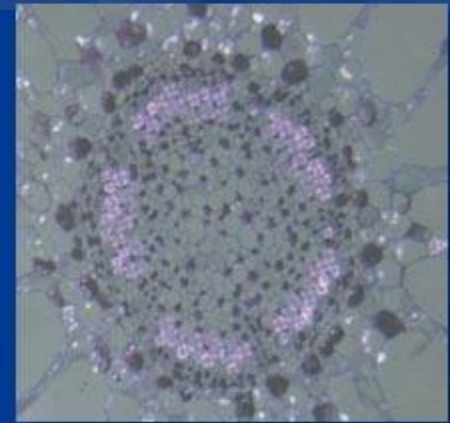


Standards in Human Language Technologies

Guergana Savova, PhD
Biomedical Informatics
Mayo Clinic

Where Discovery Begins



Overview

- **Background and motivation**
- **ISO/TC37/SC4**
- **Annotation example**
- **Architecture example**

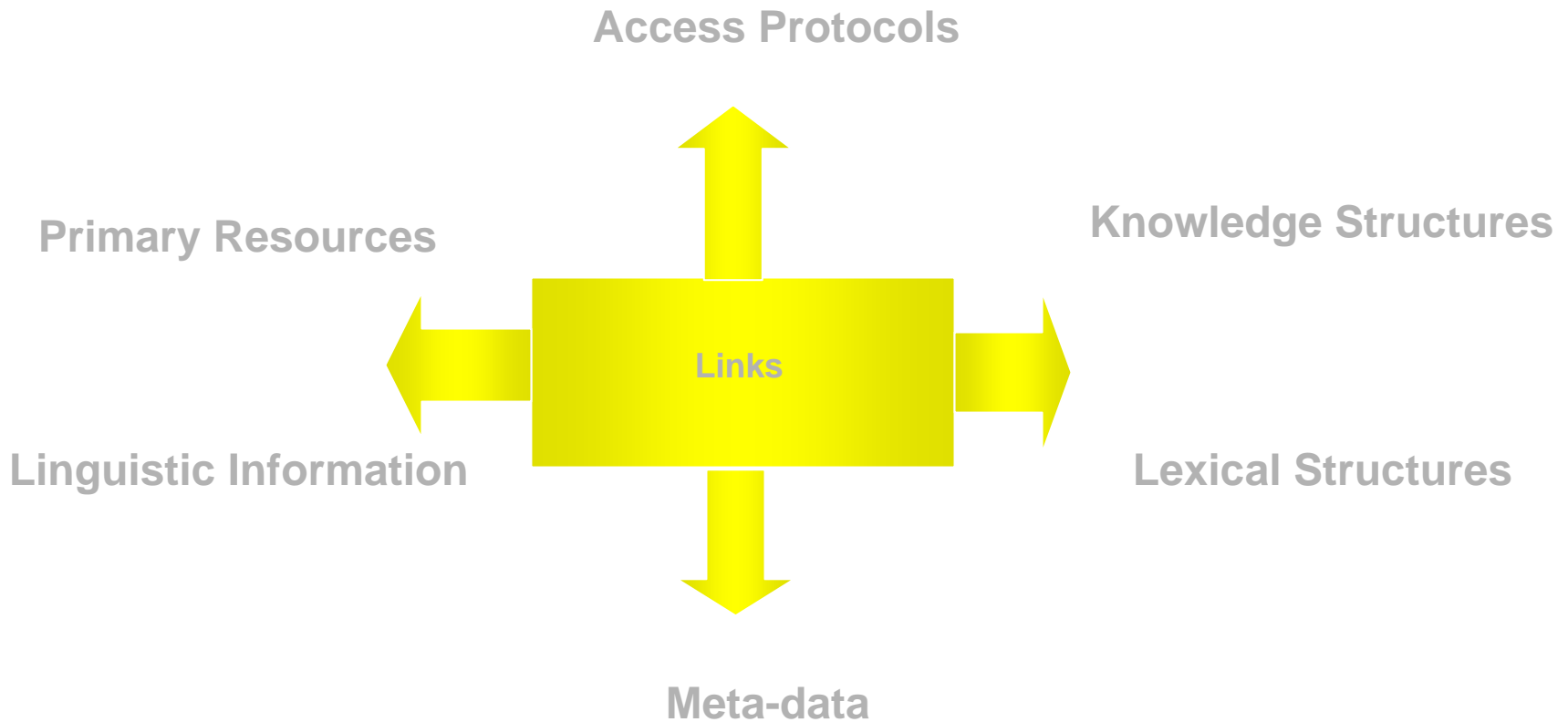
Some Background

- **Increasing need to define common practices and formats for linguistic resources to serve HLT development**
 - **Text Encoding initiative – www.tei-c.org**
 - **Corpus encoding standards**
 - **GATE**
 - **UIMA**
- **Advances in technology and a solid body of web-based standards**

Why Standards in HLT?

