Corporator: A tool for creating RSS-based specialized corpora

Cédrick Fairon
Centre de traitement automatique du langage
UCLouvain
Belgique
cedrick.fairon@uclouvain.be

Abstract

This paper presents a new approach and a software for collecting specialized corpora on the Web. This approach takes advantage of a very popular XML-based norm used on the Web for sharing content among websites: RSS (Really Simple Syndication). After a brief introduction to RSS, we explain the interest of this type of data sources in the framework of corpus development. Finally, we present Corporator, an Open Source software which was designed for collecting corpus from RSS feeds.

1 Introduction[1]

Over the last years, growing needs in the fields of Corpus Linguistics and NLP have led to an increasing demand for text corpora. The automation of corpus development has therefore became an important and active field of research. Until recently, constructing corpora required large teams and important means (as text was rarely available on electronic support and computer had limited capacities). Today, the situation is quite different as any published text is recorded, at some point of its "life" on digital media. Also, increasing number of electronic publication (textual databank, CD–ROM, etc.) and the expansion of the Internet have made text more accessible than ever in our history.

The Internet is obviously a great source of data for corpus development. It is either considered as a corpus by itself (see the WebCorp Project of Renouf, 2003) or as a huge databank in which to look for specific texts to be selected and gathered for further treatment. Examples of projects adopting the latter approach are numerous (among many Sekigushi and Yammoto, 2004; Emirkanian *et al*. 2004). It is also the goal of the WaCky Project for instance which aims at developing tools "that will allow linguists to crawl a section of the web, process the data, index them and search them"[2].

So we have the Internet: it is immense, free, easily accessible and can be used for all manner of language research (Kilgarriff and Grefenstette, 2003). But text is so abundant, that it is not so easy to find appropriate textual data for a given task. For this reason, researchers have been developing softwares that are able to crawl the Web and find sources corresponding to specific criteria. Using clustering algorithms or similarity measures, it is possible to select texts that are similar to a training set. These techniques can achieve good results, but they are sometimes limited when it comes to distinguishing between well-written texts vs. poorly written, or other subtle criteria. In any case, it will require filtering and cleaning of the data (Berland and Grabar, 2002).

One possibility to address the difficulty to find good sources is to avoid "wide crawling" but instead to bind the crawler to manually identified Web domains which are updated on a regular basis and which offer textual data of good quality (this can be seen as "vertical crawling" as opposed to "horizontal" or "wide crawling"). This is the choice made in the GlossaNet system (Fairon, 1998; 2003). This Web service gives to the users access to a linguistics based search engine for querying online newspapers (it is based on the Open Source corpus processor Unitex[3] – Paumier, 2003). Online newspapers are an interesting source of textual data on the Web because they are continuously updated and they usually publish articles reviewed through a full editorial

[1] I would like to thank CENTAL members who took part in the development and the administration of GlossaNet and those who contributed to the development of Corporator and GlossaRSS. Thanks also to Herlinda Vekemans who helped in the preparation of this paper.

[2] http://wacky.sslmit.unibo.it
[3] http://www-igm.univ-mlv.fr/~unitex/

process which ensures (a certain) quality of the text.

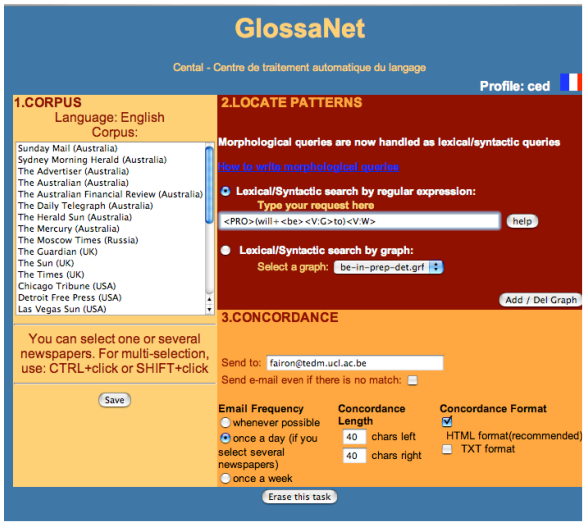

Figure 1. GlossaNet interface

GlossaNet downloads over 100 newspapers (in 10 languages) on a daily basis and parses them like corpora. The Web-based interface[4] of this service enable the user to select a list of newspapers and to register a query. Every day, the user's query is applied on the updated corpus and results are sent by email to the user under the form of a concordance. The main limitation of GlossaNet is that it works only on a limited set of sources which all are of the same kind (newspapers).

