Features of Dataset Applications for Books

Datasets Compiled From Book Examples

The remaining screenshots demonstrate how data sets can be used even outside the context of generating experimental data. The pictured data set represents a corpus of linguistic examples mined from Wiley's *Blackwell Handbook of Pragmatics*. Creating data sets from book-length publications can encompass several steps:

Text Mining In the case of linguistics, this involves locating example sentences within linguistics texts and storing them as an independent corpus.

Canonical Formatting If possible, linguistics texts should be formatted with markup allowing examples to be extracted automatically. This has the added benefit of ensuring that the dataset software can link between individual samples and their location in the book text.

Annotation Linguistic corpora are often annotated to identify structural details, beyond raw text, in each sample.