provide corpus building and analysis capabilities.

TerminoWeb's contribution is actually more at the level of the workflow that the combination of its modules allows than at the level of the strength or novelty of any particular module (except for the "informative" scoring). Such combination makes multiple corpus gathering and analysis task possible.

TerminoWeb is a bit more complex than systems such as WebCorp [24] or KWiCFinder [13] as it

[7] The notion of Web as Corpus is a current research perspective as shown by the Web as Corpus workshops often run in parallel of larger conferences (Corpus Linguistics, 2005, European Association for Computational Linguistics EACL-2006, LREC 2008).

provides an integration of multiple modules, and therefore requires a longer learning curve, but the integration also makes it quite powerful, allowing a workflow such as described in this article, to start from a source text and find valuable information from the automatically extracted terms of that source text.

Our main future work is to gather feedback from users as they experiment with the prototype. This will allow us to better understand the particular needs of different users (language learners versus translators). This will help refine our modules and refine our choice of appropriate natural language processing techniques in support of each module.

5. References

[1] Agbago, A. Barrière, C. Corpus Construction for Terminology, Corpus Linguistics Conference, Birmingham, UK, July 14-17, 2005.

[2] Aston, G. Learning with Corpora, Houston: Athelstan, 2003.

[3] Baroni, M. and Bernardini, S. BootCaT: Bootstrapping Corpora and Terms from the Web, Proceedings of LREC, 2004.

[4] Baroni, M., Kilgarriff, A., Pomikalek, J. and Pavel, R. WebBootCaT: instant domain-specific corpora to support human translators, Proceedings of the 11th Annual Conference of the European Association for Machine Translation, EAMT-2006, Norway, 2006.

[5] Barrière C. and Agbago, A. TerminoWeb: a software environment for term study in rich contexts, International Conference on Terminology, Standardization and Technology Transfer, Beijing, 103-113, 2006.

[6] Bowker, Lynne and Pearson, Jennifer. Working with Specialized Text: A Practical Guide to Using Corpora. Routledge, 2002.

[7] Bowker, Lynne. "Working Together: A Collaborative Approach to DIY Corpora". First International Workshop on Language Resources for Translation Work and Research, Gran Canaria, 2002.

[8] Cabré Castellvi, M.T., Estopa R., Palatresi, J.V. Automatic term detection: A review of current systems. In Bourigault D., Jacquemin C., L'Homme M.C. (eds) Recent advances in Computational Terminology, vol. 2, pp. 53-87, 2001.

[9] Drouin P. Term extraction using non-technical corpora as a point of leverage. Terminology 9(1), pp. 99-117, 2003.

[10] Evert, Stefan and Krenn, Brigitte (2001). Methods for the qualitative evaluation of lexical association measures. In Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics, Toulouse, France, 188-195, 2001.

[11] Ghani, R., Jones, R., Mladenic, D., Mining the web to create minority language corpora, CIKM 2001, 279-286, 2001.

[12] Greenberg, J. Automatic query expansion via lexical – semantic relationships. Journal of the American Society for Information Science and Technology, 52(5), 402 – 415, 2001.

[13] Fletcher, W.H.. Concordancing the web: promise and problems, tools and techniques, in Hundt, M. Nesselhauf, N. and Biewer, C. (Eds) Corpus Linguistics and the Web, 2007.

[14] Hoffmann, S., Evert, S., Smith, N., Lee, D. and Ylva B.P. Corpus Linguistics with BNCweb - a Practical Guide. Frankfurt am Main: Peter Lang, 2008.

[15] Hulstijn, J. H., Hollander, M. & Greidanus, T. Incidental vocabulary learning by advanced foreign language students: The influence of marginal glosses, dictionary use, and reoccurrence of unknown words. *The Modern Language Journal, 80,* 327-339, 1996.

[16] Hundt, M. Nesselhauf, N., Biewer, C. Corpus Linguistics and the Web, Amsterdam/New York, NY, 2007, VI, 2007.

[17] Kilgariff, Adam and Gregory Grefenstette, Special Issue on Web as Corpus, Computational Linguistics, 29 (3), 2003.

[18] Kilgariff, A., Rychly, P., Smrz, P., Tugwell, D. The Sketch Engine, EURALEX, Lorient, France, 2004.

[19] Laviosa, Sara. Corpus-based Translation Studies: Theory, Findings, Applications, Amsterdam: Rodopi. 2002.

[20] Maia, B., Matos, S. Corpografo V.4 – Tools for Researchers and Teachers using Comparable Corpora, LREC 2008, Marrakech, Morocco, 2008.

[21] Nagy, W. E., Herman, P., and Anderson, R. C. Learning words from context. Reading Research Quarterly, 20, 233 – 253, 1985.

[22] Olohan, Maeve. Introducing Corpora in Translation Studies. London and New York: Routledge. 2004.

[23] Pearson, P. D. and Studt, A. Effects of word frequency and contextual richness on children's word identification abilities. Journal of Educational Psychology, 67(1), 89 – 95, 1975.

[24] Renouf, A., Kehoe, A., Banerjee, J. WebCorp: an integrated system for web text search, in Hundt, M. Nesselhauf, N. and Biewer, C. (Eds) Corpus Linguistics and the Web, 2007.

[25] Smadja F. Retrieving collocations from text: Xtract, Computational Linguistics, 19(1), 134-177, 1993.

[26] Varantola, K. Disposable corpora as intelligent tools in translation, Cadernos de Traduçao IX – Traduçao e Corpora, Vol. 1, No 9 (2002),171-189, 2002.

[27] Vechtomova, O., Robertson, S., & Jones, S. Query expansion with long-span collocates. Information Retrieval, 6, 251 – 273, 2003.

[28] Zanettin, Federico, Bernardini Silvia and Stewart Dominic. Corpora in Translation Education, Manchester: St Jerome, 2003.