

Diseño de Arquitectura de Información para ToNgueLP

Ing. Abel Meneses Abad

20 de agosto de 2014

CEETI, Facultad de Eléctrica, Universidad Central Marta Abreu de Las Villas, Santa Clara, Cuba.

Resumen

Aquí se encuentran las principales herramientas para el tratamiento de corpus textuales estudiados durante mi doctorado.

Palabras clave:

corpus, corpus de texto, procesamiento de lenguaje natural

Índice

1. Introducción	1
2. Estudio de homólogos	2
2.1. AntConc	2
2.2. UAM CorpusTool	4
2.2.1. Search View	5
2.2.2. Statistics View	6
2.3. WordSmith Tools	7
2.4. RSTTools	8
2.5. Nooj	9
3. Conclusiones	9

1. Introducción

El presente resumen es un acercamiento a las herramientas de procesamiento de textos orientadas al trabajo con corpus lingüísticos. Estos deben ser anotados con información Léxica, Sintáctica o Semántica para el análisis de relaciones o el hallazgo de concordancias entre estructuras previamente anotadas.

Estas herramientas están orientadas al trabajo de los lingüistas para facilitar la búsqueda de información, estadísticas de los documentos o del corpus, así como análisis propio del procesamiento de las lenguas naturales o formales.

2. Estudio de homólogos

El estudio de homólogos es uno de los primeros procesos que se debe realizar al concebir un proyecto. En particular casi ninguna esfera del conocimiento científico ha quedado sin tratar hoy en día, por lo tanto antes de comenzar un proyecto desde cero (y luego ver que estabas descubriendo el agua tibia) debes estudiar los software similares para tomar ideas.

En particular este breve estudio está centrado en herramientas para el tratamiento de corpus lingüísticos. En particular se escogen los siguientes parámetros para compararlos:

- Existencia de una URL pública y activa.
- Licencia de software.
- Descripción de funcionalidades.
- Compatibilidad con los Sistemas operativos.
- Existencia de pantallas gráficas (y que no sean de consola).

2.1. AntConc

URL: <http://www.antlab.sci.waseda.ac.jp/index.html>

Versión: 3.2.4

Licencia: No se especifica.

Descripción: Software orientado al análisis de corpus. Muestra resumen de la cantidad de palabras, collocations, n-grams y concordancias.

Sistemas Operativos: Multiplataforma

ScreenShot:

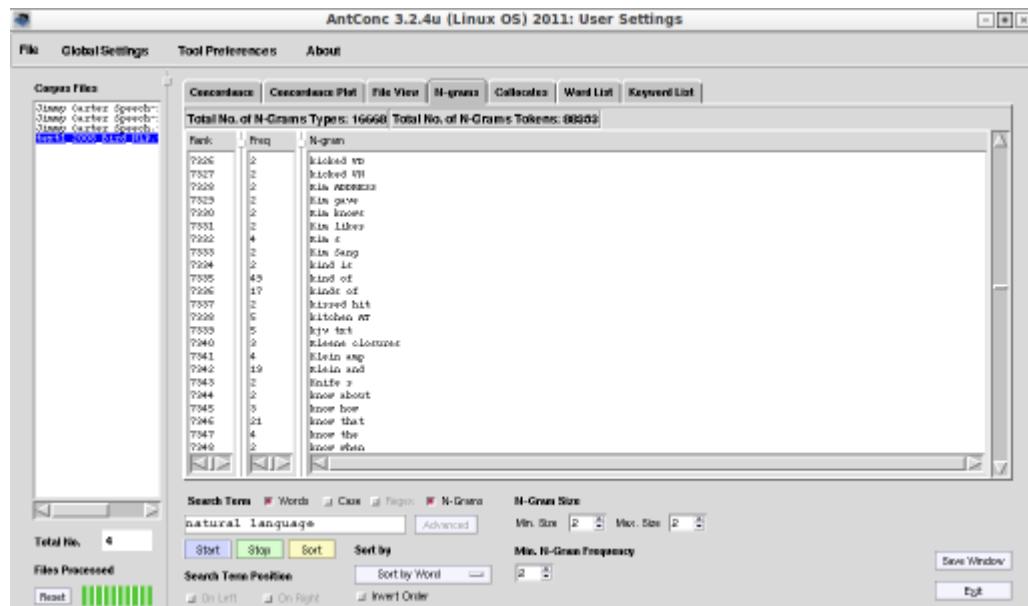


Figura 1: Captura de pantalla de AntConc 3.2.4, tomado en Ubuntu 12.04
Accedido el: 24 de junio de 2014

2.2. UAM CorpusTool

URL: <http://www.wagsoft.com/CorpusTool/index.html>

Versión 2.0 Beta5

Licencia: No se especifica.

