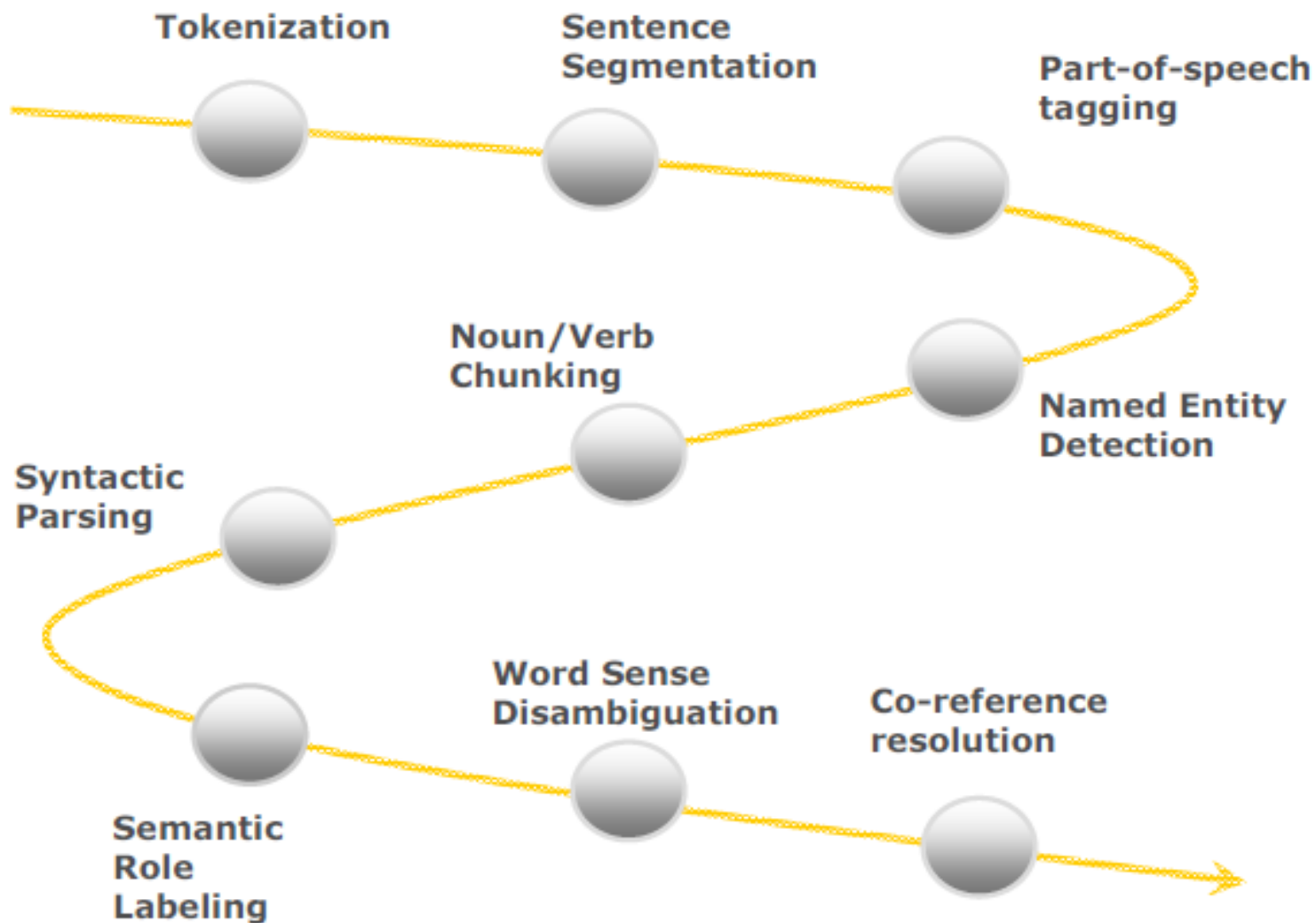


Исследование корпусов с помощью среды GATE

Ланин В.В.