In this paper we will present a new approach which takes advantage of a very popular XML-based format used on the Web for sharing content among websites: RSS (Really Simple Syndication). We will briefly explain what RSS is and discuss its possibilities of use for building corpora.

We will also present Corporator, an Open Source program we have developed for creating RSS-fed specialized corpora. This system is not meant to replace broad Web crawling approaches but rather systems like GlossaNet, which collect Web pages from a comparatively small set of homogeneous Web sites.

2 From RSS news feeds to corpora

2.1 What is RSS

RSS is the acronym for *Really Simple Syndication*[5]. It is an XML-based format used for facilitating news publication on the Web and content interchange between websites[6]. Netscape created this standard in 1999, on the basis of Dave Winer's work on the ScriptingNews format (historically the first syndication format used on the Web)[7]. Nowadays many of the press groups around the world offer RSS-based news feeds on their Web sites which allow easy access to the recently published news articles:

FR Le monde :
http://www.lemonde.fr/web/rss/0,48-0,1-0,0.html
IT La Repubblica
http://www.repubblica.it/servizi/rss/index.html
PT Público
http://www.publico.clix.pt/homepage/site/rss/default.asp
US New York Times
http://www.nytimes.com/services/xml/rss/index.html
ES El Pais :
http://www.elpais.es/static/rss/index.html
AF Allafrica.com[8]
http://fr.allafrica.com/tools/headlines/rss.html
etc.

Figure 2. Example of RSS feeds proposed by Reuters (left) and the New York Times (right)

[4] http://glossa.fltr.ucl.ac.be

[5] To be more accurate, 'r' in RSS was initially a reference to RDF. In fact, at the beginning of RSS the aim was to enable automatic Web site summary and at that time, RSS stood for *RDF Site Summary format*. But over the time this standard changed for becoming a news syndication tools and the RDF headers were removed.

[6] Atom is another standard built with the same objective but is more flexible from a technical point of view. For a comparison, see http://www.tbray.org/atom/RSS-and-Atom or Hammersley (2005).

[7] After 99, many groups were involved in the development of RSS and it is finally Harvard which published RSS 2.0 specifications under Creative Commons License in 2003. For further details on the RSS' history, see http://blogs.law.harvard.edu/tech/rssVersionHistory/

[8] AllAfrica gathers and indexes content from more than 125 African press agencies and other sources.

Figure 2 shows two lists of RSS proposed by Reuters and the New York Times respectively. Each link points to a RSS file that contains a list of articles recently published and corresponding to the selected theme or section. RSS files do not contain full articles, but only the title, a brief summary, the date of publication, and a link to the full article available on the publisher Web site. On a regular basis (every hour or even more frequently), RSS documents are updated with fresh content.

News publishers usually organize news feeds by theme (politics, health, business, etc.) and/or in accordance with the various sections of the newspaper (front page, job offers, editorials, regions, etc.). Sometimes they even create feeds for special hot topics such as "Bird flu", in Figure 2 (Reuters).

There is a clear tendency to increase the number of available feeds. We can even say that there is some kind of competition going on as competitors tend to offer more or better services than the others. By proposing accurate feeds of information, content publishers try to increase their chance to see their content reused and published on other websites (see below §2.2). Another indicator of the attention drawn to RSS applications is that some group initiatives are taken for promoting publishers by publicizing their RSS sources. For instance, the French association of online publishers (GESTE[9]) has released an Open Source RSS reader[10] which includes more than 274 French news feeds (among which we can find feeds from *Le Monde, Libération, L'Equipe, ZDNet*, etc.).

2.2 What is RSS?

RSS is particularly well suited for publishing content that can be split into items and that is updated regularly. So it is very convenient for publishing news, but it is not limited to news.

