The Word-Space Model

Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces

Magnus Sahlgren

A Dissertation submitted to Stockholm University
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy

2006

Stockholm University
Department of Linguistics
Computational Linguistics
Stockholm, Sweden

National Graduate School
of Language Technology
Gothenburg University
Gothenburg, Sweden

Swedish Institute
of Computer Science
Userware Laboratory
Kista, Sweden

ISBN 91-7155-281-2

ISSN 1101-1335
ISRN SICS-D–44–SE
SICS Dissertation Series 44

Doctoral Dissertation
Department of Linguistics
Stockholm University

ISBN Nr 91-7155-281-2
This thesis was typeset by the author using LaTeX
Printed by Universitetsservice US-AB, Sweden, 2006.

Bart: *Look at me, I'm a grad student! I'm thirty years old and I made $600 last year!*

Marge: *Bart, don't make fun of grad students! They just made a terrible life choice.*

(The Simpsons, Episode "Home Away from Homer")

Abstract

The word-space model is a computational model of word meaning that utilizes the distributional patterns of words collected over large text data to represent semantic similarity between words in terms of spatial proximity. The model has been used for over a decade, and has demonstrated its mettle in numerous experiments and applications. It is now on the verge of moving from research environments to practical deployment in commercial systems. Although extensively used and intensively investigated, our theoretical understanding of the word-space model remains unclear. The question this dissertation attempts to answer is *what kind of semantic information does the word-space model acquire and represent*?

The answer is derived through an identification and discussion of the three main theoretical cornerstones of the word-space model: the geometric metaphor of meaning, the distributional methodology, and the structuralist meaning theory. It is argued that the word-space model acquires and represents two different types of relations between words — syntagmatic or paradigmatic relations — depending on how the distributional patterns of words are used to accumulate word spaces. The difference between syntagmatic and paradigmatic word spaces is empirically demonstrated in a number of experiments, including comparisons with thesaurus entries, association norms, a synonym test, a list of antonym pairs, and a record of part-of-speech assignments.

Sammanfattning

Ordrumsmodellen använder ords distributionsmönster över stora textmängder för att representera betydelselikhet som närhet i ett mångdimensionellt rum. Modellen har funnits i över ett årtionde, och har bevisat sin användbarhet i en mängd experiment och tillämpningar. Trots att ordrumsmodellen varit föremål för omfattande forskning och användning är dess teoretiska grundvalar i stort sett outforskade. Denna avhandling syftar till att besvara frågan *vilken typ av betydelserelationer representeras i ordrumsmodellen*?

Svaret härleds genom att identifiera och diskutera ordrumsmodellens tre teoretiska grundpelare: den geometriska betydelsemetaforen, den distributionella metoden, och den strukturalistiska betydelseteorin. Avhandlingen visar att ordrumsmodellen representerar två olika betydelserelationer mellan ord — syntagmatiska eller paradigmatiska relationer — beroende på hur ordens distributionsmönster beräknas. Skillnaden mellan syntagmatiska och paradigmatiska ordrum demonstreras empiriskt i ett antal olika experiment, inklusive jämförelser med tesaurusar, associationsnormer, synonymtest, en lista med antonympar, samt ordklasstillhörighet.

Acknowledgments

"What about me? You didn't thank me!"
"You didn't do anything..."
"But I like being thanked!"
(Homer and Lisa Simpson in "Realty Bites")

Ideas are like bacteria: some are good, some are bad, all tend to thrive in stimulating environments, and while some pass by without leaving so much as a trace, some are highly contagious and extremely difficult to get out of your system. Some actually changes your whole life. The ideas presented in this dissertation have a lot in common with such bacteria: they have been carefully nourished in the widely stimulating research environments — SICS (Swedish Institute of Computer Science), GSLT (Graduate School of Language Technology), and SU (Stockholm University) — to which I am proud, and thankful, for having been part of; they have been highly contagious, and have spread to me through a large number of people, whom I have tried to list below; they have also proven impossible to get rid of. So persistent have they been that they ended up as a doctoral dissertation — if this text will work as a cure or not remains to be seen. People infected with word-space bacteria are:

First and foremost, Jussi Karlgren: supervisor extraordinaire, collaborator royale, and good friend. Thank you so much for these inspiring, rewarding, and — most importantly — *fun* years!

Secondly, Pentti Kanerva: inventor of Random Indexing and mentor. Thank you for all your tireless help and support throughout these years!

Thirdly, Anders Holst: hacker supreme and genius. Thank you for your patience in answering even the stupidest questions, and for introducing me to the magics of Guile!

Additionally, Gunnar Eriksson has read this text carefully and provided numerous invaluable comments; Magnus Boman, Martin Volk and Arne Jönsson have read and commented on parts of the text; Hinrich Schütze, Susan Dumais, Thomas Landauer and David Waltz have generously answered any questions I have had.

I salute Fredrik Olsson for being a good colleague, a good training buddy, and more than anything for being a good friend; Rickard Cöster for exciting collaborations, inspiring discussions, and most of all for all the fun; Ola Knutsson for invigorating philosophical excursions and for all the laughs; David Swanberg, who I started this word-space odyssey with, and who have remained a good friend.

My sincere thanks go to Björn Gambäck (LaTeXguru), Preben Hansen, Kristofer Franzen, Martin Svensson, and the rest of the lab at SICS; to Vicki Carleson for always hunting down the articles and books I fail to find on my own; to Mikael Nehlsen for sysadmin at SICS; to Robert Andersson for sysadmin at GSLT; to Marianne Rosenqvist for always keeping track of my coffee jug; to Joakim Nivre, Jens Allwood, and everyone at GSLT; to Martin, Magnus, Jonas, and Johnny (formerly) at KTH; and to the computational linguistics group at SU.

Lastly, but most importantly, I am utterly grateful and proud for the unwavering love and support of my family: Johanna, Ingalill and Leif. I dedicate this work to you.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

"Play along! I'll explain later."
(Moe Szyslak in "Cape Feare")

1.1 Modeling meaning

Meaning is something of a holy grail in the study of language. Some believe that if meaning is unveiled, it will bring light into the darkest realms of linguistic mystery. Others doubt its mere existence. Semanticists of many disciplines roam the great plains of linguistics, chasing that elusive, but oh-so-rewarding, thing called "meaning." Skeptics, also of many disciplines, stand by and watch their quest with agnostic, and sometimes even mocking, prejudice.

Whatever the ontological status of meaning may be, contemporary linguistics need the concept as explanatory premise for certain aspects of the linguistic behavior of language users. To take a few obvious examples, it would be onerous to try to explain vocabulary acquisition, translation, or language understanding without invoking the concept of meaning. Granted, it might prove possible to accomplish this without relying on meaning, but the burden of proof lies with the Opposition. Thus far, the exorcism of meaning from linguistics has not been successful, and no convincing alternative has been presented.

My prime concern here is not fitting meaning into the framework of linguistics. My interest here is rather the possibility of building a *computational model* of meaning. Such a computational model of meaning is worth striving for because our computational models of linguistic behavior will be incomplete without an account of meaning. If we need meaning to fully explain linguistic behavior, then surely we

need meaning to fully *model* linguistic behavior. A large part of language technology[1] today is concerned with tasks and applications whose execution typically is assumed to require (knowledge about, or proficiency in using) meaning. Examples include lexicon acquisition, word-sense disambiguation, information access, machine translation, dialogue systems, etc. Even so, remarkably few computational models of meaning have been developed and utilized for practical application in language technology. The model presented in this dissertation is one of the few viable alternatives.

Note that we seek a *computational* and not psychological model of meaning. This means that, while it *should* be empirically sound and consistent with human behavior (it is, after all, a *model*), it *does not* have to constitute a neurophysiologically or psychologically truthful model of human information processing. Having said that, human (semantic) proficiency exhibits such impressive characteristics that it would be ignorant not to use it as inspiration for implementation: it is efficient, flexible, robust, and continual. On top of all that, it is also seemingly effortless.

1.2 Difficult problems call for a different strategy

I have thus far successfully avoided the question about the meaning of "meaning," and for a good reason: it is this question that conjures the grail-like quality of the concept of meaning. As foundational as the study of meaning seems for the study of language, as elusive is the definition of the concept. In one sense, everyone knows what meaning is — it is that which distinguishes words from being senseless condensates of sounds or letters, and part of that which we understand and know when we say that we understand and know a language — but in another sense, no one seems to be able to pin down exactly what this "meaning" is. Some 2000 years of philosophical controversy should warn us to steer well clear of such pursuits.

I will neither attempt to define the meaning of "meaning," nor review the taxonomy of semantic theories. I will simply note that defining meaning seems like a more or less impossible (and therefore perhaps not very meaningful) task, and that there are many theories about meaning available for the aspiring semanticist. However, despite the abundance of meaning theories, remarkably few have proven their mettle in actual implementation. For those that have, there is usually a fair amount of "fitting circles into squares" going on; the theoretical prescriptions often do not fit observable linguistic data, which tend to be variable, inconsistent and vague. Semantics has been, and still is, a surprisingly impractical occupation.

[1] There is an abundance of terms referring to the computational study of language, including "computational linguistics," "language engineering," "natural language processing," etc. I arbitrarily choose to use the term "language technology."

In keeping with this theoretical lopsidedness, there is a long and withstanding tradition in the philosophy of language and in semantics to view the incomplete, noisy and imprecise form of natural language as an obstacle that obscures rather than elucidates meaning. It is very common in this tradition to claim that we therefore need a more exact form of representation that obliterates the ambiguity and incompleteness of natural language. Historically, logic has often been cast in this role, with the idea that it provides a more stringent and precise formalism that makes explicit the semantic information hidden in the imprecise form of natural language. Advocates of this paradigm claim that we should *not* model natural language use, since it is noisy and imprecise; instead, we should model language in the abstract.

In stark contrast to such a prescriptive perspective, proponents of descriptive approaches to linguistics argue that ambiguity, vagueness and incompleteness are essential properties of natural language that should be nourished and utilized; these properties are not signs of communicative malfunction and linguistic deterioration, but of communicative prosperity and of linguistic richness. Descriptivists argue that it would be presumptuous to believe that the single most complex communication system developed in nature could be more adequately represented by some human-made formalism. Language has the form it has because it is the most *viable* form. In the words of Ludwig Wittgenstein (1953):

> It is clear that every sentence in our language 'is in order as it is.' That is to say, we are not *striving after* an ideal, as if our ordinary vague sentences had not yet got a quite unexceptional sense, and a perfect language awaited construction by us. (§98)

The computational model of meaning discussed in this dissertation — the *word-space model* — is based *entirely* on language data, which means that it embodies a thoroughly descriptive approach. It does not rely on a priori assumptions about language (or at least it does so to a bare minimum — see further Section 1.3). By grounding the representations in actual usage data, it only represents what is *really there* in the current universe of discourse. When meanings change, disappear or appear in the data at hand, the model changes accordingly; if we use an entirely different set of data, we will end up with an entirely different model of meaning. The word-space model acquires meanings by virtue of (or perhaps despite) being based entirely on noisy, vague, ambiguous and possibly incomplete language data.

It is the overall goal of this dissertation to investigate this alternative computational path to semantics, and to examine how far in our quest for meaning such a thoroughly descriptive approach may take us.

1.3 Simplifying assumptions

It is inevitable when dealing with language in computers to make a few simplifying assumptions about the nature of the data at hand. For example, we normally will not have access to the wealth of extralinguistic information available to, and utilized by, every human. It is perhaps superfluous to point out that computers have a very limited set of senses, and even if we arguably *can* make the computer see, hear and touch, we still only have a very rudimentary knowledge how to interpret the vision, sound and tactile signal. Written text, on the other hand, is often readily available in machine-readable format (modulo issues related to encoding standards), and we have a comparably good understanding how to interpret such data. In the remainder of this dissertation, when I speak of language I speak of *written* language, unless otherwise stated.

This focus on written language admittedly undermines the behavioral consistency of the word-space model. However, it is perfectly feasible to assume that any sufficiently literate person *can* learn the meaning of a new word through reading only, as Miller & Charles (1991) observe. It is very common, at least in certain demographics (i.e. middle-class in literate areas), that people can read and write but not speak and understand foreign languages.

It also seems perfectly feasible to assume that the general word-space methodology presented here *can* be applied to data sources other than text. Having said that, it is important to point out that I *do* rely on assumptions that are text specific. For example, I use the term "word" to refer to white-space delimited sequences of letters that have been morphologically normalized to base forms (a process called *lemmatization*). Thus, when I speak of words I speak of lemmatized types rather than of inflected tokens. This notion of a word does not translate unproblematically to, e.g., speech data.

Furthermore, I assume that these lemmatized types are atomic semantic units of language. I am well aware that this might upset both linguists, who tend to see morphemes as atomic semantic units; and psychologists, who tend to argue that humans store not only morphemes and words in semantic memory, but also multi-word terms, idioms, and phrases. I will bluntly neglect these considerations in this dissertation, and merely focus on words and their meanings. It should be noted that the methodology presented in the following text can directly and unproblematically be applied also to morphemes and multi-word terms. The granularity of the semantic units is just a matter of preprocessing.

Lastly, I should point out that the methodology presented in this dissertation requires consistency and regularity of word order. Languages that utilize free word order would arguably not be applicable to the kind of distributional analysis professed in this dissertation.

1.4 Research questions

The word-space model is slowly but steadily becoming part of the basic arsenal of language technology. From being regarded almost as a scientific curiosity not more than a decade ago, it is now on the verge of moving from research laboratories to practical application; it is habitually used in information-access applications, and has begun to see employment in commercial products.

Despite its apparent viability, it remains unclear in what sense the word-space model is a model of meaning. **Does it constitute a *complete* model of the *full* spectrum of meaning, or does it only convey *specific aspects* of meaning?** If so: *which* aspects of meaning does it represent? **Is it at all possible to extract semantic knowledge by merely looking at usage data?** Surely, the practical applicability of the word-space model implies an affirmative answer to the last question, but there are neither theoretical motivations nor empirical results to indicate *what type* of semantic information the word-space model captures. Filling this void is the central goal of this dissertation.

