PRÁCTICA CON LA HERRAMIENTA GATE.

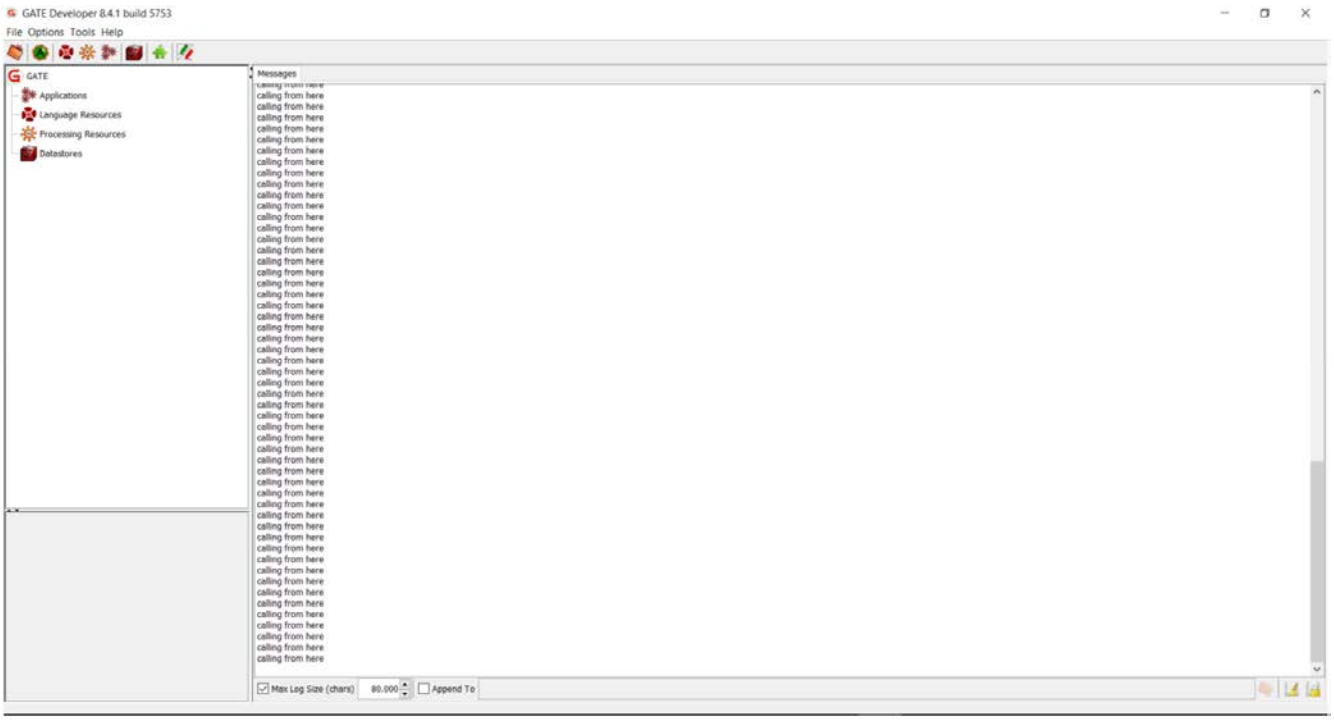
SOLUCIÓN

1. HERRAMIENTA GATE

GATE es una herramienta libre basada en Java —a diferencia de NLTK, que está programada en Python—. Sería el equivalente a NLTK+BRAT, que se ha visto en las prácticas anteriores, pero en una sola herramienta. Consta de una interfaz visual que permite crear y almacenar corpus y recursos léxicos y procesarlos usando distintas librerías que se encadenan formado cadenas de procesamiento.

1.1 DESCARGA, INSTALACIÓN Y FAMILIARIZACIÓN

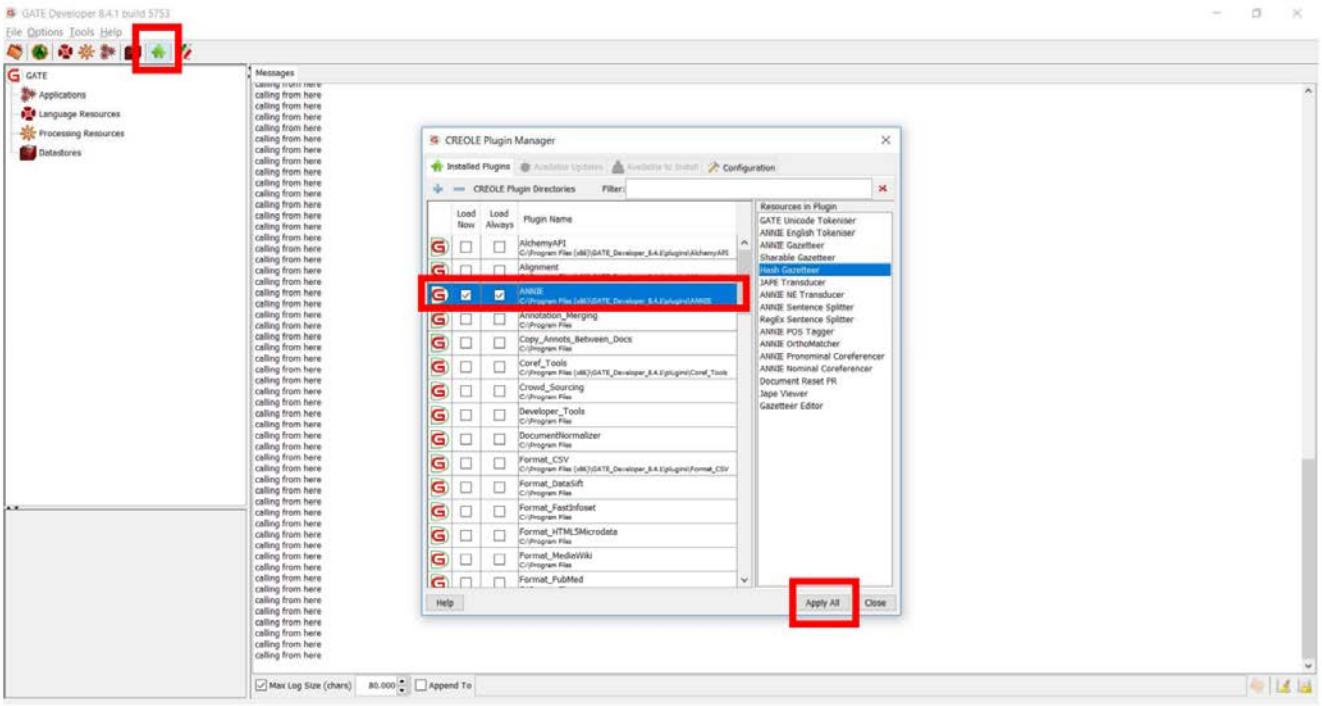
Hay que entrar en https://gate.ac.uk/download/ y proceder a la descarga e instalación del programa en nuestro sistema.