There are two main situations of use for RSS. First, on the user side, people can use an RSS enabled Web client (usually called *news aggregator*) to read news feeds. Standalone applications (like *BottomFeeder*[11] ou *Feedreader*[12]) coexist with plug-ins readers to be added to a regular Web browser. For example, *Wizz RSS News Reader* is an extension for Firefox. It is illustrated in Figure 3: the list of items provided by a RSS is displayed in the left frame. A simple click on one item opens the original article in the right frame.

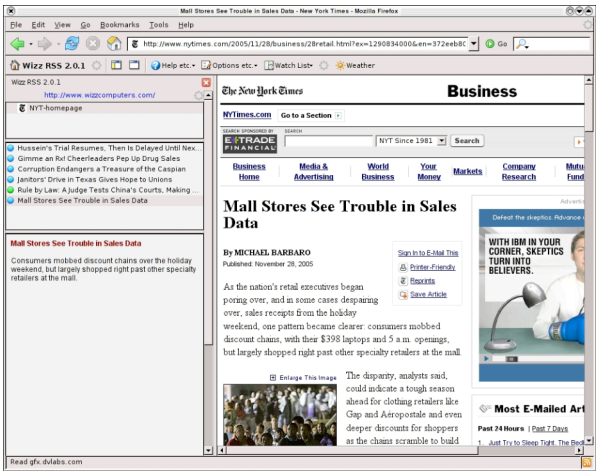

Figure 3. News aggregator plugin in Firefox

Second, on the Web administrator side, this format facilitates the integration in one Web site of content provided by another Web site under the form of a RSS. Thanks to this format, Google can claim to integrate news from 4500 online sources updated every 15 minutes[13].

2.3 How does the XML code looks like?

As can be see in Figure 4[14], the XML-based format of RSS is fairly simple. It mainly consists of a "channel" which contains a list of "items" described by a title, a link, a short description (or summary), a publication date, etc. This example shows only a subset of all the elements (tags) described by the standard[15].

```
<?xml version="1.0" encoding="iso-8859-1" ?>
<rss version="2.0">
  <channel>
    <title>NYT &gt; World Business</title>
        <item>
          <title>Russia and Ukraine Reach Compromise</title>
          <link>http://www.nytimes.com/2006/01/05/05ukraine.html</link>
          <description>The solution allowed both nations...</description>
          <author>ANDREW E. KRAMER</author>
          <pubDate>Thu, 05 Jan 2006 00:00:00 EDT</pubDate>
        </item>
    </title>
  </channel>
</rss>
```

Figure 4: example of RSS feed

2.4 Can RSS feed corpora?

As mentioned above, RSS feeds contain few text. They are mainly a list of items, but each item has a link pointing to the full article. It is therefore

[9] http://www.geste.fr
[10] AlerteInfo, http://www.geste.fr/alertinfo/home.html
[11] http://www.cincomsmalltalk.com/BottomFeeder/
[12] http://www.feedreader.com

[13] http://news.google.com
[14] This example comes from the New York Times "World Business" RSS feed and was simplified to fit our needs.
[15] It is also possible to add elements not described in RSS 2.0 if they are described in a namespace.

easy to create a kind of "greedy" RSS reader which does not only read the feed, but also download each related Web page. This was our goal when we developed Corporator, the program presented in section 3.

2.5 Why using RSS feeds?

The first asset of RSS feeds in the framework of corpus development is that they offer pre-classified documents by theme, genre or other categories. If the classification fits the researcher needs, it can be used for building a specialized corpus. Paquot and Fairon (Forthcoming), for instance, used this approach for creating corpora of editorials in several languages, which can serve as comparable corpora to the ICLE[16] argumentative essays, see section 3.1). Classification is of course extremely interesting for building specialized corpora, but there are two limitations of this asset:

- The classification is not standardized among content publishers. So it will require some work to find equivalent news feeds from different publishers. Figure 2 offers a good illustration of this: the categories proposed by Reuters and the New York Times do not exactly match (even if they both have in common some feeds like *sports* or *science*).
- We do not have a clear view on how the classification is done (manually by the authors, by the system administrators, or even automatically?).

A second asset is that RSS are updated on a regular basis. As such, an RSS feed provides a continuous flow of data that can be easily collected in a corpus. We could call this a *dynamic corpus* (Fairon, 1999) as it will grow over time. We could also use the term *monitor corpus* which was proposed by Renouf (1993) and which is widely used in the Anglo-Saxon community of corpus linguistics.

