The datasets used in teaching and testing the proposed system are both taken from outside sources (RTE datasets), and produced by hand.

The outside source datasets are taken from Stanford University, from their own paper on detecting contradictions in text (Marie-Catherine de Marneffe, Anna N. Rafferty and Christopher D. Manning. 2008. [Finding contradictions in text](http://www-nlp.stanford.edu/pubs/contradiction-acl08.pdf). *ACL-08*.), from PHEME, the dataset being available at <https://www.pheme.eu/2016/04/12/pheme-rte-dataset/> and from the SICK dataset, cited M. Marelli, S. Menini, M. Baroni, L. Bentivogli, R. Bernardi and R. Zamparelli (2014). A SICK cure for the evaluation of compositional distributional semantic models. *Proceedings of LREC 2014*, Reykjavik (Iceland): ELRA, 216-223.

For the self-produced dataset, several sentences were randomly generated via a script, then duplicated and modified by hand to create contradiction and neutral pairs.

In the case of the RTE datasets, they were formatted to fit the proposed system’s format, being that of:

ID Sentence A Sentence B Class

(The data is tab separated)

ID- The numerical ID assigned to the sentence ID.

Sentence A, Sentence B- The two sentences to be compared.

Class- In the case of training data, whether the sentence pair represent a Neutral or Contradiction result.