# Text Processing

## Line Breaks

### Issue

Cuts both ways. In some instances, line breaks provide a good sentence tokenizing tag to correctly show the end of a hypothesis. In other instances, the line break comes in the middle of the hypothesis. This comes directly from the PDF to text conversion, which places new blank lines in between lines of the hypothesis.

### Status

**Drop Line Breaks**

Currently, I do not see a way to detect when line breaks should be there, and when they are an error. Therefore, treating line breaks as tokenizing tokens vs dropping them completely will always be a tradeoff.

The general sense is that using line breaks to tokenize causes more problems them solves. So, line break vectors are dropped.

### Example Documents

* ctsm08jem.pdf
  + Line breaks are an error introduces in PDF to text conversion.
* mao10jbe
  + Line breaks are introduced between every line.
* ekks05smj.pdf
  + Line break correctly identifies end of hypothesis. The hypothesis statement does not end in a period, so during conversion, the first sentence of the next paragraph is concatenated to the end of the hypothesis.

## Hypothesis Sub-setting

### Issue

Some hypotheses have sub-parts within them, usually labeled (a), (b), (c), etc. The problem is this sub-setting method is used differently in different papers.

One method is to list items where each item could be its own hypothesis. Meaning the following hypothesis (kyr11ijsom.pdf, hypothesis 3)

*Hypothesis 3: Partner’s trust is positively related to the sharing of (a) the risk information and (b) the product information.*

Could easily be the following two hypotheses:

*Hypothesis 3a: Partner’s trust is positively related to the sharing of the risk information.*

&

*Hypothesis 3b: Partner’s trust is positively related to the product information.*

The other method simply breaks down data into chunks using letter heading breaks, but all of the information is supposed to be used together. In this case, we are not able to break out each letter heading into its own hypothesis (ck99smj.pdf, hypothesis 5).

*Hypothesis 5: The interaction of exchange conditions and interfirm cooperation will be significantly related to firm performance. Specifically, (a) the combination of unfavorable exchange conditions (i.e., low outlet-level asset specificity and geographic dispersion, and high specific knowledge) and low interfirm cooperation, and (b) the combination of favorable exchange conditions (i.e., high outlet-level asset specificity and geographic dispersion, and low specific knowledge) and high interfirm cooperation will be positively related to performance.*

### Status

**No Action**

There is currently no way to correctly identify which method of using letters to subsection a hypothesis is being used.

### Example Documents

* kyr11ijsom.pdf
  + **Can** be broken into multiple hypotheses.
* ck99smj.pdf
  + **Cannot** be broken into multiple hypotheses.

# Hypothesis Selection

## fastText Model

### Issue

The fastText model drops too many true hypotheses.

### Status

The fastText model is not used.

### Example Document

* bnk00jbr.pdf
  + Three true hypotheses, but due to hypothesis references later in the document, and unlucky table-to-text conversion, numerous other strings are identified as hypotheses at first pass.
  + The fastText model predicts Hypothesis 1 **IS NOT** a hypothesis, (0% prediction on True)
  + The fastText model predicts a later Hypothesis 1 reference **IS** a hypothesis.
  + After the true hypothesis is dropped, the reference would be picked up as accurate.