Pacific University
CommonKnowledge

Volume 9 (2009)

Interface: The Journal of Education, Community and Values

4-1-2009

An RSS Feed Analysis Application and Corpus Builder

Shereen Khoja
Pacific University

An RSS Feed Analysis Application and Corpus Builder

An RSS Feed Analysis Application and Corpus Builder

Posted on **April 1, 2009** by **Editor**

By Shereen Khoja, Pacific University Oregon

1. Introduction

This article describes a software application that downloads given RSS feeds and compiles them into a corpus. The user simply supplies RSS feed addresses and the application automatically connects to the feeds, downloads them and strips any formatting tags. The application incorporates the Expat () XML parser to identify the tags in the RSS feeds, and the user has the flexibility to define what they would like to keep and what is to be stripped [1]. The application was tested on a project to analyze Middle-Eastern blogs. Thirty-seven blogs were downloaded using the RSS Feed Analyzer and compiled into a corpus of 131,836 words. Both the RSS Feed Analyzer and corpus are freely available under the GNU General Public Licence.

2. Motivation

The motivation behind building an application to automatically download RSS feeds is the frustration that I feel working in the field of Arabic Computational Linguistics in the lack of corpus creation and analysis tools. I know that I am not alone in feeling this frustration, and researchers use various ad-hoc techniques to get through this basic step. As quoted by Al-Sulatie and Atwell [2]:

".. it is still difficult to use corpus analysis tools such as concordancers in handling Arabic text unless they are used in Arabic windows, and even so the result is not as tidy as in the case of languages with Roman script. Since our corpus will be available on the Internet we hope it would be an interesting challenge for software engineers to develop suitable analysis tools."

My research in Arabic Computational Linguistics has focused on Modern Standard Arabic, the Arabic used in most of the written forms of the text, as well as the Arabic used by newscasters on television. This form of Arabic is understood by everyone in the Arab world but it is not the

form Arabs use when speaking to each other. Instead, they use a colloquial form that differs from country to country, and in fact region to region. An Arabic speaker from Saudi Arabia would have a difficult time understanding the colloquial Arabic being spoken by a speaker from Morocco, and in these situations, both speakers will resort to the Modern Standard form of Arabic. This is known as *Diglossia* in the field of linguistics [3].

It is the colloquial form of Arabic and developing linguistic tools for this form of Arabic that is of interest to me now since research on Modern Standard Arabic is now quite prolific. My first step in investigating colloquial Arabic is to collect a corpus of Arabic blogs.

The growth in blogging has provided language researchers with a new form of writing to investigate various linguistic properties. Examples of this research include: an investigation to determine the mood of a blogger based on the text of a post [4], an analysis of the genre of blogs [5], and the development of an automated trend discovery for blogs [6].

There has been little to no research on Arabic blogs and no corpus of Arabic blogs currently exists. Compiling a corpus of Arabic blogs is the first step in analyzing the Arab blogosphere and investigating trends within that community.

3. The RSS Feed Analysis Application

An RSS Feed is an XML document used to publish frequently updated web content such as blogs. The format of these documents is mostly standardized, though there are exceptions. RSS feeds contain metadata such as information about the author and the host, and site data such as a blog item or post, and user comments on that post. An example of a blog and its RSS feed are shown in figures 1 and 2 below.