1.5 Dissertation road map

During the following 122 pages, I will explain *what* the word-space model is, *how* it is produced, and *what kind* of semantic information it contains. The "what" is the subject of Chapter 2, the "how" is the subject of Chapters 3 and 4, and the "what kind" is addressed through Chapters 5 to 15. Finally, Chapter 16 summarizes and concludes the dissertation.

The text is divided into four different parts. Part I presents the theoretical background, Part II contains the theoretical foreground, and constitutes my main contribution. Part III presents the experiments, and Part IV concludes the text.

Those who are already familiar with the word-space model and its implementations may safely skip Part I (Chapters 2 to 4). Those who are only interested in the *theoretical* contribution of this thesis can skim through Chapters 8 to 14, while those who are primarily interested in the empirical results instead should focus on these chapters. However, I want to make clear that the main contribution of this dissertation is theoretical, and that the experimental results presented in Part III should be viewed less as evidence than as indications. My advice is to read the whole thing.

Part I

Background

Chapter 2

The word-space model

"That's quite a nice model, sir."
"Model?"
(Waylon Smithers and Mr. Burns in "$pringfield")

I refer to the computational model of meaning discussed in this dissertation as the *word-space model*. This term is due to Hinrich Schütze (1993), who defines the model as follows:

> Vector similarity is the only information present in Word Space: semantically related words are close, unrelated words are distant. (p.896)

There are many different ways to produce such a computational model of semantic similarity (I will discuss three different implementations in Chapter 4). However, even if there are many different ways of arriving at a word space, the underlying theories and assumptions are the same. This fact is often obscured by the plentitude of appellations and acronyms that are used for different versions and different implementations of the underlying word-space model. The propensity for term-coining in this area of research is not only a major source of confusion, but a symptom of the theoretical poverty that permeates it. The single most important reason why researchers do not agree upon the terminology is because they fail to appreciate the fact that it is the same underlying ideas behind all their implementations.

It is one of the central goals of this dissertation to excavate the underlying semantic theory behind the word-space model, and to thereby untangle this terminological mess. I start in this chapter by reviewing the basic assumptions behind the word-space model: the theory of representation, and the theory of acquisition.

2.1 The geometric metaphor of meaning

The word-space model is, as the name suggests, a spatial representation of word meaning. Its core idea is that semantic similarity can be represented as proximity in n-dimensional space, where n can be any integer ranging from 1 to some very large number — we will consider word spaces of up to several millions of dimensions later on in this dissertation. Of course, such high-dimensional spaces are impossible to visualize, but we can get an idea of what a spatial representation of semantic similarity might look like if we consider a 1-dimensional and a 2-dimensional word space, such as those represented in Figure 2.1.

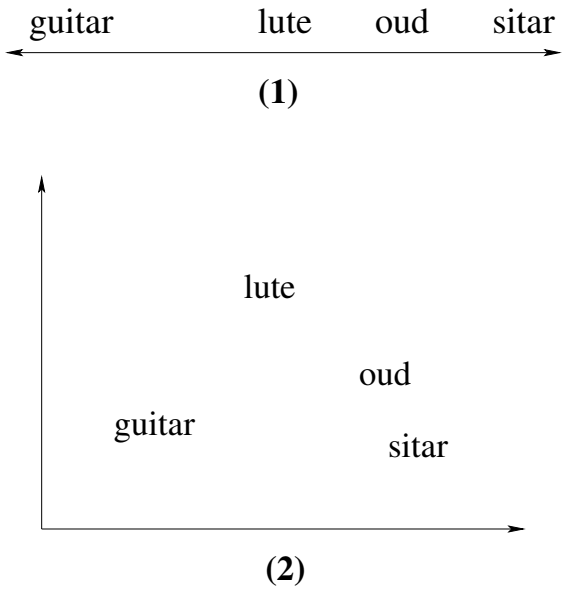

Figure 2.1: **(1)** A 1-dimensional word space, and **(2)** a 2-dimensional word space.

In these geometric representations, spatial proximity between words indicates how similar their meanings are. As an example, both word spaces in Figure 2.1 depicts *oud* as being closer to *lute* than to *guitar*, which should be interpreted as a representation of the meaning similarities between these words: the meaning (of) *oud* is more similar to the meaning (of) *lute* than to the meaning (of) *guitar*.

The use of spatial proximity as a representation of semantic similarity is neither accidental nor arbitrary. On the contrary, it seems like a very intuitive and natural way for us to conceptualize similarities, and the reason for this is obvious: we are, after all, *embodied* beings, who use our unmediated spatio-temporal knowledge of the world to conceptualize and make sense of abstract concepts. This has been pointed out by George Lakoff and Mark Johnson in a number of influential works (Lakoff & Johnson, 1980, 1999), where they argue that metaphors are the raw

materials of abstract concepts, and our basic tools for reasoning about abstract and complex phenomena. Language in general, and linguistic meaning in particular, are prime examples of such phenomena.

Lakoff and Johnson believe that our metaphorical tools for thought (to use yet another metaphor) are made up of a small number of basic, or *primary*, metaphors that are directly tied to our physical being-in-the-world. Spatial relations are salient in this respect: location, direction and proximity are basic properties of our embodied existence. This is why they also, according to Lakoff and Johnson, constitute the elements of (some of) our most fundamental metaphors.

One of the arguably most basic metaphors is the prevailing *similarity-is-proximity* metaphor: two things that are deemed to be similar in some sense are conceptualized as being *close to* or *near* each other, while dissimilar things are conceptualized as being *far apart* or *distant* from each other. This *similarity-is-proximity* metaphor is so prevalent that it is very difficult to think about similarities, let alone to talk about them, without using it (Lakoff & Johnson, 1999). This also applies to meanings: it is intuitive, if not inevitable, to use the *similarity-is-proximity* metaphor when talking about similarities of meaning. Words with similar meanings are conceptualized as being *near* each other, while words with dissimilar meanings are conceptualized as being *far apart*.

Note that the *similarity-is-proximity* metaphor presupposes another geometric metaphor: *entities-are-locations*. In order for two things to be conceptualized as being *close to* each other, they need to possess spatiality; they need to occupy (different) locations in a conceptual space. When we think about meanings as being *close to* or *distant from* each other, we inevitably conceptualize the meanings as locations in a semantic space, between which proximity can be measured. However, the *entities-are-locations* metaphor is completely vacuous in itself. Conceptualizing a sole word as a lone location in an n-dimensional space does nothing to further our understanding of the word. It is only when the space is populated with other words that this conceptualization makes any sense, and this is only due to the activation of the *similarity-is-proximity* metaphor.

Together, these two basic metaphors constitute the *geometric metaphor of meaning*:

> **The geometric metaphor of meaning**: *Meanings are locations in a semantic space, and semantic similarity is proximity between the locations.*

According to Lakoff's and Johnson's view on the embodied mind and metaphorical reasoning, this geometric metaphor of meaning is not something we can arbitrarily choose to use whenever we feel like it. It is not the product of disembodied speculation. Rather, it is part of our very existence as embodied beings. Thus,

the geometric metaphor of meaning is not based on intellectual reasoning about language. On the contrary, it is a *prerequisite* for such reasoning.

2.2 A caveat about dimensions

It might be wise at this point to obviate a few misconceptions about the nature of high-dimensional spaces.

Firstly, even though word spaces typically use more than one dimension to represent the similarities, it is still only *proximity* that is represented. It does not matter if we use one, two or six thousand dimensions, we are still only interested in how close to each other the locations in the space are. We should therefore try to resist the temptation of trying to find phenomenological correlates to the dimensions of high-dimensional word spaces. Although a 2-dimensional space adds the possibility of qualifying the similarities along the vertical axis (things can be *over* and *under* each other), and a 3-dimensional space adds depth (things can be *in front of* and *behind* each other), such qualifications are neither contained in, nor sanctioned by the *similarity-is-proximity* metaphor. As Karlgren (2005) points out, expressions such as "close in meaning" or "closer in meaning" are acceptable and widely used, whereas expressions such as "*slightly above in meaning" and "*more to the north in meaning" are not.

Why do I stress this point? Because it would lead to severe problems if we thought that we could find phenomenological correlates to higher dimensions in the word-space model. Granted, we have seen renderings of a second and third dimension, and a possible rendering of a fourth dimension might be time (things can be *before* or *after* each other), but then what? What kind of similarity does the 13th dimension represent? And what about the 666th, or the 198 604 021 003rd? One should keep in mind that the kind of spaces we normally use in the word-space model are very high-dimensional, and that it would be virtually impossible to find a phenomenological correlate to every dimension.

Secondly, high-dimensional spaces behave in ways that might seem counterintuitive to beings such as us who live in a spatially low-dimensional environment. Even the most basic spatial relations — such as proximity — behave differently in high-dimensional spaces than they do in low-dimensional ones. We can exemplify this without having to plunge too deep into mathematical terminology with the simple observation that whenever we add more dimensions to a space, there is more room for locations in that space to be far apart: things that are close to each other in one dimension are also close to each other in two, and generally also in three dimensions, but can be prohibitively far apart in 3 942 dimensions. A more mathematical example of the counterintuitive properties of high-dimensional spaces is the fact that objects in high-dimensional spaces have a larger amount

of surface area for a given volume than objects in low-dimensional spaces.[1] This is of course neither surprising nor problematic from a mathematical perspective. The lesson here is simply that we should exercise great caution about uncritically transferring our spatial intuitions that are fostered by a life in three dimensions to high-dimensional spaces.

2.3 The distributional hypothesis of meaning

We have seen that the word-space model uses the geometric metaphor of meaning as representational basis. But the word-space model is not only the spatial representation of meanings. It is also the *way* the space is built. What makes the word-space model unique in comparison with other geometrical models of meaning is that the space is constructed with no human intervention, and with no a priori knowledge or constraints about meaning similarities. In the word-space model, the similarities between words are *automatically* extracted from language data by looking at empirical evidence of real language use.

As data, the word-space model uses statistics about the distributional properties of words. The idea is to put words with similar distributional properties in similar regions of the word space, so that proximity reflects distributional similarity. The fundamental idea behind the use of distributional information is the so-called *distributional hypothesis*:

> **The distributional hypothesis***: words with similar distributional properties have similar meanings.*

The literature on word spaces abounds with formulations to this effect. A good example is Schütze & Pedersen (1995), who state that "words with similar meanings will occur with similar neighbors if enough text material is available," and Rubenstein & Goodenough (1965) — one of the very first studies to explicitly formulate and investigate the distributional hypothesis — who state that "words which are similar in meaning occur in similar contexts."

[1] As an example, visualize two nested squares, one centered inside the other. Now consider how large the small square needs to be in order to cover 1% of the area of the larger square. For 2-dimensional squares, the inner square needs to have 10% of the edge length of the outer square ($0.10 \times 0.10 = 0.01$), and for 3-dimensional cubes, the inner cube needs to have about 21% of the edge of the outer cube ($0.21 \times 0.21 \times 0.21 \approx 0.01$). To generalize, for n-dimensional cubes, the inner cube needs to have an edge length of $0.01^{\frac{1}{n}}$ of the side of the outer cube. For $n = 1\ 000$, that is 99.5%! Thus, if the outer 1 000-dimensional cube has edges 2 units long, and the inner 1 000-dimensional cube has edges 1.99 units long, the outer cube would still contain *one hundred times* more volume. This means that the vast majority of the volume of a solid in high-dimensional spaces is concentrated in a thin shell near its surface. This example was first brought to my attention in September 2005 on Eric Lippert's blog http://blogs.msdn.com/ericlippert/archive/2005/05/13/417250.aspx

The distributional hypothesis is usually motivated by referring to the *distributional methodology* developed by Zellig Harris (1909–1992). In Harris' distributional methodology, the *explanans* is reduced to a set of distributional facts that establishes the basic entities of language — phonemes, morphemes, and syntactic units — and the (distributional) relations between them. Harris' idea was that the members of the basic classes of these entities behave distributionally similarly, and therefore can be grouped according to their distributional behavior. As an example, if we discover that two linguistic entities, w_1 and w_2, tend to have similar distributional properties, for example that they occur with the same other entity w_3, then we may posit the explanandum that w_1 and w_2 belong to the same linguistic class. Harris believed that it is possible to typologize the whole of language with respect to distributional behavior, and that such distributional accounts of linguistic phenomena are "complete without intrusion of other features such as history or meaning." (Z. Harris, 1970).[2]

So where does meaning fit into the distributional paradigm? Reviewers of Harris' work are not entirely unanimous regarding the role of meaning in the distributional methodology (Nevin, 1993). On the contrary, this seems to be one of the main sources of controversy among his commentators — how does the distributional methodology relate to considerations on meaning? On the one hand, Harris explicitly shunned the concept of meaning as part of the explanans of linguistic theory:

> As Leonard Bloomfield pointed out, it frequently happens that when we do not rest with the explanation that something is due to meaning, we discover that it has a formal regularity or 'explanation.' (Z. Harris, 1970, p.785)

On the other hand, he shared with his intellectual predecessor, Leonard Bloomfield (1887–1949), a profound interest in linguistic meaning. Just as Bloomfield had done, he too realized that meaning in all its social manifestations is far beyond the reach of linguistic theory.[3] Even so, Harris was confident that his distributional methodology would be complete with regards to linguistic phenomena. The above quote continues:

> It may still be 'due to meaning' in one sense, but it accords with a distributional regularity.

