# Estrategias para la elaboración de corpus comparables a partir de la web

Isaac José González López, Pablo Gamallo Otero

Universidade de Santiago de Compostela

http://gramatica.usc.es/pln

## Resumen

Este artículo trata la creación de corpus comparables desde la web, utilizando tecnologías de reciente expansión en internet como los RSS o las bases de datos en XML, como es el caso de la Wikipedia. Permitiendo la obtención de grandes cantidades de corpus, en varios idiomas y con cierto grado de comparabilidad.

## 1. Introducción

La creación de corpus es un proceso que ha sido ampliamente documentado, pero desde la aparición y popularización de internet y de la web como su forma de exposición y consulta pública de información, esencialmente texto en sus inicios y texto y multimedia en la actualidad, se ha ampliado el horizonte de creación de corpus, trasladándose una parte importante de la creación de corpus a la web (Adam Kilgarrif, 2003).

La web como fuente de corpus tiene muchas ventajas, principalmente cuantitativas, pues la cantidad de texto a nuestra disposición a través de internet nunca antes en la historia había sido tan grande ni con un patrón de crecimiento tan elevado. Por otro lado, la creación de corpus clásica necesitaba de gran cantidad de recursos para generar poco corpus; hoy en día, la red optimiza esta tarea, ya que el software se encarga de la parte más pesada. En contraposición, la inmensidad de la red, su anarquía y heterogeneidad intrínseca, la confluencia de muchos idiomas, alfabetos y la propia complejidad técnica de la web, hacen que su transformación en corpus adquiera cierta complejidad, proporcional a las características que el corpus deba cumplir.

Las necesidades de uso del corpus marcan las características que éste debe tener y, asimismo, esas características serán los requisitos que marcarán la estrategia para su

obtención. Por tanto, las estrategias seguidas, aunque aplicables en otras situaciones, parten de unas necesidades de uso y contextuales. Dado que los corpus van a ser utilizados para extracción automática, hay varias características que tendrán que cumplir: grandes cantidades de corpus y corpus comparables (textos en varias lenguas con temática similar)(Gamallo&Pichel 2008) en diversas lenguas (gallego, portugués, español, francés e inglés). El requisito de grandes cantidades de corpus se hace crítico para el ámbito del gallego, que aunque muy presente en la red la cantidad de webs es mucho menor que la de otras lenguas con más masa de hablantes, lo que condicionará en parte las herramientas diseñadas. Relativo a la necesidad de comparabilidad, ésta deshecha de entrada la obtención "bruta" de corpus y nos encamina a la necesidad de elección de fuentes y categorización de los textos. Además, debido a la dificultad de encontrar corpus libres, tanto las herramientas como los corpus se distribuirán con licencias libres (siempre que esto sea posible).

Teniendo en cuenta las necesidades anteriores, la estrategia de creación de corpus tenderá a seleccionar fuentes textuales manualmente en vez de intentar una búsqueda bruta por enlaces en un típico proceso de barrido de la web por medio de enlaces (*crawling*) habitual en buscadores. La selección manual de fuentes permite por un lado dejar prácticamente de lado los problemas de selección por idioma en el caso de *crawling* bruto y por otro acotar la comparabilidad de los corpus. La transformación de las fuentes en la web a corpus, teniendo en cuenta que éstas han sido preseleccionadas manualmente, se puede hacer con software *ad hoc* (crear un software de verificación de patrones de texto y un explorador de enlaces para cada web) o ayudándose de algunos de los diferentes modos de presentación de la web que existen. Como se puede ver en la figura 1, las webs pueden disponer de diferentes modos de acceso para su consulta, como por ejemplo la consulta a través de un navegador, mediante tecnología RSS, en versión para impresión, en versión XML u otro tipo de base de datos, versión móvil y otros sistemas menos comunes. Teniendo en cuenta el uso extensivo que en la internet actual tiene el RSS y su facilidad para tratarlo y para convertir su contenido en texto nos hemos decantado por esta opción de manera preferente.

*Figura 1: Algunas maneras de acceder a los contenidos de la web*

Este artículo trata de exponer una aproximación a la creación de corpus con la web, aprovechando herramientas paralelas a la web, como RSS y versiones XML, en un la búsqueda de corpus multilingüe comparables.