- **Need for generic language processing components for document indexing and classification, information extraction, summarization, topic detection**
- **Mono- and multi- languages**
- **Multi-modality of data (gestures, facial expressions, speech characteristics)**

Ecology of Language Resources (after Ide and Romary)



Dimensions of Language Resource Representation

- **Rendering formats and mechanisms**
 - **SGML, XML, Lisp-like structures, annotation graphs, database format**
- **Annotation content**
 - **Categories for annotation information about linguistic phenomena and the values associated with each category**
- **General architecture principles**
 - **Pipeline architectures**

ISO Standards

- **ISO/TC37/SC4 (<http://www.tc37sc4.org/>)**
 - **establish principles and methods for creating, coding, processing and managing language resources, such as written corpora, lexical corpora, speech corpora, dictionary compiling and classification schemes.**
 - **cover the information produced by natural language processing components in these various domains.**
 - **address the needs of industry and international trade as well as the global economy regarding multi-lingual information retrieval, cross-cultural technical communication and information management.**

Annotation Standards

- **Why are they important?**
 - **The de facto definition of the linguistic phenomenon that is being modeled**
 - **Annotated data is used to train classifiers. The output of the classifiers is the training labels.**



Example: Manual Annotations

Disorder mentions annotations

- **Constructing an evaluation set for our biomedical named entity recognition system.**
 - **160 annotated clinical notes**
 - **47,975 words**
 - **249 words per note (median)**
 - **1,556 annotations**
 - **658 unique SNOMED-CT codes used**
 - **82,813 disorder concepts in SNOMED-CT**



Definition of disorders

- **Subset of SNOMED-CT corresponding to disorders by leveraging the UMLS Semantic Network.**

TUI	type name
T019	Congenital abnormality
T020	Acquired abnormality
T037	Injury or Poisoning
T046	Pathologic Function
T047	Disease or Syndrome
T048	Mental or Behavioral Dysfunction
T049	Cell or Molecular Dysfunction
T050	Experimental Model of Disease
T190	Anatomical Abnormality
T191	Neoplastic Process