[2] Harris did not exclude the possibility of other scientific studies of language. On the contrary, he explicitly states in "Distributional structure" (Z. Harris, 1970) that "It goes without saying that other studies of language — historical, psychological, etc. — are also possible, both in relation to distributional structure and independently of it." (p.775)

[3] "Though we cannot list all the co-occurrents [...] of a particular morpheme, or define its meaning fully on the basis of these." (Z. Harris, 1970, p.787)

What Harris is saying here is that even if extralinguistic factors *do* influence linguistic events, there will always be a distributional correlate to the event that will suffice as explanatory principle. Harris was deeply concerned with linguistic methodology, and he believed that linguistics as a science should (and, indeed, could) only deal with what is *internal* to language; whatever is in language is subject to linguistic analysis, which for Harris meant *distributional* analysis. This view implies that, in the sense that meaning is linguistic (i.e. has a purely linguistic aspect), it *must* be susceptible to distributional analysis:

> ...the linguistic meanings which the structure carries can only be due to the relations in which the elements of the structure take part. (Z. Harris, 1968, p.2)

The distributional view on meaning is expressed in a number of passages throughout Harris' works. The most conspicuous examples are *Mathematical Structures of Language* (p.12), where he talks about meaning being related to the combinatorial restrictions of linguistic entities; and "Distributional Structure" (p.786), where he talks about the correspondence between difference of meaning and difference of distribution. The consistent core idea in these passages is that linguistic meaning is inherently differential, and not referential (since that would require an extralinguistic component); it is *differences* of meaning that are mediated by *differences* of distribution. Thus, the distributional methodology allows us to quantify the amount of meaning difference between linguistic entities; it is the *discovery procedure* by which we can establish semantic similarity between words:[4]

> ...if we consider words or morphemes A and B to be more different in meaning than A and C, then we will often find that the distributions of A and B are more different than the distributions of A and C. In other words, difference of meaning correlates with difference of distribution. (Z. Harris, 1970, p.786)

The distributional hypothesis has been validated in a number of experiments. The earliest one that I am aware of is Rubenstein & Goodenough (1965), who compared contextual similarities with synonymy judgments provided by university students. Their experiments demonstrated that there indeed is a correlation between semantic similarity and the degree of contextual similarity between words. Almost 30 years later, Miller & Charles (1991) repeated Rubenstein's & Goodenough's experiment using 30 of the original 65 noun pairs, and reported remarkably similar results. Miller & Charles concur in that the experiments seem to support the

[4]Note that Harris talks about meaning *differences*, but that the distributional hypothesis professes to uncover meaning *similarities*. There is no contradiction in this, since differences and similarities are, so to speak, two sides of the same coin.

distributional (or, as they call it, the *contextual*) hypothesis. Other experimental validations of the distributional hypothesis include Miller & Charles (2000) and McDonald & Ramscar (2001).

2.4 A caveat about semantic similarity

As we have seen in this chapter, the word-space model is a model of *semantic similarity*. Padó & Lapata (2003) note that the notion of semantic similarity has rendered a considerable amount of criticism against the word-space model. The critique usually consists of arguing that the concept of semantic similarity is too broad to be useful, in that it encompasses a wide range of different semantic relations, such as synonymy, antonymy, hyponymy, meronymy, and so forth. The critics claim that it is a serious liability that simple word spaces cannot indicate the *nature* of the semantic similarity relations between words, and thus does not distinguish between, e.g., synonyms, antonyms, and hyponyms.

This criticism is arguably valid from a prescriptive perspective where these relations are a priorily given as part of the linguistic ontology. From a descriptive perspective, however, these relations are *not* axiomatic, and the broad notion of semantic similarity seems perfectly plausible. There are studies that demonstrate the psychological reality of the concept of semantic similarity. For example, Miller & Charles (1991) point out that people instinctively make judgments about semantic similarity when asked to do so, without the need for further explanations of the concept; people appear to instinctively understand what semantic similarity is, and they make their judgments quickly and without difficulties. Several researchers report high inter-subject agreement when asking a number of test subjects to provide semantic similarity ratings for a given number of word pairs (Rubenstein & Goodenough, 1965; Henley, 1969; Miller & Charles, 1991).

The point I want to make here is that the inability to further qualify the nature of the similarities in the word-space model is a consequence of using the distributional methodology as discovery procedure, and the geometric metaphor of meaning as representational basis. The distributional methodology only discovers differences (or similarities) in meaning, and the geometric metaphor only represents differences (or similarities) in meaning. If we want to claim that we extract and represent some particular *type* of semantic relation in the word-space model, we need to modify either the distributional hypothesis or the geometric metaphor, or perhaps even both. For the time being, we have to make do with the broad notion of semantic similarity.

Chapter 3

Word-space algorithms

"Oh, algebra! I'll just do a few equations."
(Bart Simpson in "Special Edna")

In the last chapter, we saw that the word-space model is a model of semantic similarity, which uses the geometric metaphor of meaning as representational framework, and the distributional methodology as discovery procedure. After having read the last chapter, we know what the model should look like, and we know what to put into the model. However, we do not yet know *how* to build the model; how do we go from distributional statistics to a geometric representation — a high-dimensional word space? Answering this question is the subject matter of this chapter.

3.1 From statistics to geometry: context vectors

Unsurprisingly, we get a clue to how we could proceed in order to transform distributional information to a geometric representation from Zellig Harris himself, who writes that:

> The distribution of an element will be understood as the sum of all its environments. (Z. Harris, 1970, p.775)

In this quote, Harris effectively equates the distributional profile of a word with the totality of its environments. Consider how we could go about collecting such distributional profiles for words: imagine that we have access to the data in Figure 3.1, and we want to collect distributional profiles from it.

Whereof one cannot speak thereof one must be silent

Figure 3.1: Imaginary data.

The first thing we have to decide is: what is an environment? In linguistics, an environment is called a *context*. Now, a context can be many things: it can be anything from the surrounding words to the socio-cultural circumstance of an utterance. Dictionaries often provide (at least) two different definitions of context: one specifically linguistic and one more general. A useful example is *Longman Dictionary of Contemporary English* that defines context as:

> (1) The setting of a word, phrase etc., among the surrounding words, phrases, etc., often used for helping to explain the meaning of the word, phrase, etc.
>
> (2) The general conditions in which an event, action, etc., takes place.

For the time being, it will suffice to adopt the first of these two definitions of context as the *linguistic surroundings*.[1] In this example, I define context as one preceding and one succeeding word. As an example, the context for "speak" is "cannot" and "thereof," and the context for "be" is "must" and "silent."

One way to collect this information for the example text is to tabulate the contextual information, so that for each word we provide a list of the co-occurrents of the word, and the number of times they have co-occurred:

Word	Co-occurrents
whereof	(one 1)
one	(whereof 1, cannot 1, thereof 1, must 1)
cannot	(one 1, speak 1)
speak	(cannot 1, thereof 1)
thereof	(speak 1, one 1)
must	(one 1, be 1)
be	(must 1, silent 1)
silent	(be 1)

Table 3.1: Lists of the co-occurrents.

Now, imagine that we take away the actual words, and only leave the co-occurrence counts. Also, we make each list equally long by adding zeroes in the places where we lack co-occurrence information. We also sort each list so that the co-occurrence

[1] I will discuss different notions of context at length in Chapter 7.

counts for each context come in the same places in the lists. The result would look something like this:

Word	Co-occurrents							
	whereof	one	cannot	speak	thereof	must	be	silent
whereof	0	1	0	0	0	0	0	0
one	1	0	1	0	1	1	0	0
cannot	0	1	0	1	0	0	0	0
speak	0	0	1	0	1	0	0	0
thereof	0	1	0	1	0	1	0	0
must	0	1	0	0	1	0	1	0
be	0	0	0	0	0	1	0	1
silent	0	0	0	0	0	0	1	0

Table 3.2: Lists of co-occurrence counts.

As an example, the co-occurrence-count list for "speak" is $(0, 0, 1, 0, 1, 0, 0, 0)$, and the list for "be" is $(0, 0, 0, 0, 0, 1, 0, 1)$. Such ordered lists of numbers are also called *vectors*. Formally, a vector $\vec{v}$ is an element of a vector space, and is defined by n components or *coordinates* $\vec{v} = (x_1, x_2, ..., x_n)$. The coordinates effectively describe a location in the n-dimensional space. An example of a 2-dimensional vector space with two vectors $\vec{v}_1 = (1, 2)$ and $\vec{v}_2 = (3, 2)$ is depicted in Figure 3.2.

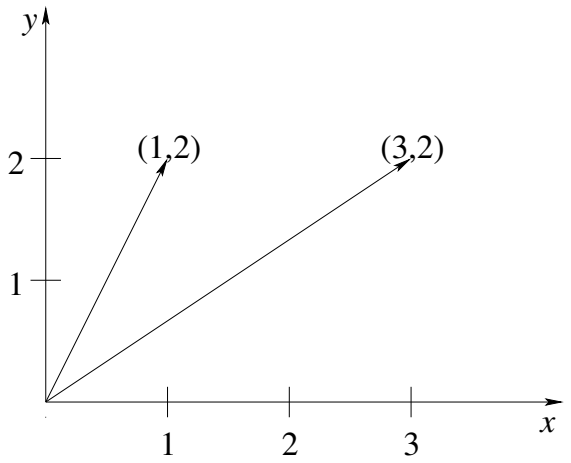

Figure 3.2: A 2-dimensional space with vectors $\vec{v}_1 = (1, 2)$ and $\vec{v}_2 = (3, 2)$.

I call vectors of co-occurrence counts such as those in Table 3.2 *context vectors*,[2] because they effectively constitute a representation of the *sum of the words' contexts*

[2]The term "context vector" has previously been used by some researchers in word-sense disambiguation to refer to a representation of a *particular* context for a word (Wilks et al., 1990;

(cf. the quote from Harris above). Another way of looking at context vectors is to say that they describe locations in context space. Thus, the concept of a context vector is the solution to our problem of how to go from distributional statistics to a geometric representation.

3.2 Probabilistic approaches

It should be noted that context vectors are not the only way to utilize distributional information. There is a large body of work in language technology that uses distributional information to compute similarities between words, but that does *not* use context vectors. Instead, this paradigm uses a probabilistic framework, where similarities between words are expressed in terms of functions over distributional probabilities (Church & Hanks, 1989; Hindle, 1990; Hearst, 1992; Ruge, 1992; Dagan et al., 1993; Pereira et al., 1993; Grefenstette, 1994; Lin, 1997; Baker & McCallum, 1998; Lee, 1999). Although these probabilistic approaches *do* rely on the distributional methodology as discovery procedure, they *do not* utilize the geometric metaphor of meaning as representational basis, and thus fall outside the scope of this venture.

A good explanation of the difference between the geometric and the probabilistic approaches is the distinction made by Ruge (1992):

> Their intentions are evaluating the relative position of two items in the semantic space in the first case, and the overlap of property sets of the two items in the second case. (p.322)

Ruge further argues that the geometric approach is psychologically more realistic, when she concludes that:

> ...the model of semantic space in which the relative position of two terms determines the semantic similarity better fits the imagination of human intuition [about] semantic similarity than the model of properties that are overlapping. (p.328–329)

3.3 A brief history of context vectors

The idea of context vectors has its earliest origins in work on feature space representations of meaning in psychology. The pioneer in this field is Charles Osgood and his colleagues, who in the early 1950s developed the *semantic differential* approach to meaning representation (Osgood, 1952; Osgood et al., 1957). In this

Schütze, 1992; Schütze & Pedersen, 1995). Note that in my use of the term, influenced by Gallant (1991a, 1991b, 2000), it refers to the *totality* of a word's contexts.

approach, words are represented by feature vectors where the elements are human attitudinal ratings along a seven-point scale for a number of contrastive adjective pairs such as "soft–hard," "fast–slow" and "clean–dirty." The idea is that such feature vectors can be used to measure the psychological distance between words. A very simplified example of the feature vectors for the words "mouse" and "rat" based on three contrastive pairs is given below:

	small–large	bald–furry	docile–dangerous
mouse	2	6	1
rat	2	6	4

Table 3.3: Feature vectors based on three contrastive pairs for the words "mouse" and "rat."

Osgood's feature-space approach was the major influence for early connectionist research that used distributed representations of meaning (Smith & Medin, 1981; Cottrell & Small, 1983; Small et al., 1988). One of the most influential heirs to the feature-space approach from this period is Waltz & Pollack (1985), who used what they call *micro-features* to represent the meaning of words. These micro-features consisted of distinctive pairs such as "animate–inanimate" and "edible–inedible," which were chosen to correspond to major distinctions that humans make about their surroundings. The set of micro-features (which were on the order of a thousand) were represented as a vector, where each element corresponded to the level of activation for that particular micro-feature. This representation was thus remarkably similar to Osgood's semantic differential approach, despite the fact that Waltz & Pollack were not directly influenced by Osgood's works.[3]

Waltz & Pollack's version of the feature-space approach was in its turn a major inspiration for Stephen Gallant, who introduced the term "context vector" to describe the feature-space representations (Gallant, 1991a, 1991b). In Gallant's algorithm, context vectors were defined by a set of manually derived features, such as "human," "man," "machine," etc. A simplified example of a manually defined context vector, such as those used in Gallant's algorithm is displayed in Table 3.4. Remnants of the feature space approach is still used in cognitive science, by, e.g., Gärdenfors (2000) under the term *conceptual spaces*.

Several researchers observed that there are inherent drawbacks with the feature-space approach (Ide & Veronis, 1995; Lund & Burgess, 1996). For example, how do we choose appropriate features? The idea with using feature spaces is that they allow us to use a *limited* number of semantic features to describe the full meanings of words. The question is which features should we use, and how can we

[3]David L. Waltz, personal communication.

	human	man	machine	politics	...
astronomer	+2	+1	-1	-1	...

Table 3.4: Manually defined context vector for the word "astronomer."

define them? Is it even theoretically possible to devise a limited set of semantic (contrastive) features that would exhaustively characterize the entire semantics of a language? How many features are enough, and how do we know when we have reached the sufficient number?

These questions imply that it would be desirable to devise automatic methods to construct feature spaces. One of the earliest examples of such methods comes from Gallant (1991b), who, in addition to the (traditional) feature vectors, used what he called *dynamic* context vectors computed from the contexts in which the words occur. In essence, Gallant's algorithm can be described as a two-step operation (Gallant, 2000):

1. A context vector is initialized for each word as a normalized random vector.

2. While making several passes through the corpus, the context vectors are changed in a manner resembling Kohonen's Self-Organizing Maps (Kohonen, 1995) to be more like the context vectors of the surrounding words.

The resulting context vectors were then used for word-sense disambiguation, by comparing them to the manually defined ones (Gallant, 1991b), and for information retrieval, by defining document vectors as the weighted sum of the context vectors of the constituent words (Gallant, 1991a, 2000).

