**Course: Advance Bio Informatics**

**Module Title: Existing Biomedical corpora**

**Module No: 129**

**MEDLINE & MeSH**

Standard corpus for measuring semantic similarity in BD.

**Biomedical Text Mining:**

Links to more than 20 different corpora and text collections, all focussed on biomedical domains & publications on biomedical corpora

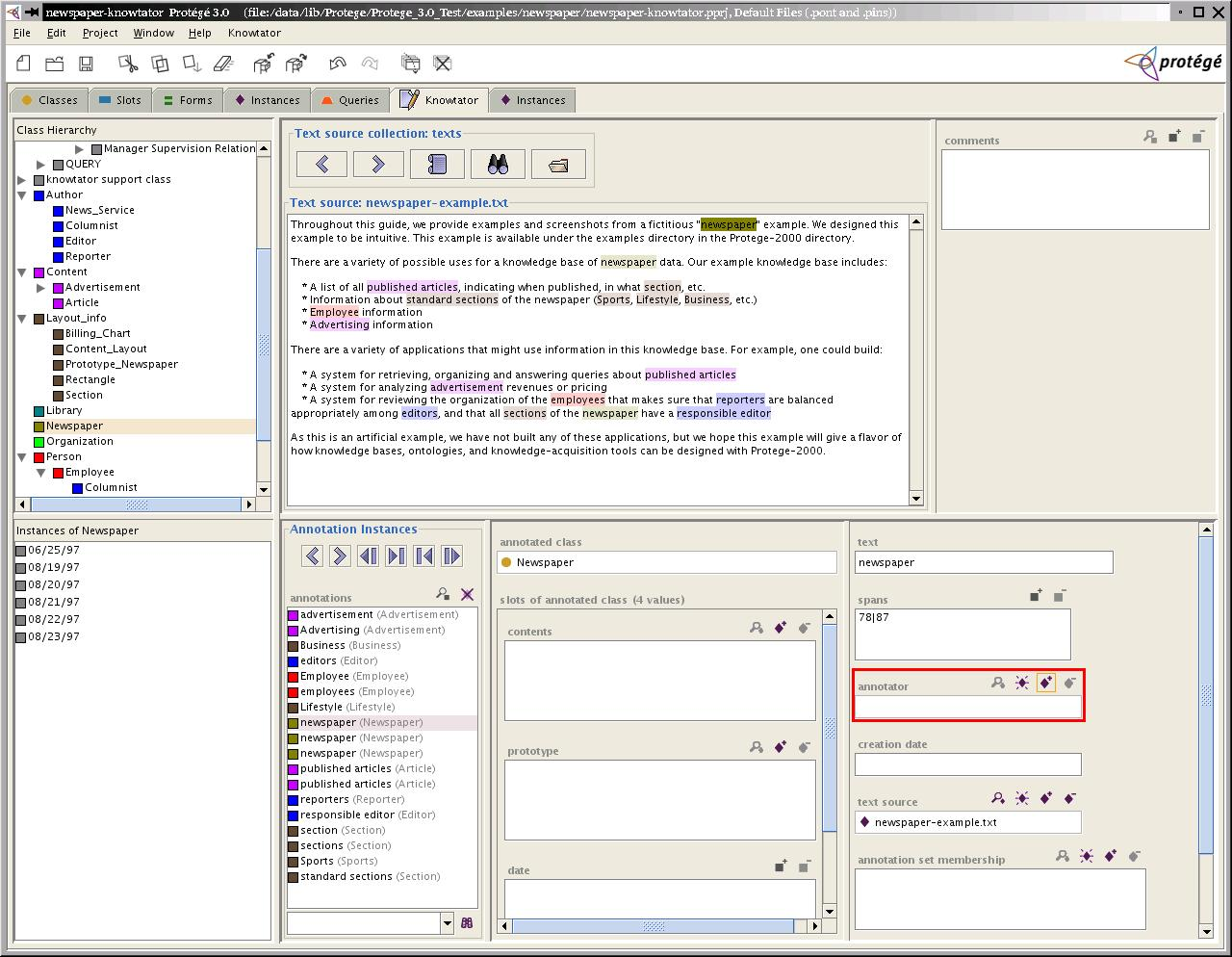
**Existing Corpora**

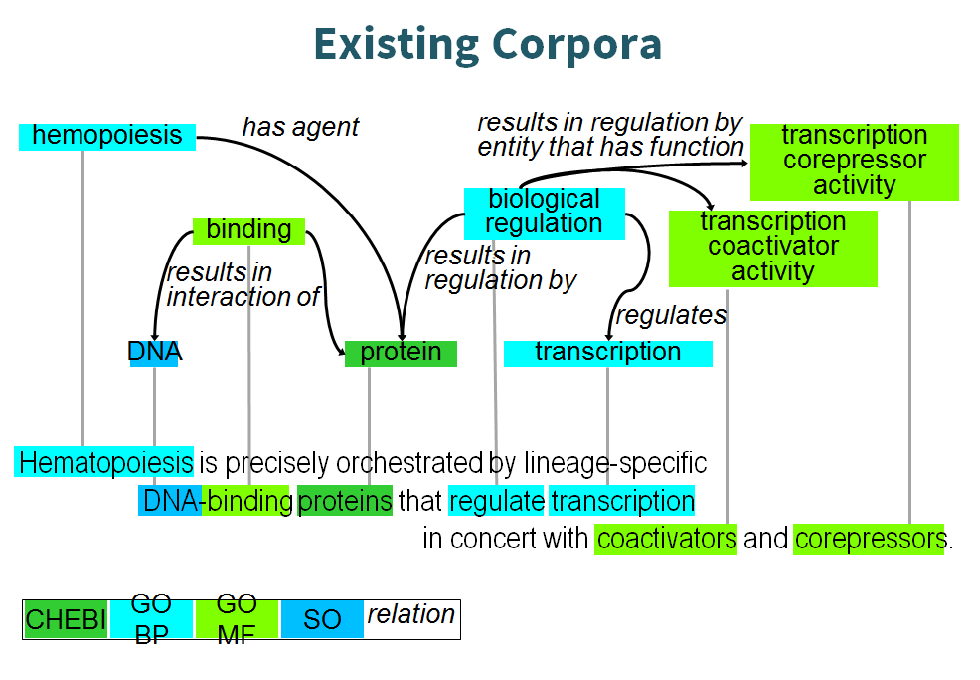
**CRAFT**

* Colorado Richly Annotated Full Text corpus
* 67 full text articles
* (+30 more reserved for future testing)
* >560,000 Tokens
* >21,000 Sentences
* ~100,000 concept annotations to  
  7 different biomedical ontologies / terminologies
* Penn Treebank markup for each sentence
* Multiple output formats available
* Integrated with UIMA

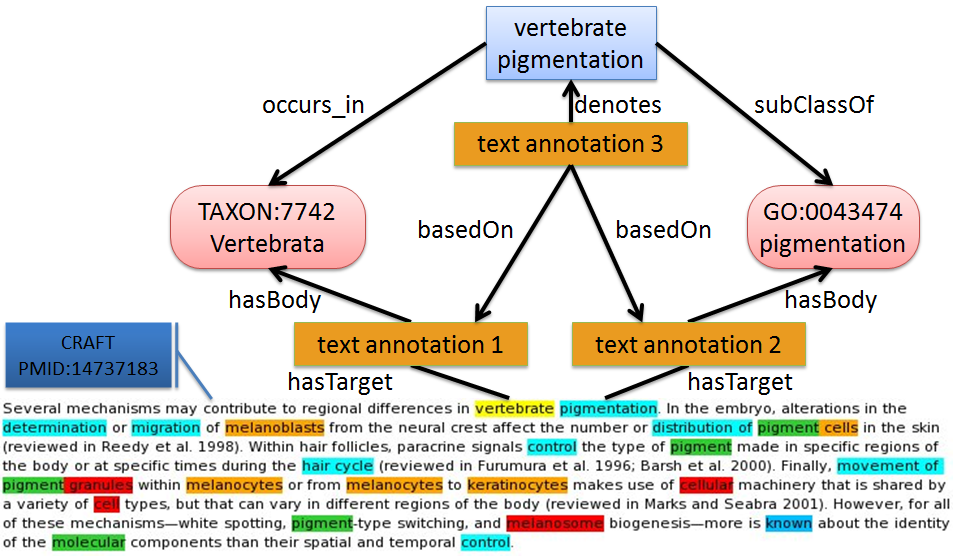
**Other Corpora:**

* Yapex
* GENETAG a corpus of 20K MEDLINE® sent. for gene/protein NER





**Information Extraction**



**GENIA**

**Domain:** Mesh term:

Human, Blood Cells, and Transcription Factors.

**Annotation:** POS, named entity, parse tree

**Penn BioIE**

**Domain:** the molecular genetics of oncology the inhibition of enzymes of the CYP450 class.

**Annotation:** POS, named entity, parse tree

**GENIA annotation**

**Linguistic annotation:** Reveals linguistic structures behind the text.

**Part-of-speech annotation:** Annotates for syntactic category of each word.

**Syntactic Tree annotation:** Annotates for syntactic structure of sentences.

**Semantic annotation:** Reveals knowledge pieces delivered by the text.

**Term annotation:** Annotates domain specific terms

**Event annotation:** Annotates events on biological entities