A third asset is that the quality of the language in one feed will be approximately constant. We know that one of the difficulties when we crawl the Web for finding sources is that we can come across any kind of document of different quality. By selecting "trusted" RSS sources, we can insure an adequate quality of the retrieved texts.

We can also note that RSS feeds comprise the title, date of publication and the author's name of the articles referred to. This is also an advantage because this information can be difficult to extract from HTML code (as it is rarely well structured). As soon as we know the date of publication, we can easily download only up to date information, a task that is not always easy with regular crawlers.

On the side of these general assets, it is also easy to imagine the interest of this type of sources for specific applications such as linguistic survey of the news (neologism identification, term extraction, dictionary update, etc.).

All these advantages would not be very significant if the number of sources was limited. But as we indicated above, the number of news feeds is rapidly and continuously growing, and not only on news portals. Specialized websites are building index of RSS feeds[17] (but we need to remark that for the time being traditional search engines such as Google, MSN, Yahoo, etc. handle RSS feeds poorly). It is possible to find feeds on virtually any domain (cooking, health, sport, education, travels, sciences) and in many languages.

3 Corporator: a "greedy" news agreggator

Corporator[18] is a simple command line program which is able to read an RSS file, find the links in it and download referenced documents. All these HTML documents are filtered and gathered in one file as illustrated in Figure 5.

Figure 5. Corporator Process

The filtering step is threefold:

- it removes HTML tags, comments and scripts;
- it removes (as much as possible) the worthless part of the text (text from ads,

[16] See Granger *et al*. (2002).

[17] Here is just a short selection: http://www.newsxs.com, http://www.newsisfree.com, http://www.rss-scout.de, http://www.2rss.com, http://www.lamoooche.com.

[18] Corporator is an Open Source program written in Perl. It was developed on the top of a preexisting Open Sources command line RSS reader named The Yoke. It will be shortly made available on CENTAL's web site: http://cental.fltr.ucl.ac.be.

links, options and menu from the original Web page)[19].

- it converts the filtered text from its original character encoding to UTF8. Corporator can handle the download of news feeds in many languages (and encodings: UTF, latin, iso, etc.)[20].

The program can easily be set up in a task scheduler so that it runs repeatedly to check if new items are available. As long as the task remains scheduled, the corpus will keep on growing.

Figure 6 shows a snapshot of the resulting corpus. Each downloaded news item is preceded by a header that contains information found in the RSS feed.

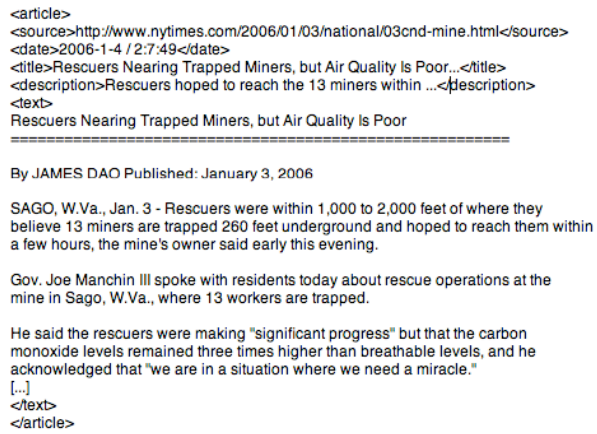

Figure 6. Example or resulting corpus

Corporator is a generic tool, built for downloading any feeds in any language. This goal of genericity comes along with some limitations. For instance, for any item in the RSS feed, the program will download only one Web page even if, on some particular websites, articles can be split over several pages: Reuters[21] for instance splits its longer articles into several pages so that each one can fit on the screen. The RSS news item will only refer to the first page and Corporator will only download that page. It will therefore insert an incomplete article in the corpus. We are still working on this issue.

3.1 Example of corpus creation

In order to present a first evaluation of the system, we provide in Figure 7 some information about an ongoing corpus development project. Our aim is to build corpora of editorials in several languages, which can serve as comparable corpora to the ICLE argumentative essays (Paquot and Fairon, forthcoming). We have therefore selected "Editorial", "Opinion" and other sections of various newspapers, which are expected to contain argumentative texts. Figure 7 gives for four of these sources the number of articles[22] downloaded between January 1st 2006 and January 31st 2006 (RSS feed names are given between brackets and URLs are listed in the footnotes). Tokens were counted using Unitex (see above) on the filtered text (*i.e.* text already cleaned from HTML and non-valuable text).