Se inicia la aplicación y se realizan los siguientes pasos:

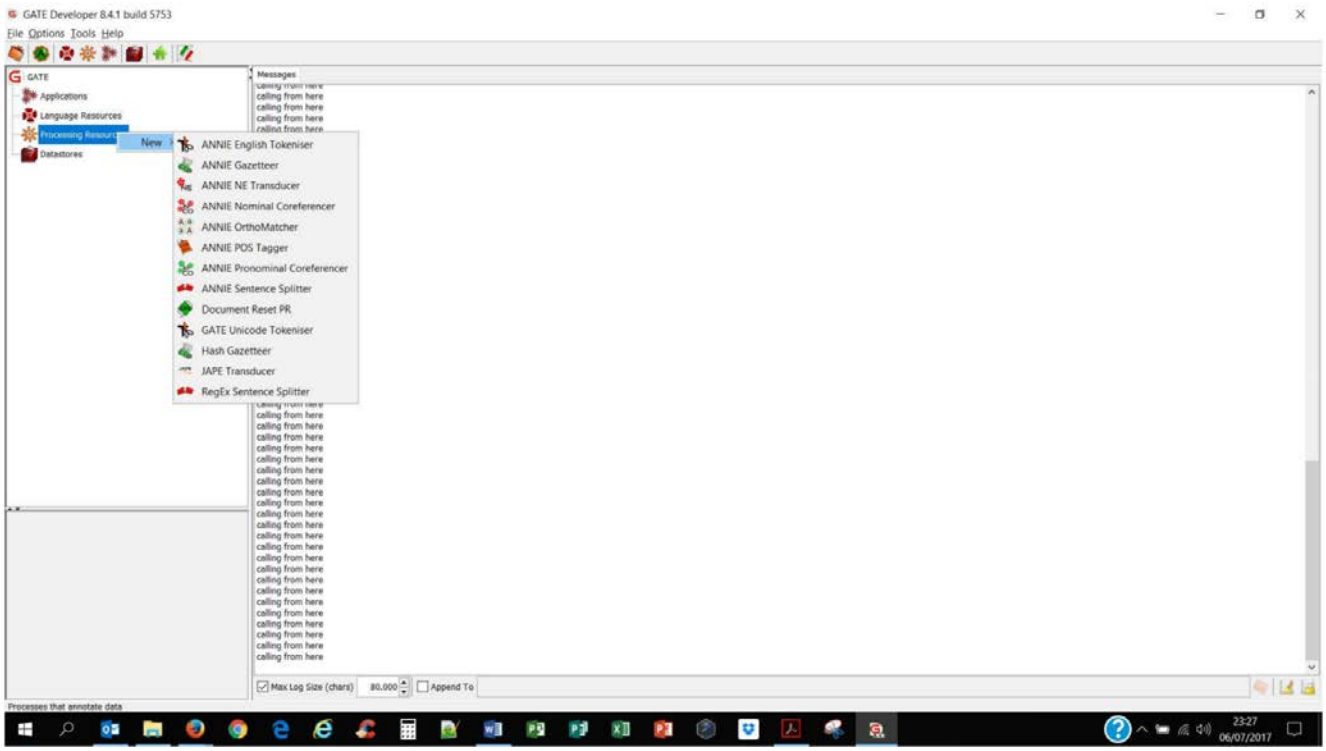
1. Cargar los 'Processing Resources' (PRs) de ANNIE:

Clic sobre el icono "puzle verde" —se muestra en la siguiente imagen—, seleccionar ANNIE y marcar el botón "Apply all". Con estos PRs se podrán construir cadenas de procesamiento NLP, encadenando unos recursos con otros y aplicándolos sobre los "Language Resources" (corpus + documentos), que se procesan y seleccionan posteriormente.

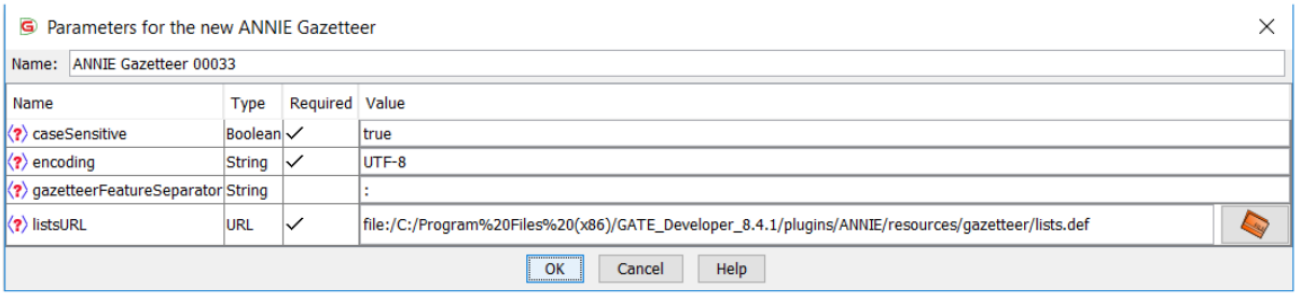


Una vez cargados todos los PRs de ANNIE, se pincha con el botón derecho sobre el icono “Processing Resources” y se seleccionan los PRs que se van a utilizar en este ejercicio:

- ‘ANNIE POS Tagger 0002A’.
- ‘ANNIE Sentence Splitter 0002B’.
- ‘ANNIE Gazetteer 0002E’.
- ‘ANNIE English Tokeniser 0002F’.
- ‘Document Reset PR 00032’.

