

**TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN**  
**ĐẠI HỌC QUỐC GIA HỒ CHÍ MINH**  
**KHOA CÔNG NGHỆ THÔNG TIN**

---

**BÁO CÁO ĐỀ TÀI TÌM HIỂU**  
**GATE TOOLS**

**Giáo viên hướng dẫn:**

- Lê Nguyễn Hoài Nam

**Sinh viên thực hiện:**

- Mssv: 1412558
- Họ và tên: Phạm Quốc Toàn

*TP Hồ Chí Minh, ngày 20 tháng 06 năm 2018*

## Mục lục

<b>I. Giới thiệu Gate Tool .....</b>	<b>1</b>
1. Khái quát về Gate .....	1
2. Các tiện ích trong Gate .....	1
3. Lợi ích của Gate.....	1
4. Tải và cài đặt Gate .....	1
<b>II. Các ứng dụng xử lý trong Gate .....</b>	<b>1</b>
1. ANNIE .....	1
2. Các thư viện chuyên xử lý ngôn ngữ khác .....	2
5. Một số Tagger và thuật toán khác.....	2
<b>III. Tạo ứng dụng.....</b>	<b>2</b>
1. Dữ liệu đầu vào của Gate.....	2
2. Tạo ứng dụng với ANNIE .....	3
3. Phân tích kết quả sau khi xử lý văn bản .....	5
<b>IV. Các thư viện khác.....</b>	<b>6</b>
1. Thêm plugin .....	6
2. Tạo Controller.....	7
<b>V. Tài liệu tham khảo .....</b>	<b>8</b>

## I. Giới thiệu Gate Tool

### 1. Khái quát về Gate

- Là một Natural Processing Language (NPL) toolkit được phát triển tại đại học Sheffield, nước Anh kể từ năm 1995.
- Là mã nguồn mở Java và được phát hành miễn phí <http://gate.ac.uk>.
- Có sự đóng góp của nhiều cộng đồng phát triển, người dùng, nhà giáo dục, sinh viên và các chuyên gia trong nhiều lĩnh vực khác nhau.
- Nhiệm vụ chính là xử lý ngôn ngữ bao gồm giọng nói, nghiên cứu ung thư, hỗ trợ quyết định, khai thác thông tin web, chú thích ngữ nghĩa ...
- Có trên Maven được sử dụng như API, Servlet cho việc phát triển web

### 2. Các tiện ích trong Gate

- Thành phần xử lý ngôn ngữ: Parser, machine learning tools, stemmers, IR tool, IE components cho nhiều ngôn ngữ khác nhau như Anh, Đức, Pháp...
- Tool Trục quang hóa, biểu diễn và thao tác văn bản(chú thích, ontologies, parse tree..)
- Tool trích xuất nhiều thông tin khác nhau: khám phá những thông tin chưa biết trước đó
- Tool Công cụ đánh giá và tính điểm chuẩn

⇒ **Các thuật toán được sử dụng trong tool là hoàn toàn tách biệt, nó được phát triển một cách độc lập từ nhiều người dùng có chuyên môn khác nhau**

### 3. Lợi ích của Gate

- Tiết kiệm thời gian quản lý văn bản và dữ liệu từ nhiều nguồn khác nhau
- Tìm link ẩn thông qua các lượng lớn thông tin đã định dạng
- Thu thập thông tin và trích xuất các nhân tố mới

### 4. Tải và cài đặt Gate

- Tải Gate tại <http://gate.ac.uk/download/>
- Yêu cầu máy phải cài đặt Java JDK phiên bản tối thiểu là 7 cho Gate ver8, để biết thêm chi tiết truy cập <https://gate.ac.uk/sale/tao/splitch2.html#x5-170002.1>
- Tuy Gate là một tool nhưng chúng ta không cần cài đặt, sau khi tải về chỉ việc giải nén và Double-Click file Gate.exe là có thể sử dụng được

## II. Các ứng dụng xử lý trong Gate

### 1. ANNIE

- ANNIE là một ứng dụng được tạo sẵn bởi GATE kèm một vài plugin cơ bản nhất giúp người dùng nhanh chóng xử lý văn bản. Nó nằm trong IE system (sử dụng các kỹ thuật phân tích để tìm các interesting patterns), được phát triển bởi đội ngũ của Gate.

