wrongly combined or out of sequence fragments of text. The essential problem is to look at all the whole page, and figure out in which order a person would read all the contained blocks. For that it is necessary to consider all the text blocks and do a topological sort of them. Figure 1.2 shows recovered reading order for a page with complex layout. All text on a page is sorted; content like *i.e.* page numbers, which is not part of the body text, is separated at a later stage.

Figure 1.2: Determining reading order. Expected reading order among the green boxes marked with arrows. Text which might not be qualified as body text in red

1.3.3 Logical layout analysis

While geometric layout analysis leaves us with a complete physical representation of a page in terms of blocks of segmented content, the next step is to somehow use that to derive a logical structure.

The essential idea is to both assign *labels* to, and figure out the logical relationship between these blocks based on an *a priori* model of a general document. These labels are meant to correspond to concepts that humans perceive as meaningful with respect to the content at hand, typical examples would be `title`, `body text`, `table`, *etc.* The relationships will be for example that a section header precedes and introduces the body text of paragraph, or that a section header belongs beneath the main title. Based on this