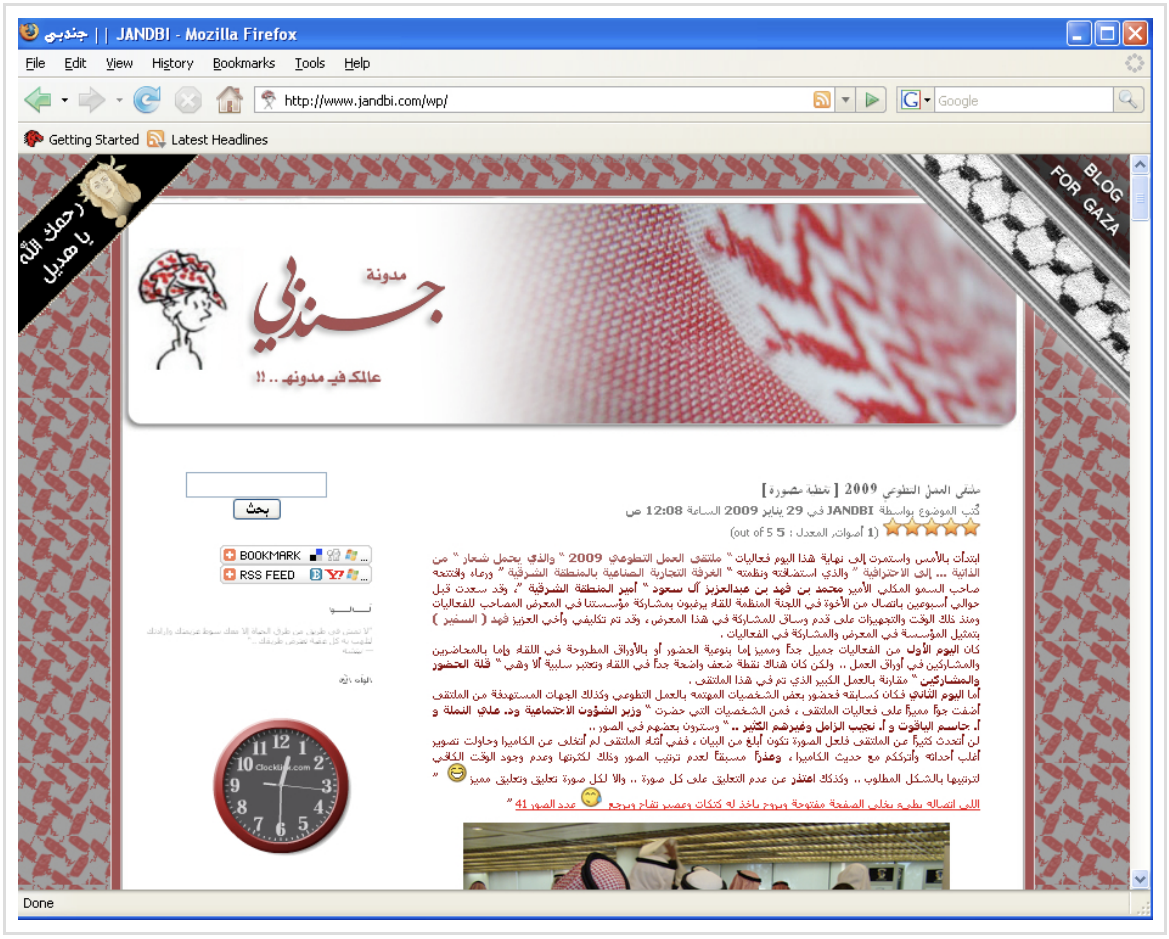

Figure 1: An Arabic Blog Site

```
<?xml version="1.0" encoding="UTF-8"?>
<rss version="2.0"
        xmlns:content="http://purl.org/rss/1.0/modules/content/"
        xmlns:wfw="http://wellformedweb.org/CommentAPI/"
        xmlns:dc="http://purl.org/dc/elements/1.1/"
        xmlns:atom="http://www.w3.org/2005/Atom"
        >

<channel>
        <title>124#&;124#& جندبي; JANDBI</title>
        <atom:link href="http://www.jandbi.com/wp/?feed=rss2" rel="self"
type="application/rss+xml" />
        <link>http://www.jandbi.com/wp</link>
        <description>مدونة .. عالك في .. جندبي .. !!</description>
```

Figure 2: The Source of an RSS Feed

The RSS Feed Analysis Application begins by reading in a file of type "*cbl*," which is an XML file that contains parameters for the application. "*cbl*" is short for Corpus Builder Language and is an XML format specifically created for the RSS Feed Analysis Application. An example of such a file is:

```
<?xml version="1.0" encoding="UTF-8"?>
<cbl version="1.0">
  <builder>
    <source>
      <outputName>out_jandbi</outputName>
      <dataAdaptor>RSS</dataAdaptor>
      <location>http://www.jandbi.com/wp/?feed=rss2</location>
      <metadata>
        <author>رائد</author>
        <gender>male</gender>
        <country>Saudi Arabia</country>
      </metadata>
    </source>
  </builder>
  <ingester databaseName="db.bin">
    <source name="out_jandbi" type="XML">
    <contentTag sourceTag="content:encoded"/>
  </source>
  </ingester>
</cbl>
```

Figure 3: An Example "cbl" File

The "*cbl*" file consists of two parts: the *Builder* and the *Ingester*, which are used by the parts of the application with the same name. The Builder uses all of the sources between the tags to download the RSS feeds. The example file above contains only one source, but many of these can be described in the "*cbl*" file. No analysis is done at this point.