Конвейер обработки ЕЯ (NLP pipeline)



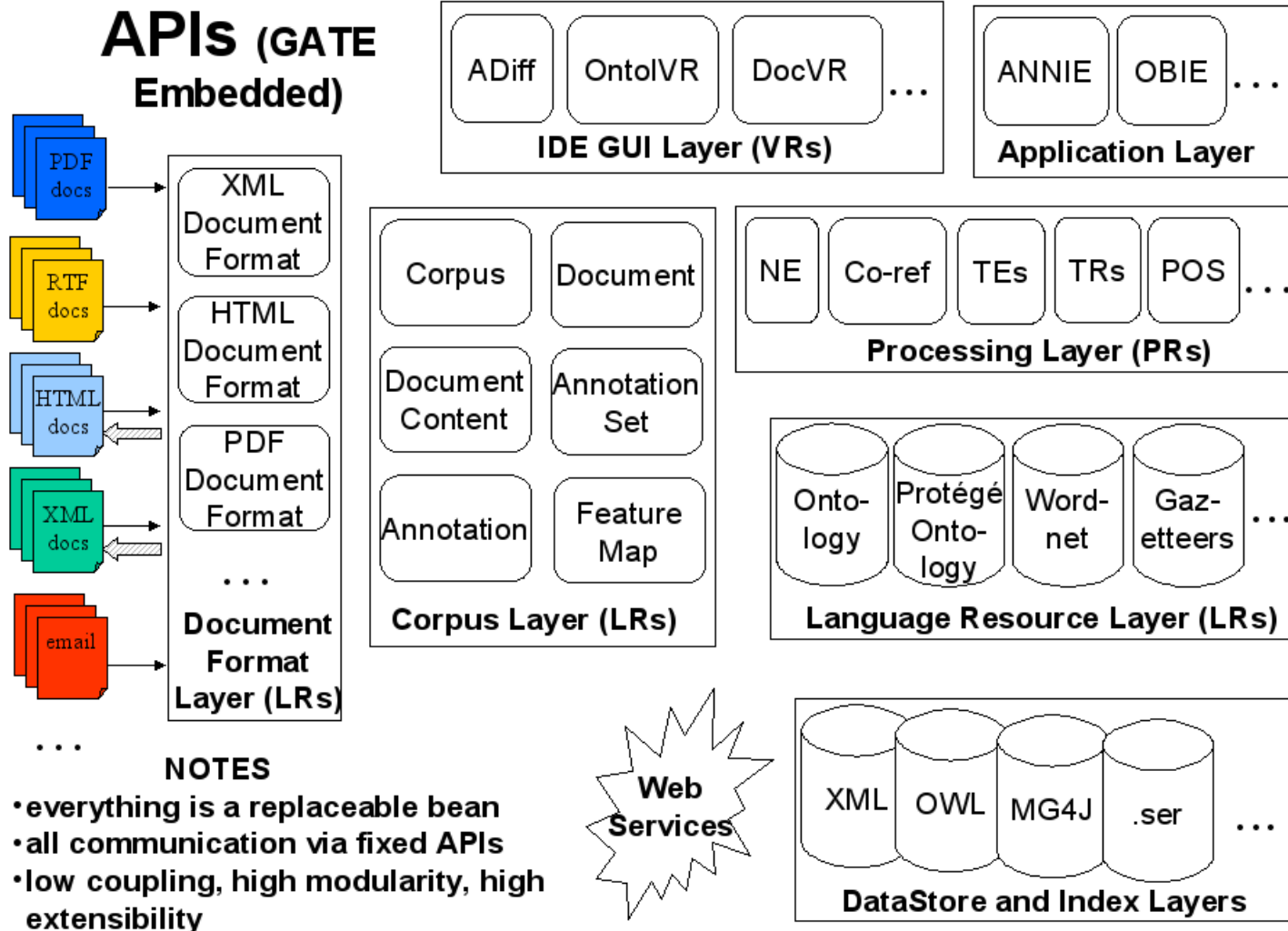
GATE - General Architecture for Text Engineering