Las fuentes de corpus han de ser bien escogidas, y de esa elección han de partir las herramientas encargadas de la creación de corpus. Tres grupos han sido definidos para crear corpus, por un lado blogs en gallego para disponer de un corpus de uso coloquial, por otro, periódicos online en diversos idiomas y en los que la comparabilidad viene dada por la propia categorización de los periódicos y por, a priori, estar en el mismo contexto lingüístico y temporal. Por último, la Wikipedia como fuente de corpus, que aporta ingentes cantidades de texto en prácticamente cualquier lengua y con una estructura más o menos estándar. Tanto los blogs, como los periódicos y la Wikipedia disponen de RSS, por lo que ésta sería en principio el modo de hacer *crawling* (es decir, el modo de exploración de documentos), pero teniendo en cuenta que la Wikipedia es ofrecida como base de datos en varias versiones de XML (con más o menos contenidos), escogeremos este formato para la Wikipedia y el RSS para blogs y

periódicos.

## 2.Las tecnologías de subscripción RSS

La tecnología RSS (*Really Simple Syndication*) se basa en ficheros xml estandarizados que se van modificando en tiempo real y que permiten monitorizar el contenido de un sitio web o parte de él (Fairon et al. 2008). Además el *parsing* es generalizable a cualquier sitio web, ya que la información está contenida en una etiqueta del RSS, separada del resto de la página; lo que supone una gran ventaja con respecto a la presentación habitual de las páginas. Se puede observar en el siguiente ejemplo de una entrada de un feed RSS, la sencillez de acceso a los datos en contraste con la que tendríamos con el texto en html de toda la web, información repetida y no delimitada.

```
<item>
    <title>Urgentemente</title>
    <link>http://calidonia.blogaliza.org/2009/12/24/urgentemente/</li
nk>
    <comments>http://calidonia.blogaliza.org/2009/12/24/urgentemente/
#comments</comments>
    <pubDate>Thu, 24 Dec 2009 11:48:39 +0000</pubDate>
    <category><![CDATA[poesia]]></category>
    <guid isPermaLink="false">http://calidonia.blogaliza.org/?
p=1465</guid>
    <description>
            É urgente o amor.
            É urgente um barco no mar.
            É urgente destruir certas palavras,
            ódio, solidão e crueldade,
            alguns lamentos,
            muitas espadas.
            É urgente inventar a alegría,
            multiplicar os beijos, as searas,
            é urgente descobrir rosas e rios
            e manhãs claras.
            Cai o silêncio nos ombros e a luz
            impura, até doer.
            É urgente o amor, é urgente
            permanecer.
            Eugénio DE ANDRADE
            Até amanhã
            Feliz 2010
    </description>
</item>
```

*Figura 2: Ejemplo de una entrada de un feed RSS*

El formato en el que los corpus serán guardados es también en XML, por lo que la estructura se parece mucho a la del feed que acabamos de ver, como se pude comprobar

en el siguiente ejemplo:

```
<item>
      <titulo>A HORA DOS AGASALLOS</titulo>
      <contido>Na recén estreada película, Anjé, Marí e todos os seus
amigos terán que traballar duro se queren que a paz volva á aldea. Os
máis vellos están incomodados entre eles, desconfiados. A culpa de
todo tena un estranxeiro, que tivo un accidente de avión nas montañas
do redor. Prometeu diñeiro e sona para quen lle atope o avión.Pero as
cousas hanse enguedellar aínda máis cando o reloxo máxico de
Olentzero desapareza. Sen el non poderá parar o tempo e, por iso, non
dará repartido os agasallos na noite de Noiteboa.Fontes: Novas da
Xunta de Galicia e Observatorio da Lingua Galega</contido>
      <url>http://biblospazos.blogspot.com/2008/12/hora-dos-
agasallos.html</url>
      <data>Thu Jan 15 18:26:43 +0100 2009</data>
</item>
```

