Parsing citations in academic papers

The aim of this semester project is to design and evaluate a citations parsing pipeline for scientific papers. It will be done in collaboration with Frontiers, which will provide real-world data.
Frontiers is a leading open-access publisher of academic journals. They process large amounts of unstructured information (in textual form) to build knowledge databases.
This semester project includes three distinct parts:

1. When an author has just submitted a manuscript, we would like to parse its citations (the "bibliography" part of an article). Whereas this is relatively simple for manuscript submitted in Latex format, this is less straightforward for manuscripts written in Word. This part might include the development of a statistical or machine learning model to identify the bibliographical region of an article.
2. The second part deals with parsing and normalization of the cited article. There are numerous rules to build citations, and this should be approached in a statistical way (e.g. finding linguistic patterns in citations). For example, in the citation "*Yan, E. (2014). Research dynamics: Measuring the continuity and popularity of research topics. Journal of Informetrics, 8(1), 98-110*", one has to understand the different parts; e.g. that *2014* is the publishing date, and that "*Journal of Infometrics*" is not part of the title, but is the name of the journal.
3. The last part deals with the matching of the citation to some database of published articles. This could rely on a simple vector space model algorithm (e.g. tf-idf), with proper text preprocessing and postprocessing.

All three parts will make use of existing software or algorithms whenever possible.