**Course: Advance Bio Informatics**

**Module Title: Corpus Design**

**Module No: 125**

**Corpus Design**

* Written text / speech
* Purposefully collected
* Structured
* Annotated
* Accessible via computer
* Corpus ~ text archive
* format: XML
* code several layers of information
* publicity: write specific papers about corpus, publicise its availability
* accuracy & consistency of the annotations
* usability of the data
* General purpose vs specialized
* Dynamic (monitor) vs static
* Representativeness and balance
* Size
* Storage and access
* Permission
* Text capture & markup
* Organizations