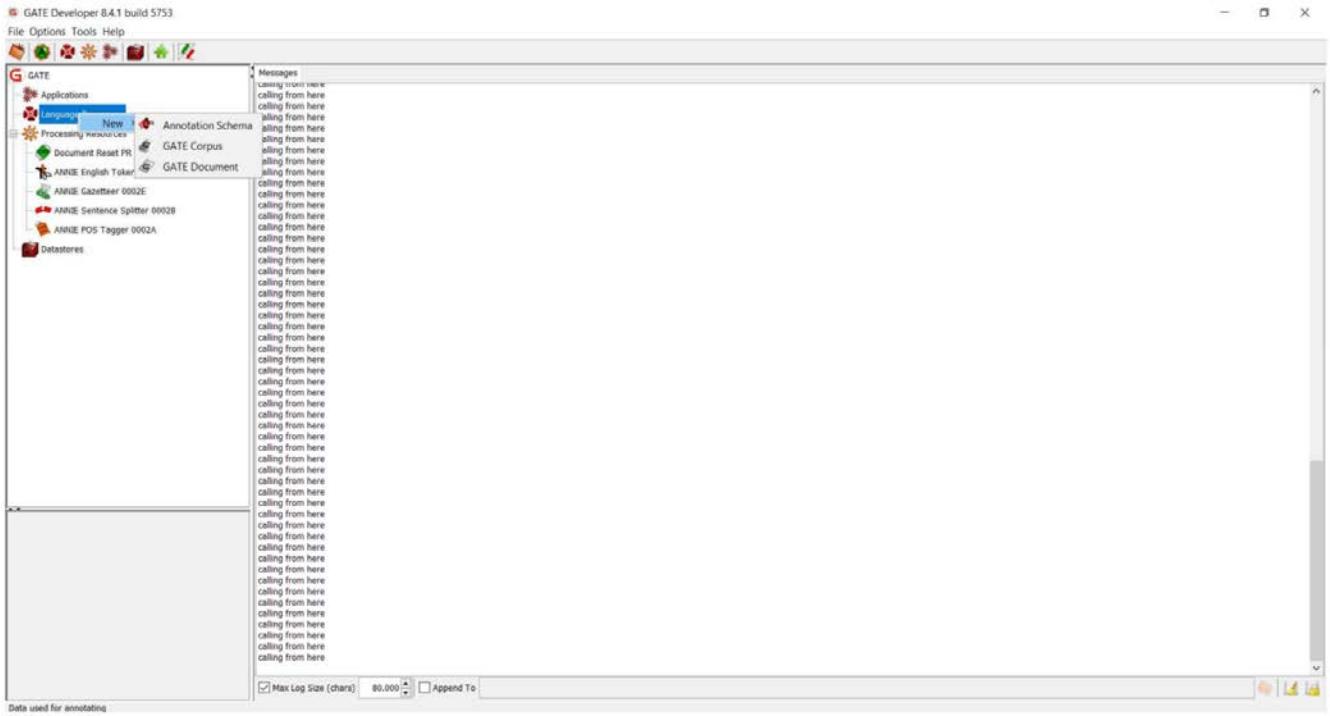


En cada ventana que se abre, hay que pinchar sobre al botón “OK” para finalizar su incorporación.

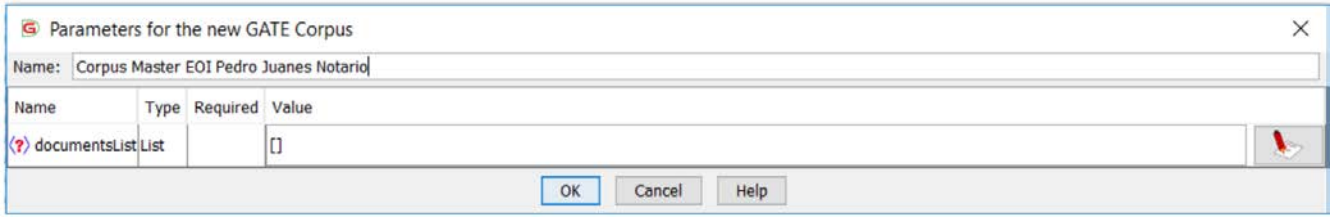


2. Crear un corpus y un documento de prueba:

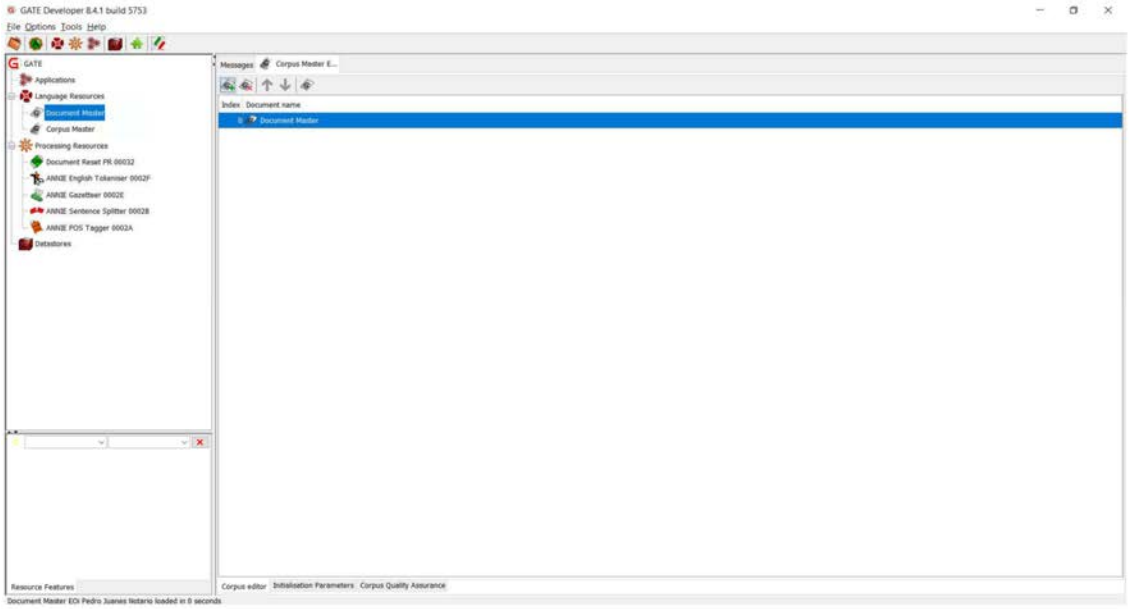
Hacer clic derecho sobre “Language Resources” y después “New Corpus” para añadir un corpus al sistema.



Se nombra y se pulsa sobre el botón “OK”. Se crea el corpus.



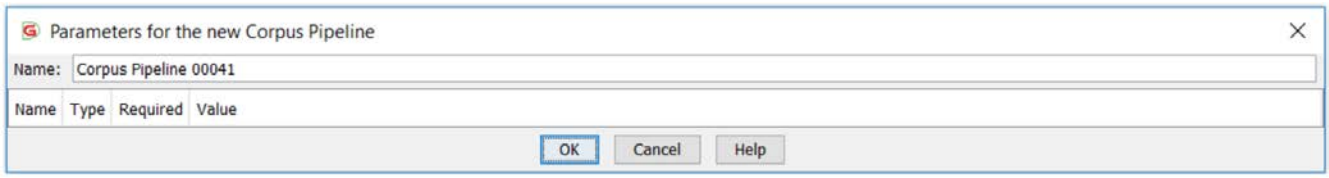
Sobre este corpus, se crea un documento que se utilizará en el proceso de análisis mediante los PRs de ANNIE. Para ello, hay que hacer clic con el botón derecho sobre “Language Resources” y después “New Document”, se nombra y se arrastra a la parte derecha de la pantalla para añadirlo al corpus creado con anterioridad. De esta manera, se pueden incluir varios documentos en un mismo corpus.