Descripción: Software orientado a la anotación de corpus lingüísticos. Permite crear esquemas XML en el que son guardados los datos del corpus.

Sistemas Operativos: Windows & MacOSX

ScreenShot:

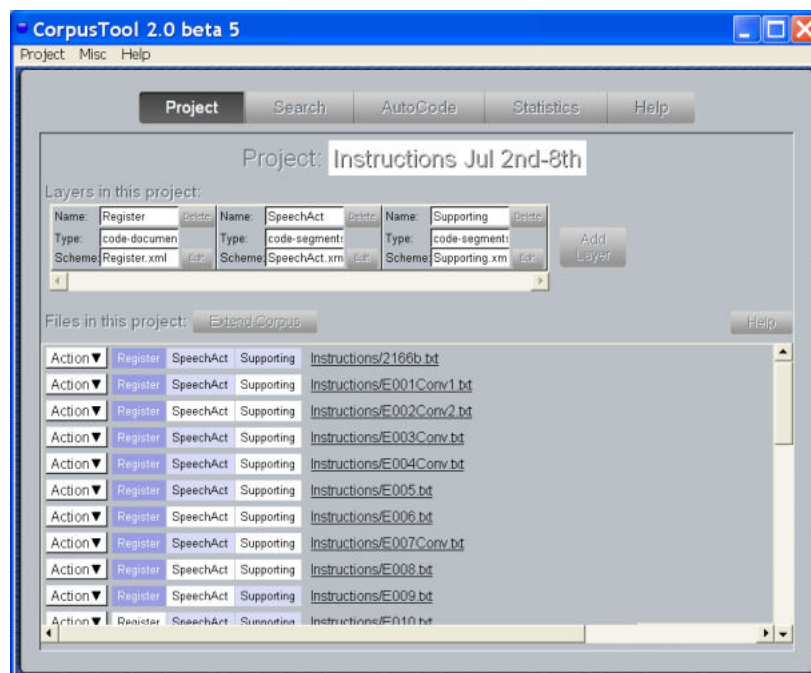


Figura 2: Captura de pantalla de UAM CorpusTool 2.0 beta5, tomado en Ubuntu 12.04