Other early attempts at deriving context vectors automatically from the contexts in which words occur include Wilks et al. (1990), Schütze (1992), Pereira et al. (1993), and Niwa & Nitta (1994). The arguably most influential work from this period comes from Hinrich Schütze (1992, 1993), who builds context vectors (which he calls "term vectors" or "word vectors") in precisely the manner described in Section 3.1 above: co-occurrence counts are collected in a words-by-words matrix, in which the elements record the number of times two words co-occur within a set window of word tokens. Context vectors are then defined as the rows or the columns of the matrix (the matrix is symmetric, so it does not matter if the rows or the columns are used). A similar approach is described by Qiu & Frei (1993), with the difference that they use a words-by-documents matrix to collect the co-occurrence counts.

3.4 The co-occurrence matrix

The approach pioneered by Schütze and Qiu & Frei has become standard practice for word-space algorithms: data is collected in a matrix of co-occurrence counts, and context vectors are defined as the rows or columns of the matrix. Such a matrix of co-occurrence counts is called a *co-occurrence matrix*, and is normally denoted by F (for frequency). As we have already seen, the matrix can either be a words-by-words matrix $w \times w$, where w are the word types in the data, or a words-by-documents matrix $w \times d$, where d are the documents in the data. A cell f_{ij} of the co-occurrence matrix records the frequency of occurrence of word i in the context of word j or in document j. As the attentive reader will have noticed, we have already seen an example of a words-by-words co-occurrence matrix in Table 3.2.

Those versed in the field of information retrieval will recognize words-by-documents matrices as instantiations of the *vector-space model* developed by Gerald Salton and colleagues in the 1960s within the framework of the SMART information-retrieval system (Salton & McGill, 1983). In the traditional vector-space model, a cell f_{ij} of matrix F is the *weight* of term i in document j.[4] The weight is usually composed of three components (Robertson & Spärck Jones, 1997):

$$f_{ij} = \textsc{tf}_{ij} \cdot \textsc{df}_i \cdot \textsc{s}_j$$

where $\textsc{tf}_{ij}$ is some function of the frequency of term i in document j (TF for *term frequency*), $\textsc{df}_i$ is some function of the number of documents term i occurs in (DF for *document frequency*), and $\textsc{s}_j$ is some normalizing factor, usually dependent on document length (S for *scaling*).

The point of the first component $\textsc{tf}_{ij}$ is to indicate how important term i is for document j. The idea is that the more often a term occurs in a document, the more likely it is to be important for identifying the document. The observation that frequency is a viable indicator of the quality of index terms originates in the work of Hans Peter Luhn in the late 1950's (Luhn, 1958).

The second component $\textsc{df}_i$ indicates how discriminative term i is. The idea is that terms that occur in few documents are better discriminators than terms that occur in many. The arguably most common version of the document frequency measure is to use the *inverse document frequency* (IDF), originally established by Karen Spärck Jones (1972), and computed as:

$$\textsc{idf} = \log \frac{D}{\textsc{df}_i}$$

[4]Note that I use "term" instead of "word" here. The reason is that multi-word terms are often used in information retrieval, so it is more natural to speak of "terms" than "words" in this context.

where D is some constant — usually the total number of documents (or a function thereof) in the document collection.

The third component S_j is normally a function of the length of document j, and is based on the idea that a term that occurs the same number of times in a short and in a long document should be more important for the short one. That is, we do not want long documents to end up at the top of the ranking list in an information-retrieval system merely because they are long. There are many variations of document-length normalization, ranging from very simple measures that just count the number of tokens in a document (Robertson & Spärck Jones, 1997), to more complex ones, such as pivoted document-length normalization (Singhal et al., 1996).

Most information-retrieval systems in use today implement some version of this type of combinatorial weight, known as the TFIDF family of weighting schemes (Salton & Yang, 1973). This is true also for word-space algorithms that use a words-by-documents co-occurrence matrix. However, word-space algorithms that use a words-by-words co-occurrence matrix normally do *not* use TFIDF-weights (the exception being Lavelli et al. (2004), who I will return to in Chapter 15).

Words-by-words co-occurrence matrices are instead typically populated by simple frequency counting: if word i co-occurs 16 times with word j, we enter 16 in the cell f_{ij} in the words-by-words co-occurrence matrix. The co-occurrences are normally counted within a *context window* spanning some — usually small — number of words. Remember from Section 3.1 that we used a window consisting of only the immediately preceding word and the immediately succeeding word when populating the matrix in Table 3.2.

Note that if we count co-occurrences symmetrically in both directions within the window, we will end up with a symmetric words-by-words co-occurrence matrix in which the rows equals the columns. However, if we instead count the co-occurrences in only one direction (i.e. the left or right context only), we will end up with a *directional* words-by-words co-occurrence matrix. In such a directional co-occurrence matrix, the rows and columns contain co-occurrence counts in different directions: if we only count co-occurrences with preceding words within the context window, we will end up with a co-occurrence matrix in which the rows contain left-context co-occurrences, and the columns contain right-context co-occurrences; if we only count co-occurrences with succeeding words within the context window, we will end up with the transpose: the rows contain right-context co-occurrences, while the columns contain left-context co-occurrences. We can refer to the former as a left-directional words-by-words matrix, and to the latter as a right-directional words-by-words matrix.

Table 3.5 demonstrates a right-directional words-by-words co-occurrence matrix for the example data in Section 3.1. Note that the row and column vectors for the words are different. The row vector contains co-occurrence counts with words

Word	Co-occurrents							
	whereof	one	cannot	speak	thereof	must	be	silent
whereof	0	1	0	0	0	0	0	0
one	0	0	1	0	0	1	0	0
cannot	0	0	0	1	0	0	0	0
speak	0	0	0	0	1	0	0	0
thereof	0	1	0	0	0	0	0	0
must	0	0	0	0	0	0	1	0
be	0	0	0	0	0	0	0	1
silent	0	0	0	0	0	0	0	0

Table 3.5: Directional words-by-words co-occurrence matrix.

that have occurred to the right of the words, while the column vector contains co-occurrence counts with words that have occurred to their left. I will discuss the use of context windows more thoroughly in Section 7.4.

3.5 Similarity in mathematical terms

Now that we know how to construct context vectors — we collect data in a co-occurrence matrix and define the rows or columns as context vectors — we may ask what we can use them for? What should we do with the context vectors once we have harvested them?

As I mentioned in Section 3.1, an n-dimensional vector effectively identifies a location in an n-dimensional space. However, the locations by themselves are not particularly interesting — knowing that the word "massaman" is located at the coordinates $(31, 5, -34, 17, -8, -21, 67)$ in a real-valued 7-dimensional space does not tell us anything (other than its location in the 7-dimensional space, that is). Rather, it is the *relative* locations that are interesting — knowing that the words "massaman" and "panaeng" are closer to each other than to the word "norrsken" is precisely the kind of information we are interested in. The principal feature of the geometric metaphor of meaning is not that meanings can be represented as locations in a (semantic) space, but rather that *similarity* between (the meaning of) words can be expressed in spatial terms, as *proximity* in (high-dimensional) space. As I pointed out in Chapter 2, the *meanings-are-locations* metaphor is completely vacuous without the *similarity-is-proximity* metaphor.

Now, the context vectors do not only allow us to go from distributional information to a geometric representation, but they also make it possible for us to *compute* (distributional, semantic) proximity between words. Thus, the point of the context

vectors is that they allow us to define (distributional, semantic) similarity between words in terms of vector similarity.

There are many ways to compute the similarity between vectors, and the measures can be divided into *similarity* measures and *distance* measures. The difference is that similarity measures produce a high score for similar objects, whereas distance measures produce a low score for the same objects: large similarity equals small distance, and conversely. A similarity measure can therefore be seen as the inverse of a distance measure. Generally, we can transform a distance measure $\text{dist}(x, y)$ into a similarity measure $\text{sim}(x, y)$ by simply computing:

$$\text{sim}(x, y) = \frac{1}{\text{dist}(x, y)}$$

The arguably simplest vector similarity metric is the scalar (or *dot*) product between two vectors $\vec{x}$ and $\vec{y}$, computed as:

$$\text{sim}_{\text{S}}(\vec{x}, \vec{y}) = x \cdot y = x_1 y_1 + x_2 y_2 + ... + x_n y_n$$

Another simple metric is the Euclidian *distance*, which is measured as:

$$\text{dist}_{\text{E}}(\vec{x}, \vec{y}) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

The Euclidean distance is a special case of the general Minkowski metric:

$$\text{dist}_{\text{M}}(\vec{x}, \vec{y}) = \left(\sum_{i=1}^{n} |x_i - y_i|^N \right)^{\frac{1}{N}}$$

with $N = 2$. If we let $N = 1$, we get the City-Block (or Manhattan) metric, and if we let $N \to \infty$, we get the Chebyshev distance. An illustration (inspired by Chávez & Navarro (2000)) of the differences between a few Minkowski metrics is given in Figure 3.3.

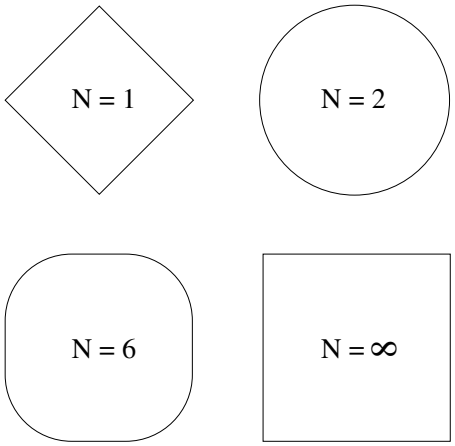

Figure 3.3: The same distance to the center for a number of Minkowski metrics with different N.

As Widdows (2004) points out, these measures are not ideal to use for word-space algorithms. The reason is that the scalar product favors frequent words (i.e. words with many and large co-occurrence counts will end up being too similar to most other words), while Minkowski metrics have the opposite problem: frequent words will end up being *too far* from the other words. A solution to this problem is to factor out the effects of vector length,[5] which can be done by normalizing the vectors by their length (or *norm*), given by:

$$| x | = \sqrt{x \cdot x}$$

A convenient way to compute normalized vector similarity is to calculate the cosine of the angles between two vectors $\vec{x}$ and $\vec{y}$, defined as:

$$\text{sim}_{\text{COS}}(\vec{x}, \vec{y}) = \frac{x \cdot y}{| x || y |} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2} \sqrt{\sum_{i=1}^{n} y_i^2}}$$

Note that the cosine measure corresponds to taking the scalar product of the vectors and then dividing by their norms. The cosine measure is the most frequently utilized similarity metric in word-space research, and the one I will use throughout this dissertation. It is attractive because it provides a fixed measure of similarity — it ranges from 1 for identical vectors, over 0 for orthogonal vectors,[6] to -1 for vectors pointing in the opposite directions. It is also comparatively efficient to compute (Widdows, 2004).

[5] A *long* vector is a vector with large values. As an example, consider the vectors in Figure 3.2. Vector $\vec{v}_2$ is longer than $\vec{v}_1$ because it has larger values.

[6] The word "*orthogonal*" comes from the Greek words "ortho", which means right, and "gonia", which means angle. Thus, orthogonal means being at right angles, just like two streets crossing each other. Formally, two vectors $\vec{x}$ and $\vec{y}$ are orthogonal if their scalar product is zero.

Chapter 4

Implementing word spaces

"Of course, it's so simple! Wait, no it's not. It's needlessly complicated!"
(Homer Simpson in "The Computer Wore Menace Shoes")

In the last chapter, I built a bridge between the distributional methodology and the geometric metaphor of meaning. The bridge turned out to be the concept of a context vector, which allows us to go from distributional statistics to a geometric representation. In this chapter, I will discuss a couple of problems with word-space algorithms, and look at how different implementations of the word-space model handle them.

4.1 The problem of very high dimensionality

When writing an algorithm that implements the word-space model, choosing vector similarity metric is not the only design decision we have to make. Another — important — decision is how to handle the potentially very high dimensionality of the context vectors. The problem is that the word-space methodology relies on statistical evidence to construct the word space — if there is not enough data, we will not have the required statistical foundation to build a model of word distributions. At the same time, the co-occurrence matrix will become prohibitively large for any reasonably sized data, which seriously affects the scalability and efficiency of the algorithm. This presents us with the following delicate dilemma: on the one hand, we need as much data as we can get our hands on in order to build a truthful model of language use; on the other hand, we want to use as little data as possible because our algorithms will become computationally prohibitive otherwise.

4.2 The problem of data sparseness

Another potential problem with the word-space methodology is that a majority of the cells in the co-occurrence matrix will be zero. For all co-occurrence matrices used in the experiments reported in Part III, more than 99% of the entries are zero. This is because only a fraction of the co-occurrence events that are possible in the matrix will actually occur, regardless of the size of the data. Only a tiny amount of the words in language are distributionally promiscuous; the vast majority of words only occur in a very limited number of contexts. This phenomenon is well known, and is an example of the general *Zipf's law* (Zipf, 1949).

4.3 Dimensionality reduction

The solution to these predicaments is usually spelled *dimensionality reduction*. The point of dimensionality reduction here is to represent high-dimensional data in a low-dimensional space, so that both the dimensionality and sparseness of the data are reduced, while still retaining as much of the original information as possible. In most implementations of word-space algorithms, dimensionality reduction is performed by first collecting the co-occurrence matrix, and then applying some form of dimensionality-reduction technique to it. The result of such an operation is a much denser and lower-dimensional space, in which the dimensions are typically no longer identified with the words or documents in the data.

The arguably simplest form of dimensionality reduction in word-space research is to filter out words and documents based on either linguistic or statistical criteria. Linguistic criteria can consist of, for example, removing documents based on stylistic criteria (Karlgren, 1999), or removing words that belong to certain grammatical classes. This latter approach is more widely known as *part-of-speech* filtering, and normally consists of removing words that belong to closed grammatical classes,[1] since they are assumed to have little or no semantic meaning. An apparent drawback with part-of-speech filtering is that it only removes a very small fraction of the vocabulary; the vast majority of words in language belong to open grammatical classes, so the effect of using part-of-speech filtering as dimensionality reduction is modest at best.