Figure 7 shows that the amount of text provided for a given section (here, *Opinion*) by different publishers can be very different. It also illustrates the fact that it is not always possible to find corresponding news feeds among different publishers: *Le Monde*, for instance, does not provide its editorials on a particular news feed. We have therefore selected a rubric named *Rendez-vous* in replacement (we have considered that it contains a text genre of interest to our study).

Le Monde[23] (*Rendez-vous*)
58 articles
90,208 tokens
New York Times[24] (*Opinion*)
220 articles
246,104 tokens
Washington Post[25] (*Opinion*)
95 articles
137,566 tokens
El Pais[26] (*Opinión*)
337 articles
399,831 tokens

Figure 7. Download statistics: number of articles downloaded in January 2006

[19] This is obviously the most difficult step. Several options have been implemented to improve the accuracy of this filter : *delete text above the article title*, *delete text after pattern X*, *delete line if matches pattern X, etc*.

[20] It can handle all the encodings supported by the Perl modules Encode (for information, see Encode::Supported on Cpan). Although, experience shows that using the Encode can be complicated.

[21] http://today.reuters.com

[22] This is the number of articles recorded by the program after filtering. It may not correspond exactly to the number of articles really published on this news feed.

[23] www.lemonde.fr/rss/sequence/0,2-3238,1-0,0.xml

[24] www.nytimes.com/services/xml/rss/nyt/Opinion.xml

[25] www.washingtonpost.com/wp-dyn/rss/index.html#opinion

[26] www.elpais.es/rss/feed.html?feedId=1003

3.2 Towards an online service

Linguists may find command line tools hard to use. For this reason, we have also developed a Web-based interface for facilitating RSS-based corpus development. GlossaRSS provides a simple Web interface in which users can create "corpus-acquisition tasks". They just choose a name for the corpus, provide a list of URL corresponding to RSS feeds and activate the download. The corpus will grow automatically over time and the user can at any moment log in to download the latest version of the corpus. For efficiency reasons, the download managing program checks that news feeds are downloaded only once. If several users require the same feed, it will be downloaded once and then appended to each corpus.

Figure 8. Online service for building RSS-based corpora

This service is being tested and will be made public shortly. Furthermore, we plan to integrate this procedure to GlossaNet. At the moment, GlossaNet provides language specialists with a linguistic search engine that can analyze a little more than 100 newspapers (as seen in Figure 1, users who register a linguistic query can compose a corpus by selecting newspapers in a predefined list). Our goal is to offer the same service in the future but on RSS-based corpora. So it will be possible to create a new corpus, register a linguistic query and get concordance on a daily or weekly basis by email. There is no programming difficulty, but there is a clear issue on the side of "scalability" (at the present time, GlossaNet counts more than 1,300 users and generates more than 18,800 queries a day. The computing charge would probably be difficult to cope with if each user started to build and work on a different corpus). An intermediate approach between the current list of newspapers and an open system would be to define in GlossaNet some thematic corpora that would be fed by RSS from different newspapers.

3.3 From text to RSS-based speech corpora

The approach presented in this paper focuses on text corpora, but could be adapted for collecting speech corpora. In fact RSS are also used as a way for publishing multimedia files through Web feeds named "podcasts". Many medias, corporations or individuals use podcasting for placing audio and video files on the Internet. The advantage of podcast compared with streaming or simple download, is "integration". Users can collect programs from a variety of sources and subscribe to them using a podcast-aware software which will regularly check if new content is available. This technology has been very successful in the last two years and has been rapidly growing in importance. Users have found many reasons to use it, sometimes creatively: language teachers, for example, have found there a very practical source of authentic recordings for their lessons.
Regarding corpus development, the interest of podcasting is similar to the ones of text-based RSS (categorization, content regularly updated, etc.). Another interesting fact is that sometimes transcripts are published together with the podcast and it is therefore a great source for creating sound/text corpora[27].

Many portals offer lists of poscast[28]. One of the most interesting ones, is Podzinger[29] which not only indexes podcasts metadata (title, author, date, etc.), but uses a speech recognition system for indexing podcast content.

