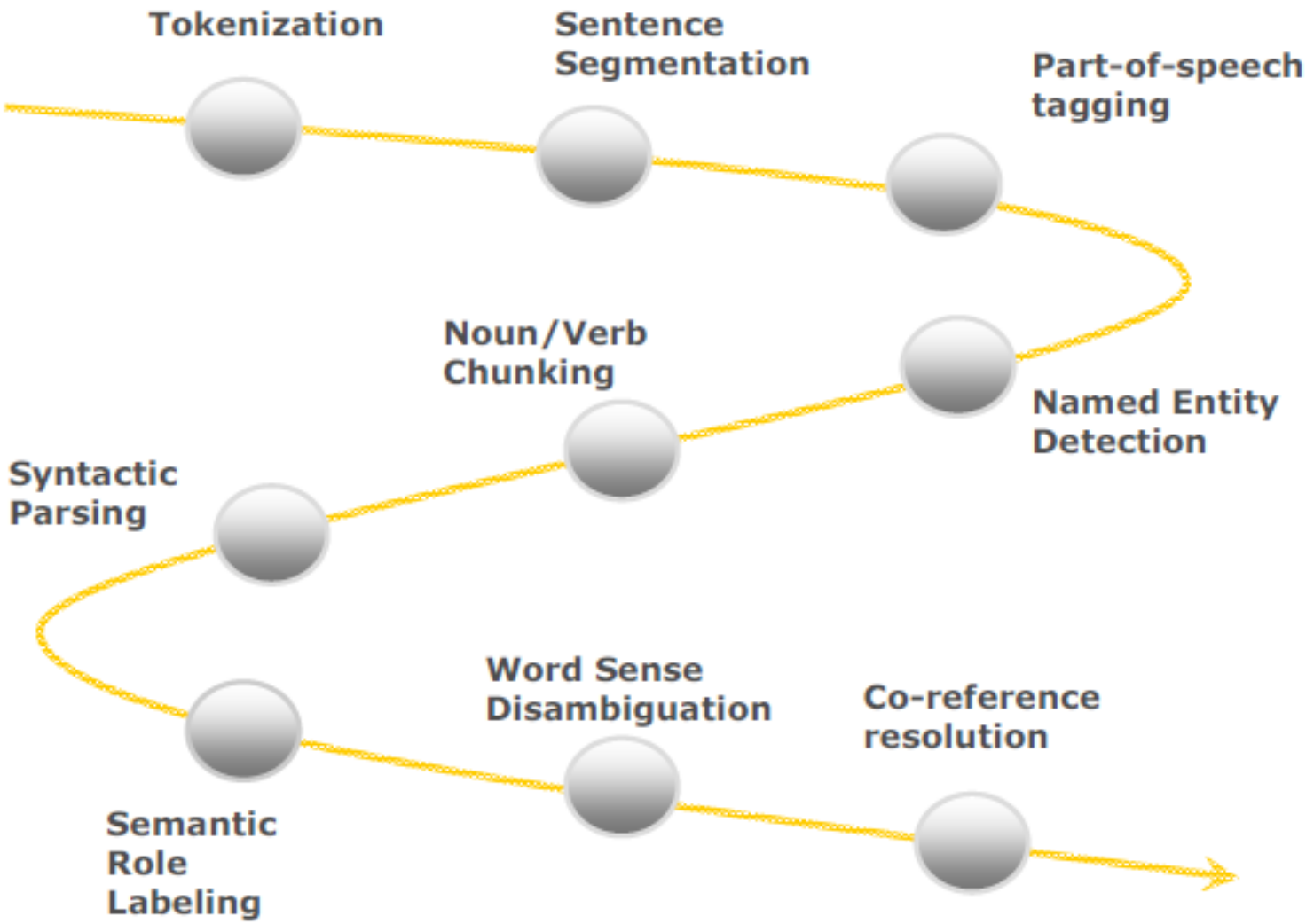
Исследование корпусов с помощью среды GATE

Ланин В.В.