Words can also be filtered out based on statistical criteria, by removing words with unfavorable statistical properties. In the simplest case, this means removing words with very high and very low frequency of occurrence. Note that removing

[1] *Closed* grammatical classes are classes that very seldom acquire new members. Examples include adpositions, pronouns, conjunctions, and determiners. *Open* grammatical classes, on the other hand, are ever-changing, and constantly drop, replace, and add new members. Examples of such classes are nouns, adjectives, and verbs.

high-frequency words is comparable to using part-of-speech filtering, since words with high frequency tend to belong to closed grammatical classes. More sophisticated statistical criteria includes the TFIDF, and different variants and mixtures of the *Poisson* distribution (Damerau, 1965; Harter, 1975; Katz, 1996).[2] While statistical filtering typically removes more words than using linguistic criteria, and thus achieves a more distinct dimensionality-reduction effect, it also tends to remove a fair amount of words that belong to the open grammatical classes. The problem with this is that it becomes more or less arbitrary which words are excluded.

4.4 Latent Semantic Analysis

A radically different approach to dimensionality reduction is exemplified by the incomparably most renowned word-space implementation: Latent Semantic Analysis (LSA) (Landauer & Dumais, 1997). LSA was developed under the name Latent Semantic Indexing (LSI) (Dumais et al., 1988; Deerwester et al., 1990) in the late 1980s as an extension to the traditional vector-space model in information retrieval. The terms LSI and LSA have since come to be used more or less synonymously in the literature, but whereas "LSI" is used primarily in the context of information retrieval, "LSA" is used for the more general application of these ideas.[3] I thus use "LSA" in this dissertation.

The development of LSA was motivated by the inability of the vector-space model to handle synonymy — a query about "boats" will not retrieve documents about "ships" in the standard vector-space model. LSA solves this problem by reducing the original high-dimensional vector space into a much smaller space (but still relatively large; usually a few hundred dimensions), in which the original dimensions that represented words and documents have been condensed into a smaller set of "latent" dimensions that collapses words and documents with similar context vectors. This alleviates the problem with synonymy when retrieval is performed in the reduced space.

The dimensionality reduction is accomplished by using a statistical dimensionality-reduction technique called Singular Value Decomposition (SVD). The mathematical details of SVD need not concern us here (but see, for example, Berry et

[2]The Poisson distribution is a discrete probability distribution that can be used to describe the distributional behavior of function words (i.e. words belonging to the closed grammatical classes). Formally, the probability that a function word w will occur exactly k times in a text is given by $P_w(k) = e^{-\lambda_w} \frac{\lambda_w^k}{k!}$ where the parameter λ_w is the average number of occurrences of w per text: $\lambda_w = \frac{f}{T}$ where f is the total number of occurrences of w in the collection, and T is the total number of texts in the collection.

[3]Susan Dumais and Thomas Landauer, personal communication. Apparently, the patent for the original application of an information-retrieval system based on these ideas refers to it as "LSA."

al. (1995) for more details); for the present study it is sufficient to note that SVD is a matrix factorization technique that decomposes, or factorizes,[4] the original matrix into several (three when using SVD) smaller matrices, which can be multiplied to reproduce the original one. These smaller matrices contain the linearly independent factors of the original matrix (in the case of SVD, they are called "singular vectors" and "singular values").[5] If the smallest factors are disregarded when multiplying the smaller matrices, the result will be an *approximation* of the original co-occurrence matrix. This process is called *truncated* SVD, and is the favored dimensionality-reduction method in LSA.[6]

The applicability of LSA for information retrieval is well documented (Deerwester et al., 1990; Dumais, 1993; Dumais et al., 1997; Jiang & Littman, 2000).[7] The rationale for using SVD in an information-retrieval setting is obvious: words with similar co-occurrence patterns are grouped together, alleviating problems with synonymy, and allowing for the retrieval of documents that need not contain any of the query words. However, one may ask what the benefit of using SVD for the word-space model would be (in addition to the purely computational advantages of dimensionality reduction)?

One of the arguments for using SVD when constructing the word space is that the resulting reduced space not only contains "surface" co-occurrence relations (i.e. those contained in the original words-by-documents matrix), but, supposedly, also latent relations that reflect higher-order co-occurrences. Thus, the argument goes, when using truncated SVD to restructure the data, the word space will not only group together words that properly co-occur in a document, but also words that occur in *similar contexts* (Landauer et al., 1998). The intended effect is that the truncated SVD induces relations between rows (or between columns) that are similar to the same *other* rows (or columns) in the original co-occurrence matrix.

[4]Factorization is a mathematical process where an object is decomposed into the product of other objects, called factors. Thus, a matrix factorization is the right side of the equation $F = M_1 M_2 \dots M_m$.

[5]Linear independence means that none of a set of vectors can be written as a linear combination of the other vectors. As an example, the vectors (0,0,1), (0,1,0) and (1,0,0) are linearly independent, while the vectors (1,0,0), (0,0,1) and (1,0,1) are not, since the third of these last vectors can be written as a combination of the other two.

[6]It is interesting to note that Osgood and colleagues already in 1957 (Osgood et al., 1957) — roughly 30 years before the advent of LSA — mentioned the use of factor analysis to uncover orthogonal dimensions in the semantic space.

[7]It should be noted that there are critical voices as well. For example, Isbell & Viola (1998) demonstrate experimentally that truncated SVD generates an approximation that is sub-optimal with regard to a given information-retrieval problem, and Husbands et al. (2001) observe that LSA performs sub-optimally on the sizeable TREC collection. Dumais (2004) notes that "the reasons for the inconsistent performance of LSA are not clear and require further research." Perhaps one of the reasons for this lack of clarity is a dearth of understanding of the semantic properties of the word-space representations?

As an example, if w_1 and w_2 never occur in the same documents, but sometimes in the same documents as w_3, they will get similar vectors after the application of truncated SVD.

To summarize, LSA uses

- a words-by-documents matrix;
- entropy-based weighting of the co-occurrences, e.g. according to the formula (Dumais, 1993):

$$f_{ij} = (\log(\mathrm{TF}_{ij}) + 1) \times (1 - \left(\sum_j \frac{p_{ij} \log p_{ij}}{\log D} \right))$$

  where D is the total number of documents in the collection, $p_{ij} = \frac{\mathrm{TF}_{ij}}{f_i}$ and f_i is the frequency of term i in the whole document collection;
- truncated SVD to reduce and to restructure its dimensionality;
- the cosine measure to compute vector similarities.

4.5 Hyperspace Analogue to Language

An inherently different word-space implementation is the Hyperspace Analogue to Language (HAL) (Lund et al., 1995), which, in contrast to LSA, was developed specifically for word-space research, and was explicitly influenced by Schütze's paper from 1992.

HAL uses a words-by-words co-occurrence matrix, which is populated by counting word co-occurrences within a directional context window 10 words wide. The co-occurrences are weighted with the distance between the words, so that words that occur next to each other get the highest weight, and words that occur on opposite sides of the context window get the lowest weight. The result of this operation is a directional co-occurrence matrix in which the rows and the columns represent co-occurrence counts in different directions. We have already seen an example of such a directional co-occurrence matrix in Table 3.5 in the previous chapter.

Each row–column pair (i.e. the left and right-context co-occurrences) is then concatenated to produce a very-high-dimensional context vector, which has a dimensionality two times the size of the vocabulary. If such very-high-dimensional vectors prove to be too costly to handle, HAL reduces their dimensionality by computing the variances of the row and column vectors for each word, and discarding the elements with lowest variance, leaving only the 100 to 200 most variant vector elements. It should be noted that this dimensionality-reduction step is not

essential to HAL; it is only used whenever computational efficiency becomes an issue (Burgess et al., 1998). Thus, HAL uses

- a directional words-by-words matrix;
- distance weighting of the co-occurrences;
- concatenation of row–column pairs;
- *if necessary*: discarding low-variant dimensions;
- normalization of the vectors to unit length;
- a Minkowski metric to compute vector similarities.

4.6 Random Indexing

Yet another inherently different word-space implementation is Random Indexing (RI) (Kanerva et al., 2000; Karlgren & Sahlgren, 2001; Sahlgren, 2005), which was developed at the Swedish Institute of Computer Science (SICS) based on Pentti Kanerva's work on sparse distributed memory (Kanerva, 1988). RI is motivated first and foremost by the problem of high dimensionality in other word-space implementations. As we have seen, while dimensionality reduction *does* make the resulting lower-dimensional context vectors easier to compute with, it *does not* solve the problem of initially having to collect a potentially huge co-occurrence matrix. Even implementations that use powerful dimensionality reduction, such as SVD, need to initially collect the words-by-documents or words-by-words co-occurrence matrix. RI targets the problem at its source, and removes the need for the huge co-occurrence matrix.

Furthermore, RI represents a novel way of conceptualizing the construction of context vectors. Instead of first collecting co-occurrences in a co-occurrence matrix and then extracting context vectors from it, RI *incrementally accumulates* context vectors, which can then, if needed, be assembled into a co-occurrence matrix. This methodology can be used to assemble both a words-by-documents and a words-by-words co-occurrence matrix. The ability to use both types of contexts is another thing that makes RI unique in word-space research.

RI accumulates context vectors in a two-step operation:

1. Each context (i.e. each document or each word type) in the text is assigned a unique and randomly generated representation called an *index vector*. In RI, these index vectors are sparse, high-dimensional, and ternary, which means that their dimensionality r is on the order of thousands, and that they consist

of a small number (ϵ) of randomly distributed non-zero elements (as many +1s as –1s). Each word also has an initially empty context vector of the same dimensionality r as the index vectors.

2. The context vectors are then accumulated by advancing through the text one word token at a time, and adding the context's (the surrounding word types' or the current document's) r-dimensional index vector(s) to the word's r-dimensional context vector. When the entire data has been processed, the r-dimensional context vectors are effectively the sum of the words' contexts.

If we then want to construct the equivalent of a co-occurrence matrix, we can simply collect the r-dimensional context vectors into a matrix of order $w \times r$, where w (as before) is the number of unique word types, and r is the chosen dimensionality of the vectors. Note that the dimensions in the RI vectors are randomly chosen, and thus do not represent any kind of context (which is the case in the original co-occurrence matrix) — they constitute a *distributed* representation. Furthermore, r is chosen to be much smaller than the size of the vocabulary and the number of documents in the data, which means that RI will accumulate (roughly) the same information in the $w \times r$ matrix as other word-space implementations collect in the $w \times w$ or $w \times d$ co-occurrence matrices, but that $r \ll d, w$.

The methodology described above can also be used to produce a words-by-words or a words-by-documents co-occurrence matrix by using *unary* index vectors of the same dimensionality n as the number of contexts. These unary index vectors have a single 1 in a different position for each context.[8] Mathematically, these n-dimensional unary vectors are orthogonal, whereas the r-dimensional random index vectors are only *nearly* orthogonal. This near-orthogonality of the random index vectors is they key to the RI methodology. Since there are many more nearly orthogonal than truly orthogonal directions in a high-dimensional space (Kaski, 1999), choosing random directions — as we do when generating the index vectors — can get us very close to orthogonality. This means that the r-dimensional random index vectors can be seen as *approximations* of the n-dimensional unary vectors. Consequently, the r-dimensional context vectors produced by RI can be interpreted as approximations, in the sense that their mutual similarities are (nearly) equal, of the n-dimensional context vectors extracted from the co-occurrence matrix.

The near-orthogonality of random directions in high-dimensional spaces is exploited by a number of dimensionality-reduction techniques that includes methods such as Random Projection (Papadimitriou et al., 1998), Random Mapping (Kaski,

[8] As an example, imagine we want to assemble context vectors equivalent to those in a words-by-words co-occurrence matrix. Also imagine we have 1 000 word types in our data. Then the first word type would have a 1 000-dimensional index vector with a single 1 in the first position, the second word type would have a single 1 in the second position, and so on.

1999), and Random Indexing. These methods rest on the same insight — the Johnson–Lindenstrauss lemma (Johnson & Lindenstrauss, 1984), which states that if we project points in a vector space into a randomly selected subspace of sufficiently high dimensionality, the distances between the points are approximately preserved. Thus, the dimensionality of a given matrix F can be reduced to F' by multiplying it with (or projecting it through) a random matrix R:

$$F'_{w\times r} = F_{w\times d} R_{d\times r}$$

Obviously, the choice of the random matrix R is an important design decision for dimensionality-reduction techniques that rely on the Johnson–Lindenstrauss lemma. As we saw above, if the d random vectors in matrix R are orthogonal, so that $R^T R = I$, then $F' = F$. If the random vectors are nearly orthogonal, then $F' \approx F$ in terms of the similarity of their rows. RI uses the following distribution for the elements of the random index vectors:

$$r_{ij} = \begin{cases} +1 & \text{with probability } \frac{\epsilon/2}{r} \\ 0 & \text{with probability } \frac{r-\epsilon}{r} \\ -1 & \text{with probability } \frac{\epsilon/2}{r} \end{cases}$$

where r is the dimensionality of the vectors, and ϵ is the number of non-zero elements in the random index vectors.

To sum up this section, RI is unique in the following four ways:

1. It is incremental, which means that the context vectors can be used for similarity computations even after just a few examples have been encountered. By contrast, other word-space implementations, by and large, require the entire data to be sampled and represented in the co-occurrence matrix before similarity computations can be performed.

2. It uses fixed dimensionality, which means that new data do not increase the dimensionality of the vectors. Increasing dimensionality can lead to significant scalability problems in other word-space implementations.

3. It uses implicit dimensionality reduction, since the fixed dimensionality is much lower than the number of contexts in the data. This leads to a significant gain in processing time and memory consumption as compared to word-space implementations that employ computationally expensive dimensionality-reduction techniques. As an example, the complexity of computing an SVD is on the order of $O(wzd)$ (under the assumption that the data are sparse), where w is the size of the vocabulary, d is the number of documents, and z is the number of non-zero elements per column (Papadimitriou

et al., 1998). Performing a random projection of the original (sparse) data — i.e. forming a $w \times r$ random matrix and projecting the original $w \times d$ matrix through it — is $O(zrw)$ (Papadimitriou et al., 1998; Bingham & Mannila, 2001), where r is the dimensionality of the vectors. By contrast, producing context vectors with RI is only $O(wr)$, since the method is not reliant on the initial construction of the co-occurrence matrix.