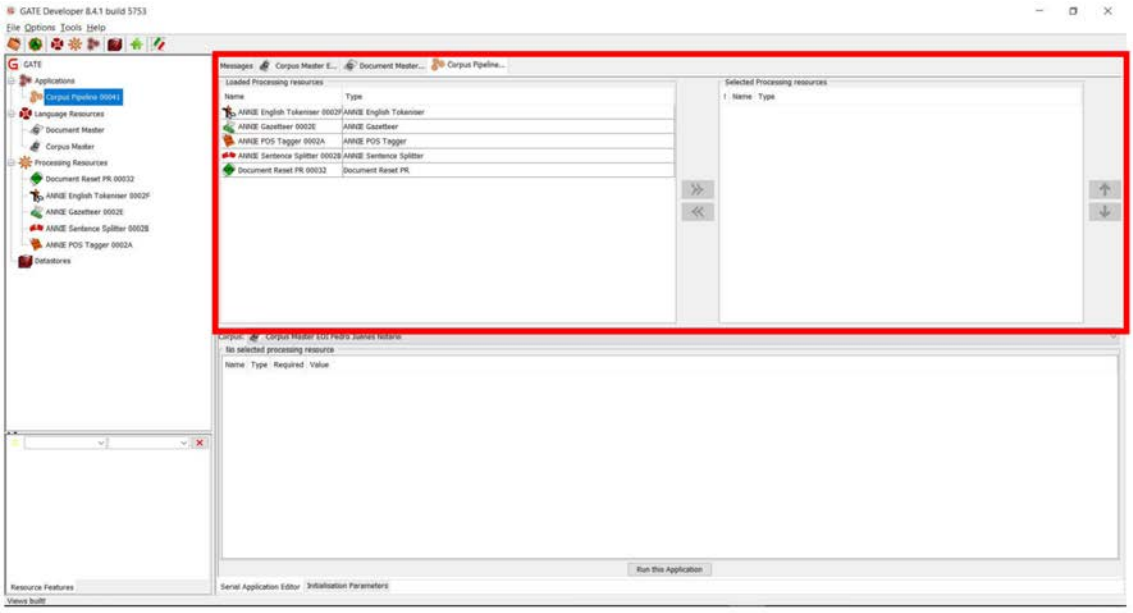
Sobre este documento, se hace doble clic y se completa el contenido en la ventana que se abre —en inglés, ya que ANNIE realiza el análisis en este idioma—. En este caso, se recoge un artículo del *New York Times* (http://nyti.ms/2uNVMiC) comentando la última visita de Donald Trump a Europa.

3. Crear una cadena de procesamiento y analizar el corpus creado:

Hacer clic derecho sobre “Applications”, seleccionar “Create New Application” y marcar “Corpus Pipeline”, se nombra y se pulsa el botón “OK”.



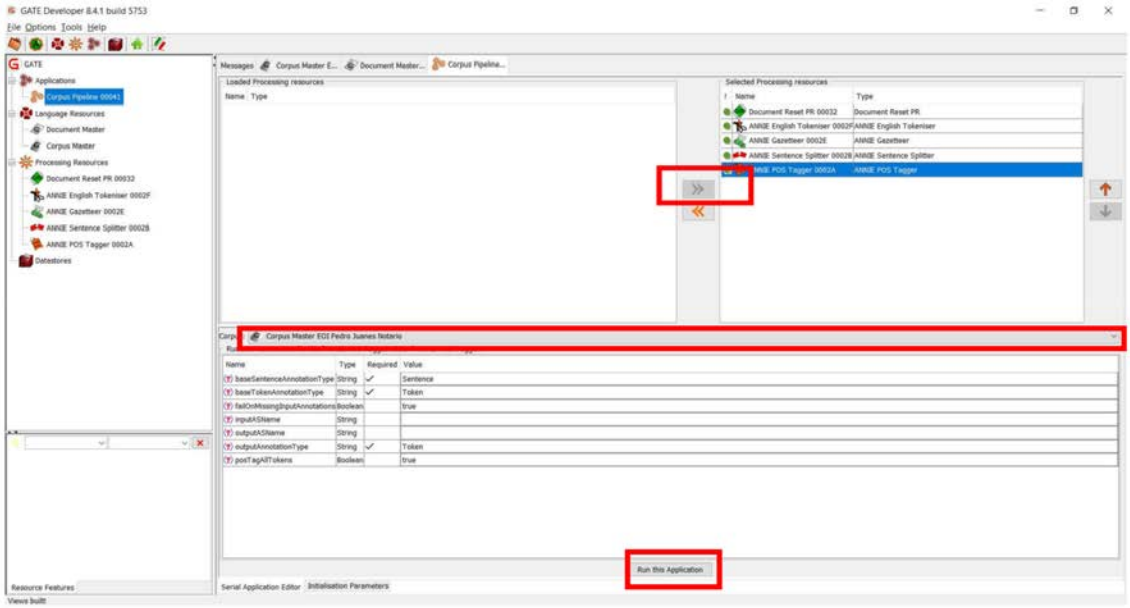
Hacer doble clic sobre “Corpus Pipeline” y el sistema mostrará la siguiente pantalla:



Ahora hay que incorporar los PRs de ANNIE en el orden adecuado. Para ello, mediante la “flecha verde derecha”, hay que ir agregándolos a la parte izquierda en el siguiente orden:

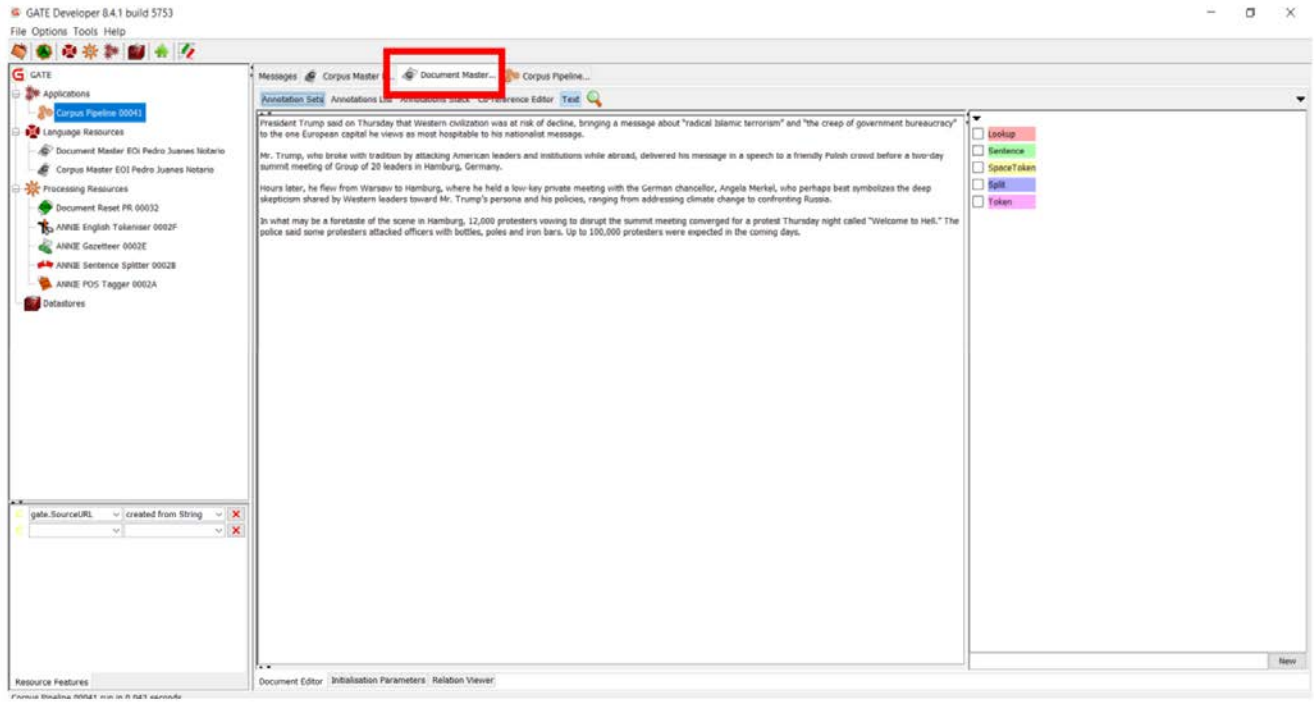
1. ‘Document Reset PR 00032’.
2. ‘ANNIE English Tokeniser 0002F’.
3. ‘ANNIE Gazetteer 0002E’.
4. ‘ANNIE Sentence Splitter 0002B’.
5. ‘ANNIE POS Tagger 0002A’.