- 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 сообщество и документация

Архитектура GATE



Основные понятия GATE

- Аннотация
- Языковые ресурсы
 - Документ
 - Корпус
- Процессинговые ресурсы
- Конвейер (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

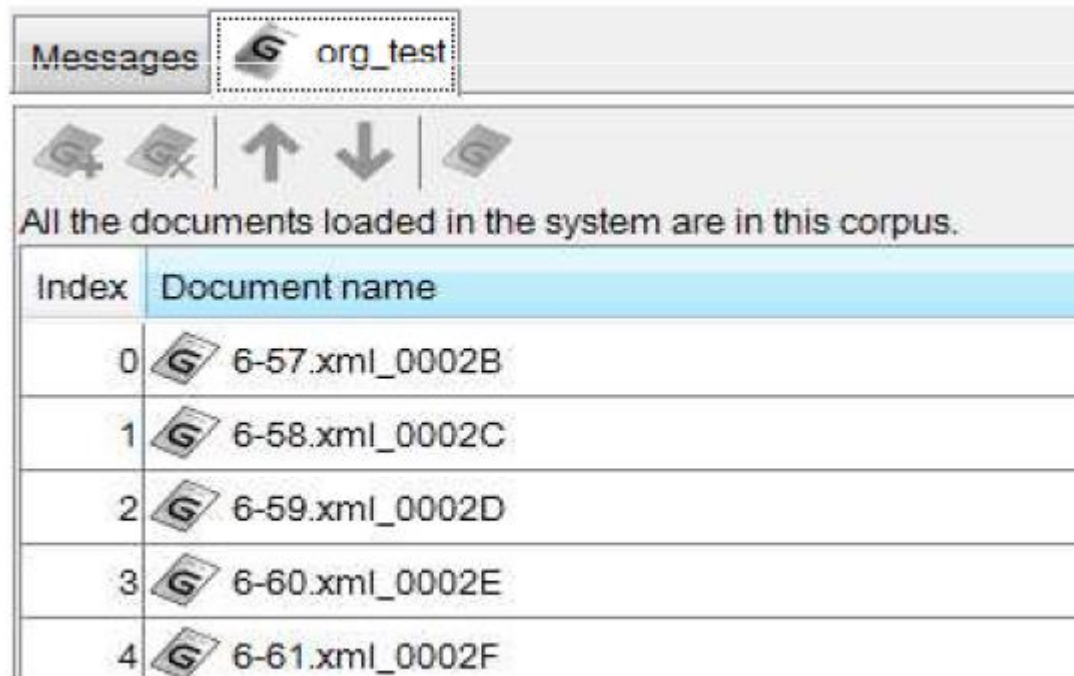
Sen. Ben Nelson, D-Neb., says he wouldnt serve on the debt super committee if asked. Now that **Congress** has passed a bill to raise the debt limit and address the deficit, leaders have two weeks to choose delegates to a new "super committee" that will recommend further deficit and debt reduction ideas. At least one lawmaker has taken himself out of the running for the 12-member committee, while other congressmen mull over who'd be a good fit for the "super" group. Congressional leaders will choose three **House** Democrats, three **Senate** Democrats, three **House** Republicans and three **Senate** Republicans. They'll have to consider which members could survive the political liability that comes with making hard decisions ahead of the **2012** elections. They'll also have to decide whether to choose members that are typically loyal to party ideology or are more interested in compromise. Once the group is selected, they have until **Thursday** to draft a plan to create \$1.2 billion in savings. Seven of the 12 members would have to approve the plan to send it to **Congress**. The full **Congress** can then either approve the plan or allow across-the-board cuts to security and entitlement programs to kick in.








- Имя
- Текст (контент)
- Аннотации
- Атрибуты

Языковые ресурсы: корпус






- набор размеченных документов с поддержкой общих операций над ними





Messages

  |   | 

All the documents loaded in the system are in this corpus.

