The identification of collocations, defined in this project as recurrent word sequences displaying varying degrees of semantic idiosyncrasy and syntactic fixedness, plays a crucial role in various fields, both descriptive/theoretical and applied (from discourse and genre analysis to specialised lexicography). However, current approaches to the definition and identification of collocations do not always prove satisfactory. On the one hand, statistical approaches that rely on frequency of co-occurrence of two or more words tend to produce inaccurate results, i.e. mixing intuitively relevant collocations and less interesting word combinations (Evert and Krenn 2001). On the other hand, approaches employing very fine-grained semantic criteria are difficult to implement computationally (Sag et al. 2002). My research project aims at combining such approaches, applying them to the task of identifying phraseology in a specific variety of English, namely the English used in institutional University Websites.

In particular, the goal of my research is to devise methods that integrate frequency information with semantically informed criteria for the task of automatically classifying English word sequences. Focusing on either member of a sequence, it might be possible to group the words co-occurring with it, i.e. the potential collocates, into a number of semantically homogenous classes (e.g., of synonyms, hyponyms, etc.). The number of different words in each semantic class could then be taken as a clue to the greater or lesser degree of “collocativeness” of the word under study, and ultimately of the pair. Indeed, it is often hypothesized that a peculiar property of collocations is non-substitutability (the replacement of words in a sequence with other semantically related words is limited; Manning and Schütze 1999), and hence the more a word tends to co-occur with semantically related words, the less collocationally restricted it is likely to be. On the contrary, semantically isolated co-occurring words might be good candidate collocates.

The method will be applied to a a specialised corpus of English, which will have to be set up specifically for this project. While the method to be developed should be largely independent of the sub-language analysed, the research goal is to also provide a descriptive account of the phraseological regularities found in a specific subset of the English language. An area of study has been identified as being particularly promising, i.e. the variety of English produced in institutional University settings in countries in which English is an official language, and published on the Web. There are two main reasons for this choice: first, identifying phraseological regularities in this variety of English would be a first step towards the description of the language produced in comparable settings by "English as Lingua Franca" writers and translators (e.g., those responsible for publishing Web content for Italian Universities; on ELF see e.g. Jenkins 2006 and Seidlhofer 2005). Secondly, from a more applied perspective, a descriptive account of the collocational properties of "native" institutional English in University Websites would be of extreme value in producing writing aids for writers and translators working in this field. Given the increasing pressure on Italian Universities to attract foreign students, one would expect the production of Web content in English to be soon becoming a priority, making the provision of the descriptive account proposed in this study particularly timely.

References