- Sử dụng ngôn ngữ quy tắc hành động kiểu mẫu hữu hạn JAPE (Java Annotation Patterns Engine) để xử lý các từ vựng.
- Ứng dụng ANNIE bao gồm 1 tập PRs (Processing Resources) như là:
  - English Tokeniser: Tách văn bản thành các từ có nghĩa dựa vào khoảng trắng.
  - Sentence Splitter: Phân văn bản thành các câu dựa vào dấu chấm, \t, \r.
  - POS tagger: Gán nhãn loại từ cho từng vệt dựa vào JAPE
  - Gazetteer: Xây dựng wordlist để thực hiện Named Entity. Kết quả là tập Named Entity được đánh dấu Name entity tagger:
  - Orthographic coreference: Bổ sung tên cho thực thể được tìm thấy bởi NE transducers

## **2. Các thư viện chuyên xử lý ngôn ngữ khác**

- Gate là một tool mà tập hợp nhiều bộ thư viện xử lý ngôn ngữ khác nhau. Do đó, người dùng có thể cùng lúc tạo nhiều ứng dụng với các bộ thư viện ANNIE, OpenNLP, Stanford\_CoreNLP ... để so sánh hiệu năng và tính đúng đắn đối với bộ dữ liệu chúng ta cần xử lý.
- Về cơ bản, các bộ thư viện đều có các class để xử lý Tokenizer, Sentence Splitter, POS tagger. Nhưng do nhiều người đóng góp cùng như văn hóa dùng ngôn ngữ mà kết quả đạt được của các thuật toán là khác nhau.

## **5. Một số Tagger và thuật toán khác**

- Ngoài một số thư viện chuyên xử lý ngôn ngữ, Gate còn tiếp nhận sự đóng góp từ cộng đồng một số thư để phục vụ việc phân số một số dữ liệu nhất định như:
  - Các Tagger: Tagger\_MetaMap, Tagger\_Date, Tagger\_Number, Tagger\_Chemistry...
  - Các thư viện: Machine Learning, Wordnet...

# **III. Tạo ứng dụng**

## **1. Dữ liệu đầu vào của Gate**

Từ thanh menu, chọn File -> New Language Resource -> Gate Document

Parameters for the new GATE Document

Name:

Name	Type	Required	Value
collectRepositioningInfo	Boolean	✓	false
encoding	String		
markupAware	Boolean	✓	true
mimeType	String		
preserveOriginalContent	Boolean	✓	false
sourceUrl	URL	✓	<input type="text"/>
sourceUrlEndOffset	Long		
sourceUrlStartOffset	Long		

OK Cancel Help

Chọn lần lượt các file văn bản cần xử lý, hoặc dán URL của trang web chứa văn bản. Sau đó File -> New Language Resource -> Gate Copus, để tạo con trỏ xử lý một lúc nhiều văn bản khác nhau cùng lúc.

Parameters for the new GATE Corpus

Name:

Name	Type	Required	Value
documentsList	List		<input type="text"/>

OK Cancel Help

List of gate.Document

demo2.txt\_0000C

Add Remove

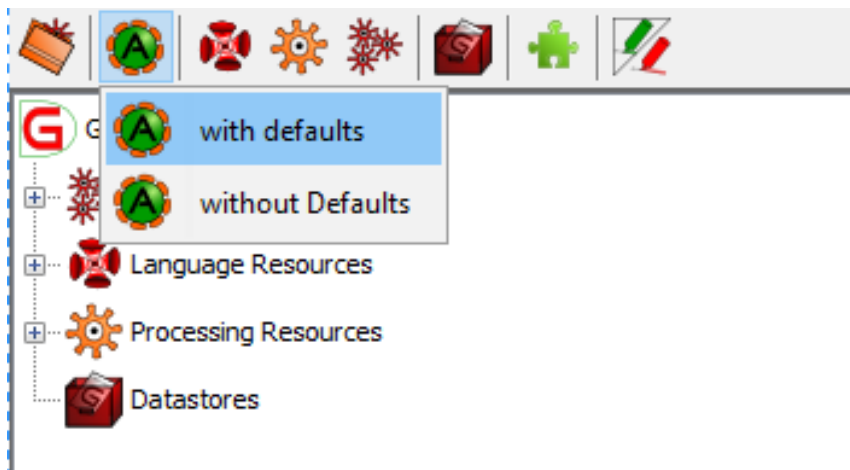
demo1.txt\_0000D

demo2.txt\_0000C

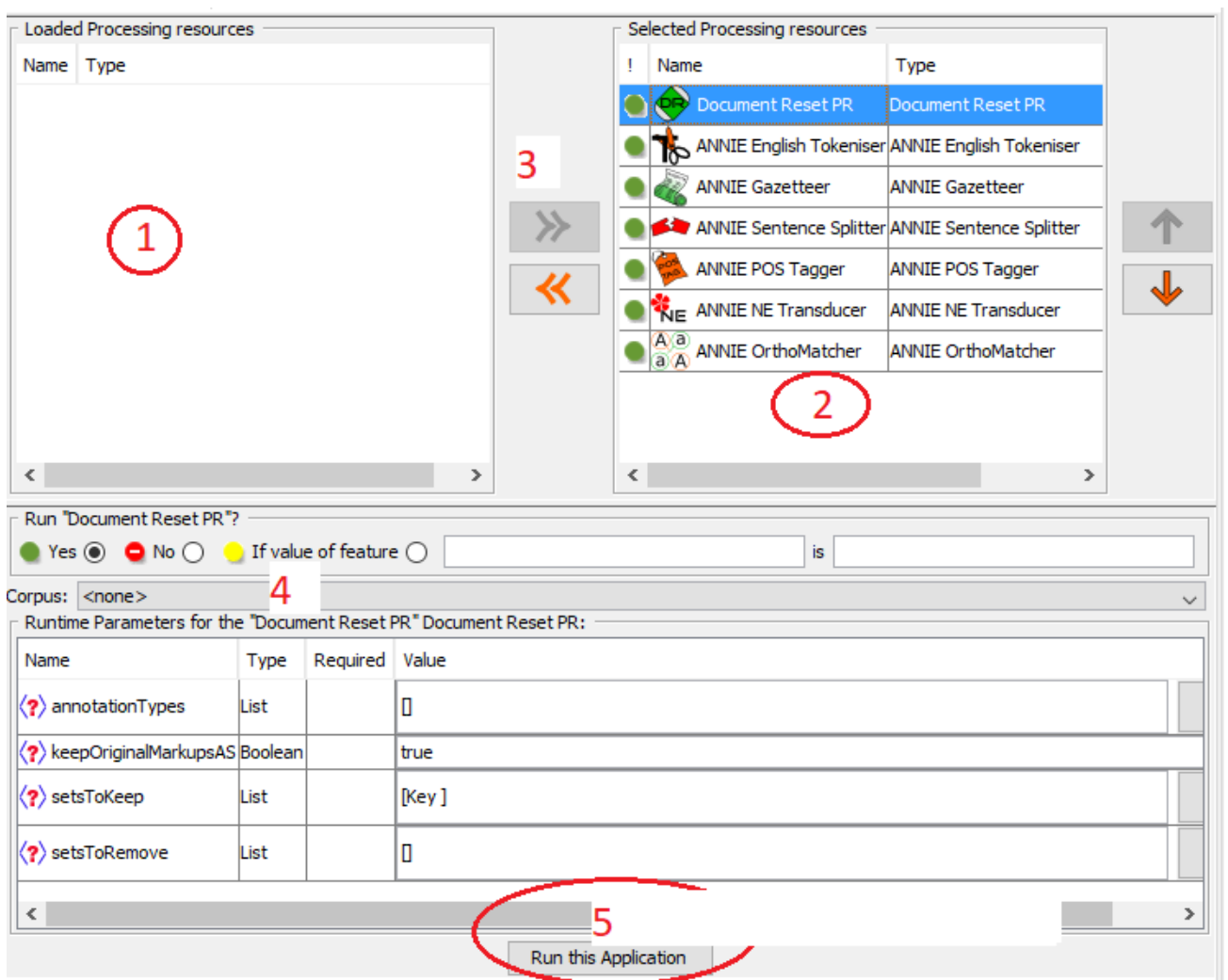
OK Cancel

Hình trên mô tả thao tác thêm văn bản vào Copus. Ta thêm các văn bản vào Copus bằng lệnh Add và Remove chúng nếu không dùng nữa.