4. It is comparably robust with regards to the choice of parameters. Other word-space implementations, such as LSA, are very sensitive to the choice of dimensionality for the reduced space. For RI, the choice of dimensionality is a trade-off between efficiency and performance (random projection techniques perform better the closer the dimensionality of the vectors is to the number of contexts in the data (Kaski, 1999; Bingham & Mannila, 2001)). In previous experiments, we have shown that the performance of RI reaches a stable level when the dimensionality of the vectors become sufficiently large (Sahlgren & Cöster, 2004; Sahlgren & Karlgren, 2005a).

Part II

Setting the scene

Chapter 5

Evaluating word spaces

"I typed 'pathetic clown' into a search engine, and your name popped right up!"
(Small girl to Krusty the Clown in "Insane Clown Poppy")

In the two previous chapters, I explained how to build a word space. I first introduced the general word-space algorithms, and the central notion of a context vector. I then discussed a couple of problems with the word-space methodology, and identified dimensionality reduction as the solution to these problems. I also reviewed three inherently different implementations of the word-space model: LSA, HAL and RI.

Now, imagine that we have produced a word space by using any of the implementations described in the previous chapter. How can we determine whether it is a *good* word space? How can we evaluate it? Do we even know *what* it is we should evaluate? The notion of evaluation is the subject matter of this chapter.

5.1 Reliability

As we saw in Chapter 4, a central concern in word-space research is how to go from the original context space to a reduced, more compact, representation. We saw that dimensionality reduction can be performed in a number of different ways, and that it typically involves a number of parameters that have to be optimized in order to arrive at a feasible approximation. We also saw that the implementations differ with regards to which weighting scheme, which type of context, and which similarity metric they use. Obviously, each of these factors influence the performance of the resulting word space, so parameter optimization is a main focus in word-space evaluations. This typically means optimizing the mentioned factors with regards to the performance of the word space in some specific experiment.

The overall important property of such experiments is *reliability*, and concerns the question how consistent the experiment is. Does it produce roughly the same result when repeated, or is there a large variation in the results? As an example, consider an IQ test in which a test subject's results fluctuates between moronic and genius level. Would we say that such a test is useful? Probably not. It is difficult, if not impossible, to draw any conclusions from comparisons using a test whose results are inherently inconsistent.

In the following sections, I provide a brief overview over the most common evaluation schemes in current word-space research. I begin with a somewhat terse presentation of my own previous experiments (Sections 5.2 through 5.4), before turning to a very brief and simplified summary of other evaluations of word spaces (Section 5.5).

5.2 Bilingual lexicon acquisition

In Sahlgren (2004) and Sahlgren & Karlgren (2005a), we used RI to acquire bilingual lexica from aligned parallel data. This was done by first assigning one index vector to each alignment in the parallelized bilingual data. In our experiments, we used document-aligned data. This means that the corresponding documents (i.e. the alignments) in both languages were represented by the same index vector. Context vectors were then produced by adding a document's index vector to the context vector for a word every time the word occurred in that document. The result was that if two words in opposite languages occurred in exactly the same aligned documents, their context vectors became identical. By the same token, if two words in opposite languages occurred in *mostly* the same aligned documents, their context vectors became *similar*.

We could then extract the h most similar words in the opposite language to any word in the data. This was done by computing the similarity between the context vector of a given word, and the context vectors of *all* words in the opposite language. By discarding all but the h words whose context vectors were most similar to that of the given word in the other language, we effectively compiled a bilingual lexicon.

We applied this methodology to English–German and Swedish–Spanish parallel data, and evaluated the resulting bilingual lexica by comparing them to online lexical resources. For every word in our data that was also found in the online lexical resources, we counted the number of words that were listed as translations in both the acquired bilingual lexica and in the online lexical resource. Around 60% of the translations in the acquired bilingual lexica were also listed in the online lexical resources. This result reflected the need for linguistic preprocessing,[1] and

[1]Plausible translations were rejected because of morphological and orthographical variation.

the importance of using topically compatible data and lexica. More than anything, these experiments demonstrated the need for devising more pertinent evaluation methodologies for word-space research.

5.3 Query expansion

In a number of experiments (Sahlgren & Karlgren, 2002; Sahlgren et al., 2003, 2002; Karlgren, Sahlgren, Järvinen, & Cöster, 2005; Karlgren, Sahlgren, & Cöster, 2005), we have used RI for both monolingual and bilingual query expansion, for bilingual query translation, and for monolingual query-word selection.

Query expansion is the adding of words to a query in order to improve its recall or precision (depending on how the added words are used by the retrieval engine). When using word-space methodology for this purpose, we first select a number of query words. Then, for each query word, we extract the h most similar words (possibly in another language) from the word space, and add these words to the query. When using word-space methodology for query translation, we use a bilingual word space produced from parallel data, as described in Section 5.2 above. For each query word, we then we extract only the *most* similar word in the other language, and use this as translation. Finally, we use the word-space methodology for query-word selection by first selecting a small number of salient query words, and then adding words to the query based on how similar their context vectors are to the context vectors of the selected query words.

We have used word-space methodology in monolingual settings for query expansion in Swedish, French, and Italian (Sahlgren et al., 2003), and also for French query word selection (Karlgren, Sahlgren, & Cöster, 2005). In bilingual settings, we have used word-space methodology for query expansion and query translation using French–English and Swedish–English (Sahlgren & Karlgren, 2002), English–Japanese (Sahlgren et al., 2002), and Swedish–French (Karlgren, Sahlgren, Järvinen, & Cöster, 2005) data. The results have ranged from "reasonably good" (Karlgren, Sahlgren, Järvinen, & Cöster, 2005; Karlgren, Sahlgren, & Cöster, 2005), over "decidedly mixed" (Sahlgren & Karlgren, 2002; Sahlgren et al., 2003) to "downright catastrophic" (Sahlgren et al., 2002). However, the performance bottleneck has not so much been the word-space methodology, as it has been an effect of poor linguistic preprocessing and failure to properly tune the baseline retrieval engines.[2]

The admittedly mixed results in using word spaces for query expansion clearly

[2]Information-retrieval systems are sensitive to a number of factors, including morphological variation, compounding phenomena, and the scoring function used by the retrieval engine. In some of the mentioned experiments, we did not handle these factors, and did not optimize the scoring function. Such neglects gravely affect the retrieval performance.

demonstrate that we have yet to understand how to properly utilize the word-space methodology in information-access applications. I submit the hypothesis that this impasse is due to the fact that we do not (yet) have a clear picture of what kind of semantic information is captured by the word-space model.

5.4 Text categorization

Text categorization is the task of assigning a given text to one or more of a set of predefined categories. In Sahlgren & Cöster (2004), we used word-space representations — what we called *bag-of-concepts* (BoC) — to improve the performance of a Support Vector Machine (SVM) (Vapnik, 1995) classifier.

The idea was to represent documents in the Reuters-21578 test collection, not as the sum of the words in the documents as is normally done, but as the sum of the *context vectors* of the words in the documents. We then used the resulting TFIDF-weighted vectors as input to the SVM classifier. The categorization results were compared to those reached using standard *bag-of-words* (BoW) representations.[3] The results were that the BoW reached a slightly better result than the BoC over all 90 categories (82.77% vs. 82.29%), but the situation was reversed when only counting the ten largest categories (88.09% vs. 88.74%). We also demonstrated how the performance could be improved from 82.77% to 83.91% over all 90 categories, and from 88.74% to 88.99% over the ten largest categories, by combining the BoW and BoC representations.

Although this experiment showed that the performance of an SVM classifier can be improved by using word-space representations, it is difficult to determine the reason for this improvement. We would of course like to believe that the reason is the added semantic content of the BoC representations, but it could just as well be a purely statistical effect due to the denser nature of the BoC vectors. This kind of experiment simply involves too many factors to allow us to audibly single out the reason for the impact of the word-space representations.

5.5 Compacting 15 years of research

It would neither make much sense, nor be practically feasible, to review every single evaluation of word spaces from the last 15 years. Instead, I will summarize the last 15 years of word-space evaluation by identifying five main evaluation schemes.[4]

[3] Bag-of-words refers to a representation of a text in which all syntactic and grammatical structure has been discarded, and only the mere word tokens are retained. Sort of like pouring all the words in a text into a bag, and then shaking it.

[4] A more thorough survey specifically focused on evaluations and applications of LSA is Dumais (2004).

- **Information retrieval** (Deerwester et al., 1988, 1990; Gallant, 1991a; Dumais, 1993; Caid et al., 1995; Schütze & Pedersen, 1995; Dumais et al., 1997; Schütze & Pedersen, 1997; Jiang & Littman, 2000). These tests normally use an information-retrieval test collection, and compare the performance of a search engine using standard bag-of-words representations, and the word-space representations. As I explained in Section 4.4, the idea is that the word-space representations can be used to alleviate problems with synonymy in the retrieval phase.

- **Synonym tests** (Landauer & Dumais, 1997; Levy et al., 1998; Karlgren & Sahlgren, 2001; Turney, 2001; Rapp, 2003). These tests are normally designed as multiple-choice tests, in which the task is to select the synonym to a given word out of a number of provided alternatives. When using the word-space model to solve this task, vector similarities between the context vector of the alternatives and the context vector of the given word are computed, and the most similar alternative is chosen as synonym. I will discuss this kind of evaluation scheme thoroughly in Chapter 12.

- **Word-sense disambiguation** (Gallant, 1991b; Schütze, 1992, 1993, 1998). In these evaluations, disambiguated data is used to build context vectors for each of the different senses of a word. These "sense vectors" are then used for disambiguation by comparing them to "occurrence vectors" produced for each unique occurrence of an ambiguous word. The sense whose sense vector is most similar to the occurrence vector is chosen as the sense for that occurrence.

- **Lexical priming data** (Lund et al., 1995; Lund & Burgess, 1996; McDonald & Lowe, 1998; Lowe, 2000). The idea here is to compare distances between context vectors in word space to reaction times collected in lexical decision tasks. In such tasks, human subjects are asked to decide, as quickly as possible, whether a given string of letters is a word or not. The decision is facilitated when the subjects are exposed to a semantically related word (called a *prime*) before the decision is made. In psychology, lexical priming effects are seen as evidence for the strength between concepts in human semantic memory. In word-space research, a high correlation between distances in word space and the reaction times is taken to indicate that the word space approximates human semantic memory.

- **Knowledge assessment** (Wolfe et al., 1998; Rehder et al., 1998). The general idea when using word spaces for knowledge assessment is to produce text vectors by summing the context vectors of the constituent words, and then using these text vectors to compute similarity between a "target" text

and a test text on some subject. This general methodology can be used for a number of knowledge-assessment tasks. For example, student essays can be automatically graded by comparing them to "gold standard" texts written by experts, to capture a student's level of background knowledge on some subject, or to match students with appropriate instructional texts.

5.6 Rethinking evaluation: validity

Having reviewed the most common evaluation schemes in word-space research, we can note that many of them are not so much evaluations as *applications* of word spaces. In many of these experiments, it is not primarily the *semantic properties* of the word space that is in focus, but rather the question whether the word space *can be used* for some particular application. Even if parameter optimization is an integral part of such experiments, they do not focus on how the parameters influence the semantic properties of the word spaces.

Optimizing the performance of the word-space model with regards to some particular task is not the focus of interest here. Rather, it is the question whether the word-space model constitutes a viable computational model of meaning that is the focus of interest in this dissertation. In what sense is it *meaning* we acquire and represent in the word space, and which parameters determine this information? Evaluating whether, and in what way, the word space is a viable model of meaning requires not only reliable tests, but *valid* ones.

Validity addresses the question whether the test measures what it is supposed to measure. Again, consider the IQ test: does that test really measure intelligence, or is it something else that we measure when using it? Without venturing into this highly explosive territory, we can at least maintain that the validity of the IQ test has been subject to much heated debate.

So how do we measure the semantic proficiency of the word-space model? How do we decide if a given word space is a viable representation of meaning? Consider the evaluation schemes mentioned above: do they measure this? Surely, some of them do, to some extent. But none of them licenses the conclusion that a word space is a pertinent representation of *all* aspects of meaning. Of course, it would not make much sense to claim that our tests should measure meaning in general, since we obviously cannot (or at least thus far have successfully avoided to) provide a definite answer to the question what meaning *is*.

The inability (or hesitation) to define "meaning" is the source of the problem with defining pertinent evaluation methodologies for measuring the semantic proficiency of the word-space model. If we want to measure meaning, we have better first provide a theory of what meaning is, since it is hopelessly futile to try to measure something without knowing *what* to measure. Note that the distributional

hypothesis is *not* a theory of meaning: it is a *discovery procedure* for excavating meaning similarities, and it does not make any ontological commitments about the nature of meaning. What we have is a theory of representation and a theory of acquisition, but we have *no* theory of meaning.

The lesson in this chapter is that we should be very cautious with making claims about the semantic nature of the word-space representations based solely on empirical evidence. Unless we can base our claims on a theory of meaning, such interpretations require a considerable leap of faith. The point is that there can be no evidence unless there is a case: *if we do not have a hypothesis, we have nothing to verify.*

Chapter 6

Rethinking the distributional hypothesis

"This can't be what it looks like! There's gotta be some other explanation."
(Bart Simpson in "Bart of Darkness")

The last chapter concluded that we need a theory of meaning in order to determine whether the word-space model is a viable computational model of meaning. Now, the distributional hypothesis, as motivated by the works of Zellig Harris, is a strong methodological claim with a weak semantic foundation. It states that *differences* of meaning correlate with *differences* of distribution, but it neither specifies *what kind* of distributional information we should look for, nor *what kind* of meaning differences it mediates. This does not necessarily mean that the use of the distributional methodology as discovery procedure in the word-space model is not well motivated by Harris' distributional approach. On the contrary, it is; but if we want to uncover the *nature* of the differences, we need to thoroughly understand the differential view on meaning.