Source contains a name for the output file, a data adaptor, a location, and metadata. The output file name is used as the name of an XML file produced by the builder that lists the location of the downloaded RSS feed and the source of the feed. The data adaptor is the type of input used to build the corpus. Currently the only acceptable type is RSS, but the application is designed in a modular way so that handlers for other adapters such as email or message boards could be added. The location is the URI (Uniform Resource Identifier) of the RSS feed. Finally, the metadata can be anything that the user desires. In the example I have here, I am specifying the author, gender and country of the blogger. This information is not found within the RSS feed and must be added manually.

The Builder

The builder uses the information in the "*cbl*" file to produce three files for each specified source. These files are the "document" file shown below and two other files: the data file is a copy of the raw RSS feed, and the info file contains information obtained from the server such as the date downloaded, the date last modified and the content type.

```
<?xml version="1.0" encoding="UTF-8"?>
<Doc>
  <DocumentDataFile>.\s3e8.data.raw</DocumentDataFile>
  <DocumentInfoFile>.\s3e8.info.raw</DocumentInfoFile>
  <SourceLocation>http://www.iandbi.com/wp/?feed=rss2</SourceLocation>
  <UnitLocation>http://www.iandbi.com/wp/?feed=rss2</UnitLocation>
</Doc>
```

Figure 4: An Example Document File

The Ingester

The Ingester analyzes the downloaded RSS feed to produce the corpus. The second part of the "*cbl*" file contains an ingester tag, which contains the information used to build the corpus. This includes the source that points to the document file produced by the Builder. The ingester tag also contains a content tag that lists the tags that are going to be used to build the corpus. In the "*cbl*" file shown in Figure 3 above, only one tag is used, which is the "content:encoded" tag. Many of these tags could be listed, for example if the user wanted the corpus to contain the title of the article as well as the content. The RSS analysis application is developed in such a way as to give the user flexibility in what they want downloaded and stored in the corpus.

Running the Ingester on the "*cbl*" and "document" file produces the raw text that is used to build the corpus. Each word is placed on its own line to aid in the analysis. A sample of the output is shown here:

وقد
سعدت
قبل
حوالي
أسبوعين
باتصال
من
الأخوة
في
اللجنة
المنظمة
للقاء
يرغبون
بمشاركة
مؤسستنا
في
المعرض
المصاحب
للفعاليات
ومنذ
ذلك

Figure 5: The Arabic Corpus Produced by the RSS Feed Analysis Application

The Corpus

A corpus of Arabic blog was compiled using the RSS Feed Analysis Application.

Before the corpus could be compiled, I needed to determine the blogs that would be included in the corpus. Since I was also interested in the community of Arab bloggers, I decided to compile my corpus around an incident that seemed to unite and shock Arab bloggers from many different countries. This incident was the death of female Saudi Arabian journalist and blogger, Hadeel Alhodaif, who died at the age of twenty-five after slipping into a coma for a month [7]. Hadeel fought for a freer media in Saudi Arabia, and her death made national and international news. After Hadeel slipped into a coma, her father posted a notice on her blog and many Arabic bloggers posted notices and prayers on their blogs.

Blogs were chosen to be included in the blog by searching for the name (Hadeel Alhodaif) in Arabic on Google Blog Search [8].

Thirty seven blogs were collected and the corpus contained 131,836 words. A third of these blogs were written by female bloggers, and the bloggers were from seven countries. The divide is shown below.