## 2. Tạo ứng dụng với ANNIE



Chọn biểu tượng “A” trên tranh công cụ với “with defaults” để tạo ứng dụng mặc định ANNI được Gate tích hợp sẵn.



Vùng 1: Nơi chứa chức năng xử lý dữ liệu văn bản, bên dưới là thư viện đã được import sẵn và khởi tạo. Chúng được gọi là controller.

Vùng 2: Danh sách các controller được chọn cho ứng dụng.

Vùng 3: Điều khiển thêm, xóa controller của ứng dụng

Vùng 4: Chọn Copus(con trỏ chứa tập các văn bản)

Vùng 5: Bấm Run để khởi chạy ứng dụng

### 3. Phân tích kết quả sau khi xử lý văn bản

The screenshot shows the ANNIE application interface. The 'Annotations List' tab is active, displaying a table of annotations for a given text snippet. The text snippet is: "one people, tree books. there are two books. Here is attached map. Today is 10/05/2018. my dog also likes eating sausage. My name is Kuga. Mr. Join who retired last week died. I would like to go swimming. 5.5\*4^5. Ms.Kuga, her wife, died after this day. Golan Heights (CNN) Israel claims it struck almost all of Iran's military capabilities in Syria after what it says was an Iranian missile attack on the Golan Heights. In the most direct confrontation between Israel and Iran to date, the regional enemies exchanged fire for hours late Wednesday. The extended barrage of fire comes amid soaring tensions between Israel and Iran, rivals battling for regional influence, and less than two days after the United States withdrew from the deal to curb Iran's nuclear program. Health care delivery IE systems have been designed to summarise medical patient records by extracting diagnoses, symptoms, physical findings, test results, and therapeutic treatments. These systems can be used to assist health care providers with quality assurance studies, or to support insurance processing, where each patient encounter must be categorised for reimbursement purposes."

The table of annotations is as follows:

Type	Set	Start	End	Id	Features
Date		77	87	1192	{kind=date, rule=DateNumSlashDot, ruleFinal=DateOnlyFinal}
Person		133	137	1212	{NMRule=Unknown, kind=PN, matches=[1195, 1212], rule=Unknown}
Person		139	147	1193	{gender=male, kind=personName, rule=PersonTitle, ruleFinal=PersonFinal, surname=Join, title=Mr.}
Date		160	169	1194	{kind=date, rule=ModifierDate, ruleFinal=DateOnlyFinal}
Person		215	222	1195	{gender=[null], kind=personName, matches=[1195, 1212], rule=PersonTitle, ruleFinal=PersonFinal, surname=Kug}
Date		245	253	1196	{kind=date, rule=ModifierDate, ruleFinal=DateOnlyFinal}
Location		254	267	1197	{locType=city, matches=[1197, 1202], rule=Location1, ruleFinal=LocFinal}
Organization		269	272	1198	{orgType=company, rule=GazOrganization, ruleFinal=OrgFinal}
Location		273	279	1199	{locType=country, matches=[1199, 1203, 1206], rule=Location1, ruleFinal=LocFinal}
Location		311	315	1200	{locType=country, matches=[1200, 1204, 1207, 1209], rule=Location1, ruleFinal=LocFinal}
Location		343	348	1201	{kind=locName, locType=country, rule=InLoc1, ruleFinal=LocFinal}
Location		405	418	1202	{locType=city, matches=[1197, 1202], rule=Location1, ruleFinal=LocFinal}
Location		461	467	1203	{locType=country, matches=[1199, 1203, 1206], rule=Location1, ruleFinal=LocFinal}
Location		472	476	1204	{locType=country, matches=[1200, 1204, 1207, 1209], rule=Location1, ruleFinal=LocFinal}
Date		532	546	1205	{kind=date, rule=GazDate, rule2=EarlyDate, ruleFinal=DateOnlyFinal}
Location		613	619	1206	{locType=country, matches=[1199, 1203, 1206], rule=Location1, ruleFinal=LocFinal}
Location		624	628	1207	{locType=country, matches=[1200, 1204, 1207, 1209], rule=Location1, ruleFinal=LocFinal}
Location		703	716	1208	{locType=country, rule=Location1, ruleFinal=LocFinal}
Location		748	752	1209	{locType=country, matches=[1200, 1204, 1207, 1209], rule=Location1, ruleFinal=LocFinal}

Sau khi Run ứng dụng, ta click vào tệp văn bản mún xem kết quả:

Click Annotation Sets để hiện thị các tập đối tượng mà ứng dụng phát hiện được, nó hiện thị bên lề phải và được tô các màu khác nhau. Chúng ta sẽ tick vào nó để được highlight tương ứng bên văn bản giúp tìm nhanh ra các từ hoặc cụm từ liên quan.