Accedido el: 26 de junio de 2014

2.2.1. Search View

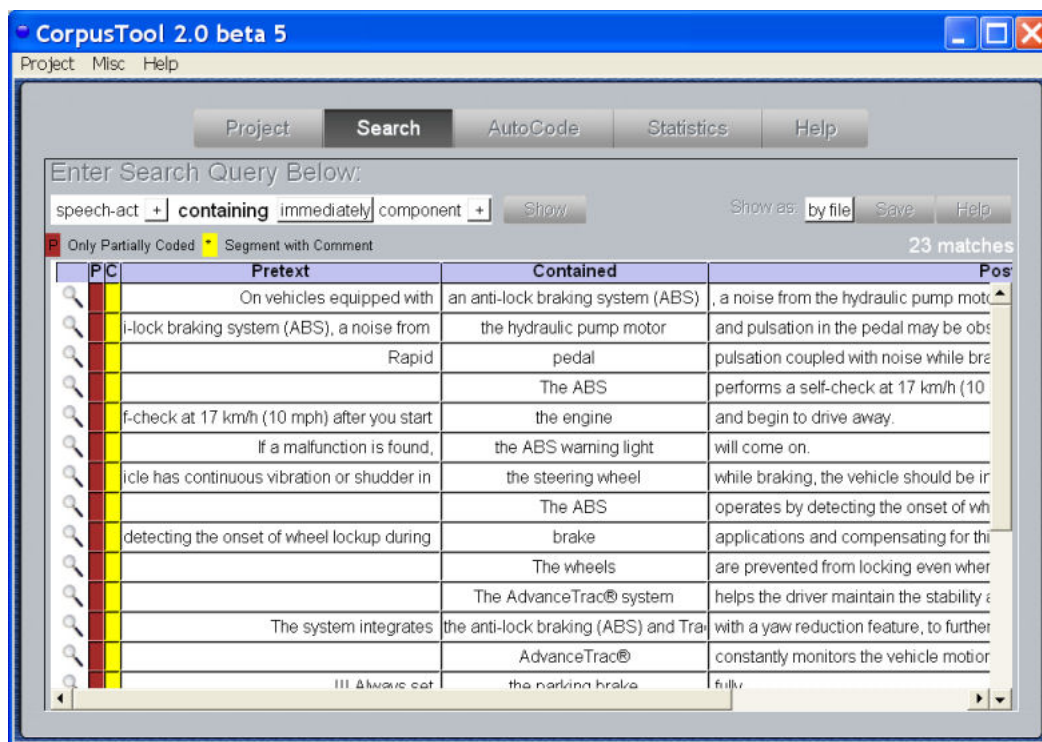
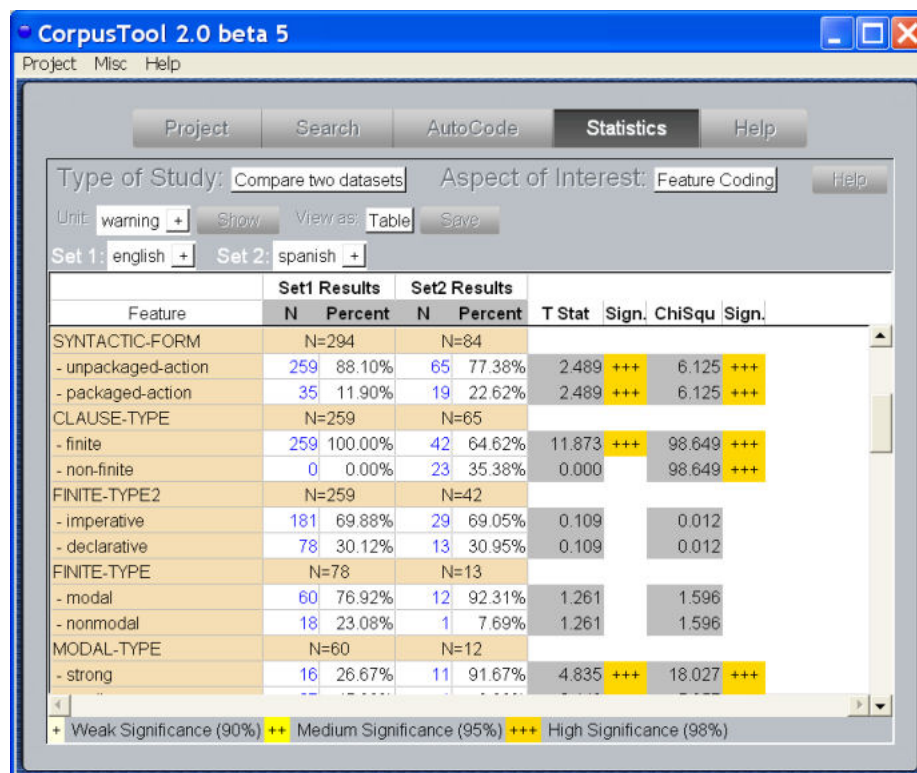


Figura 3: Captura de pantalla de UAM CorpusTool 2.0 beta5, tomado en Ubuntu 12.04