6.1 The origin of differences: Saussure

The differential view on meaning that Harris assumes in his distributional methodology does not originate in his theories. Rather, it is a consequence of his theoretical ancestry. Although Harris' primary source of inspiration was American structuralist Leonard Bloomfield, the origin of the differential view on meaning goes back even further, to Swiss linguist Ferdinand de Saussure. Saussure (1857–1913) was one of the fathers of modern linguistics, and an important inspiration for the Bloomfieldian structuralism that developed into the distributional paradigm.

Saussure's posthumously published collection of lecture notes — *Cours de linguistique générale* — is widely regarded as one of the most influential works throughout the history of linguistics. It was his ideas that established the structural study of language, and gave rise to the sciences of structuralism and semiotics.

It should be noted that Saussure himself never used the term "structuralism" in the *Cours de linguistique générale*.[1] Saussure-exegesis is complicated by the fact that the primary source of information (the *Cours*) consists of a collection of lecture notes that have been edited and published posthumously. Attributing a view to the Saussure of the *Cours* is not the same thing as attributing a view to the putative theorist behind the actual lectures. I am well aware of this concern. When I speak of Saussure, I speak of the ideas presented in the *Cours*, and of the structuralism they gave rise to.

Saussure saw language as a system of signs, where a sign is an inseparable dichotomy between a *signifier* and a *signified*. The signifier is an "image acoustique" (normally translated as "sound image" or "sound pattern"), which should be understood as the psychological impression of the sound, and not as an actual physical sound. The signified is the psychological concept related to the sound impression. The dyadic nature of the sign is depicted in Figure 6.1.

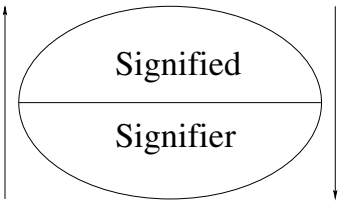

Figure 6.1: The Saussurian sign.

The relation between signifier and signified is purely arbitrary, in the sense that there is no natural or internal connection between the concept and the sound impression. As an example, nothing in the sequence /dog/ (or, to prove the point, /hund/ in Swedish, /gǒu/ in Mandarin, or /žwala/ in Chechen) motivates the concept of a dog, and vice versa. The arbitrariness of the sign is a very important principle in structuralist theory, since it guarantees that no extralinguistic factors influence the constitution of signs. Without it, structuralism could never succeed.

An unfortunate side effect of the arbitrariness principle is that it makes it seem as if we all could have different concepts associated with different sound impressions. If nothing in the sequence /dog/ determines what concepts I connect with it, then there is nothing that guarantees that my concept of a dog is not the same as your concept of a cat. This is a notorious philosophical nightmare that

[1] In fact, as Harris (2001) points out, the term "structure" only occurs three times in the *Cours*, and never in a structuralist sense.

would render communication extremely difficult, if not impossible. Saussure has an ingenious solution to this potential problem.

According to Saussure (1916/1983), the association between signifier and signified can only be established by a linguistic community as a matter of *convention*:

> ...the arbitrary nature of the sign enables us to understand more easily why it needs social activity to create a linguistic system. (p.157/111)

Even though Saussure saw language as an essentially *social* phenomenon, he was not interested in questions about its social foundation, and about *how* the association between signifier and signified gets established by the linguistic community. Such questions, he argued, belong to the field of *diachronic* linguistics, while he was mainly concerned with *synchronic* linguistics.[2] Once the connection between signifier and signified has been fixed by convention, the sign finds its role within the language system. It is then up to the structuralist to describe how the sign *functions* within the language system.

The structuralist is only interested in the *structure* of language (hence the term "structuralism"), and not in the actual *use* of language. Saussure believed that the object of study for linguistics should be the abstract principles of language as a system, referred to as *langue* (in English "language"), since they are constitutive for any individual utterance, referred to as *parole* (in English "speech"). Saussure illustrated the idea using chess as an analogy. Chess is defined by the rules of the game together with the pieces and the board. Individual moves and actual games of chess are only interesting to the participants, and are not essential to (and may even obscure) the definition of the game. In the same manner, individual utterances are certainly interesting to the language users, but are not essential for (and may even obscure) the description of the language system.

To continue the chess analogy, the individual pieces of the game are identified by their functional differences; the king moves one step at a time in any direction, while bishops move diagonally as many steps as desired. Similarly in la langue, signs are identified by their functional differences. Saussure uses the term *valeur* (in English "value") to describe the function of a sign. This is arguably the most important concept in structuralist theory, since it is a sign's valeur that defines its role within the language system. Valeurs are defined purely differentially, so that a sign has a valeur only by virtue of being *different* from the other signs. Such a differential view on the functional distinctiveness of linguistic elements highlights the importance of the system as a whole, since differences (i.e. valeurs) cannot exist in isolation from the system itself. A single isolated sign cannot enter into difference relations, since there are no other signs to differ from. In this view, the system itself becomes an interplay of functional differences:

[2]Diachronic (or historical) linguistics studies language over a large period of time, while synchronic linguistics studies language at a given point in time.

> In the language itself, there are only differences. (Saussure, 1916/1983, p.166/118)

6.2 Syntagma and paradigm

The concept of valeur corresponds to the idea of a thoroughly linguistic aspect of meaning. Consider the difference between the French word "mouton" and the English word "sheep." These words may be said to have the same extralinguistic (i.e. referential) meaning, but they do not have the same valeur, since English makes a distinction between "mutton" and "sheep" that is not available in French. Thus, the functional differences between the signs within la langue is the key to the idea of linguistic meaning, and Saussure divides these functional differences into two kinds: *syntagmatic* and *paradigmatic* relations.[3]

Syntagmatic relations concern positioning, and relate entities that co-occur in the text; it is a relation *in praesentia*. This relation is a linear one, and applies to linguistic entities that occur in sequential combinations. One example is words that occur in a sequence, as in a normal sentence like "I am hungry." Syntagmatic relations are combinatorial relations, which means that words that enter into such relations can be combined with each other. A *syntagm* is such an ordered combination of linguistic entities. For example, written words are syntagms of letters, sentences are syntagms of words, and paragraphs are syntagms of sentences.

Paradigmatic relations, on the other hand, concern substitution, and relates entities that do *not* co-occur in the text; it is a relation *in absentia*. Paradigmatic relations hold between linguistic entities that occur in the same context but not at the same time, like the words "hungry" and "thirsty" in the sentence "I am [hungry|thirsty]". Paradigmatic relations are substitutional relations, which means that linguistic entities have a paradigmatic relation when the choice of one excludes the choice of another. A *paradigm* is thus a set of such substitutable entities.

The syntagmatic and paradigmatic relations are usually depicted as orthogonal axes in a 2-dimensional grid:

[3]The word "syntagmatic" (French "syntagmatique") comes from the Greek word "suntagmatikos," which means *arranged*, or *put in order*. The word "paradigmatic" (French "paradigmatique") comes from the Greek word "paradeigmatikos," which means *serving as a model*, from the Greek word "paradeigma," which means *example*. It should be noted that Saussure himself never used the word "paradigmatique." According to Harris (2001), it was Hjelmslev who coined the term *paradigmatic relations* as a substitute for Saussure's "rapports associatifs." I use the term "paradigmatic" because it is now the accepted term in structuralist theory.

	Paradigmatic relations Selections: "x or y or..."			
Syntagmatic relations Combinations: "x and y and..."	she	buys	green	paint
	he	eats	blue	clay
	they	paint	red	colour

6.3 A Saussurian refinement

The Saussurian notion of valeur as functional difference along the syntagmatic and paradigmatic axes is the origin of the differential view on meaning so prevalent in structuralist theories. Although Harris was arguably more directly influenced by the works of Bloomfield than of Saussure, the latter's structuralist legacy is foundational for both Bloomfield's and Harris' theories, and the differential view on meaning is decidedly foundational for the distributional hypothesis. Armed with this new-found theoretical insight and terminology, we may answer the questions from the beginning of this chapter: *what kind* of distributional information should we look for, and *what kind* of meaning differences does it mediate?

A Saussurian refinement of the distributional hypothesis not only clarifies the semantic pretensions of the word-space model, but it also elucidates the distributional methodology. As we have seen in this chapter, words have a syntagmatic relation if they co-occur, and a paradigmatic relation if they share neighbors. Thus, we should be able to populate the word-space model with syntagmatic relations if we collect information about which words tend to co-occur, and with paradigmatic relations if we collect information about which words tend to share neighbors. Instead of talking about unqualified semantic similarities mediated by unspecified distributional patterns, we can now state concisely that:

> **The refined distributional hypothesis***: A word-space model accumulated from co-occurrence information contains syntagmatic relations between words, while a word-space model accumulated from information about shared neighbors contains paradigmatic relations between words.*

Chapter 7

Syntagmatic and paradigmatic uses of context

"Each letter is as important as the one that preceded it. Maybe more important! No, as important."
(Homer Simpson in "Homer to the Max")

In the last chapter, I excavated the origin of the differential view on meaning, and used the structuralist dichotomy of syntagmatic and paradigmatic relations to refine the distributional hypothesis. A consequence of the reformulation is that it becomes clear that the semantic properties of the word space is determined by the choice of context. This might seem as a trivial statement, but the fact is that very few studies have investigated the effects of using different contexts to generate word spaces. The other word-space parameters, on the other hand, have been meticulously scrutinized. Examples include Nakov et al. (2001), who studied the effects of using different weighting schemes in LSA; Bingham and Mannila (2001), who investigated the effects of using different dimensionality-reduction techniques; and Weeds et al. (2004), who studied the effects of using different similarity measures for computing distributional similarity. By contrast, the impact of using different types of context has been severely neglected. The only previous investigation of the impact of different contexts on the word-space model that I am aware of is by Lavelli et al. (2004), to which I will return in Chapter 15. Out of all the word-space parameters, the choice of context is by far the least studied one.

This is highly remarkable for at least two reasons. Firstly, different word-space algorithms and implementations use different notions of context to assemble the context vectors. As we saw in Chapter 4, LSA uses a words-by-documents matrix, HAL uses a directional words-by-words matrix, and RI can be used to approximate both of these representations. Despite their different uses of context, most word-space implementations claim to arrive at the same kind of meaning representations.

Secondly, it follows from the theoretical analysis of the foundations of the distributional hypothesis in the last chapter that syntagmatic and paradigmatic relations between words should be discoverable by using co-occurrence information and information about shared neighbors, respectively. Thus, a qualitative comparison between these different uses of context should be able to divulge this difference.

Before an empirical investigation of the difference between using co-occurrences and shared neighbors for accumulating context vectors can be performed, we need to know more concisely what these different uses of context entail, what their differences are, and how they can be used to build word spaces. Answering these questions is the subject matter of this chapter.

7.1 Syntagmatic uses of context

As I explained in the last chapter, a syntagmatic relation holds between words that co-occur. The prime example of co-occurrence events is collocations, such as "hermetically sealed," where the first part "hermetically" very seldom occurs without the second part "sealed." Collocations are probably the most obvious examples of syntagmatically related words, because the parts of the collocation tend to occur next to each other, without any intervening words. However, syntagmatically related words can also be defined as words that co-occur within the same text region, with a (possibly large) number of words between them. In the same sense as "co-occurrence" and "matrix" constitute a syntagmatically related word pair in a number of places throughout this dissertation, we could say that *any* two words in this paragraph (or section, or chapter, or part, or even dissertation) constitute a syntagmatically related word pair.

I will refer to the use of co-occurrence events for building the word space as a *syntagmatic use of context*. There is at least one parameter that applies to such syntagmatic use of context:

1. A syntagmatic use of context can be characterized by the size of the context region within which co-occurrences are counted.

7.2 The context region

When constructing context vectors from a syntagmatic use of context, we let the n context regions c in the data define the n dimensions of the word space. Thus, a context vector $\vec{v}$ in a syntagmatic word space has the following constitution:

$$\vec{v} = (c_1, c_2, ..., c_n)$$

As I pointed out in Section 3.5, this representation does not mean anything in vacuo. It is the *similarity* between two context vectors constructed in this way that is of the syntagmatic type.

The context region can be anything from a very small sequence of words to an entire text region. In practice, larger text regions are favored by word-space algorithms. The reason for this seems to be primarily that the word-space algorithms that prefer a syntagmatic use of context, such as LSA, hail from the information-retrieval community, where a document is a natural context of a word. To see why, consider the information-retrieval universe, in which *documents* and *words* are two of the most basic elements. Documents are assumed to represent topical units (and consequently also topical *unities*), whereas words are seen as *topic indicators*, whose distribution is governed by a limited number of *topics*. In the standard type of information retrieval, this is as far as the metaphor goes, and elements of the universe (e.g. queries and documents) are matched based on word overlap, without utilizing the topics. In more sophisticated, LSA-type, information retrieval, the topics constitute the fundamental ontology, and all elements in the universe — such as words and documents — can be grouped according to them.[1] In that way, queries and documents can be matched according to their topicality, without necessarily having to share vocabulary. Note that in both types of information retrieval, documents constitute the natural context of words.

This is a perfectly feasible simplification of textual reality when viewed from an information-retrieval perspective. However, information retrieval is an artificial problem, and a "document" in the sense of a topical unit–unity is an artificial notion that hardly exists elsewhere; before the advent of library science, the idea that the content of a text could be expressed with a few index terms must have seemed more or less appalling. In the "real" world, content is something we *reason about*, *associate to*, and *compare*. The uncritical assimilation by word-space algorithms of the information-retrieval community's conception of context is unfortunate, since the simplification is uncalled for, and may even be harmful, outside the information-retrieval universe. In the world beyond information-retrieval test collections (which tend to consist of text types for which the metaphor actually makes sense, such as short newswire articles or downloaded web pages), text (and, in the big picture, language) is a continuous flow where topics intertwine and overlap. In this complex structure, finding a correlate to the information-retrieval notion of a "document" is at best an arbitrary choice. As Ruge (1992) notes:

[1] Proponents of factor-analytic dimensionality-reduction techniques tend to argue that these techniques uncover the topics in the form of latent dimensions, or independent or principal components (Bingham, 2003).