Конвейер обработки ЕЯ (NLP pipeline)



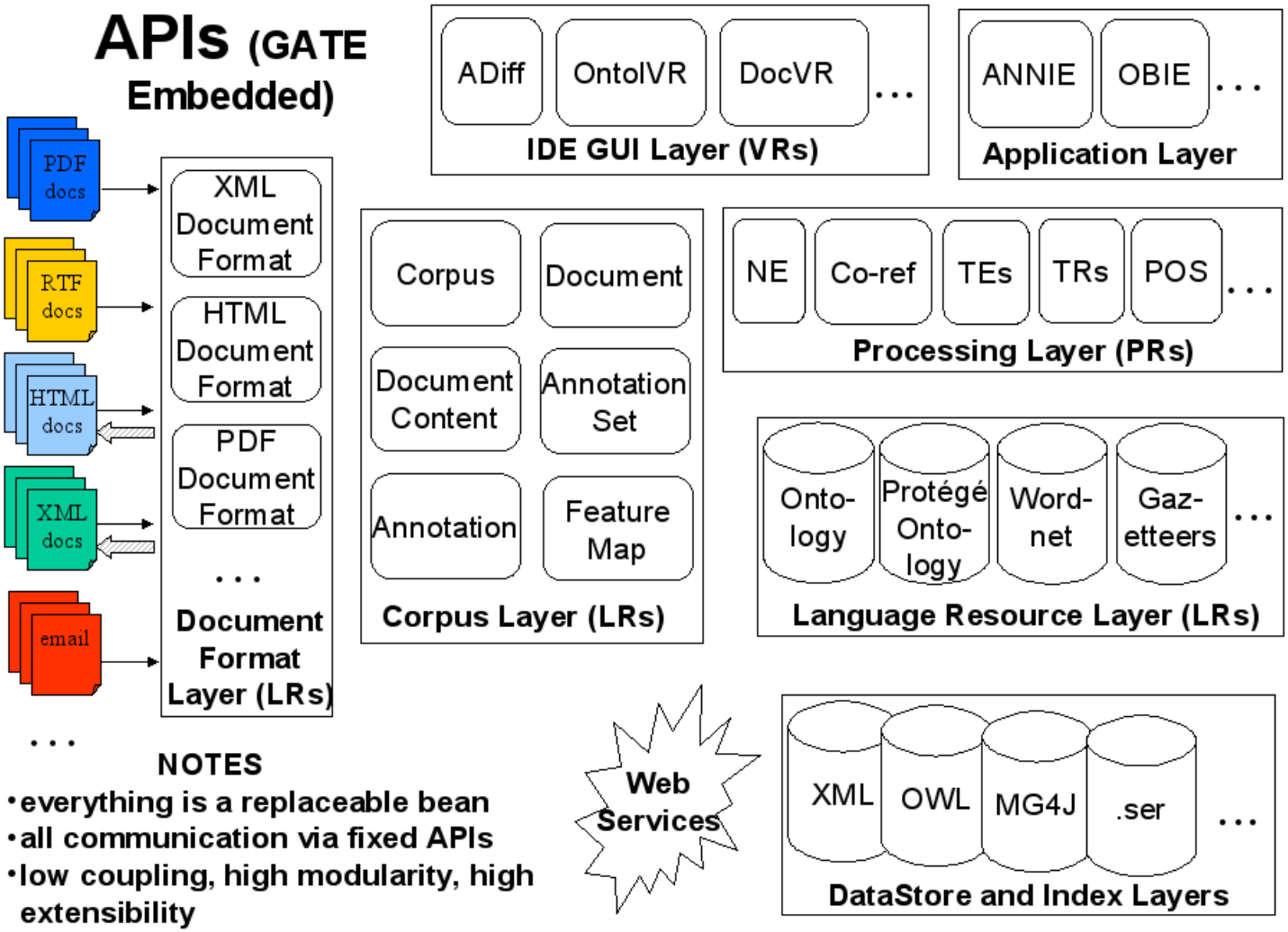
GATE - General Architecture for Text Engineering

- GATE система обработки естественного языка с открытым исходным кодом, использующая наборы компонентов на языке Java.
- Система изначально была разработана в Университете Шеффилда и развивается с 1995 г.