Accedido el: 26 de junio de 2014

2.2.2. Statistics View



CorpusTool 2.0 beta 5

Project Misc Help

Project Search AutoCode **Statistics** Help

Type of Study: Compare two datasets Aspect of Interest: Feature Coding Help

Unit: warning Show View as: Table Save

Set 1: english Set 2: spanish

Feature	Set1 Results		Set2 Results		T Stat	Sign.	ChiSqu	Sign.
	N	Percent	N	Percent				
SYNTACTIC-FORM	N=294		N=84					
- unpackaged-action	259	88.10%	65	77.38%	2.489	+++	6.125	+++
- packaged-action	35	11.90%	19	22.62%	2.489	+++	6.125	+++
CLAUSE-TYPE	N=259		N=65					
- finite	259	100.00%	42	64.62%	11.873	+++	98.649	+++
- non-finite	0	0.00%	23	35.38%	0.000		98.649	+++
FINITE-TYPE2	N=259		N=42					
- imperative	181	69.88%	29	69.05%	0.109		0.012	
- declarative	78	30.12%	13	30.95%	0.109		0.012	
FINITE-TYPE	N=78		N=13					
- modal	60	76.92%	12	92.31%	1.261		1.596	
- nonmodal	18	23.08%	1	7.69%	1.261		1.596	
MODAL-TYPE	N=60		N=12					
- strong	16	26.67%	11	91.67%	4.835	+++	18.027	+++

+ Weak Significance (90%) ++ Medium Significance (95%) +++ High Significance (98%)

Figura 4: Captura de pantalla de UAM CorpusTool 2.0 beta5, tomado en Ubuntu 12.04

Accedido el: 26 de junio de 2014

2.3. WordSmith Tools

URL: <http://www.lexically.net/wordsmith/>

Versión: 0.6

Licencia: No se especifica.

Descripción: Permite el análisis de corpus textuales extrayendo el listado de palabras, anotar el texto, realizar análisis de concordancia, etc.

Sistemas Operativos: Windows

ScreenShot:

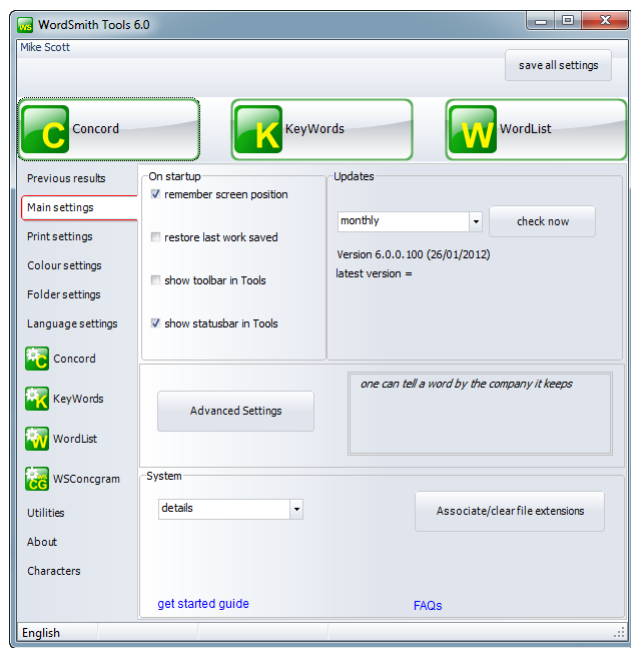


Figura 5: Captura de pantalla de WordSmith Tools 6.0, tomado en Ubuntu 12.04
Accedido el: 26 de junio de 2014

2.4. RSTTools

URL: www.wagsoft.com/RSTTool/index.html

Versión: 0.6

Licencia: No se especifica.

Descripción: Permite la segmentación de textos manualmente, marcar luego la relación entre estos segmentos, hacer anotaciones del discurso y generar estadísticas sobre el texto y lo construido.

Sistemas Operativos: Multiplataforma

ScreenShot:

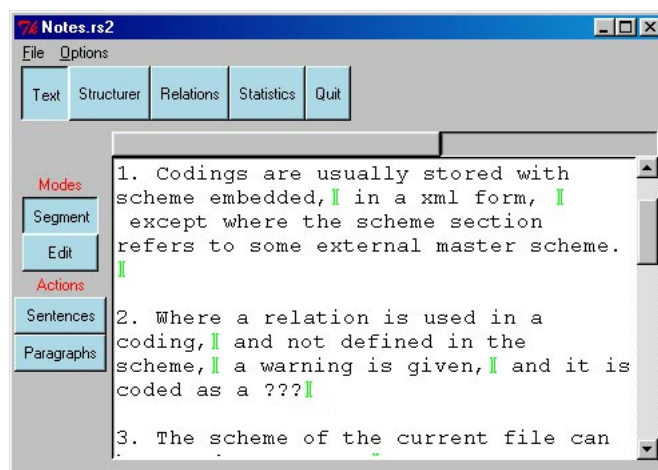


Figura 6: Captura de pantalla de RSTTools 3.0, tomado en Ubuntu 12.04
Accedido el: 26 de junio de 2014

2.5. Nooj

URL: <http://www.nooj4nlp.net/pages/nooj.html>

Versión: No se especifica.

