PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE

FACULTAD DE LETRAS

Métodos de Investigación en Letras

César Antonio Aguilar

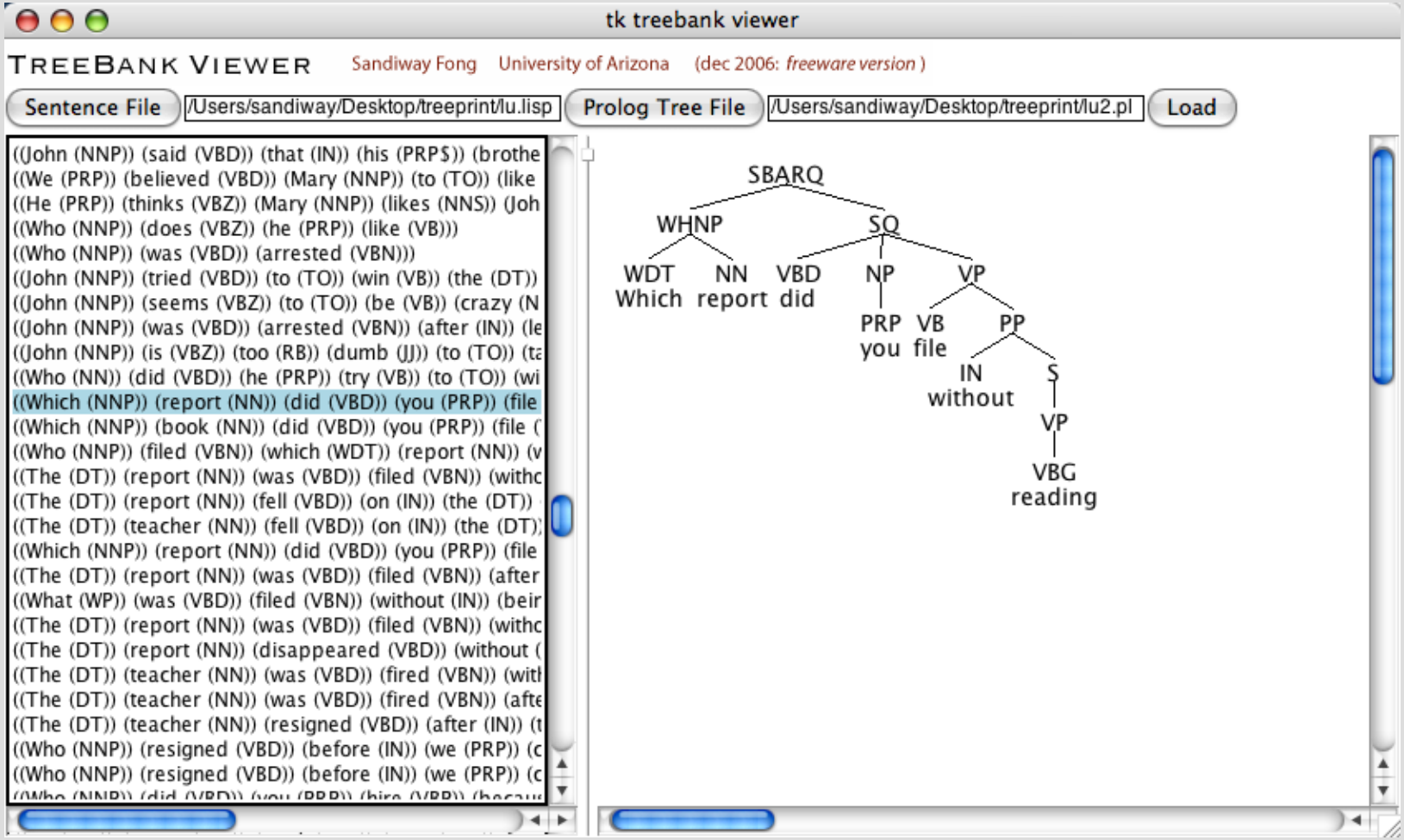
Facultad de Lenguas y Letras

31/10/2017

caguilara@uc.cl

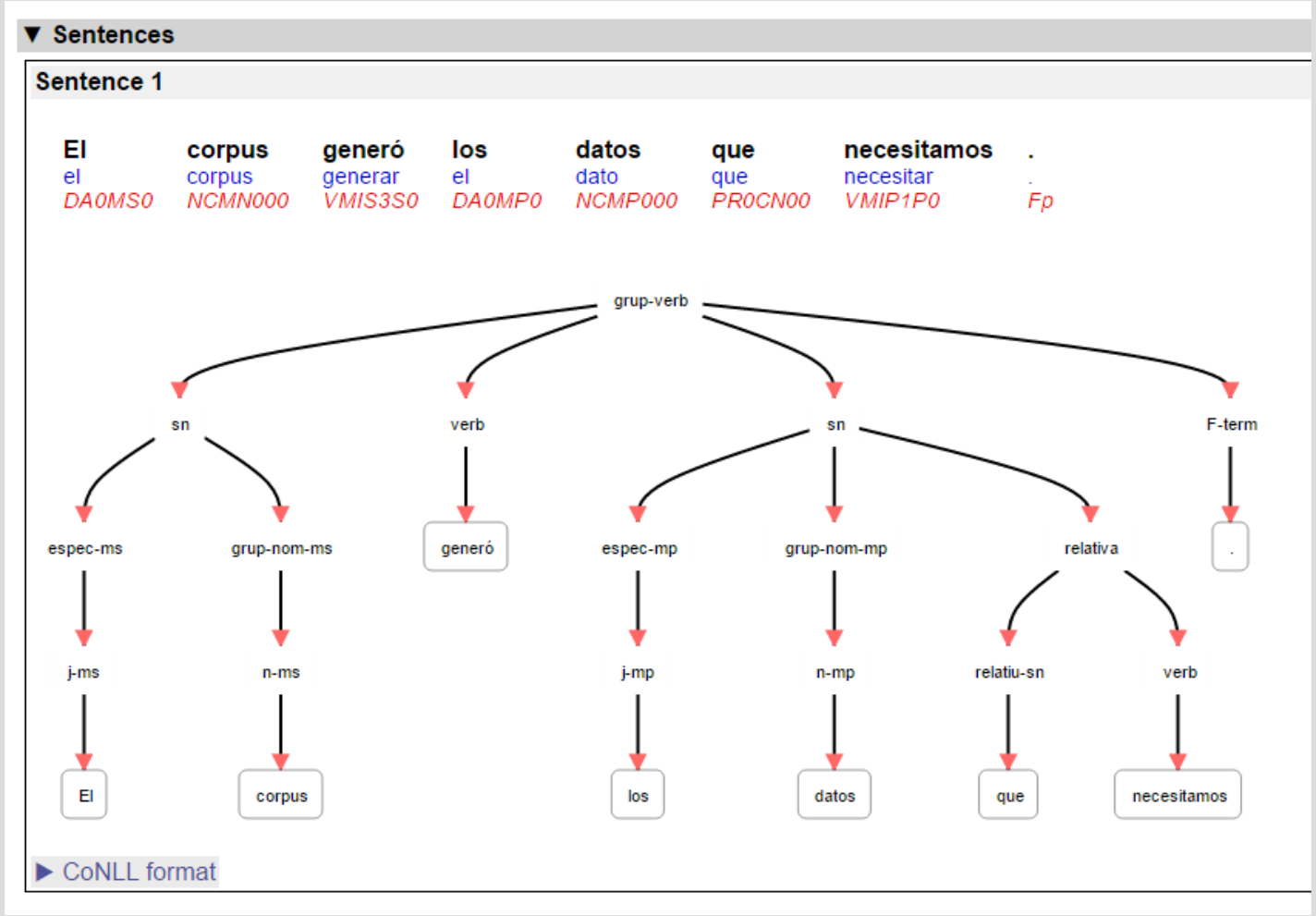
Síntesis de la clase pasada (1)

En la clase pasada vimos de manera general el uso de corpus anotados gramaticalmente, los cuales son útiles para hacer múltiples tareas.



Síntesis de la clase pasada (2)

Esta clase de corpus, como pudieron comprobar, nos ayudan a hacer una búsqueda más detallada de patrones regulares que, suponemos, pueden ser observados en un corpus textual.



Buscando *huellas* en un corpus (1)

Un corpus anotado nos permite identificar **patrones recurrentes de construcción**, ya sea de palabras, de sintagmas o frases, e incluso de oraciones.

De hecho, teniendo en cuenta esta clase de recursos, podemos observar, de forma general, **patrones que describen fenómenos lingüísticos muy específicos, p. e.: identificar la temática de un de una colección de documentos sin tener que leerlos uno por uno.**

Veamos algunos ejemplos al respecto.

En Estados Unidos, cada vez que un nuevo Presidente es electo, éste da lo que se conoce como un **Discurso de Investidura Presidencial**, por medio del cual el nuevo Presidente delinea, de manera general, cuáles serán los ejes temáticos de su gobierno, qué problemas tratará de resolver, qué se puede esperar de su periodo en comparación otros, etc.

Buscando *huellas* en un corpus (2)

Todos estos discursos actualmente se encuentran disponibles en el siguiente sitio electrónico:

www.presidency.ucsb.edu/inaugurals.php

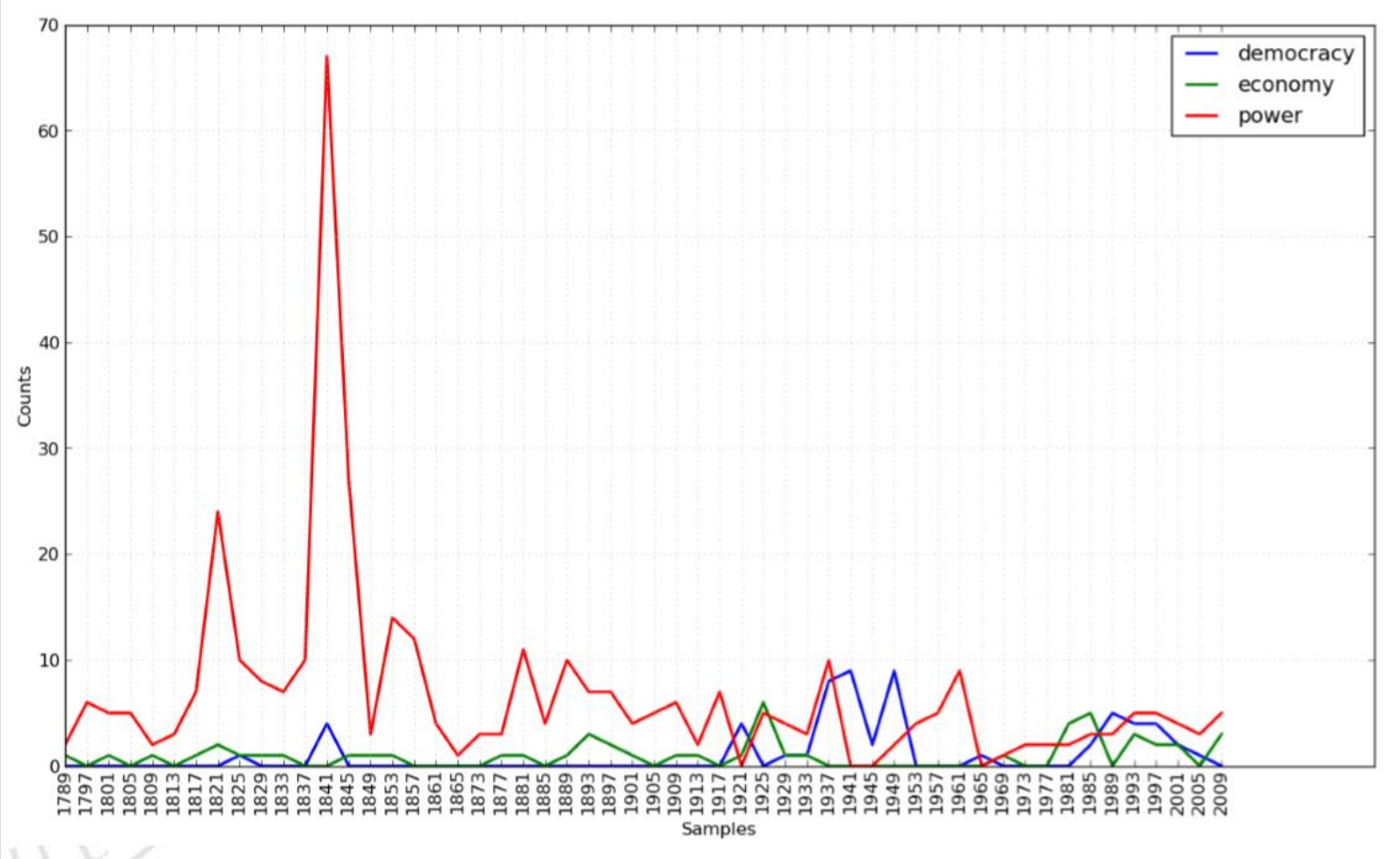
www.ucsb.edu

Este link los lleva al portal del **American Presidency Project (APP)**, el cual les brinda acceso a una amplia colección de documentos históricos sobre los Presidentes de Estados Unidos, desde **George Washington** (nombrado Presidente en 1789), hasta **Barak Obama** (hasta su segundo periodo, iniciado en 2013).

Este portal es administrado por la **Universidad de California, campus Santa Barbara**

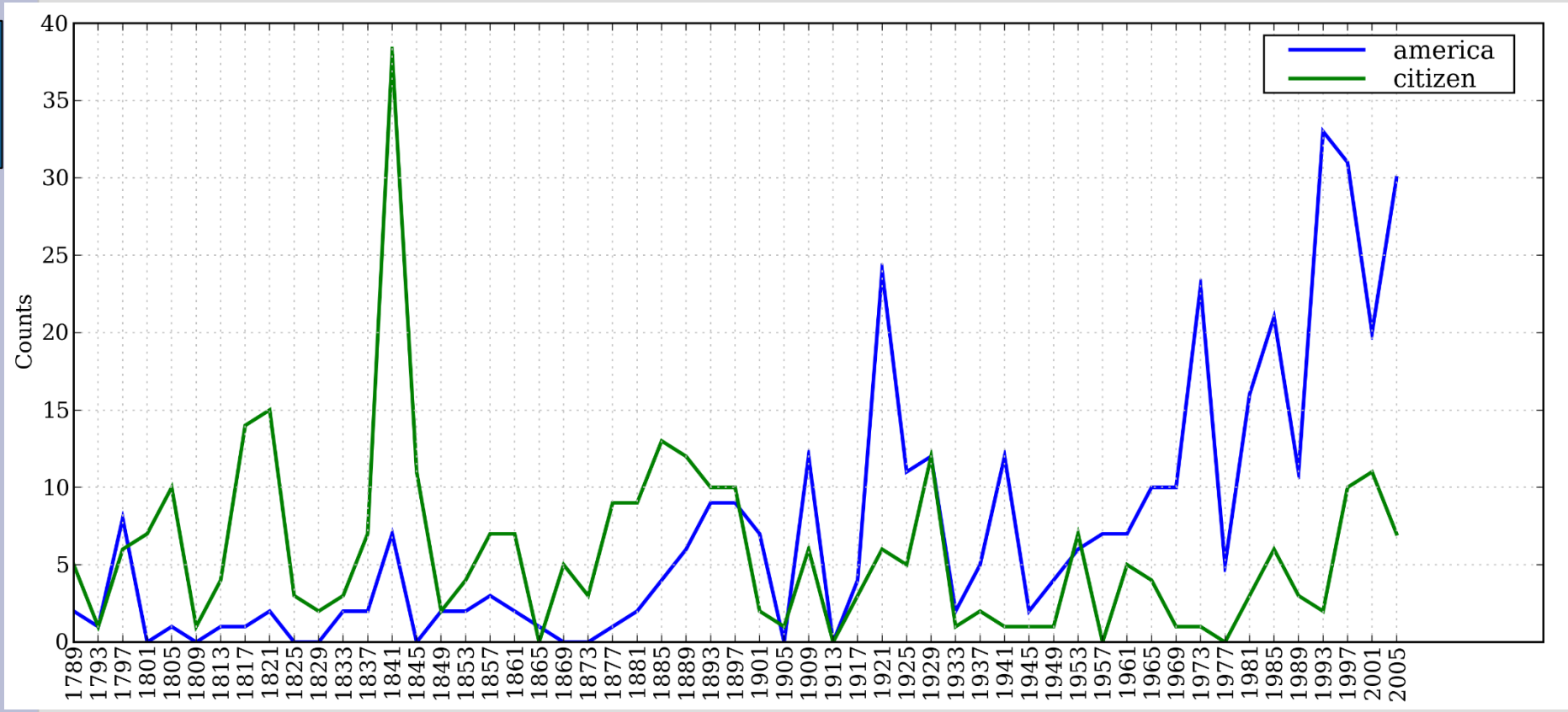
Buscando *huellas* en un corpus (3)