Index	Document name
0	 6-57.xml_0002B
1	 6-58.xml_0002C
2	 6-59.xml_0002D
3	 6-60.xml_0002E
4	 6-61.xml_0002F















Конвейер обработки

Messages  7-59.xml_0009C  ANNIE

Loaded Processing resources

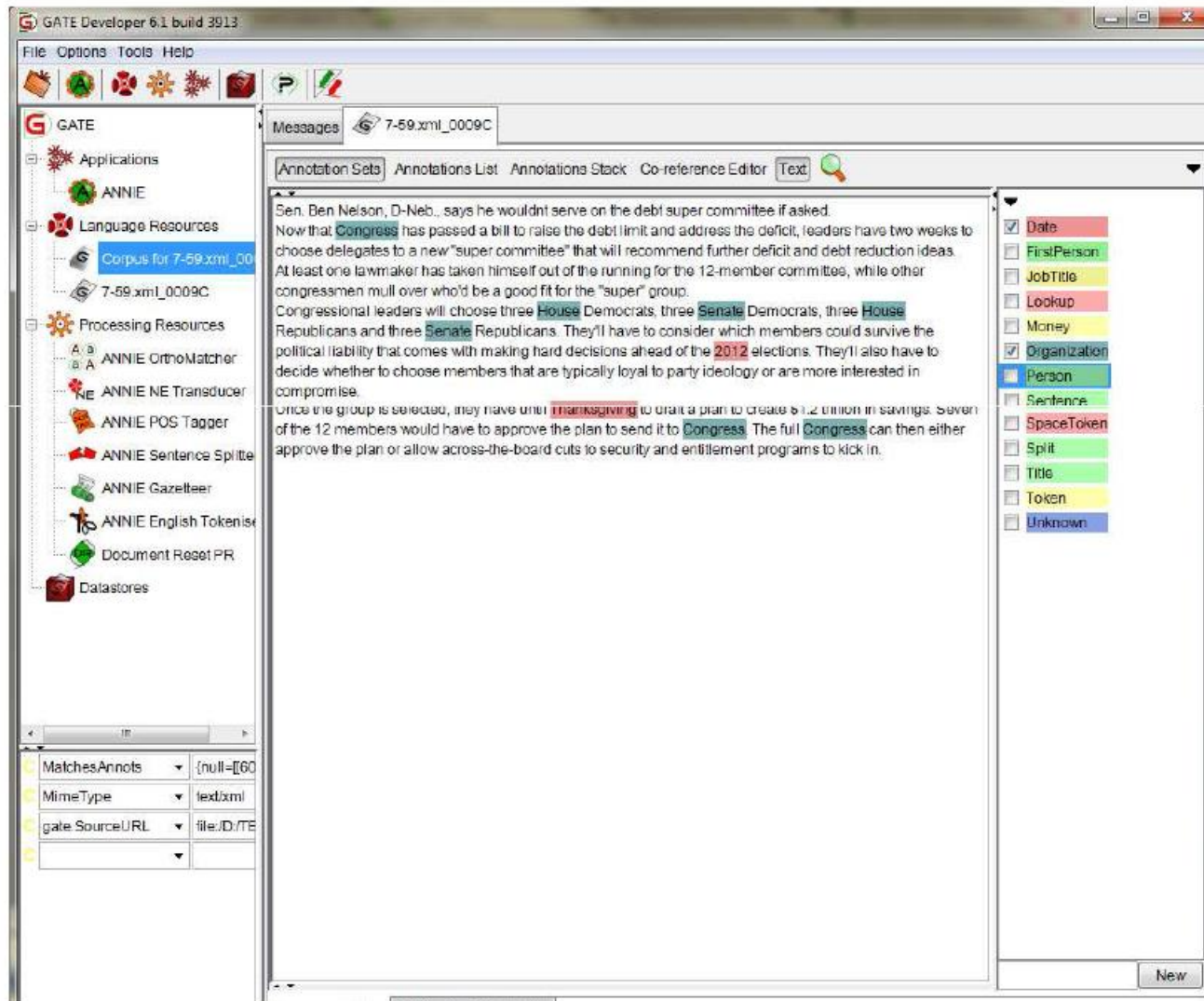
Name	Type
------	------

Selected Processing resources

!	Name
	 Document Reset PR
	 ANNIE English Tokeniser
	 ANNIE Gazetteer
	 ANNIE Sentence Splitter
	 ANNIE POS Tagger
	 ANNIE NE Transducer
	 ANNIE OrthoMatcher