Licencia: No se especifica.

Descripción: Permite extraer datos de corpus textuales, adicionalmente realizar el etiquetado de estos atendiendo a funciones lingüísticas, permite realizar desambiguación de palabras

Sistemas Operativos: Windows. Permite su uso en Linux a través de Mono.

ScreenShot:

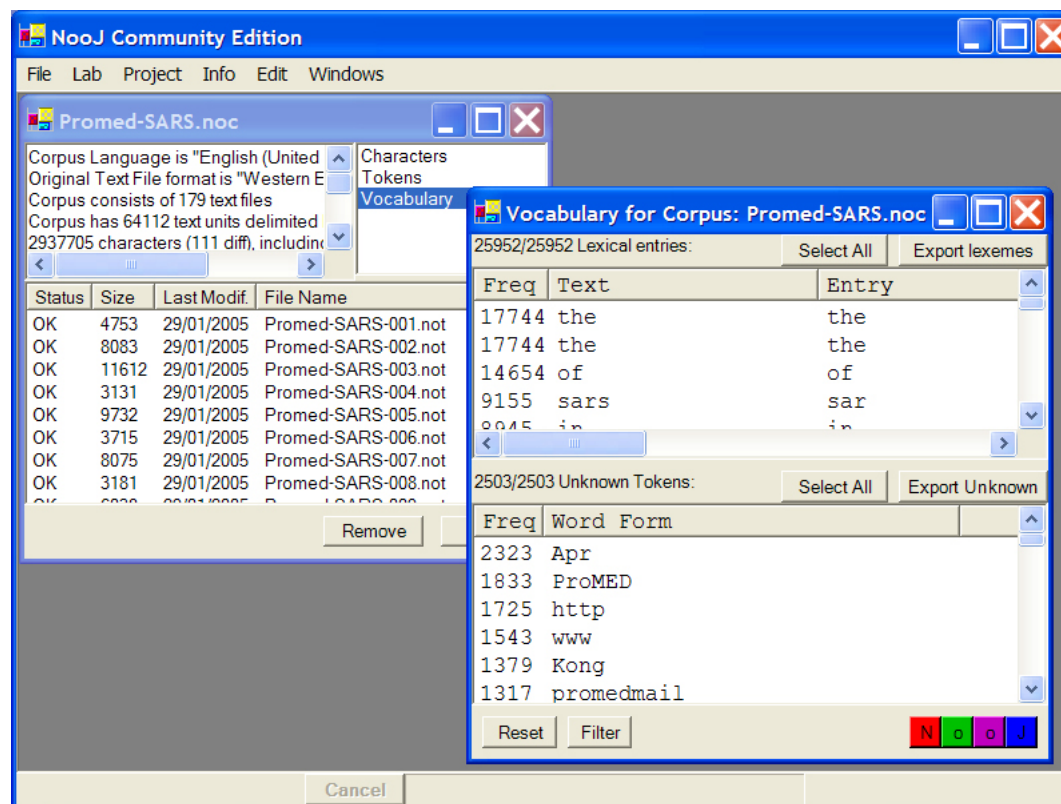


Figura 7: Captura de pantalla de NooJ Community Edition, tomado en Ubuntu 12.04

Accedido el: 26 de junio de 2014

3. Conclusiones

Existen un gran número de aplicaciones para este tipo de fin.

En la URL <http://www.uow.edu.au/~dlee/software.htm> se pueden en-

contrar una lista más grande de herramientas que se utilizan para el tratamiento de corpus.

...

Cosas que faltan:

- Adicionar la empresa que le da soporte a cada herramienta.
- Lista de funcionalidades por cada una de las aplicaciones.