Si analizamos este corpus, podemos obtener mucha información sobre diferentes aspectos políticos relacionados con los Presidentes estadounidenses, con sus partidos políticos, con su auditorio, etc.
Veamos un ejemplo: ¿cuál de los siguientes temas ha tenido mayor importancia en estos discursos: *democracia*, *economía* o *poder*?



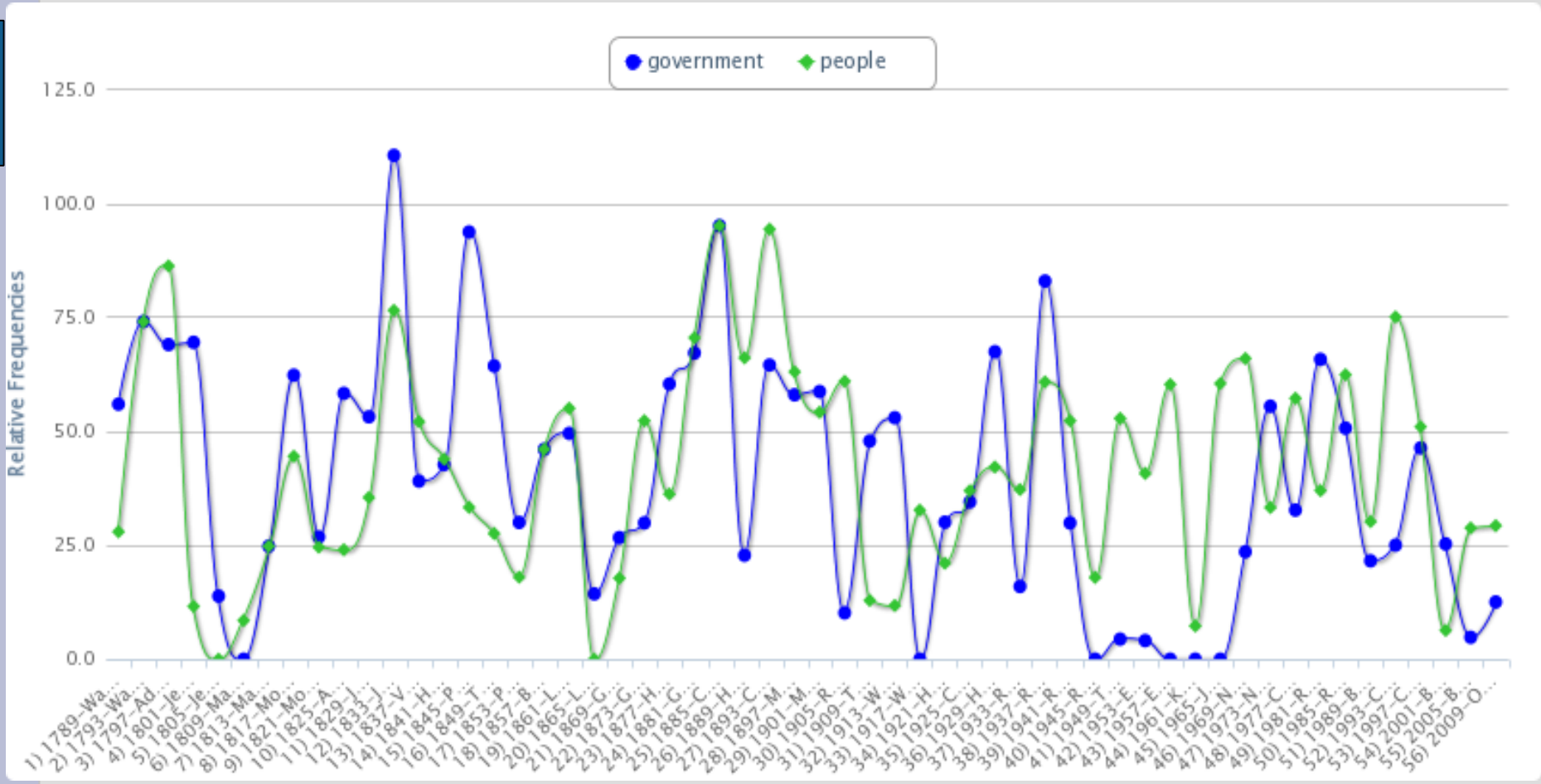
Buscando *huellas* en un corpus (4)