*Figura 3: Ejemplo del formato del corpus*

Algunos generadores de contenido en la red prefieren que el acceso RSS a su contenido sea parcial, para no perder los ingresos publicitarios de las visitas en sus webs; esta práctica está convertida en norma para los periódicos digitales en los que su RSS tiene un brevísimo resumen del contenido. Para solventar este inconveniente, el software deberá acceder a la versión web de los periódicos y extraer la parte de la web que interesa ser añadida al corpus.

```
<item>
    <title>Sube el número de nuevas hipotecas por primera vez en dos
años</title>
    <link>http://www.xornal.com//artigo/2010/01/26/economia/sube-
numero-viviendas-hipotecadas-primera-vez-
anos/20100126112656009 48.html</link>
    <description><![CDATA[El tipo de inter&eacute;s medio de los
pr&eacute;stamos hipotecarios se situ&oacute; en el 4,09%, un 26,7%
inferior al de un a&ntilde;o atr&aacute;s]]></description>

    <guid
isPermaLink="true">http://www.xornal.com//artigo/2010/01/26/economia/
sube-numero-viviendas-hipotecadas-primera-vez-
anos/20100126112656009 48.html</guid>
    <author><![CDATA[Xornal.com]]></author>
    <pubDate><![CDATA[Tue, 26 Jan 2010 11:26:56 +0100]]></pubDate>
</item>
```

*Figura 4: Ejemplo de entrada RSS resumida de un periódico*

# 3. Wikipedia

La Wikipedia es descargable en su totalidad en ficheros XML, en distintas versiones en la que se pude escoger que cantidad de metadatos descargar. Como se puede ver en el ejemplo de una entrada de la Wikipedia que se muestra en la figura 5; la forma de acceso a los datos es también simple y muy eficiente, con la diferencia respecto a los RSS y a la web al uso, que el texto en vez de tener un formato plano, html o xhtml, tiene un formato propio de la Wikipedia, que ha de ser convertido a texto sin formato.

```
<page>
    <title>Arqueoloxía</title>
    <id>3</id>
    <revision>
      <id>1310468</id>
      <timestamp>2009-10-06T02:42:14Z</timestamp>
      <contributor>
        <username>SieBot</username>
        <id>2109</id>
      </contributor>
      <minor />
      <comment>bot Engadido: [[ku:Arkeolojî]]</comment>
      <text xml:space="preserve">{{Historia en progreso}}

A '''arqueoloxía''' é a [[ciencia]] que estuda as [[arte|artes]],
[[monumento|monumentos]] e [[obxecto]]s da [[antigüidade|
antigüidade]], especialmente a través dos seus restos. O nome ven do
[[lingua grega|grego]] ''archaios'', &quot;vello&quot; ou
&quot;antigo&quot;, e ''logos'', &quot;ciencia&quot;,
&quot;saber&quot;.

[...]

[[zh:□□□]]
[[zh-yue:□□]]</text>
    </revision>
  </page>
```

*Figura 5: Ejemplo del formato XML de la Wikipedia*

Siendo un formato tan rico en posibilidades, la construcción de corpus se antoja el primer paso para explotar las posibilidades de extracción de información de la Wikipedia, por lo que el formalismo que describimos a continuación contendrá algunos campos que exceden el objetivo de éste artículo (Clark et al. 2009). En la figura 6 podemos observar el código XML del corpus generado a partir de la Wikipedia. Los campos *title*, *category* y *plaintext* son los necesarios relativos al uso de este fichero como corpus comparable, siendo el apartado *category* lo que puede aportar información

del ámbito del texto y así agruparlo con otros textos de categoría similar en caso de búsqueda de alta comparabilidad. El resto de campos pertenecen a la extracción de características de cada entrada que posibiliten otros usos, como la traducción del término, búsqueda de artículos relacionados o enlazados, uso del formato de la Wikipedia y otros.

```
<article>
      <title>Arqueoloxía</title>
      <category>Arqueoloxía</category>
      <related>Antropoloxía, Arqueoloxía industrial, Arqueoloxía
submarina</related>
      <links>ciencia, arte|artes, monumento|monumentos, obxecto,
antigüidade|antigüidade, lingua grega|grego, cultura, estudo,
psicolóxico, condutistas, antropoloxía, idade de pedra, Idade Media,
Arqueoloxía industrial, Antropoloxía, Arqueoloxía industrial,
Arqueoloxía submarina</links>
      <translations>#   Arqueologia Arqueología Archaeology
      Archéologie Arqueologia Arkeologia  #     Archeologia
      Archeologie Археология  Αρχαιολογία</translations>
      <plaintext>A arqueoloxía é a ciencia que estuda as artes,
monumentos e obxectos da antigüidade, [...] o que se coñece como
Arqueoloxía industrial.</plaintext>
      <wikitext>{{Historia en progreso}}

A '''arqueoloxía''' é a [[ciencia]] que estuda as [[arte|artes]],
[[monumento|monumentos]] e [[obxecto]]s da [[antigüidade|
antigüidade]], [...]
[...]
[[yi:ארכעאלאגיע]]
[[zh:□□□]]
[[zh-yue:□□]]</wikitext>
</article>
```

*Figura 6: Formato del corpus generado a partir de la Wikipedia*

# 4.Las herramientas de generación de corpus

Para conseguir extraer corpus de los tres tipos de fuentes que hemos definido se han desarrollado tres herramientas: AgregadorSimple, RSScrawler y CorpusPedia. Cada una de ellas responde a unas necesidades diferentes de procesado de texto, como se puede ver en la Figura 7.