GATE - General Architecture for Text Engineering

- Среда для решения лингвистических задач
- Открытый исходный код, java
- Набор готовых инструментов
- Модули расширения (PlugIns)
- Графический интерфейс пользователя
- Облачная версия (платная)
- Большое сооющество и документация

Архитектура GATE



Основные понятия GATE

- Аннотация
- Языковые ресурсы
  - Документ
  - Корпус
- Процессинговые ресурсы
- Конвеер (Pipeline)

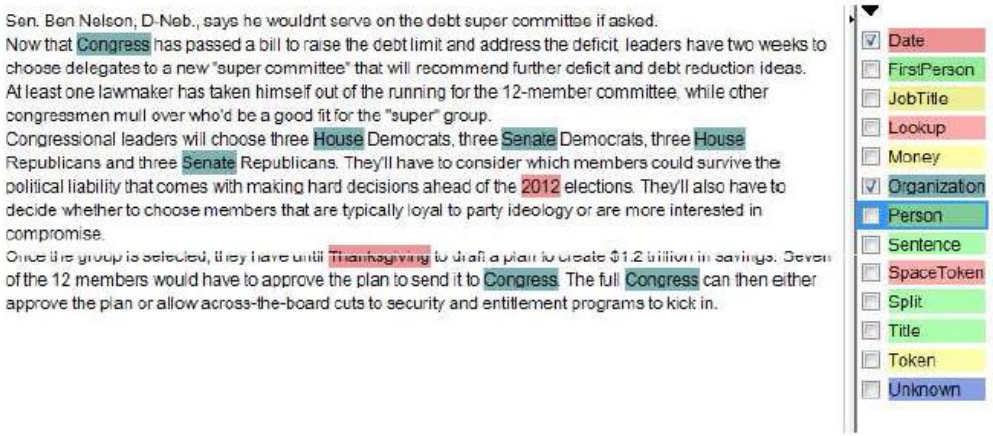
Аннотация

- Универсальный способ представления информации о тексте
- Всегда соответствует некоторый отрезок текста
- Документ может иметь неограниченное количество аннотаций
- Всегда типизирована
- Может иметь атрибуты, набор которых расширяем

Языковые и процессинговые ресурсы

- Language Resource (LR): хранение информации (документы, корпуса, онтологии, словари)
- Processing Resource (PR): изменение LR (токенизаторы, стеммеры, морфология, синтаксис, выделение объектов, поиск отношений...)