Otro caso: ¿en qué periodos históricos se han hecho mayores alusiones a los ciudadanos (*citizens)*, y en cuáles se ha priorizado más a la nación *(America*)? Veamos:



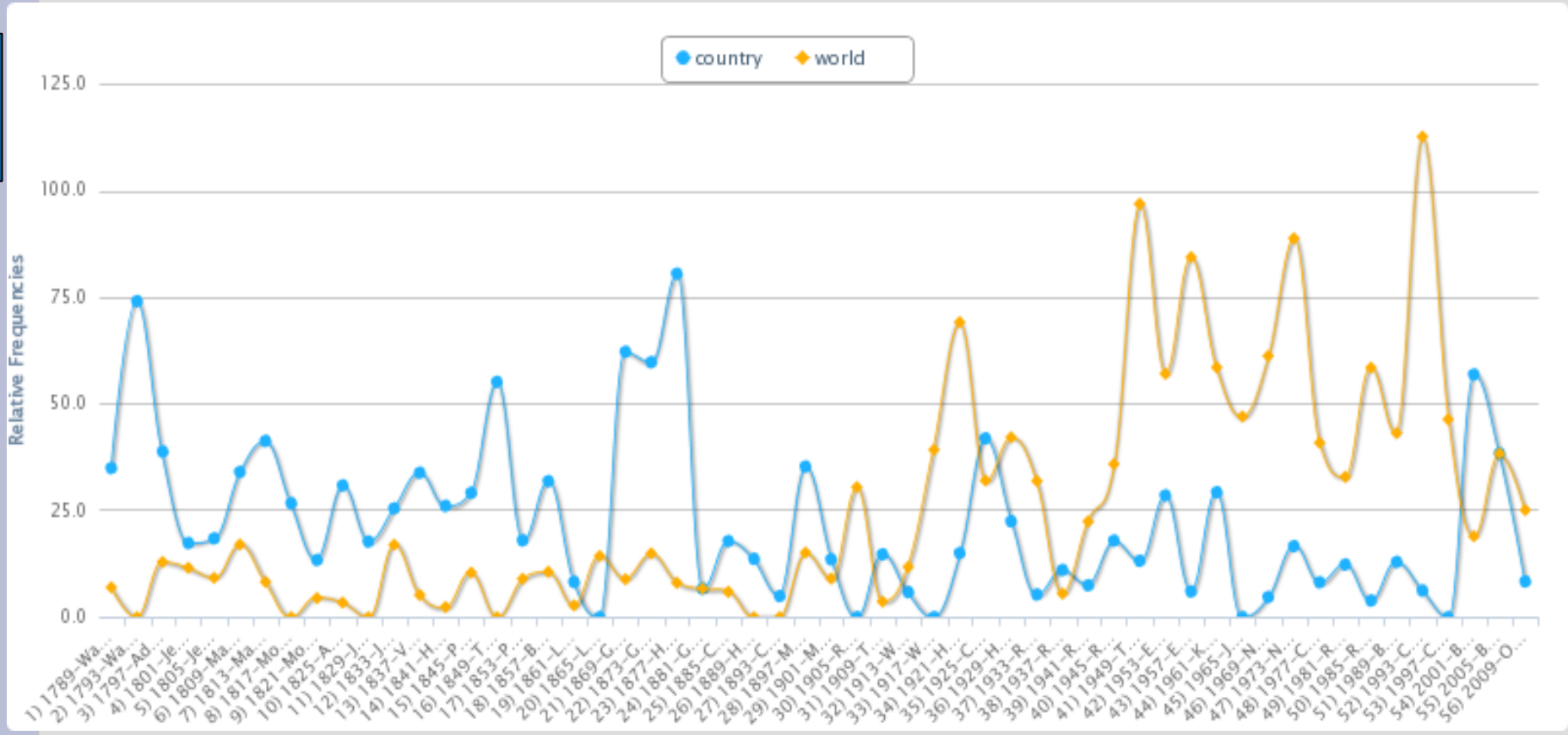
Buscando *huellas* en un corpus (5)

Otra tendencia interesante que emerge analizando este corpus es la siguiente: para algunos Presidentes, la relación entre *Gobierno* y *Personas* es equilibrada, para otros, el *Gobierno* tiene un mayor peso, y para otros, la prioridad son las *Personas*.