Navigation buttons: >>, <<, ↑, ↓

Графический интерфейс пользователя



Java Transducer

Java – язык регулярных выражений над

- аннотациями с использованием логических
- операторов

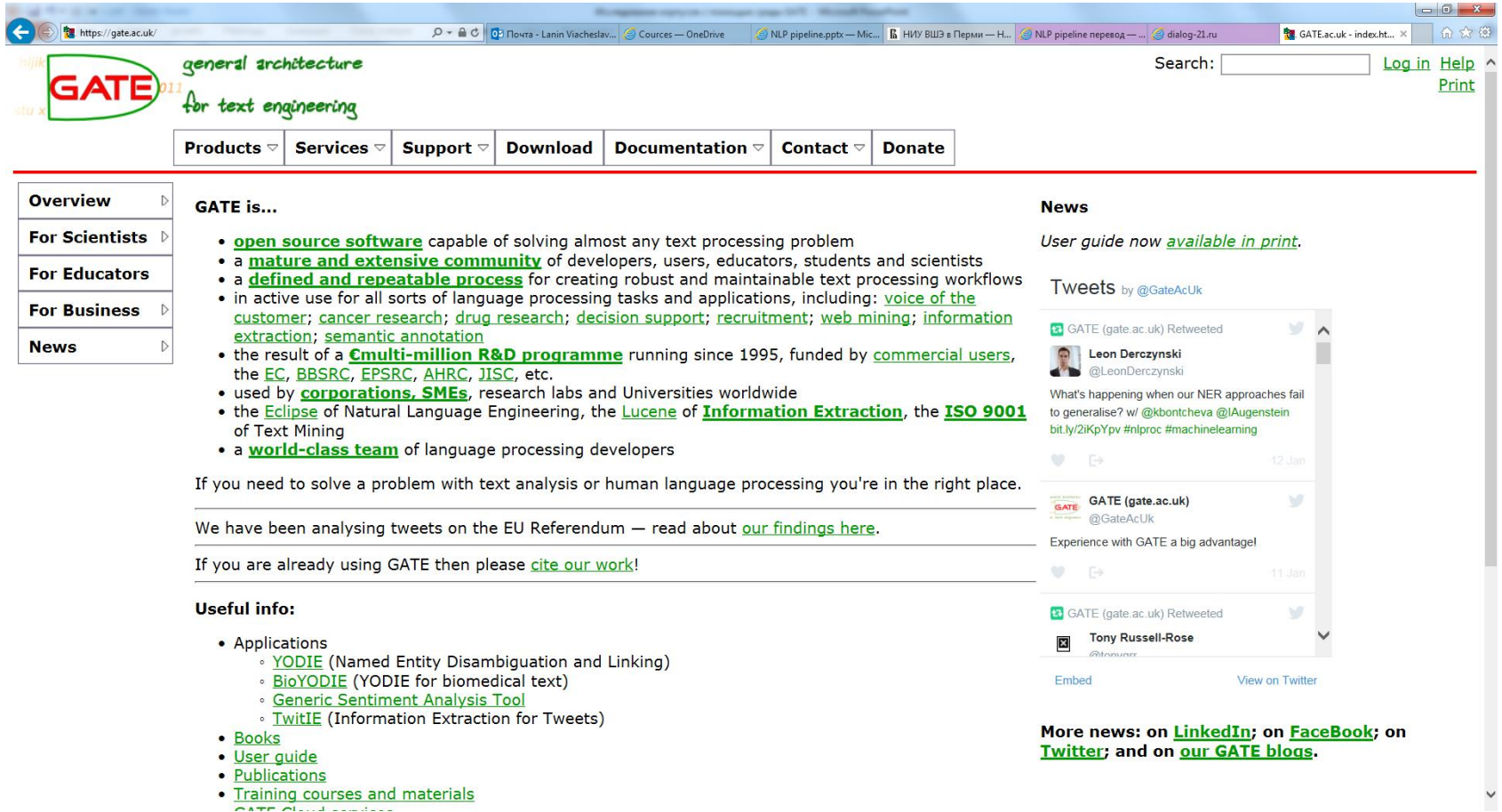
Правило состоит из левой и правой частей

