This paper presents work on the application of corpus linguistic techniques to the pragmatic analysis of large text collections. Our particular focus is on the automatic identification of apologies in blog data, and we begin by describing how we built the Birmingham Blog Corpus: a collection of texts downloaded from popular blog hosting sites (Kehoe & Gee 2012; http://www.webcorp.org.uk/blogs). Our analyses in this paper are based on an 80 million word sub-section of the corpus, which includes both blog posts and reader comments on these posts.

A common approach in pragmatic analysis is to search for Illocutionary Force Indicating Devices (IFIDs), or specific words which may be indicative of a particular speech act (Searle 1979). In our case the speech act is apology and the initial list of IFIDs, taken from a thesaurus and from previous studies, includes words such as sorry, pardon and excuse. As we go on to illustrate, however, there are limitations to this approach. There may be instances in the corpus where apparent apology IFIDs are not used to apologise (e.g. ‘a sorry state’) and, at the other end of the spectrum, apologies which do not contain an apology IFID (so-called ‘hidden manifestations’ of apologies – Kohnen 2007). With these limitations in mind, we turn to the linguistic principle of collocation – the frequent co-occurrence of particular word pairs – to help refine and expand our initial list of apology IFIDs.

Previous attempts at the pragmatic analysis and annotation of corpora have focused primarily on relatively small collections of spoken data, often relating to specific discourse situations such as courtroom dialogues (Archer 2005) or customer service telephone calls (Leech and Weisser 2003). Our work is novel in being a large-scale analysis of a written text type, but one which has been shown to exhibit some of the features traditionally associated with spoken discourse. The fact that the Birmingham Blog Corpus facilitates the study of reader responses to texts (and responses to responses) in the form of comments means that it offers new possibilities for pragmatic analysis.

References