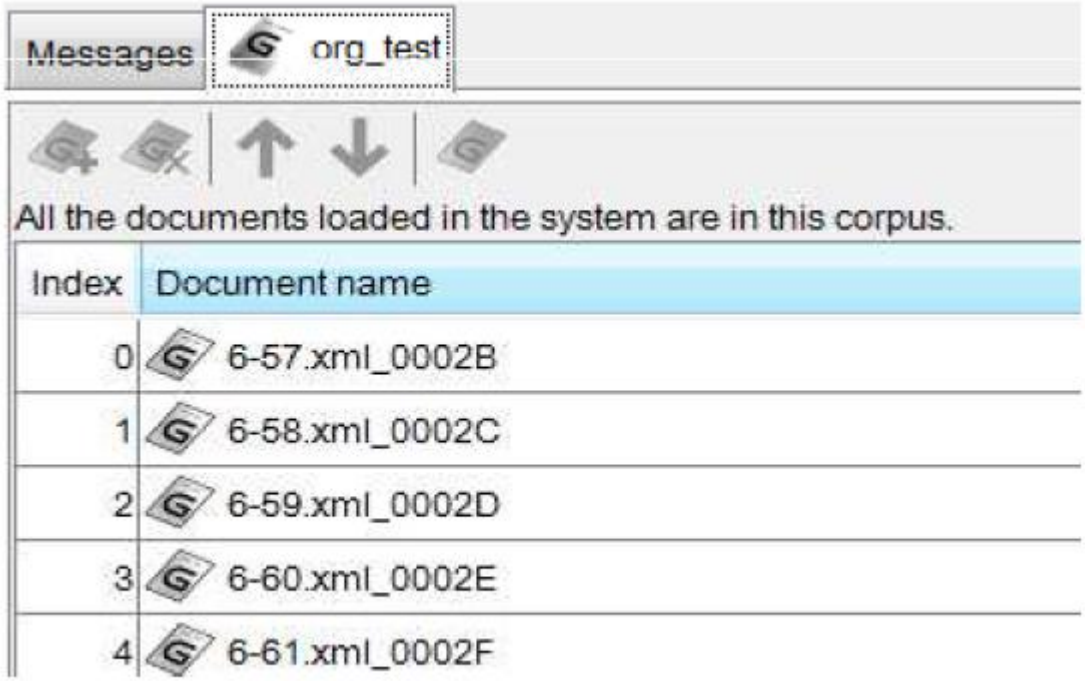
Языковые ресурсы: документ



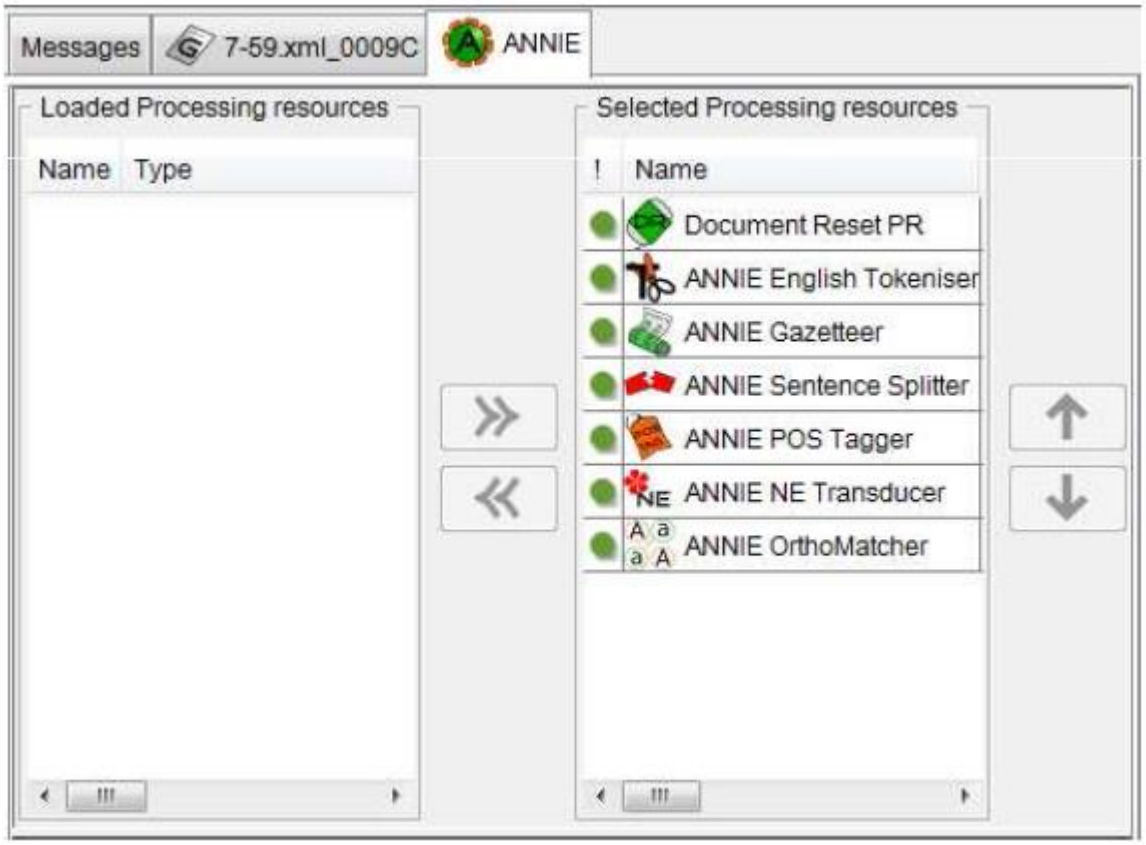
- Имя
- Текст (контент)
- Аннотации
- Атрибуты

