ground information' from 'statements of the particular aim of the current paper', to take an example from Teufel's work (Teufel and Moens, 2002). While text zoning has been mainly applied to scientific texts so far, one can also find this technique applied to other domains where it is relevant, for example email messages (Lampert et al., 2009), or job ads (Gnehm and Clematide, 2020).

The number of zones considered varies but is generally around ten or less (7 for example in (Teufel, 1999) and (Guo et al., 2011), or 8 in (Gnehm, 2018)). Zone annotation is generally performed by a group of experts at the sentence level (more rarely at the paragraph level). Inter-annotator agreement on the task is generally high: (Guo et al., 2011) for example reports a score of 0.85 for Cohen's κ (Cohen, 1960).

Once a representative corpus is available, it is possible to train a classifier for the task. Features considered are generally low level (unigrams, bigrams, sometimes specific terms also receive a specific weight) (Teufel and Moens, 2002) but higher level features are also sometimes considered (like syntactic relations in (Guo et al., 2011)). Contextual information (like the previous zone) is also often taken into consideration, since a specific zone tends to appear in typical positions in scientific abstracts. As for training, most recent ML techniques have been explored, from Naive Bayes (Teufel, 1999) to LSTM (Gnehm, 2018), through CRF and SVM (Guo et al., 2011). In this last paper, the authors also investigate semi-supervised learning and active learning, in order to reduce the amount of data needed for training, which often constitutes a bottleneck for the task. More recently, large language models like Bert have also been explored (Gnehm and Clematide, 2020), but they require large amount of data for training.

Here our goal is partly the same as the one in these previous studies. However, our corpus is very different since we analyze theater reviews, which may not be as regular as scientific papers. In our context, zones are important to determine whether the critic is addressing acting, staging or the general setting of the play (we use the rather neutral term 'text zoning' instead of 'argument zoning', since the zones we consider do not always correspond to arguments). Analyzing the overall organization of theater reviews will also make it possible to determine whether these have a rather fixed structure or not, if reviews in newspapers differ a lot from those directly written for blogs on the Web.

3 Corpus Creation

To answer these questions, the first step consisted in collecting the necessary data to create two subcorpora. The first subcorpus is made of journalistic reviews only, while the second one is based on digital theater reviews written by bloggers.

3.1 Subcorpus 1: Journalistic Theater Reviews

The first subcorpus was created thanks to the online database *Theatre Record*. *Theatre Record* is a biweekly paper magazine which reprints in full all the national drama reviews of the productions in London and its regions. Its archives were digitized in 2019 and each newspaper published since 1981 is now available online (in PDF format).

All the newspapers issues have the same characteristics. For each of the shows, a certain number of reviews is given as well as a series of details on the production, such as the cast, the credits and the photographs. The theater in which the play was performed as well as the opening and the closing dates of the show are also indicated. Most of the newspapers represented in this database are well-known among the general public: *The Times, The Guardian, The Independent*, etc. Out of the 84 newspapers available on *Theatre Record*, we have selected 32 of them in total. A number of sources had to be removed. Since this corpus focuses on printed newspapers, online news websites had to be excluded. We also removed newspapers whose reviews were not about London performances and all the newspapers which had a too limited number of reviews.

3.2 Subcorpus 2: Digital Theater Reviews

The second subcorpus is based on 18 English blog platforms whose authors' publications deal with London plays only. The content of these websites was extracted using webscraping techniques. These 18 blog platforms have the following characteristics: they have no printed equivalent, their content is entirely free and their authors are not paid for their activity. They are either run by one person, or by multiple authors.

The selection of these blogs was made according to the top 10 most popular British theater blogs established by Vuelio in 2020. A majority of them also came from the platform *MyTheatreMates*. All