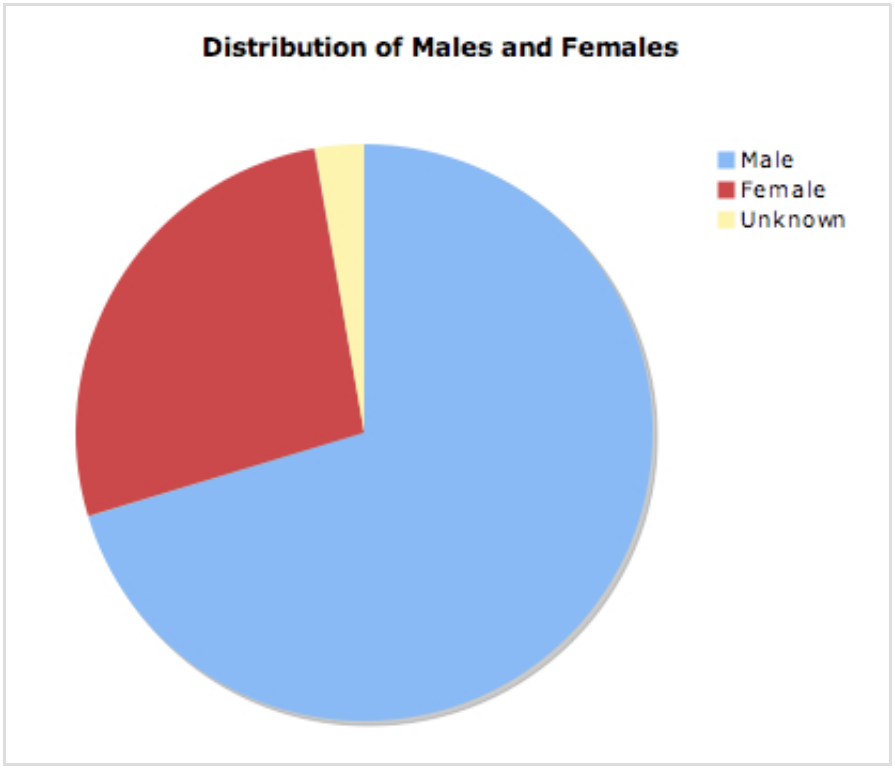

Figure 6: Distribution of Male and Female Bloggers

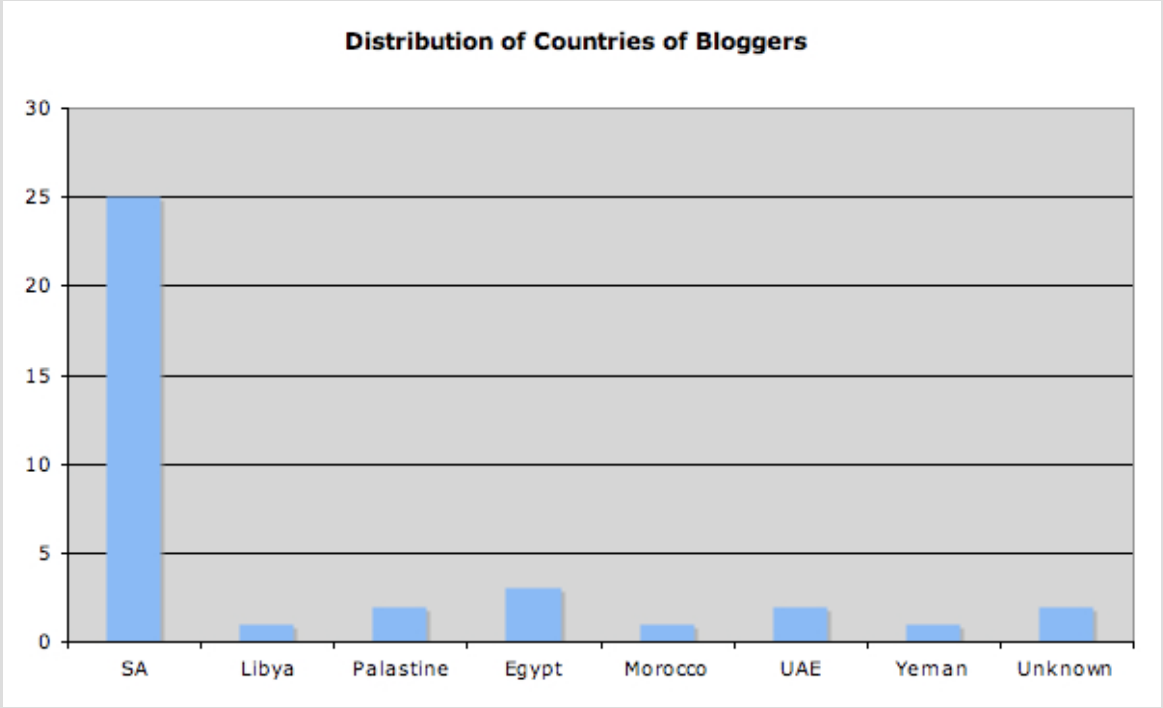

Figure 7: Countries of Origin of the Bloggers

The corpus could also be analysed language specific patterns. For example, do the bloggers use Modern Standard Arabic colloquial Arabic or a mixture of both? Sociolinguists may be interested in how bloggers from different countries or genders use the language.