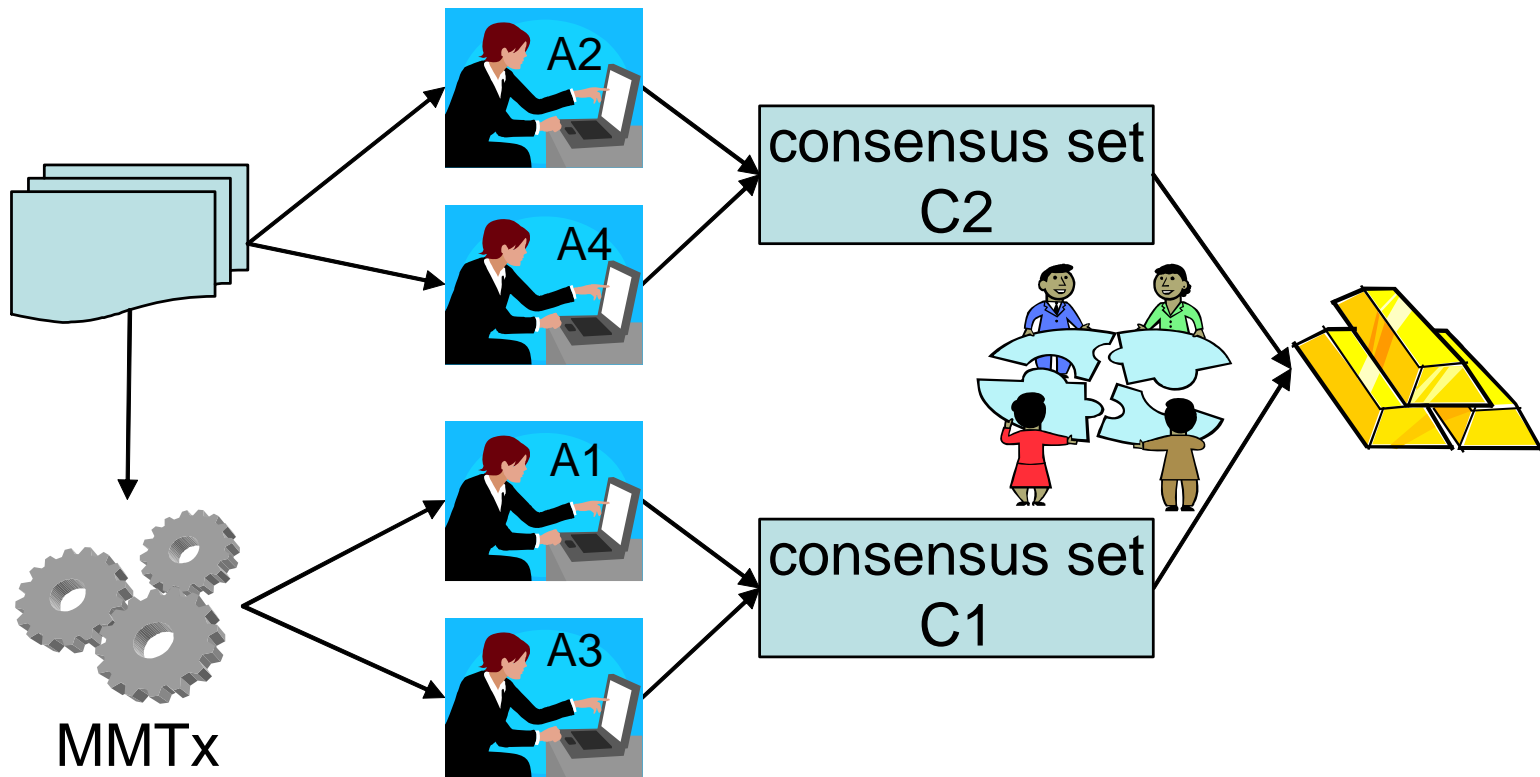
Annotation schema

- **Single named entity type called *disorder* with four properties, *span*, *concept code*, *context*, and *status*.**

span	character offsets corresponding to the mentioned disorder
concept code	SNOMED-CT code corresponding to disorders
context	<i>current, history of, family history of, or unrelated to patient</i>
status	<i>confirmed, possible, negated, or unrelated to patient</i>

Workflow

- Knowtator: <http://knowtator.sourceforge.net/>





Annotation guidelines

- 1) *A mentioned disorder should be assigned the most specific concept code named by the span of text.*
- 2) *Annotate all mentions of disorders in each note.*
- 3) *A disorder is defined as any concept that appears in the subset of SNOMED-CT that has been provided.*
- 4) *There should be only one annotation per mentioned disorder.*

IAA results

Compared attributes

compared annotation sets	spans exact	spans overlap	spans overlap +			
			concept	context	status	concept + context + status
A1, A2, A3, A4	75.7%	87.9%	72.7%	79.0%	80.9%	62.5%
C1, C2	81.4%	90.9%	81.7%	84.8%	86.0%	74.6%
C1, MMTx	42.3%	47.0%	42.3%	n/a	n/a	n/a
C2, MMTx	38.2%	44.1%	37.3%	n/a	n/a	n/a
GS, MMTx	39.8%	45.7%	39.5%	n/a	n/a	n/a

Annotation schemas: examples

- **Temporal and event expressions**
 - http://fofoca.mitre.org/annotation_guidelines/2005_timex2_standard_v1.1.pdf
 - <http://timeml.org/site/>
- **Coreference resolution**



Example: Architecture and workflow

Unstructured Information Management Architecture (UIMA)

