English:

This thesis deals with the detection of three rhetorical figures based on repetition of words: chiasmus (“**Fair** is **foul**, and **foul** is **fair**.”), epanaphora (“**Poor old European** Commission! **Poor old European** Council.”) and epiphora (“This house is **mine**. This car is **mine**. You are **mine**.”). For a computer, locating all repetitions of words is trivial, but locating just those repetitions that achieve a rhetorical effect is not. How can we make this distinction automatically?

First, we propose a new definition of the problem. We observe that rhetorical figures are a graded phenomenon, with universally accepted prototypical cases, equally clear non-cases, and a broad range of borderline cases in between. This makes it natural to view the problem as a ranking task rather than a binary detection task. We therefore design a model for ranking candidate repetitions in terms of decreasing likelihood of having a rhetorical effect, which allows potential users to decide for themselves where to draw the line with respect to borderline cases.

Second, we address the problem of collecting annotated data to train the ranking model. Thanks to a selective method of annotation, we can reduce by three orders of magnitude the annotation work for chiasmus, and by one order of magnitude the work for epanaphora and epiphora. In this way, we prove that it is feasible to develop a system for detecting the three figures without an unsurmountable amount of human work.

Finally, we propose an evaluation scheme and apply it to our models. The evaluation reveals that, even with a very incompletely annotated corpus, a system for repetitive figure detection can be trained to achieve reasonable accuracy. We investigate the impact of different linguistic features, including length, n-grams, part-of-speech tags, and syntactic roles, and find that different features are useful for different figures. We also apply the system to four different types of text: political discourse, fiction, titles of articles and novels, and quotations. Here the evaluation shows that the system is robust to shifts in genre and that the frequencies of the three rhetorical figures vary with genre.

Swedish:

Denna avhandling behandlar tre retoriska figurer som bygger på upprepning av ord: kiasm (“Om inte **Muhammed** kan komma till **berget** får **berget** komma till **Muhammed**.”), anafor (“**Det är inte** rimligt. **Det är inte** hållbart. **Det är inte** rättvist.”), och epifor (“Den här stugan **är min**. Den här bilen **är min**. Du **är min**.”). En dator kan lätt identifiera upprepningar av ord i en text, men att urskilja enbart de upprepningar som har en retorisk effekt är svårare. Hur kan vi få datorer att göra detta?

För det första föreslår vi en ny definition av problemet. Vi noterar att retoriska figurer är ett graderbart fenomen, med prototypiska fall å ena sidan och klara icke-fall å andra sidan; däremellan finns ett brett spektrum av gränsfall. Detta gör det naturligt att se problemet som en uppgift som gäller rangordning snarare än binär klassificering. Vi skapar därför en modell för att rangordna repetitioner efter sannolikheten att de har en retorisk effekt. Därigenom tillåts systemets användare att själva avgöra hur gränsfall ska hanteras.

För det andra försöker vi undvika tänkbara svårigheter med att samla in annoterade data för att träna modellen för rangordning. Genom att använda en selektiv metod kan vi reducera mängden annoteringsarbete tusenfalt för kiasm och tiofalt för anafor och epifor. Det är alltså möjligt att utveckla ett system för att identifiera de aktuella retoriska figurerna utan en stor mängd manuell annotering.

Slutligen föreslår vi en metod för utvärdering och tillämpar den på våra modeller. Utvärderingen visar att vi även med en korpus där få exempel är annoterade kan träna ett system för identifiering av repetitiva figurer med godtagbart resultat. Vi undersöker effekten av olika särdrag som bygger på t.ex. längd, n-gram, ordklasser och syntaktiska roller. En slutsats är att olika särdrag är användbara i olika grad för olika figurer. Vi prövar också systemet på fyra olika texttyper: politisk diskurs, skönlitteratur, titlar på artiklar och romaner, samt citat. Utvärderingen visar att systemet är robust vad gäller genreskillnader. Vi ser även att figurernas frekvens varierar över olika genrer.