Cuando ya estén incorporados todos estos procesos y en este orden, el usuario deberá hacer clic sobre el botón “Run this Application”, teniendo seleccionado el corpus anteriormente creado (“Corpus Master”) para que trabaje sobre él.



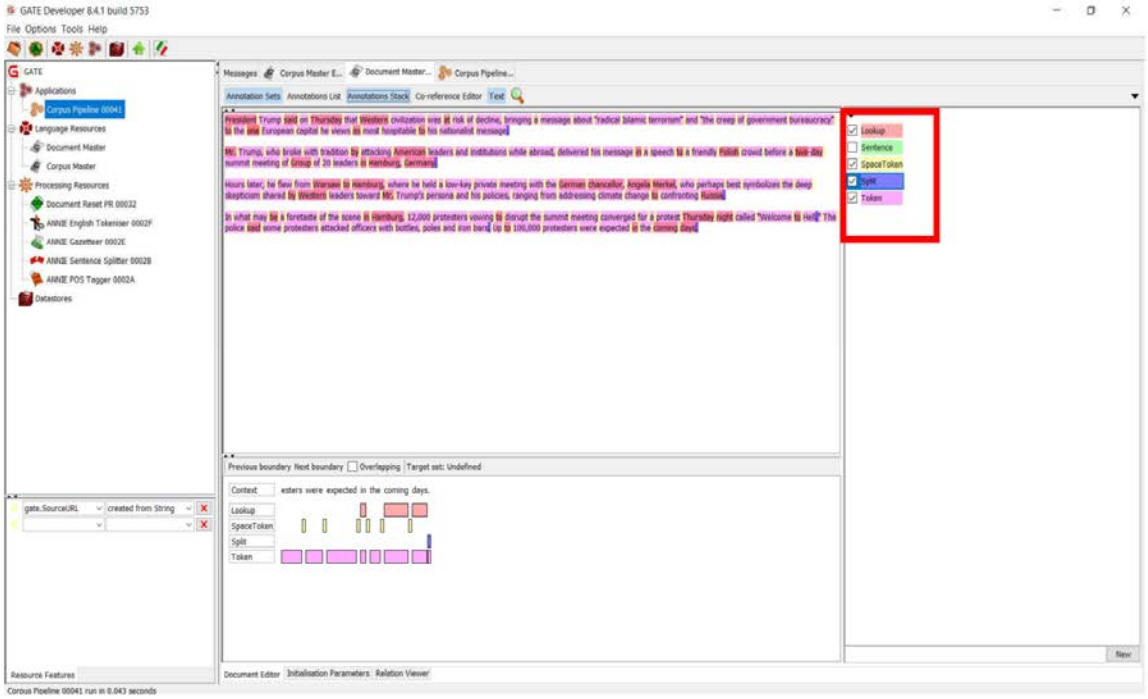
4. Visualizar el resultado del proceso:

Si se pincha sobre la pestaña “Documento...”, el sistema muestra el texto contenido en el mismo. Si nos situamos en la pestaña “Annotation Sets”, se podrá ver el resultado del análisis generado por el proceso lanzado con anterioridad.



Pinchando sobre los diferentes tipos de anotaciones que existen en ANNIE (‘Lookup’, ‘Sentence’,

‘SpaceToken’, ‘Split’ y ‘Token’), se podrá ver cómo el sistema ha realizado esa anotación sobre el texto del documento.


GATE Developer 8.4.1 build 5753
File Options Tools Help
GATE
Applications
Corpus Pipeline 00041
Language Resources
Document Master
Corpus Master
Processing Resources
Document Reset PR 00032
ANNIE English Tokeniser 0002F
ANNIE Gazetteer 0002E
ANNIE Sentence Splitter 0002B
ANNIE POS Tagger 0002A
Datastores
Messages
Corpus Master E...
Document Master...
Corpus Pipeline...
Annotation Sets
Annotations List
Annotations Stack
Co-reference Editor
Text
President Trump said on Thursday that Western civilization was at risk of decline, bringing a message about "radical Islamic terrorism" and "the creep of government bureaucracy"
Mr. Trump, who broke with tradition by attacking American leaders and institutions while abroad, delivered his message in a speech to a friendly Polish crowd before a two-day
Hours later, he flew from Warsaw to Hamburg, where he held a low-key private meeting with the German chancellor, Angela Merkel, who perhaps best symbolizes the deep
In what may be a foretaste of the scene in Hamburg, 12,000 protesters vowing to disrupt the summit meeting converged for a protest Thursday night called "Welcome to Hell." The
Lookup
Sentence
SpaceToken
Split
Token
Previous boundary
Next boundary
Overlapping
Target set: Undefined
Context
esters were expected in the coming days.
gate.SourceURL
created from String
Resource Features
New
Document Editor
Initialisation Parameters
Relation Viewer
Corpus Pipeline 00041 run in 0.043 seconds

5. Análisis del funcionamiento de cada proceso de ANNIE:

Hay que salvar el documento, el corpus y la aplicación creada en el PC. El orden de ejecución de los procesos, tal y como se ha señalado antes, debe ser el siguiente; las acciones que realiza cada uno de ellos se definen a continuación:

1. 'Document Reset'. Resetea el procesamiento realizado anteriormente en los documentos pertenecientes al corpus a analizar. Si se elimina de la cadena, el texto mantiene los datos que son producto del análisis anterior y se suman a los del proceso que se lanza.

2. 'ANNIE English Tokeniser'. Tokeniza el texto -tiene que ser en inglés— contenido en el documento/s del corpus. Genera la marca 'Token' y la marca 'SpaceToken'. Si se elimina este proceso, el sistema no puede realizar los siguientes (salvo 'ANNIE Gazetteer'), ya que se encarga de descomponer en unidades tokens y prepararlo para su análisis sintáctico y semántico.

3. 'ANNIE Gazetteer'. Trabaja como "diccionario" y localiza en el texto los nombres de entidades tales como países, ciudades, nombre de personajes, organizaciones, días de la semana, etc. y los categoriza e incluso genera una jerarquía (major type/minor type). Crea la marca 'Lookup'.

4. 'ANNIE Sentence Splitter'. Divide el texto en párrafos/frases. Genera la marca 'Split' (tipo 'internal'/punto seguido o aparte o 'external'/salto de párrafo) y la marca 'Sentence' (frase).

5. 'ANNIE POS Tagger'. Sobre los tokens y frases encontradas, este proceso etiqueta parte del discurso como una anotación para cada palabra o símbolo de puntuación. Este proceso utiliza un léxico y un conjunto de reglas predeterminado —tomados de un gran corpus procedente del periódico *The Wall Street Journal*—. Puede ser modificado manualmente.