Click Annoations List: hiện thị chi tiết các thuộc tính, thành phần của đối tượng được chúng ta tick bên phải.

Kết quả chạy với ANNIE with Default sẽ cho ta kết quả chung, thông thường được thực thi từ các thư viện xử lý cơ bản nhất như : English Tokenizer, POS, Sentence Splitter, Named Entity.

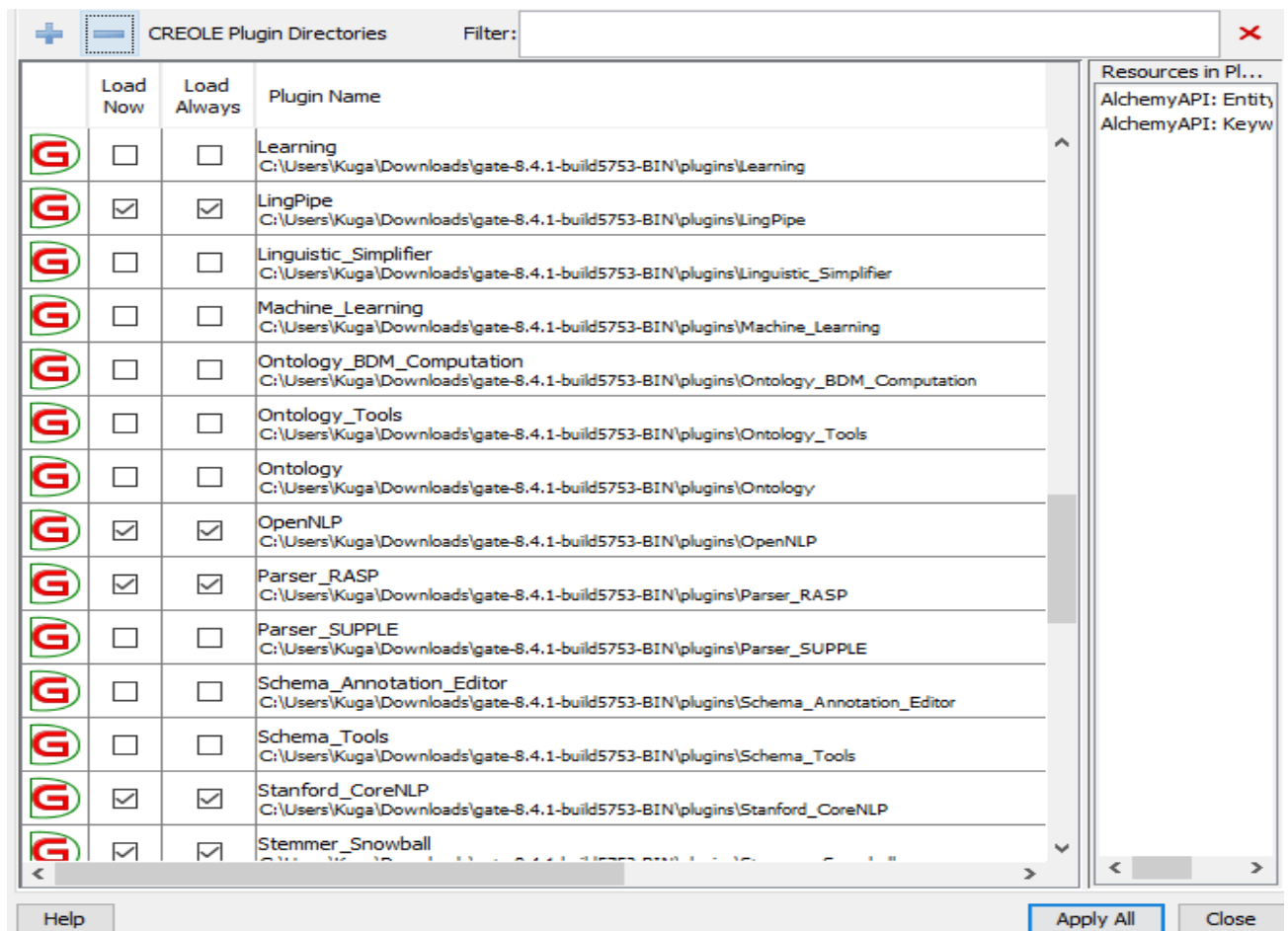
## IV. Các thư viện khác

### 1. Thêm plugins

Ngoài việc xây dựng ANNIE với một vài thư viện có sẵn trong GATE. Chúng ta cũng có thể ứng dụng riêng của mình với nhiều bộ thư viện cùng xử lý một chức năng cho văn bản. Để thêm các thư viện, Plugins này ta thực hiện như sau. File -> Manage CREOLE Plugins hoặc chọn Icon dấu “+” trên menu



Kết quả cho ra một giao diện cho phép ta tick chọn các thư viện cần thiết.




