Introduction

Human readers have the extraordinary capability to infer the meaning of a word from text, even if they have never heard of the word before, and there are no visual cues present to support its interpretation. Take for example sentences (1) – (3):

(1) **Fomalhaut** staat op 24 lichtjaren van de zon, en is 15 keer zo helder.
    Fomalhaut stands on 24 light years of the sun and is 15 times as bright.

(2) Zelfs een rijpe **kumquat** smaakt nog tamelijk zuur.
    Even a ripe kumquat tastes still rather sour
    *Even a ripe kumquat still tastes rather sour.*

(3) Op een statig deuntje danste men de **pavane**.
    On a stately tune danced one the pavane
    *On a stately tune, they danced the pavane.*

Even if we have never heard of words like **Fomalhaut, kumquat** and **pavane** before, we can still make reasonable assumptions about their meaning. The surrounding words allow us to assume that **Fomalhaut** is probably a star, that **kumquat** is some kind of edible organic object – most likely a fruit – and that a **pavane** is some kind of dance. We are able to infer the meaning of words because the context gives us cues about their semantic content. Likewise, we might be able to use the context of unknown words to infer their meaning automatically.
Of course, we can only get at the meaning of *Fomalhaut* because we are already familiar with the other words in the sentence, such as *lichtjaar* ‘light year’ and *zon* ‘sun’. Likewise, we know what kind of objects might be ripe and taste sour, and we have a general idea of what it is that can be danced. The situation becomes more complicated if we want to use a computer to automatically infer the meaning of words: unlike humans, a computer does not have any a priori knowledge of words whatsoever. For this reason, most work on the acquisition of semantics from text has focused on semantic similarity. Determining how similar a word is to other known words is much easier than determining what its actual meaning is. It is hard to extract the meaning of *Fomalhaut* from scratch, but it is much easier to determine that *Fomalhaut* appears in the same contexts as words like *Sirius* and *Betelgeuse*. Likewise, the word *kumquat* will appear in similar contexts as words like *orange*, *lemon* and *apple*, and *pavane* will be found in locations similar to words like *bourrée*, *gigue* or *polka*.

The semantic similarity of a word can thus be determined by accumulating its different contexts in a large corpus, and comparing those contexts to the contexts of other words. If two words have similar contexts, they are likely to be semantically related. Likewise, if two words share only few or no contexts at all, they are probably semantically unrelated. This process of determining a word’s semantics by looking at the way it is distributed in texts is called *distributional similarity*, and it will be the main foundation of this dissertation.

The extraction of semantics from text is a very broad and extensive subject. It therefore makes sense to clearly demarcate what will be our object of research, and also mention what this dissertation will not be about. The main subject of this dissertation is the *lexico-semantic* extraction of *nouns* from large-scale written corpora. The three italicized words are important here. By *lexico-semantic* extraction, we mean that we restrict ourselves to the extraction of semantic information on the lexical level: we are interested in extracting the semantics of single words, as they might be described in a dictionary. Of course, the lexical level is only a part of the vast domain that is semantics. Humans combine words into sentences to form complex meanings. Moreover, the combination of words has an influence on the meaning of the combined parts. These phenomena belong to the domain of *compositional semantics*, and they are beyond the scope of this dissertation.

Secondly, this dissertation mainly investigates the extraction of *nouns*. It goes without saying that there are many more word classes – adjectives, verbs and even function words – that deserve our attention with regard to the determination of their semantics. Verbs, in particular, make up an interesting word class with complex semantic behaviour. This dissertation, however, is limited to the extraction
of nouns, and the lexico-semantic extraction of other word classes is again largely beyond its scope.¹

Thirdly, we will investigate the lexical semantics of nouns by looking at their distribution in large written text corpora of newspaper texts. It is very well possible that the semantics present in these corpora is very different from the semantics to be found in spoken corpora, or in sources that go beyond the realm of linguistics. The semantics that we investigate in this dissertation, however, will be the semantics of words as they appear in written newspaper texts.

With these reservations set aside, we can introduce the three main research questions that will be investigated in this dissertation. First of all, we want to determine what kind of semantic similarity is captured by different context models. We will present three different groups of models (each based on a different notion of context), and we will quantitatively investigate what kind of semantic similarity is captured by them. At the same time, we will investigate the usefulness of the models (and the resulting lexico-semantic resources) in a number of applications, notably multi-word expression extraction and word sense discrimination.

Secondly, we want to investigate the applicability of dimensionality reduction methods for semantic similarity extraction. Dimensionality reduction – or factorization – is the collective name for mathematical techniques that try to reduce an abundant number of overlapping features to a limited number of independent, informative dimensions. We will discuss a number of dimensionality reduction methods, and quantitatively evaluate whether they are beneficial for semantic similarity extraction. Again, their usefulness is also investigated in two applications: word sense discrimination and selectional preference induction.

Thirdly, we will investigate the use of three-way methods for lexico-semantic information extraction. Up till now, most research on the extraction of lexical semantics from text has focused on two-way methods, in which two-way co-occurrences (e.g. terms × documents) are used. Co-occurrences need not be limited to two ways, though; it is easy to think of entities that occur in three (or more) ways (e.g. verbs × subjects × direct objects). We will present the mathematical machinery to deal with multi-way co-occurrences, and test the usefulness of three-way methods in an application, namely the induction of selectional preferences.

The outline of this dissertation is as follows. The first part provides a theoretic framework – grounding the distributional similarity theorem and setting up a formal framework for a computational implementation. This includes a thorough overview of two dimensionality reduction algorithms – singular value decompos-

¹Note, however, that we will touch upon the extraction of verb semantics in the last chapter, where we discuss the extraction of selectional preferences.
tion and non-negative matrix factorization – and a discussion of three-way methods. The second part investigates the three different groups of models and their various model parameters, and provides a quantitative evaluation of their ability to extract semantic similarity. The final part of this dissertation provides a number of applications, in which the different models and techniques presented in the first part are applied.