2. GRAMÁTICA JAPE

Una gramática JAPE consiste en un conjunto de fases, cada una de las cuales consta de un conjunto de reglas/patrón de acción. Las fases se ejecutan secuencialmente y constituyen una cascada de transductores de estados finitos sobre anotaciones. Los del lado izquierdo (LHS) de las reglas se componen de una descripción del patrón de anotación. Los del lado derecho (RHS) se componen de instrucciones de manipulación de anotación. En esta práctica hay que hacer lo siguiente:

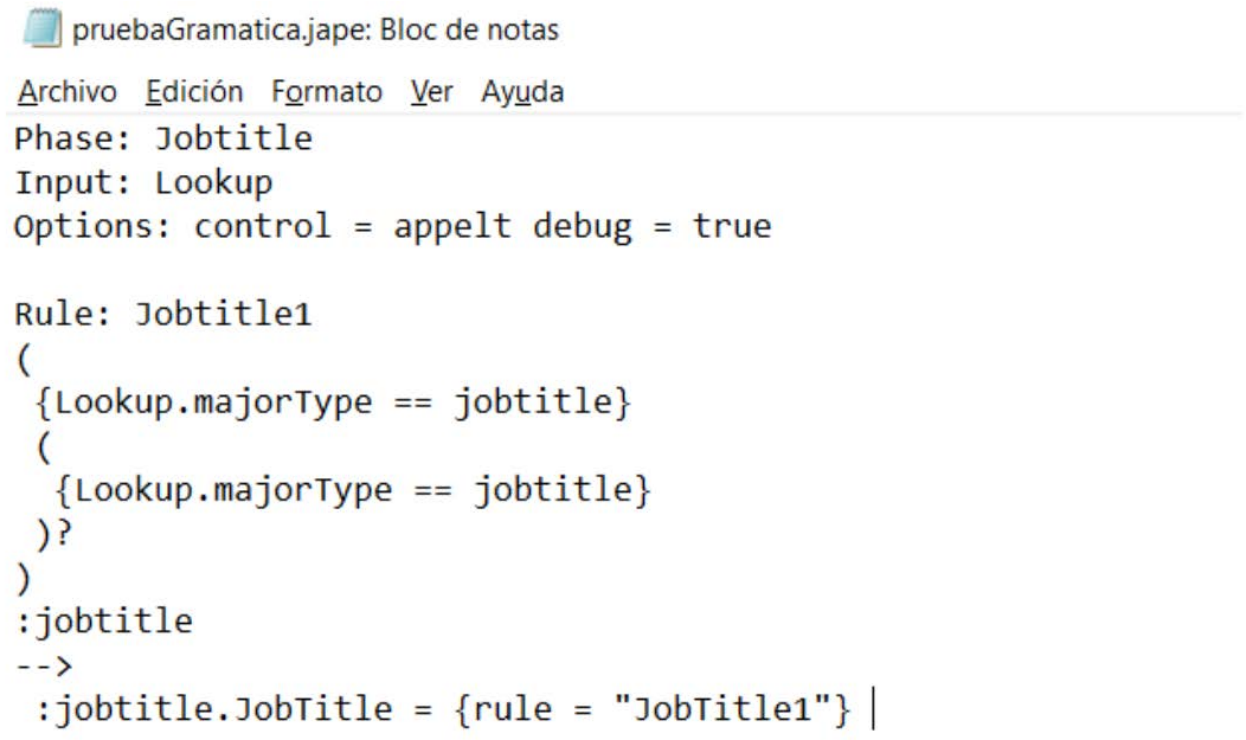
1. Leer el capítulo contenido en la siguiente URL:

https://gate.ac.uk/sale/tao/splitch8.html#chap:jape

2. Crear un fichero de texto plano y nombrarlo de la siguiente forma:

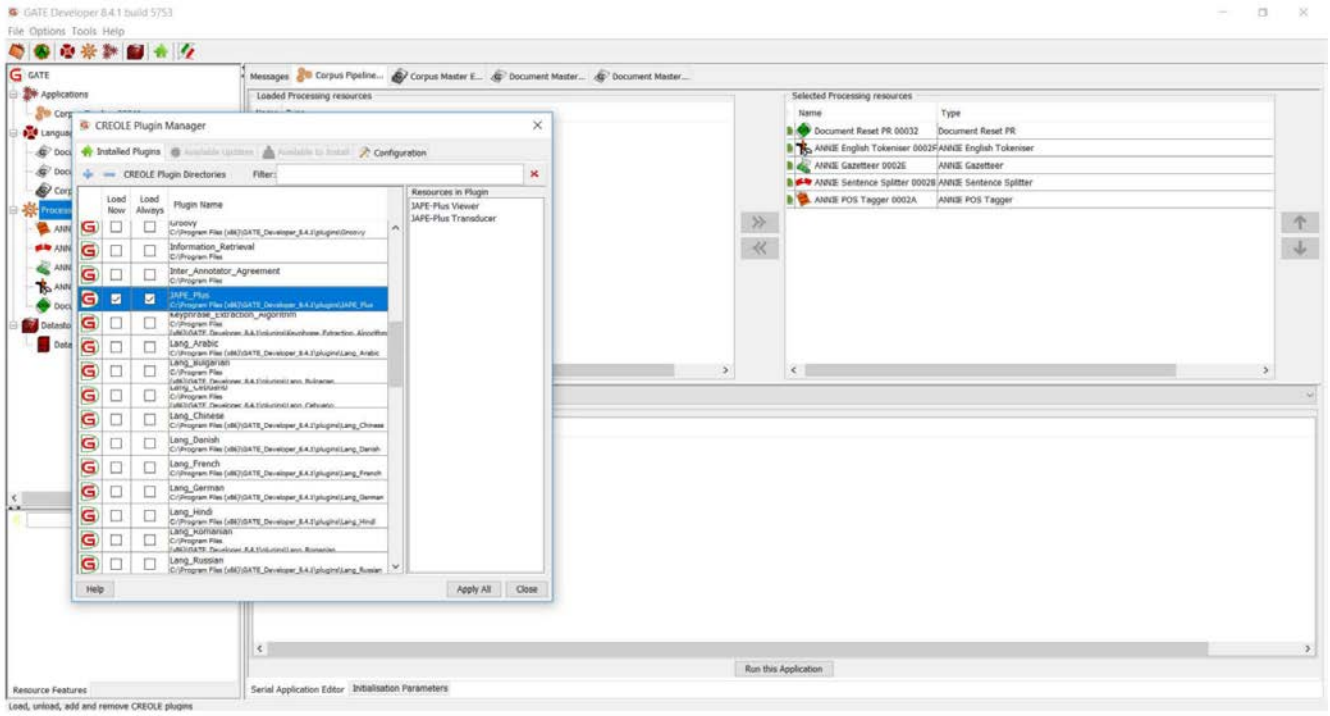
"pruebaGramatica.jape"

Este fichero contendrá el primer ejemplo del tutorial online que comienza por "Phase: Jobtitle" ... y termina por "rule = jobtitle1"). A continuación, se muestra el fichero creado y su contenido.