Языковые ресурсы: корпус

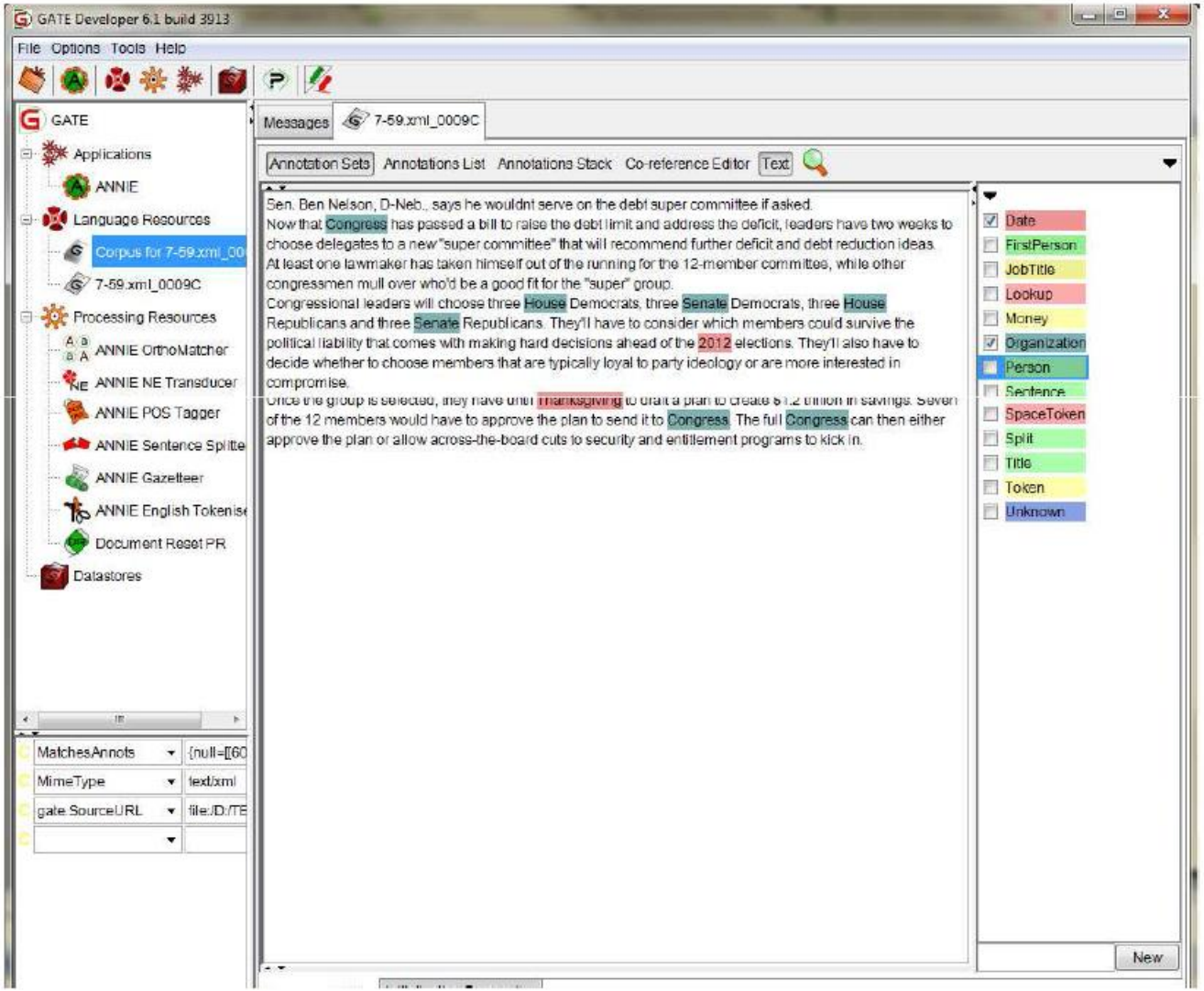
- набор размеченных документов с поддержкой общих операций над ними



Конвейер обработки



Графический интерфейс пользователя



Jape Transducer

Jape – язык регулярных выражений над

- аннотациями с использованием логических
- операторов

Правило состоит из левой и правой частей

- Левая часть – jape-шаблон
- Правая часть – действие (стандартное jape-действие или код на java)

Правила организованы в фазы

- Фазы могут включать в себя другие фазы, т.е. можно строить иерархию любой сложности

https://gate.ac.uk/