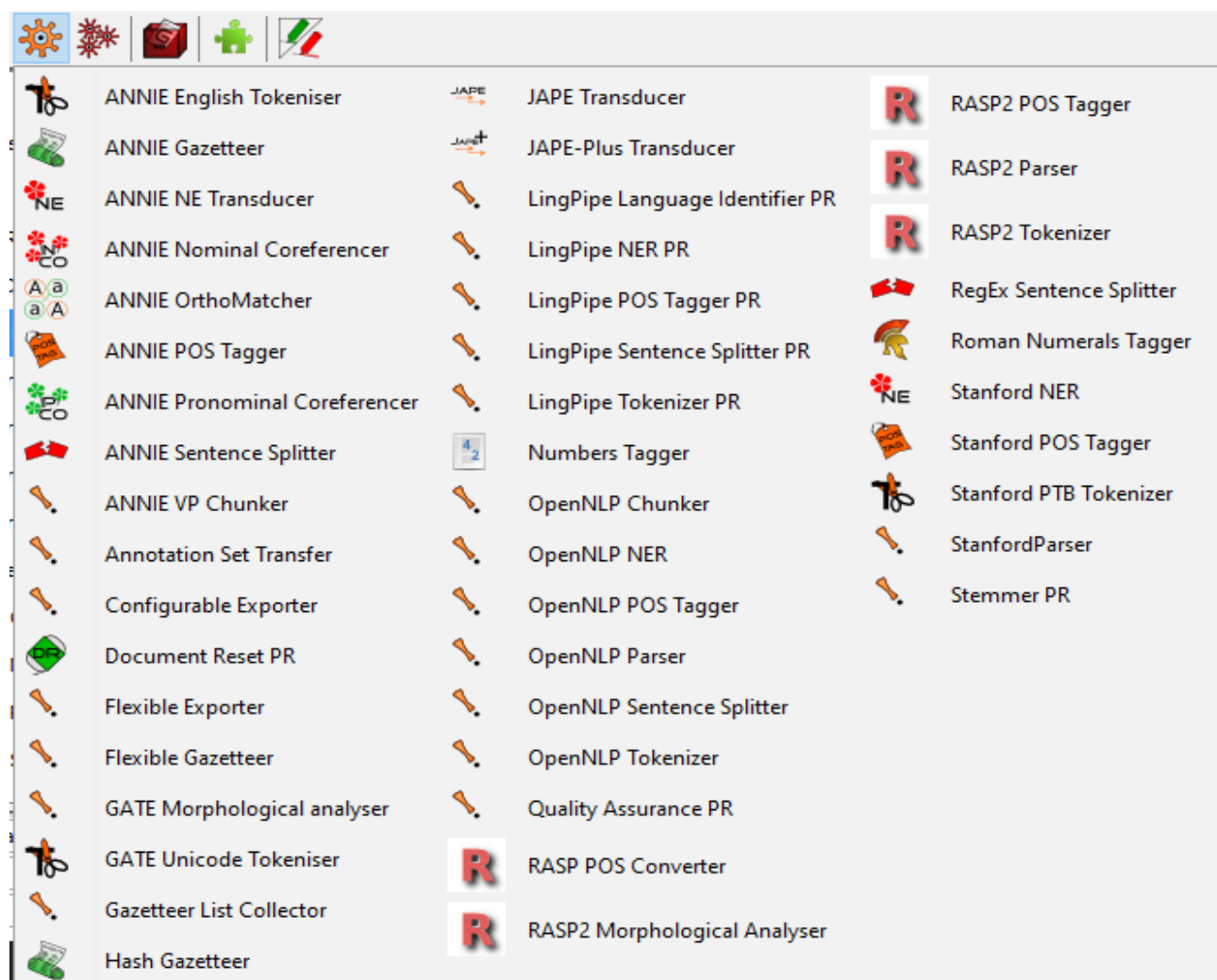


Sau khi tick chọn xong ta chọn Apply All. Lúc này thư viện đã được import vào GATE và chúng ta cần khởi tạo Controller tương ứng cho thư viện.

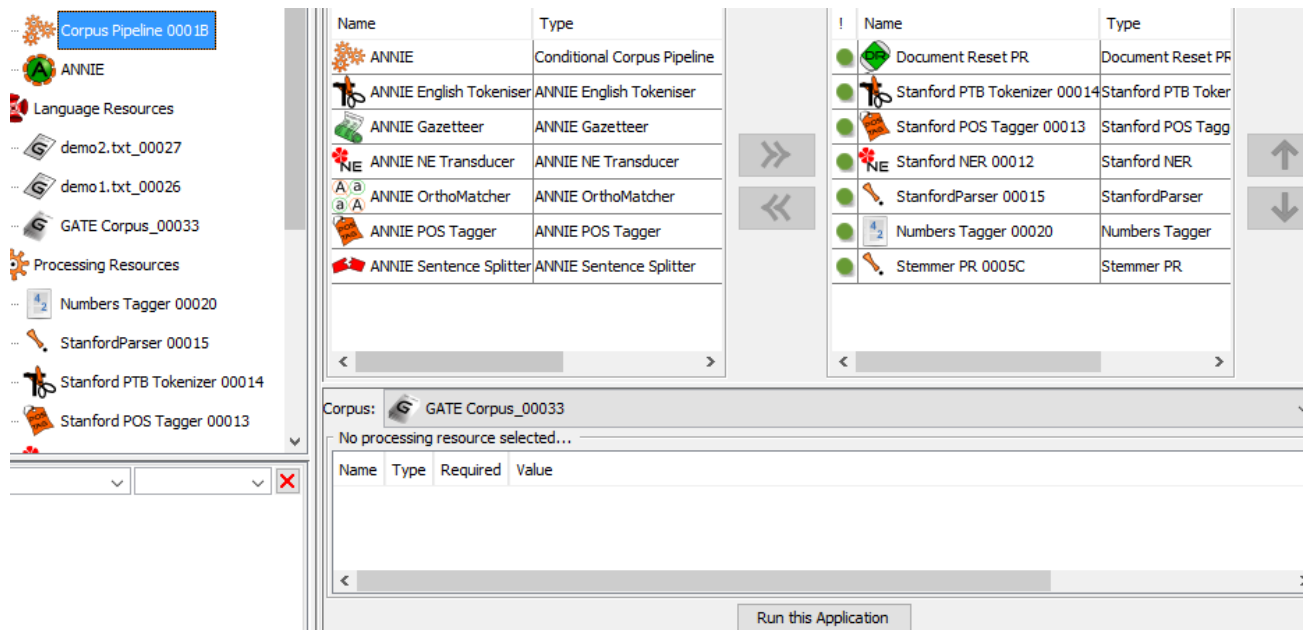
## 2. Tạo Controller

Trên thanh menu, chọn File - > New Processing Resource, chọn controller mình cần, sau đó chúng sẽ xuất hiện bên sidebar bên trái của GATE.

Ngoài ra chúng ta có thể click vào biểu tượng  trên menu và chọn controller cho nhanh.



Để tạo ứng dụng, trên menu chọn File -> New Application -> Corpus Pipeline



Thao tác sử dụng giống với tạo ứng dụng ANNIE đã đề cập ở mục III của tài liệu này.

## V. Tài liệu tham khảo

<https://www.slideshare.net/dianamaynard/text-analysis-in-gate>

<https://gate.ac.uk/wiki/quick-start/GATE-Tutorial-2010-Gate4.0.pdf>

<https://gate.ac.uk/demos/movies.html>