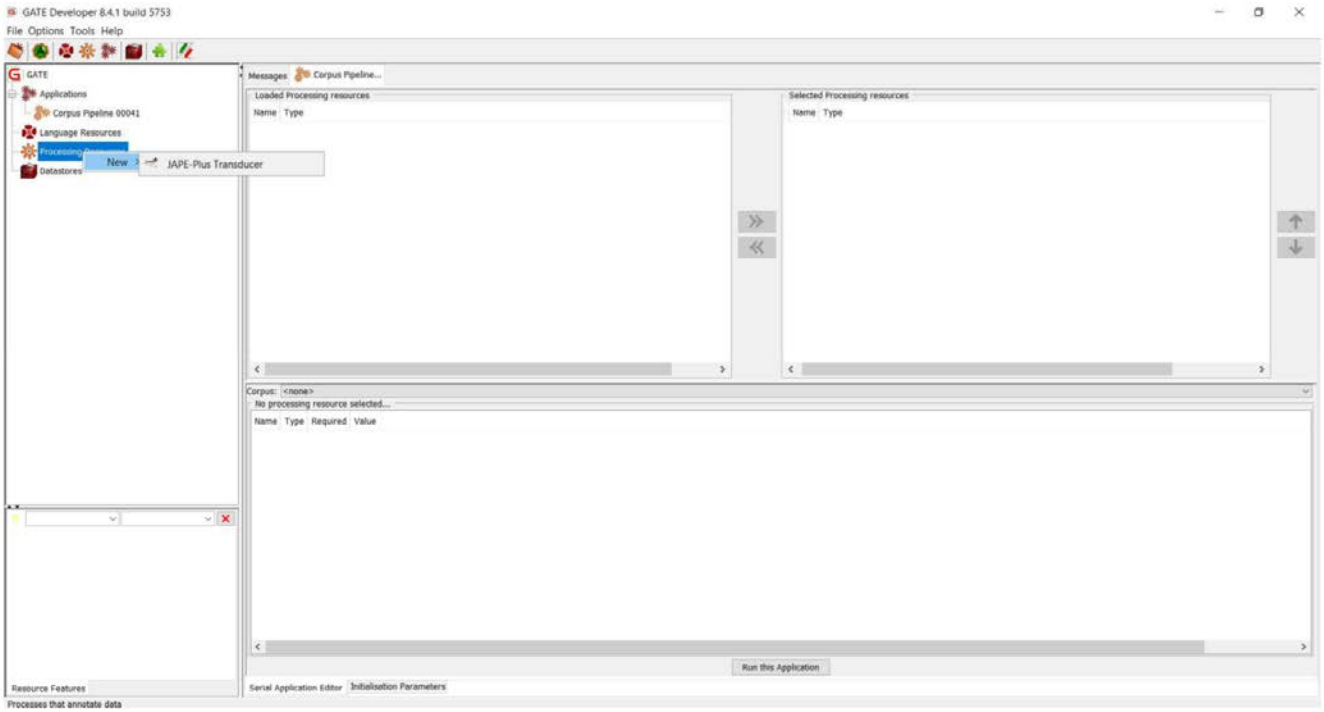


3. Cargar el Plugin “Jape Plus” en GATE de la siguiente forma:

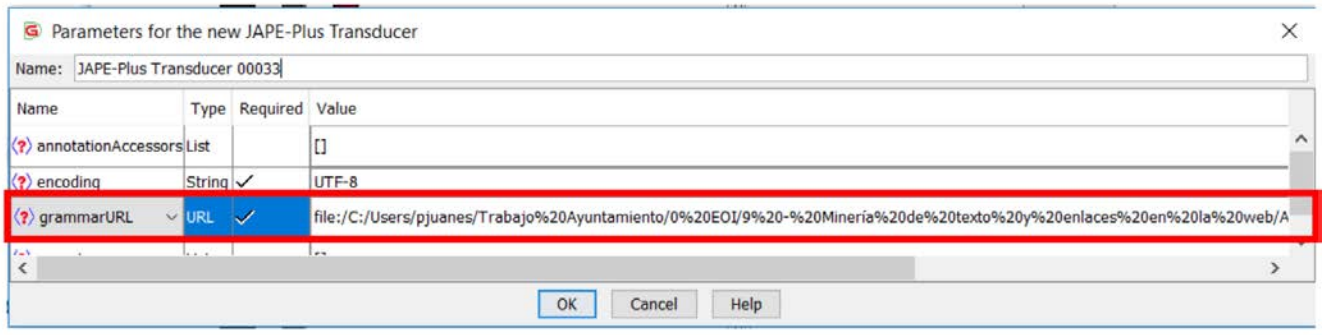
Hacer clic sobre el icono “puzle verde”, seleccionar “JAPE_Plus” y marcar el botón “Apply all”. Con este PR se podrá añadir una gramática y aplicarla sobre los “Language Resources” (corpus + documentos) que se procesaron y seleccionaron en el día 1 del ejercicio.



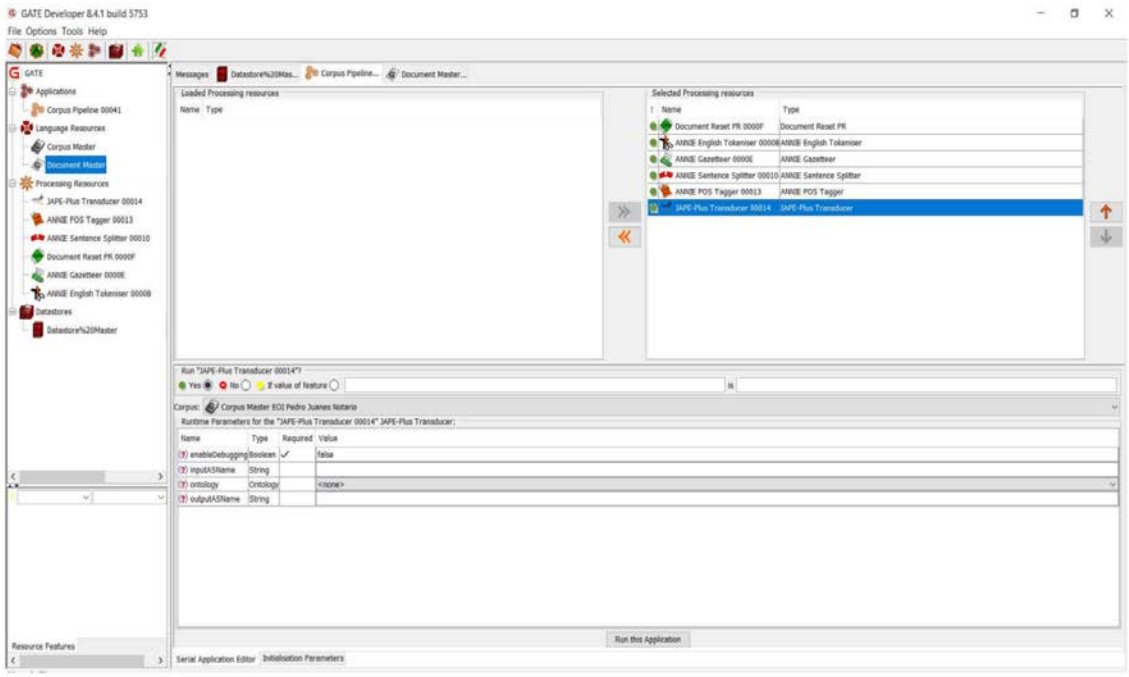
Con el botón derecho del ratón hay que señalar “Processing Resources” y seleccionar “JAPE-Plus Transducer”.



