AIMed

* 1000 Medline abstracts discussing human genes/proteins, manual curation
* XML file: 1955 sentences

<https://pubmed.ncbi.nlm.nih.gov/15811782/>

IEPA

* Medline? => Human and/or pathgen? Didn’t find much information – just a conference proceedings abstract
* XML file: 486 sentences

<https://pubmed.ncbi.nlm.nih.gov/11928487/>

HPRD50

* Corpus appears to be a union of LLL B. subtilis and Human genes, ids extracted programmatically using a vocabulary
* Looks like the relation extraction annotation is programmatic based on parser trees?
* XML file: 145 sentences

<https://academic.oup.com/bioinformatics/article/23/3/365/236564>

BioInfer

* Corpus of biomedical sentences deeply annotated for purposes including PPI protein and relationship retrieval, including compound entities (ie. relationships include interactions between other interacting proteins, etc)
* Annotation is comprehensive and is based on ontologies and “Link Grammar”  
  (complex: original 1100 sentences took 15 man months)
* I don’t see any restrictions (species or anything else) applied to the corpus source – maybe a more recent paper has something about this

*Addendum*: Looks like this is almost entirely human, human pathogens and other conditions

* XML file: 1100 sentences

<https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-8-50>

<http://mars.cs.utu.fi/smbm2008/files/smbm2008proceedings/smbmpaper_19.pdf>

LLL

* Annotations were restricted to B. subtilis protein interacts with gene. I suspect therefore that these are all regulatory interactions and therefore “enzyme” type. We don’t expect any structural type annotations here.
* Protein and gene ideas were restricted to a prepared vocabulary list – not NER
* XML file: 77 sentences

<http://genome.jouy.inra.fr/texte/LLLchallenge/>