It would require only minor technical adaptation to enable Corporator to deal with podcasts, something that will be done shortly. Of course, this will only solve the problem of collecting sound files, not the problem of converting these files into speech data useful for linguistic research.

4 Conclusion

Corpora uses and applications are every year more numerous in NLP, language teaching, corpus linguistics, etc. and there is therefore a growing demand for large well-tailored corpora. At the same time the Internet has grown enormously, increasing its diversity and its world

[27] It is even possible to find services that do podcast transcripts (http://castingwords.com).
[28] http://www.podcastingnews.com, http://www.podcast.net, etc.
[29] http://www.podzinger.com

wide coverage. It is now an ideal "ground" for finding corpus sources. But these assets (size, diversity) is at the same time an issue for finding good, reliable, well-written, sources that suit our needs. This is the reason why we need to develop intelligent source-finder crawlers and other softwares specialized in corpus collection. Our contribution to this effort is to bring the researchers' attention to a particularly interesting source of text on the Internet: RSS news feeds. The main interest of this source is to provide classified lists of documents continuously updated and consistent in terms of language quality.

To build specialized corpora with a traditional crawler approach, the process will probably consist in retrieving documents (using a search engine as starting point) and then sorting the retrieved documents and selecting the ones that pass some kind of validity tests. With RSS-based corpus, the approach is different and could be summarized as follows: **do not sort a list of retrieved documents, but retrieve a list of sorted documents**. This is of course only possible if we can find RSS-feeds compatible with the theme and/or language we want in our corpus.

References

Berland, Sophie and Natalia Grabar. 2002. Assistance automatique pour l'homogénéisation d'un corpus Web de spécialité. In *Actes des 6èmes Journées internationales d'analyse statistique des données textuelles (JADT 2002)*. Saint-Malo.

Fairon, Cédrick. 1999. Parsing a Web site as a corpus. In C. Fairon (ed.). *Analyse lexicale et syntaxique: Le système INTEX*, Lingvisticae Investigationes Tome XXII (Volume spécial). John Benjamins Publishing, Amsterdam/Philadelphia, pp. 327-340.

Granger, Sylviane, Estelle Dagneaux and Fanny Meunier (eds). 2002. *The International Corpus of Learner English*. CD-ROM and Handbook. Presses universitaires de Louvain, Louvain-la-Neuve.

Hammersley, Ben. 2005. *Developing Feeds with RSS and Atom*. O'Reilly, Sebastopol, CA.

Kilgarriff, Adam and Gregory Grefenstette. 2003. Introduction to the Special Issue on the Web as Corpus. *Computational Linguistics*, Vol. 29(3): 333-348.

Paquot, Magali and Cédrick Fairon. (forthcoming). Investigating L1-induced learner variability: Using the Web as a source of L1 comparable data.

Paumier, Sébastien. 2003. *De la reconnaissance de formes linguistiques à l'analyse syntaxique*, Ph.D., Université de Marne-la-Vallée.

Renouf, Antoinette. 1993. 'A Word in Time: first findings from the investigation of dynamic text'. In J. Aarts, P. de Haan and N. Oostdijk (eds), *English Language Corpora: Design, Analysis and Exploitation*, Rodopi, Amsterdam, pp. 279-288.

Renouf, Antoinette. 2003. 'WebCorp: providing a renewable energy source for corpus linguistics'. In S. Granger and S. Petch-Tyson (eds), *Extending the scope of corpus-based research: new applications, new challenges*, Rodopi , Amsterdam, pp. 39-58.

Sekiguchi, Youichi and Kazuhide Yamamoto. 2004. 'Improving Quality of the Web Corpus'. In *Proceedings of The First International Joint Conference on Natural Language Processing (IJCNLP-04)*, pp. 201-206.

Emirkanian Louisette, Christophe Fouqueré and Fabrice Issac. 2004. Corpus issus du Web : analyse des pertinences thématique et informationnelle. In G. Purnelle, C. Fairon and A. Dister (eds), *Le Poids des mots. Actes des 7èmes Journées internationales d'analyse statistique des données textuelles (JADT 2004)*, Presses universitaires de Louvain, Louvain-La-Neuve, pp. 390-398.