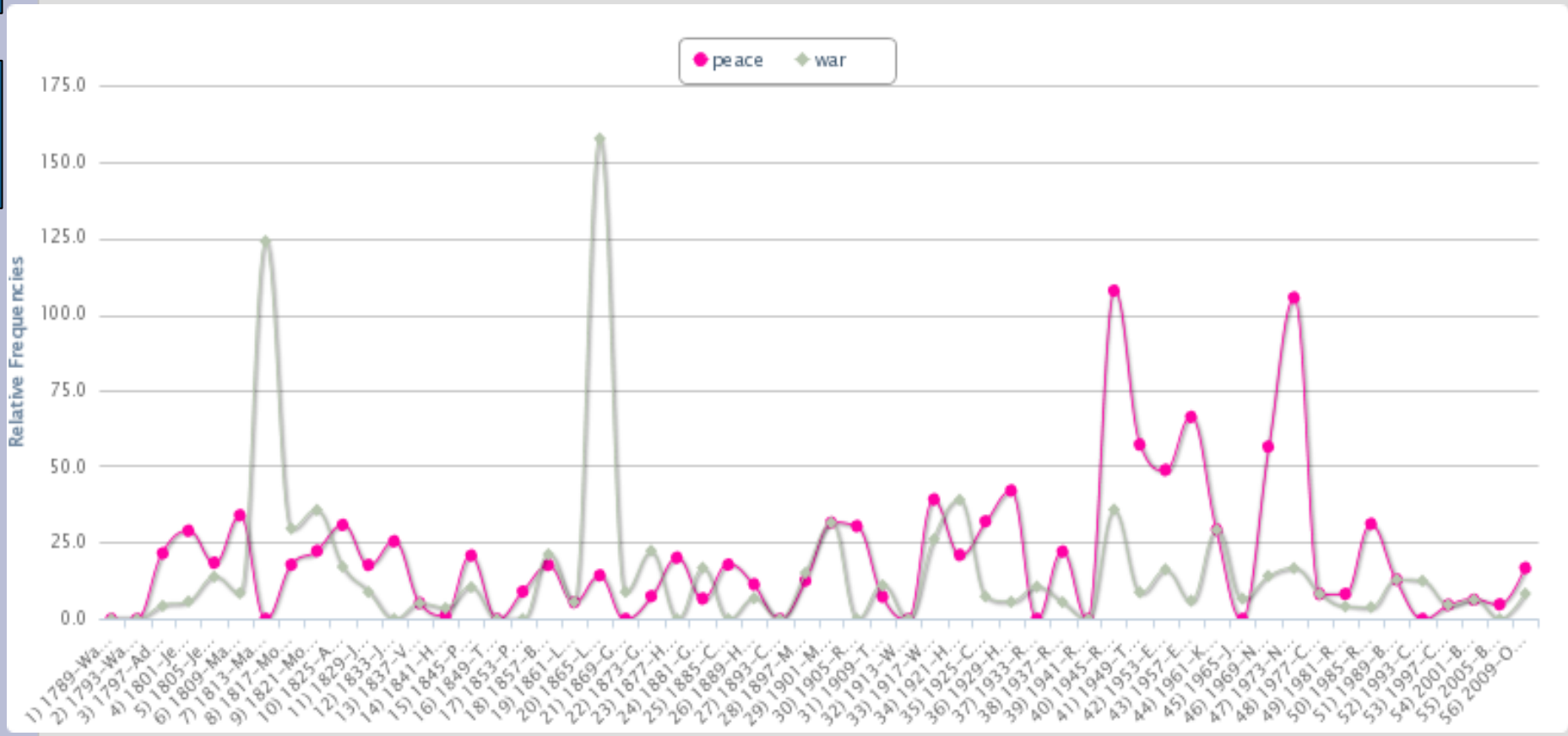
Buscando *huellas* en un corpus (6)

De la misma manera, podemos ver en qué periodos se prioriza la relación que tiene Estados Unidos con el resto del mundo, y en qué periodos se enfatiza una actitud nacionalista:



Buscando *huellas* en un corpus (7)

Finalmente, ¿en qué periodos se han priorizado la paz, y en qué momentos se han tenido que enfatizar las acciones bélicas?



Deduciendo el estilo de Charles Dickens (1)

El corpus anterior nos ofreció una serie de datos que nos permiten inferir qué rasgos temáticos caracterizan a los autores de cada uno de estos discursos, e igualmente podríamos deducir las reacciones que pueden provocar en las personas que los escuchan.

Ahora bien, podríamos igualmente realizar un análisis similar a las obras de un autor único, y tratar entender cómo es su estilo a partir de sus rasgos lingüísticos.

Para ello, vamos a revisar un corpus conformado por varias novelas de Charles Dickens (1812-1870), el cual cuenta con etiquetas gramaticales.

