# corpora

## Table of contents

## 1. Annotated Corpora

- Genia (University of Tokyo)
  - 2000 abstracts from Medline
  - handed annotations for biological terms
  - articles with MeSH terms: human, blood cell and trascription factor
- Genia Treebank
  - A collection of parsed Medline abstracts (using an HPSG approach).
  - **NOT** Manually verified
  - on the web site they release 200, but actually we have a CD which contains 100000. (The CD was distributed at BioNLP04, COLING, Geneva)
- Craven's IE Data Sets There are three datasets, focusing on the relations described below. The labelling was done using a completely automated method.
  - subcellular-localization(PROTEIN, LOCATION) The relation tuples were gathered from the (now defunct) Yeast Proteome Database (YPD).
  - disease-association(GENE, DISEASE). The relation tuples were gathered from the Online Mendelian Inheritance in Man (OMIM) database.
  - protein-interaction(PROTEIN, PROTEIN). This data was collected from the MIPS Comprehensive Yeast Genome Database.
- PASTA Corpora (University of Sheffield)
  - Annotated corpus for baseline evaluation (gzipped): 52 abstracts http://www.dcs.shef.ac.uk/nlp/pasta/corpora/keys_ne.dev.gz
  - Annotated corpus for blind evaluation (gzipped): 61 abstracts. http://www.dcs.shef.ac.uk/nlp/pasta/corpora/keys_ne.bli.gz
- Medstract Corpus (Brandeis University): for two target applications: acronym identification, and entity anaphora resolution.
- corpus by Nigel Collier
- PASBio, N. Collier
- Integrated Annotation of Biomedical Text at Pennsylvania University
  - started in 2003
  - integrates different types of annotations: syntactic (Treebank), predicate-argument structure (Propbank), domain entitites and co-reference.
  - first results are expected in early 2004 ( **Check** )