- <http://www.alphaworks.ibm.com/tech/uima>

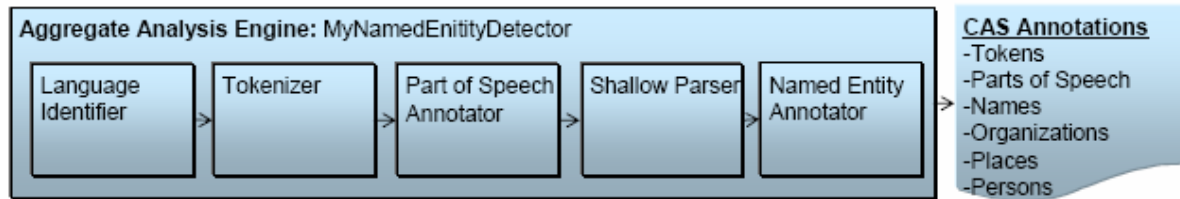


Figure 3: Sample Aggregate Analysis Engine

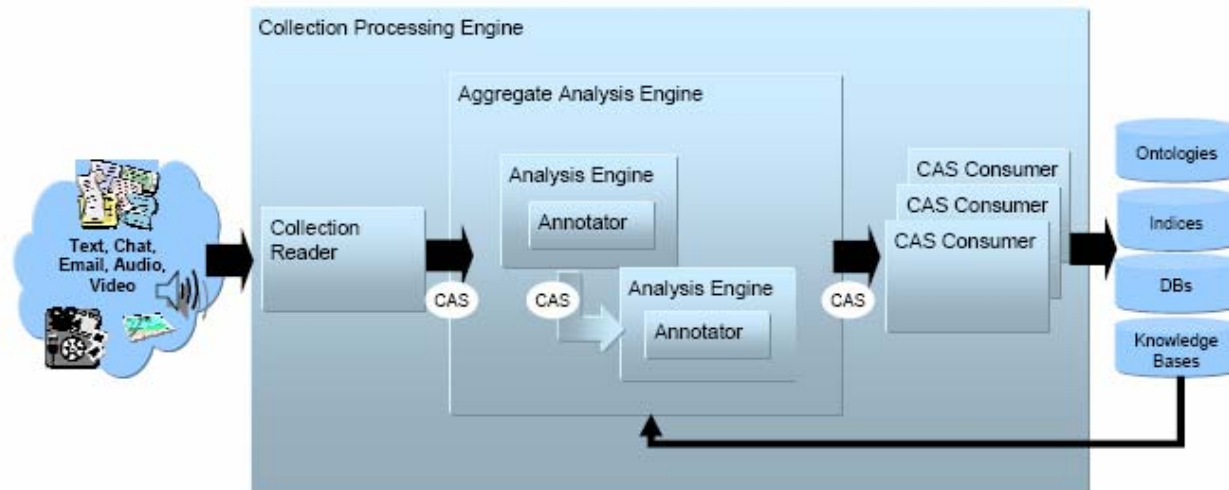


Figure 5: High-Level UIMA Component Architecture from Source to Sink

General Architecture for Text Engineering (GATE)

- <http://gate.ac.uk/>

Gate 3.0-alpha build 1667

File Options Tools Help

Gate

- Applications
 - ANNIE_003BC
- Language Resources
 - document
- Processing Resources
 - ANNIE OrthoMatcher_0
 - ANNIE NE Transducer_0
 - ANNIE POS Tagger_0C
 - ANNIE Sentence Splitte
 - ANNIE Gazetteer_003C
 - ANNIE English Tokenis
 - Document Reset PR_0
- Data stores

Messages document

Annotations

Text

```

') }

FT.com | TotalSearch | Global Archive | Print

document.write(getAdHTML('ban',468,60));

Return to Article | Print this Page

Airlines take over running of air traffic control

FT.com site, Jul 27, 2001
BY KEVIN DONE, AEROSPACE CORRESPONDENT

Seven UK airlines including British Airways, Virgin Atlantic, BMI
British Midland and EasyJet,
air traffic control system,
most controversial public-pr

Completion of the National A
critical time for the government as it tries to push through the
PPP for the London Underground.

The sale to a strategic investor of a 46 per cent stake in Mats is
the first time in Europe that management control of en route air
traffic services has passed into private hands.

It has been carried out despite a pledge by Labour before the 1997
general election that UK air was "not for sale."

Under the terms of the deal, which was approved by the European
competition authorities in May, the government has retained a 49
per cent stake and a golden share, while a 5 per cent stake is to
be allocated to Mats' 5,700 staff.

```

Annotation Sets

- Organization
- Original markups
 - a
 - b
 - body
 - br
 - head
 - html
 - img
 - link
 - p
 - script
 - table
 - td
 - title
 - tr

Co-reference Editor

Annotations

Document Editor Initialisation Parameters OLD Document Editor

document loaded in 0.094 seconds

More...

- **Interoperability of data**
 - **Interchangeable components**
- **Intraoperability of data**
 - **components can talk with each other**
- **Questions?**

References

- **Ide, N., Romary, L. (2007). Towards International Standards for Language Resources. In Dybkjaer, L., Hensen, H., Minker, W. (Eds.), *Evaluation of Text and Speech Systems*, Springer, 263-84.**