- Левая часть – java-шаблон
- Правая часть – действие (стандартное java-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/



The screenshot shows the GATE website homepage. At the top, there is a navigation bar with a search box and links for 'Log in', 'Help', and 'Print'. Below this is a main navigation menu with categories: Products, Services, Support, Download, Documentation, Contact, and Donate. The main content area is divided into three columns. The left column contains a sidebar with links for Overview, For Scientists, For Educators, For Business, and News. The middle column features the 'GATE is...' section, which lists key features and achievements of the project. The right column contains a 'News' section with a tweet from Leon Derczynski and another from GATE (@GateAcUk) about the EU Referendum findings. At the bottom right, there is a section for 'More news' with links to LinkedIn, Facebook, Twitter, and GATE blogs.

general architecture
GATE for text engineering

Search: [Log in](#) [Help](#) [Print](#)

[Products](#) [Services](#) [Support](#) [Download](#) [Documentation](#) [Contact](#) [Donate](#)

Overview **For Scientists** **For Educators** **For Business** **News**

GATE is...

- [open source software](#) capable of solving almost any text processing problem
- a [mature and extensive community](#) of developers, users, educators, students and scientists
- a [defined and repeatable process](#) for creating robust and maintainable text processing workflows
- in active use for all sorts of language processing tasks and applications, including: [voice of the customer](#); [cancer research](#); [drug research](#); [decision support](#); [recruitment](#); [web mining](#); [information extraction](#); [semantic annotation](#)
- the result of a [multi-million R&D programme](#) running since 1995, funded by [commercial users](#), the [EC](#), [BBSRC](#), [EPSRC](#), [AHRC](#), [JISC](#), etc.
- used by [corporations](#), [SMEs](#), research labs and Universities worldwide
- the [Eclipse](#) of Natural Language Engineering, the [Lucene](#) of [Information Extraction](#), the [ISO 9001](#) of Text Mining
- a [world-class team](#) of language processing developers

If you need to solve a problem with text analysis or human language processing you're in the right place.

We have been analysing tweets on the EU Referendum — read about [our findings here](#).

If you are already using GATE then please [cite our work!](#)

Useful info:

- Applications
 - [YODIE](#) (Named Entity Disambiguation and Linking)
 - [BioYODIE](#) (YODIE for biomedical text)
 - [Generic Sentiment Analysis Tool](#)
 - [TwitIE](#) (Information Extraction for Tweets)
- [Books](#)
- [User guide](#)
- [Publications](#)
- [Training courses and materials](#)
- [GATE Cloud services](#)

News

User guide now [available in print](#).

Tweets by [@GateAcUk](#)

[GATE \(gate.ac.uk\)](#) Retweeted [@LeonDerczynski](#)

What's happening when our NER approaches fail to generalise? w/ [@kbontcheva](#) [@IAugenstein](#) [bit.ly/2iKpYpv](#) #nlproc #machinelearning

12 Jan

[GATE \(gate.ac.uk\)](#) [@GateAcUk](#)

Experience with GATE a big advantage!

11 Jan

[GATE \(gate.ac.uk\)](#) Retweeted [@tonyrussellrose](#)

[Embed](#) [View on Twitter](#)

More news: on [LinkedIn](#); on [FaceBook](#); on [Twitter](#); and on [our GATE blogs](#).