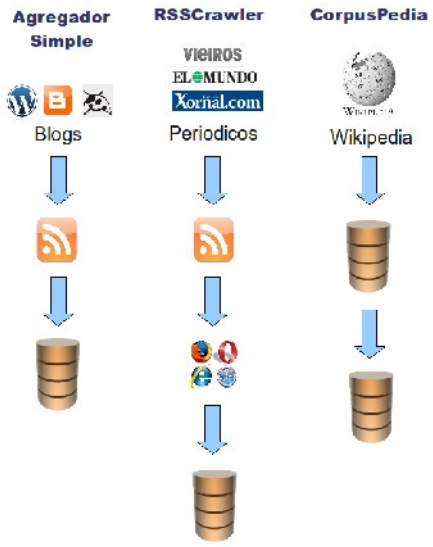


*Figura 7: Los tres métodos de obtención de corpus de las herramientas*

El AgregadorSimple a partir de una lista de urls de ficheros RSS genera el corpus; éste es organizado de tal modo que cada fichero del corpus se corresponde a una fuente RSS, pudiendo así crear subgrupos de corpus comparables.

El RSSCrawler obtiene, para cada fuente RSS, una lista de urls web en la que está el texto a extraer. Después, con la ayuda de los ficheros de configuración, identificará qué parte de la web se debe añadir al corpus. Existe pues, una configuración *ad hoc* para cada fuente RSS, en la que se debe indicar el nombre, el/los idiomas de la fuente, lo/los RSS fuente y las rutas XPATH para extraer el contenido.

CorpusPedia se encarga de descargar automáticamente la Wikipedia en los idiomas requeridos y aplicar después un proceso de conversión del XML descargado al XML formato del corpus.

## 5. Los corpus

Los corpus generados con el AgregadorSimple para blogs en gallego, a partir de 2691 fuentes RSS, ha generado 23 millones de palabras en aproximadamente un año de ejecución, con crecimiento a razón de un millón de palabras por mes.

Los corpus generados con RSScrawler han generado en 11 semanas un corpus de 6 millones de palabras en gallego y 25 millones en español.

CorpusPedia, a diferencia de los corpus anteriores no se va actualizando con ejecuciones diarias, sino que genera todo el corpus una sola vez, que en el momento actual es de aproximadamente 20 millones de palabras para la versión en gallego de la Wikipedia, 120 millones en la versión portuguesa y 180 en la versión en español.

## 6. Conclusiones y trabajo futuro

La expansión en la red de tecnologías orientadas a la suscripción (RSS) o la abertura de las bases de datos (Wikipedia), posibilitan nuevos métodos de creación de corpus a partir de la web, que son más eficientes y potentes que los tradicionales. Asimismo permiten crear corpus comparables a partir de la extracción de la o las temáticas de cada porción de texto.

La ampliación de características de las herramientas de generación de corpus presentas se debería encaminar a posibilitar la creación de subconjuntos del corpus total que sean más comparables, a partir de la categorización actual de cada entrada del corpus, permitiendo así tener corpus más pequeños pero mucho más comparables.

## 7. Referencias


Kilgarrif, Adam & Gregory Grefenstette (2003) “Introduction to the Special Issue on the Web as Corpus”, Computational Linguistics, 29, (333-347)

Clark M, Ian Ruthven & Patrik O’Brian Holt (2009), “The Evolution of Genre in Wikipedia”, *JLCL,* vol 24 (1), (1-22)

Fairon Cédrick, K´evin Mac´e & Hubert Naets (2008), “GlossaNet 2: a linguistic search engine for RSS-based corpora”, *Proceedings of LREC 2008. Workshop WAC4.*, Marrakesh.

Gamallo P. & J-R. Pichel (2008) "Learning Spanish-Galician Translation Equivalents Using a Comparable Corpus and a Bilingual Dictionary", *LNCS*, vol. 4919, Springer-Verlag, (423-433)