The central topic of this thesis is the automatic extraction of definitions from text. Definition extraction can play a role in various applications including the semi-automatic development of glossaries in an eLearning context, which constitutes the main focus of this dissertation. A glossary provides definitions for the most important terms that are discussed in a text.

The semi-automatic extraction approach presented in this study consists of two phases. As a first step, a method entirely based on lexico-syntactic patterns has been used to distinguish between definitions and non-definitions. A corpus consisting of 600 definitions has been employed to identify recurrent definition patterns. Since many of these patterns are not unique to definitions, a second step was employed to reduce the number of non-definitions identified. It has been investigated whether other textual characteristics can contribute to the correct classification of definitions, in addition to the lexico-syntactic patterns. The properties that have been examined vary from the importance of the defined word (phrase) within a text to the layout of the definition. Machine learning techniques have been employed to identify which are the most relevant (combinations of) definition properties.

The results of this dissertation are relevant for researchers in linguistics and lexicography as well as for the development of language technology applications.