Deduciendo el estilo de Charles Dickens (2)

1. Accedan a la siguiente liga:

http://corpora.linguistik.uni-erlangen.de/demos/CQP/CQPDemo/frames-cqp.html

Se trata de un corpus en inglés construido a partir de algunas novelas escritas por Charles Dickens (1812-1870)

CWB
The IMS Open Corpus Workbench

Este mini-corpus fue elaborado por el *Institute for Natural Language Processing* de la Universidad de Stuttgardt:

http://cwb.sourceforge.net/demos.php

Deduciendo el estilo de Charles Dickens (3)

Este corpus ha sido desarrollado por tres investigadores alemanes de la **Universidad de Stuttgart:** Ulrich Heid, Helmut Schmid y Oliver Christ.

La idea de este corpus era evaluar una serie de herramientas computacionales, capaces de reconocer patrones léxicos y sintácticos en una gran colección de documentos.

www.ims.uni-stuttgart.de/index.en.html

Deduciendo el estilo de Charles Dickens (4)

Busquemos algunos personajes de las novelas de Dickens:

1. En *The Posthumous Papers of the Pickwick Club* existen dos que se apellidan *Weller*. ¿Quiénes son? ¿Qué expresión regular necesitan para identificarlos?

2. Siguiendo el mismo método, distingan cuántos miembros tiene la familia *Wardle*, en la misma novela.

3. Basándonos en *Oliver Twist*, ¿qué tienen que ver *Artful Dodger*, *Jack Dawkins* y *Fagin* con la palabra *pickpocket*?

4. En *David Copperfield*, ¿cómo puedo saber que la palabra *strong* puede ser un sustantivo, un adjetivo o un apellido?

5. Finalmente, en *Great Expectations*, ¿quién es el *convict*, y quién es el *benefactor?*

Deduciendo el estilo de Charles Dickens (5)

Algunos tips: en la siguiente clase las veremos con más detalle, pero consideren que algunas búsquedas pueden dar mejores resultados si ocupan **etiquetas de partes de la oración** (Ing. *POST*). El corpus Workbench emplea etiquetas basadas en el proyecto **PennTree Bank**. Las pueden consultar en la siguiente liga:

http://corpora.linguistik.uni-erlangen.de/demos/CQP/CQPDemo/frames-cqp.html

Pensando en estas etiquetas, ustedes pueden formular patrones como los siguientes:

```
gentleman

un*ness

Oliver Twist

said NP

as A as NP

PP=from PP=to

It|it was [ Det ] N of { A } N
```

¡OJO!: el signo " | ", en el buscador del Workbench equivale a disyunción (*x OR y*)

Deduciendo el estilo de Charles Dickens (6)

Veamos algunos resultados. Respecto a la primera pregunta, primero formulamos un patrón usando combinaciones de palabras con etiquetas gramaticales, p.e.:

NP (frase nominal) + **Weller**

```
[pos="NP"] "Weller"
```

En la siguiente lámina se muestran algunos resultados.

Deduciendo el estilo de Charles Dickens (7)

1 - 20 Go **[1039 matches]** tokens = word/POS phrases = off context = sentence Apply

1. Master Humphrey's Clock, Chapter 3

context '" You$_{PP}$ have$_{VBP}$ not$_{RB}$ told$_{VBN}$ me$_{PP}$., '" said$_{VBD}$ I$_{PP}$., '" anything$_{NN}$ about$_{IN}$ **Sam$_{NP}$ Weller$_{NP}$** .$_{SENT}$

2. Master Humphrey's Clock, Chapter 3

context '" And$_{CC}$ **Mr.$_{NP}$ Weller$_{NP}$** senior$_{JJ}$?$_{SENT}$

3. Master Humphrey's Clock, Chapter 3

context '" Old$_{NP}$ **Mr.$_{NP}$ Weller$_{NP}$** ., '" returned$_{VBD}$ Mr.$_{NP}$ Pickwick$_{NP}$., '" is$_{VBZ}$ in$_{IN}$ no$_{DT}$ respect$_{NN}$ more$_{RBR}$ altered$_{JJ}$ than$_{IN}$ Sam$_{NP}$., unless$_{IN}$ it$_{PP}$ be$_{VB}$ that$_{WDT}$ he$_{PP}$ is$_{VBZ}$ a$_{DT}$ little$_{RB}$ more$_{RBR}$ opinionated$_{JJ}$ than$_{IN}$ he$_{PP}$ was$_{VBD}$ formerly$_{RB}$., and$_{CC}$ perhaps$_{RB}$ at$_{IN}$ times$_{NNS}$ more$_{RBR}$ talkative$_{JJ}$.$_{SENT}$

4. Master Humphrey's Clock, Chapter 3

context He$_{PP}$ spends$_{VBZ}$ a$_{DT}$ good$_{JJ}$ deal$_{NN}$ of$_{IN}$ his$_{PP\$}$ time$_{NN}$ now$_{RB}$ in$_{IN}$ our$_{PP\$}$ neighbourhood$_{NN}$., and$_{CC}$ has$_{VBZ}$ so$_{RB}$ constituted$_{VBD}$ himself$_{PP}$ a$_{DT}$ part$_{NN}$ of$_{IN}$ my$_{PP\$}$ bodyguard$_{NN}$., that$_{IN}$ when$_{WRB}$ I$_{PP}$ ask$_{VBP}$ permission$_{NN}$ for$_{IN}$ Sam$_{NP}$ to$_{TO}$ have$_{VB}$ a$_{DT}$ seat$_{NN}$ in$_{IN}$ your$_{PP\$}$ kitchen$_{NN}$ on$_{IN}$ clock$_{NN}$ nights$_{NNS}$ ((supposing$_{VBG}$ your$_{PP\$}$ three$_{CD}$ friends$_{NNS}$ think$_{VBP}$ me$_{PP}$ worthy$_{JJ}$ to$_{TO}$ fill$_{VB}$ one$_{CD}$ of$_{IN}$ the$_{DT}$ chairs$_{NNS}$)) ., I$_{PP}$ am$_{VBP}$ afraid$_{JJ}$ I$_{PP}$ must$_{MD}$ often$_{RB}$ include$_{VB}$ **Mr.$_{NP}$ Weller$_{NP}$** too$_{RB}$.$_{SENT}$