Future Work

The RSS Feed Analysis Application is the first step in developing a multi-purpose corpus creation and analysis application. This application would aid researchers in compiling large corpora from multiple sources and provide basic analysis modules such as stemmers, taggers, and concordancers. Initially these will be targeted for the Arabic language, and support for other languages will be added as required.

The application will be provided to the community for feedback on the various modules that could be incorporated.

Summary

This article described an RSS Feed Analysis Application that is freely available to researchers under the GNU General Public License and can be obtained by contacting the author. The application will run on Windows and can be easily modified to run on Linux and Mac OS X. All the users need to do is use the Corpus Builder Language to specify the RSS feeds they wish to download and the tags they would like included in the corpus.

The application could be used to compile a corpus of RSS of any language simply and quickly for analysis. For example, linguists could investigate multilingual web-based discourse as described in

Using Corpora in Discourse Analysis [9].

Furthermore, the 131,836 word corpus is also freely available to the community and can be obtained by contacting the author.

I am grateful to the Berglund Center for Internet Studies, Pacific University Oregon, for supporting this work.

References

[1] http://expat.sourceforge.net/

[2] Al-Sulaiti, Latifa; Atwell, Eric. 2004. Designing and Developing a Corpus of Contemporary Arabic. In Proceedings of the sixth TALC conference. Granada, Spain, p.92

[3] Walters, K. 1996. Diglossia, linguistic variation, and language change in Arabic. In Perspectives on Arabic linguistics VIII: Papers from the Eighth Annual Symposium on Arabic Linguistics, M. Eid (ed.), 157-197. Amsterdam: Benjamins.

[4] Mishne, G. and de Rijke, M. 2006. Capturing global mood levels using blog posts. In Computational Approaches to Analyzing Weblogs: Papers from the 2006 AAAI Spring Symposium. Menlo Park, Calif.: AAAI Press, pp. 145-152.

[5] Herring, S.C., Scheidt, L.A., Bonus, S., and Wright, E. 2004. Bridging the gap: A genre analysis of weblogs. Proceedings of the 37th Annual Hawaii International Conference on System Sciences (HICSS '04). Los Alamitos, Calif.: IEEE Press.

[6] Glance, N., Hurst, M., and Tomokiyo, T. 2004. BlogPulse: Automated trend discovery for weblogs. Proceedings of the First Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics.

[7] http://www.timesonline.co.uk/tol/news/world/middle_east/article3961731.ece

[8] http://blogsearch.google.com/

[9] Baker, P. 2006. Using Corpora in Discourse Analysis. London – New York: Continuum.

This entry was posted in Uncategorized by **Editor**. Bookmark the **permalink [http://bcis.pacificu.edu/interface/?p=3602]** .

6 THOUGHTS ON "AN RSS FEED ANALYSIS APPLICATION AND CORPUS BUILDER"

Zielona Góra
on **January 30, 2014 at 7:10 AM** said:

Thank you for sharing great information. Your internet site is quite cool. I am impressed by the information that you've on this site. It reveals how nicely you understand this subject. Bookmarked this web site page, will occur back for additional articles.

naija
on **January 30, 2014 at 1:57 PM** said:

upset to your huge analysis, but I'm incredibly loving the post, and hope this, and also the very good review some other people have written, will assist you decide if it is the correct option for you.

telewizja
on **February 1, 2014 at 3:53 AM** said:

you have any? Kindly permit me realize so that I may well just subscribe. Thanks. likewise conceive so , perfectly written post! .

social network
on **February 2, 2014 at 3:27 PM** said:

it is often good to see these details in your post, i was searching precisely the same but clearly there was hardly any appropriate resource, thanx now i've the connection that we wanted my research.

dating online
on **February 3, 2014 at 1:32 AM** said:

I have mastered some significant items via your web site post. One other subject I wish to talk about is that there are lots of games obtainable on a industry created especially for toddler age children. They include pattern acceptance, colors, family pets, and shapes. These generally focus on familiarization as an selection to memorization. This keeps little kids engaged without having sensing like they're studying. Thanks

temat
on **February 3, 2014 at 1:55 AM** said:

Hurrah, that is certainly what I was seeking for, what a information! produce the following at this website, thanks admin of this website.