程式處理步驟簡化如下:

1. 將原語料庫的行列排版重新整理 --> 去除多餘的空行、把同一筆長句合併起來。

2. 抓出每筆長句中，含有「誰」的短句，可能有一個以上。

3. 去除沒有意義的標點符號，如上句首上引號等。

4. 比較並移除重複之短句。