En la ventana que se presenta a continuación, en el apartado “grammarURL”, hay que seleccionar el fichero que se ha creado en el paso anterior “pruebaGramatica.jape” y pulsar “OK” para finalizar su incorporación al pipeline.



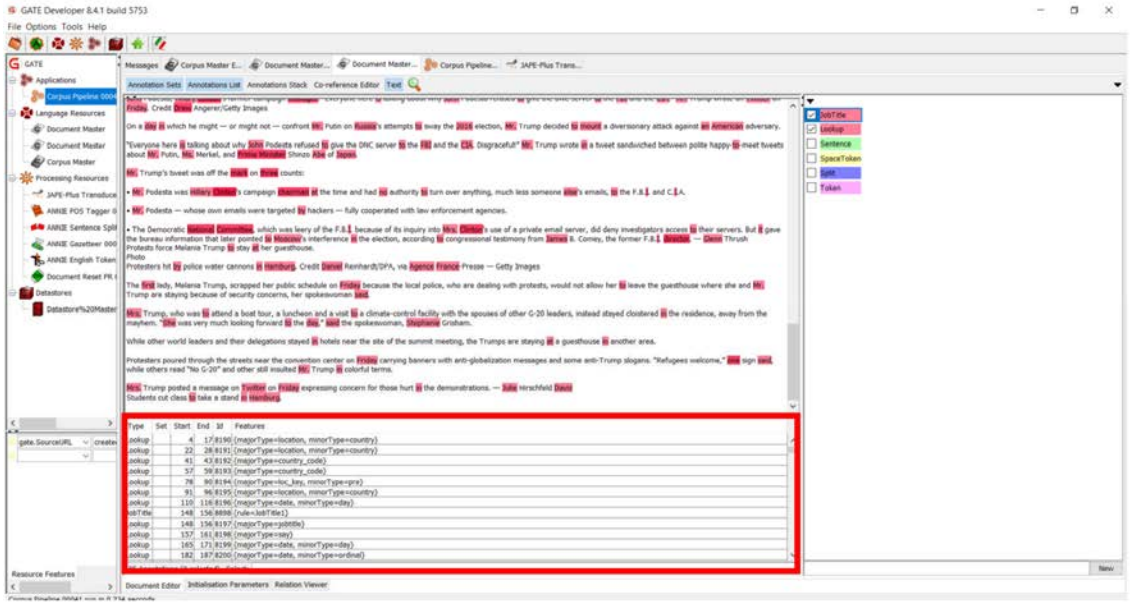
A continuación, se añade este nuevo proceso al final del pipeline que se conformó en el ejercicio del día 1, se ejecuta y se analiza su resultado sobre el corpus/documento que existía previamente.



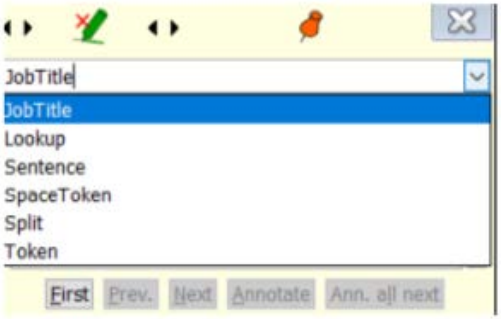
La gramática cargada contiene la siguiente sintaxis:

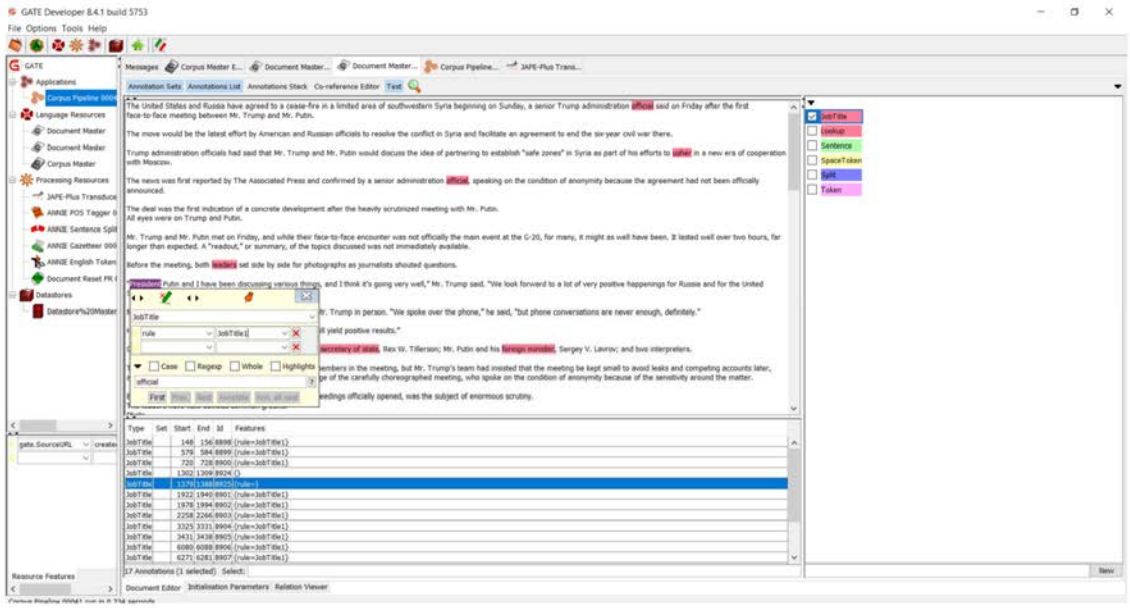
```
Phase: Jobtitle
Input: Lookup
Options: control = appelt debug = true
Rule: Jobtitle1
(
{Lookup.majorType == jobtitle}
(
{Lookup.majorType == jobtitle}
)?
)
:jobtitle
-->
:jobtitle.JobTitle = {rule = "JobTitle1"}
```

Esta gramática añade la anotación ‘JobTitle’ (profesión) que afecta al elemento ‘Lookup’, creado anteriormente mediante el proceso ‘ANNIE Gazetteer’, identificando este elemento en el texto del documento. A continuación, se muestra el resultado y el detalle de cada elemento en la parte inferior de la pantalla (recuadro rojo).



Si el usuario desea asignar una nueva anotación de este tipo —o de cualquier otro—, debe situarse sobre la palabra en concreto y hacer un doble clic para que aparezca el menú contextual de anotación. Entonces, hay que seleccionar la anotación dentro del combo existente y definir sus atributos.





3. CONCLUSIONES

Tal y como se señalaba en los apuntes, GATE es una herramienta que permite realizar tanto procesamiento y análisis NLTK como anotaciones sobre el texto, en un mismo entorno de trabajo. Su dificultad y usabilidad es adecuada y, desde el punto de vista del autor y sin profundizar más en la misma porque no es objeto de este ejercicio, es una buena herramienta para trabajar, de forma completa, en el análisis de textos.