5. Master Humphrey's Clock, Chapter 3

context That$_{DT}$ gentleman$_{NN}$ made$_{VBD}$ a$_{DT}$ suitable$_{JJ}$ reply$_{NN}$ with$_{IN}$ a$_{DT}$ cordiality$_{NN}$ and$_{CC}$ frankness$_{NN}$ peculiarly$_{RB}$ his$_{PP\$}$ own$_{JJ}$., and$_{CC}$ glanced$_{VBD}$ with$_{IN}$ a$_{DT}$ smile$_{NN}$ towards$_{IN}$ two$_{CD}$ persons$_{NNS}$ behind$_{IN}$ the$_{DT}$ door$_{NN}$., whom$_{WP}$ I$_{PP}$ had$_{VBD}$ not$_{RB}$ at$_{IN}$ first$_{RB}$ observed$_{VBN}$., and$_{CC}$ whom$_{WP}$ I$_{PP}$ immediately$_{RB}$ recognised$_{VBD}$ as$_{IN}$ Mr.$_{NP}$ **Samuel$_{NP}$ Weller$_{NP}$** and$_{CC}$ his$_{PP\$}$ father$_{NN}$.$_{SENT}$

6. Master Humphrey's Clock, Chapter 3

En total, logramos obtener 1039 aciertos, pero requerimos saber si todos pertenecen a la misma obra, o se encuentran dispersos en varios textos.

Deduciendo el estilo de Charles Dickens (8)

Veamos entonces su distribución:

A Christmas Carol	0.00 ppm [f = 0]
David Copperfield	0.00 ppm [f = 0]
Dombey and Son	0.00 ppm [f = 0]
Great Expectations	0.00 ppm [f = 0]
Hard Times	0.00 ppm [f = 0]
Master Humphrey's Clock	1737.41 ppm [f = 93]
Nicholas Nickleby	0.00 ppm [f = 0]
Oliver Twist	0.00 ppm [f = 0]
Our Mutual Friend	0.00 ppm [f = 0]
Sketches by BOZ	0.00 ppm [f = 0]
A Tale of Two Cities	0.00 ppm [f = 0]
The Old Curiosity Shop	0.00 ppm [f = 0]
The Pickwick Papers	2563.02 ppm [f = 949]
Three Ghost Stories	0.00 ppm [f = 0]

Deduciendo el estilo de Charles Dickens (9)

Ahora, consideremos su frecuencia:

8 unique strings found among 1042 matches.

f	matching string
875	mr. weller
67	sam weller
54	samuel weller
42	mrs. weller
1	lawyer.--"tony weller
1	man weller
1	master weller
1	parent weller

Con todos estos datos, volvamos a la pregunta original: ¿cuántos personajes tienen el apellido *Weller* en *The Pickwick Papers*?

Respuesta: por lo menos 2: ***Tony Weller*** y ***Samuel (Sam) Weller***.

Deduciendo el estilo de Charles Dickens (10)

Hay varias formas de visualizar los datos que nos muestra este corpus, lo que nos permite hacernos nuevas preguntas, p. e.: ¿con qué otros personajes de Dickens se relacionan los Weller? Una respuesta la podemos deducir revisando el siguiente fragmento:

1006. The Pickwick Papers, Chapter 22

context inquired$_{VBD}$ [Mr.$_{NP}$ Peter$_{NP}$ Magnus$_{NP}$] .$_{SENT}$ '" [[The$_{DT}$ old$_{JJ}$] '" un$_{JJ}$] means$_{NNS}$ [a$_{DT}$ turnpike-keeper$_{NN}$., gen'l'm'n,'$_{NN}$] observed$_{VBD}$ [Mr.$_{NP}$ **Samuel$_{NP}$ Weller$_{NP}$**] ., [in$_{IN}$ [explanation$_{NN}$]] .$_{SENT}$ '" Oh$_{UH}$., '" said$_{VBD}$ [Mr.$_{NP}$ Pickwick$_{NP}$] ., '" [I$_{PP}$] see$_{VBP}$.$_{SENT}$

Según este ejemplo, en el capítulo 22, en la línea 1006, hay una mención a dos personajes concretos: ***Peter Magnus*** y ***Mr. Pickwick***.

¿Qué otra clase de preguntas pueden formularse que sean susceptibles de ser analizadas dentro de este corpus?

Gracias por su atención

Blog del curso:
http://cesaraguilar.weebly.com/presentacioacuten.html