> Inside of a large context (e.g. a whole document) there are lots of terms not semantically compatible. In large contexts nearly every term can co-occur with every other; thus this must not mean anything for their semantic properties. (p.318)

So what would be a more linguistically justified definition of context in which to collect syntagmatic information? Perhaps a clause or a sentence, since they seem to be linguistic universals; clauses and sentences, or at least the functional equivalent to such entities (i.e. some sequence delimited by some kind of delimiter), seem to exist in every language — spoken as well as written or signalled. Thus, it would be possible to argue for its apparent linguistic reality as context. Sentences have been used to harvest co-occurrences by, e.g., Rubenstein & Goodenough (1965), Miller & Charles (1991), and Leacock et al. (1996).

Another possibility would be to use a smaller context region consisting of only a couple of consecutive words, as in the example with collocations. However, a serious problem with using such a small context to collect syntagmatic information is that very few words — basically only collocations — co-occur often within a small context region. In fact, as, e.g., Picard (1999) points out, the majority of terms *never* co-occur. The smaller the context regions are that we use to collect syntagmatic information, the poorer the statistical foundation will be, and consequently the worse the sparse-data problem will be for the resulting word space.

7.3 Paradigmatic uses of context

Paradigmatically related words are words that *do not themselves* co-occur, but whose surrounding words are often the same. One example of such paradigmatically related words is different adjectives that modify the same nouns — e.g. "bad" and "good" in "bad news," "good news." As with syntagmatically related words, paradigmatic relations need not only consist of words that share the same immediately preceding or succeeding neighbor or neighbors. The paradigmatic relation may just as well be defined as words that share some, or several, of the s preceding or succeeding neighbors.

I will refer to the use of surrounding words for building the word space as a *paradigmatic use of context*. There are at least three parameters that apply to the characterization of such paradigmatic uses of context:

1. The size of the context region within which paradigmatic information is collected.

2. The position of the words within the context region.

3. The direction in which the context region is extended (preceding or succeeding neighbors).

7.4 The context window

When constructing context vectors from a paradigmatic use of context, we let the n word types w in the data define the n dimensions of the word space. Thus, a context vector $\vec{v}$ in a paradigmatic word space has the following constitution:

$$\vec{v} = (w_1, w_2, ..., w_n)$$

Again, this representation is vacuous by itself. It is the *similarity* between two context vectors constructed in this way that is of the paradigmatic type.

The most common way to collect paradigmatic information is to define a context window of some size and extension, within which the information is collected. As an example, imagine the following two highly imaginary word sequences:

```
bla bla bla blo bli bli bli
bla bla bla ble bli bli bli
```

Notice that `blo` and `ble` constitute a paradigmatically related word pair in this example, and that it would suffice to look at the immediately preceding and succeeding words to establish this — what we call a 1+1-sized context window. In the same manner, a 2+2-sized context window would consist of the two preceding and the two succeeding words, and a 5+3-sized window of the five preceding and the three succeeding words, and so on. Naturally, nothing precludes us from using an entire sentence (or, for that matter, an entire text) as context window, but — as we will soon see — most researches favor the use of statically-sized context windows. I will use the term *focus word* to refer to the word in the middle of the context window (i.e. the word whose context we are currently analyzing). The context window is normally advanced one word at a time until the entire data has been processed — a so-called *sliding* context window.

As noted in the previous section, we also need to account for the *position* of the words within the context windows, since paradigmatically related words may also be defined as words that share *some* of the s surrounding words. For example, imagine that the word `bli` in the example above could take any of a number of different modifiers, so that one sequence would be arbitrarily realized as `bla blo e bli`, and the other as `bla ble r bli`. In this case, we would like to exclude the arbitrary modifiers from the context windows. One way to accomplish this is to use a null-weight for that position in the context window, so that the configuration for, e.g., a 1+2-sized window would be `1 + 0 1`, where `0` means that the word is ignored. This would then be realized in the example as:

```
bla blo e bli -> blo:  (bla) + (0 bli)
bla ble r bli -> ble:  (bla) + (0 bli)
```

It is also possible to *weight* the positions in the context window more finely, instead of just using binary values. I will discuss a couple of different weighting schemes for the positions in the context windows in Section 8.5.

The million-euro question regarding context windows is their size: how many words to the left and to the right should we count? There have been many suggestions in the literature. For example, Schütze (1992) uses a window size of 1 000 *characters*, with the argument that a few long words are possibly better than many short words, which tend to be high-frequency function words. Yarowsky (1992) uses 100 words, while Gale et al. (1994) uses 50 words to the left and 50 words to the right, with the argument that this kind of large context is useful for "broad topic classification." Schütze (1998) uses a 50-word window, whereas Schütze & Pedersen (1997) uses a context window spanning 40 words. Niwa & Nitta (1994) uses a 10+10-sized window, and as we saw in Chapter 3, the HAL algorithm uses a directional 10-word window. Black et al. (1988) uses narrow windows spanning 3–6 words, Church & Hanks (1989) used 5 words, and Dagan et al. (1993) uses a window spanning 3 words, when ignoring function words.

As we can see, examples of window sizes range from 100 words to just a couple of words. There is very seldom a theoretical motivation for a particular window size. Rather, the context window is often seen as just another experimentally determinable parameter. Levy et al. (1998) is a good example of this viewpoint:

> These and other technical and practical questions can only be answered by careful and time-consuming experimentation. (p.4 in the offprint.)

There seems to be some evidence for the feasibility of using a fairly small context window. Kaplan (1955) asked people to identify the sense of a polysemous word when they were shown only the words in its immediate vicinity. They were almost always able to determine the sense of the word when shown a string of five words — i.e. a 2+2-sized context window. This experiment has been replicated with the same result by Choueka & Lusignan (1985). Our previous experiments (Karlgren & Sahlgren, 2001) also indicate that a narrow context window is preferable for acquiring semantic information.

Despite these (admittedly somewhat summary) arguments for the feasibility of narrow context windows, we should heed the cautious words of Miller & Leacock (1998):

> ...we still don't understand how to extract an adequate contextual representation from the local context. [...] perhaps we should look at context more broadly. (p.156)

7.5 What is the difference?

For the sake of clarity, I will demonstrate the different uses of context with a concrete example. Consider the following text:

Die Welt ist alles was der Fall ist

Was der Fall ist die Tatsache ist das Bestehen von Sachverhalten

Das logische Bild der Tatsache ist der Gedanke

Figure 7.1: Example text.

Now imagine that we collect this text into two co-occurrence matrices; one from a syntagmatic use of contexts, collected within the given sentences; one from a paradigmatic use of contexts, also collected within the given sentences. I only include content words in this example. The result from the co-occurrence counting is displayed in Tables 7.1 and 7.2.

	c_1	c_2	c_3
Welt	1	0	0
alles	1	0	0
Fall	1	1	0
Tatsache	0	1	1
Bestehen	0	1	0
Sachverhelten	0	1	0
logische	0	0	1
Bild	0	0	1
Gedanke	0	0	1

Table 7.1: Words-by-documents co-occurrence matrix.

Obviously, the matrix constructed from a paradigmatic use of context is much richer in information, which means that it has a more robust statistical foundation. Several researchers have noted this difference between the two approaches; Schütze & Pedersen (1997) argue that a paradigmatic use of context (what they call *lexical* co-occurrences) is both quantitatively (because it provides a better statistical foundation) and qualitatively (because the fact that two words occur close to each other is, they argue, likely to be more significant than the fact that they

	w_1	w_2	w_3	w_4	w_5	w_6	w_7	w_8	w_9
Welt (w_1)	0	1	1	0	0	0	0	0	0
alles (w_2)	1	0	1	0	0	0	0	0	0
Fall (w_3)	1	1	0	1	1	1	0	0	0
Tatsache (w_4)	0	0	1	0	1	1	1	1	1
Bestehen (w_5)	0	0	1	1	0	1	0	0	0
Sachverhelten (w_6)	0	0	1	1	1	0	0	0	0
logische (w_7)	0	0	0	1	0	0	0	1	1
Bild (w_8)	0	0	0	1	0	0	1	0	1
Gedanke (w_9)	0	0	0	1	0	0	1	1	0

Table 7.2: Words-by-words co-occurrence matrix.

occur in the same document) better than a syntagmatic use, and Picard (1999) observes that a syntagmatic use of context can only be used for very frequent words, while a paradigmatic use may be applied for *all* words.

It now becomes obvious why word-space algorithms that prefer a syntagmatic use of context favor statistical dimensionality-reduction techniques such as SVD that smooths the original co-occurrence counts. As we saw in Section 4.4, the application of factor-analytical dimensionality-reduction techniques to the matrix alleviates the problem with data sparseness. Furthermore, it has the effect of grouping together words that do not necessarily co-occur, but that occur in similar contexts — in other words, it *approximates* paradigmatic relations. This means that SVD, and related dimensionality-reduction techniques, can be viewed as "poor man's" paradigmatic relations. This is a very useful tactic in information retrieval, where we are forced to work with a syntagmatic words-by-documents matrix. However, in word-space research, it seems futile *not* to use a paradigmatic words-by-words matrix, if that is the kind of relations one wants to model. As Schütze (1992) notes, it is unnecessary to apply SVD (or any other related dimensionality-reduction technique) when using paradigmatic contexts, since they are already dense, and of course already contain paradigmatic relations between words.[2]

[2]Some researchers, such as Lemaire & Denhière (2004), claim that *only* algorithms that use SVD take advantage of higher order co-occurrences, which they define as relations that hold between words that do not co-occur, but that occur in similar contexts. As we saw in Section 7.3, this essentially equals paradigmatic relations. Lemaire's and Denhière's claim stems from a lack of theoretical underpinnings, and is not compatible with the analysis in this section.

7.6 And what about linguistics?

Any linguist that is exposed to this discussion will no doubt frown upon the lack of linguistic sophistication in the use of contexts. We are, after all, throwing away basically everything we know about language when we are simply counting surface (co-) occurrences. Some word-space paradigms, such as LSA and HAL, are openly agnostic towards linguistics, even though the latter includes at least the bare minimum of linguistic knowledge about word order into the directional co-occurrence matrix by letting the row and column vectors for a word represent co-occurrences from the right and left contexts, respectively. However, remember from Chapter 4 that this directional information is thrown away in the final stages of the algorithm when the row and column vectors are concatenated. The only experiment I am aware of that exploits the directional information in a directional words-by-words co-occurrence matrix is Schütze & Pedersen (1993), to which I will return in Section 15.6.

Considering the two different uses of context discussed in this chapter, a paradigmatic use of context is arguably more linguistically sophisticated than a syntagmatic use of context, since a context window at least captures some rudimentary information about word order. But why pretend that linguistics never existed — why not use linguistic knowledge explicitly?

There have been a few attempts at using linguistic knowledge when creating the word space. In Karlgren & Sahlgren (2001), we increased the performance of RI on a synonym test by using lemmatized data. We also experimented with adding part-of-speech tags to the words, thus performing grammatical disambiguation. However, adding part-of-speech information *decreased* the performance for all sizes of the context window, except when using a minimal 1+1-sized window. Wiemer-Hastings & Zipitria (2001) also noticed a decrease in performance of LSA when they added part-of-speech tags to the words, and Widdows (2003) noted that adding part-of-speech information improves the representation for common nouns, but not for proper nouns or finite present-tense verbs when enriching the WordNet[3] taxonomy. The reason for the decrease in performance is that adding part-of-speech information increases the number of unique words in the data, thus aggravating the sparse-data problem.

A more sophisticated approach to utilizing linguistic information is Padó & Lapata (2003), who uses syntactically parsed data to build contexts that reflect the dependency relations between the words. Their approach is inspired by the works of Strzalkowski (1994) and Lin (1997, 1998a, 1998b), who also used parsed data to compute distributional similarity between words. In Lin's experiments, words were represented by the frequency counts of all their *dependency triplets*. A dependency

[3]http://wordnet.princeton.edu/

triplet consists of two words and the grammatical relationship between them in a sentence, such as `(have subj I)` from the sentence "I have angst." Similarity between words was then defined as (Lin, 1998b):

$$\mathrm{sim}_{\mathrm{LIN}}(w_1, w_2) = \frac{2 \times I(G(w_1) \cap G(w_2))}{I(G(w_1)) + I(G(w_2))}$$

where G_w is the set of features possessed by w, and $I(G)$ is the amount of information contained in a set of features G, calculated as $-\sum_{g \in G} \log P(g)$, where $P(g)$ is the probability of feature g.

Other attempts at using linguistic information for computing distributional similarity between words include Hindle (1990), who used predicate–argument structure to determine the similarity of nouns; Hearst (1992), who extracted hyponyms by using lexical–syntactic templates; Ruge (1992), who used head–modifier relations for extracting similar words; and Grefenstette (1992a, 1992b, 1993), who also used syntactic context to measure similarity between words.

On the downside, such linguistically refined notions of context require a non-negligible amount of preprocessing, and tend to suffer from sparse data (Schütze, 1998). Furthermore, empirical evidence for the supremacy of linguistically refined contexts are still very scarce. Much more research is needed in order to determine the viability of, e.g., dependency relations for building word spaces.

Part III

Foreground

Chapter 8

Experiment setup

"In fact, I made a graph. I make lots of graphs."
(Lisa Simpson in "Homr")

The last chapter explained what syntagmatic and paradigmatic uses of context are, what the differences are between them, and how they can be applied to assemble word spaces. Having clarified both the terminology and methodology, we can now turn to a qualitative empirical comparison between these different uses of contexts. This chapter describes the experiment setup: the data, the preprocessing, the transformation and weighting schemes, the word-space implementation, the software, the tests, and the evaluation metrics.

8.1 Data

The primary data I use in these experiments is the Touchstone Applied Science Associates (TASA) corpus,[1] which consists of high-school level English texts on a number of different topics such as language arts, health, economics, science, social studies, and business. I use this data for two reasons. The first is that the TASA corpus is divided into sections spanning approximately 150 words. This means that I can use both *sentences* and *sections* as context regions for the syntagmatic uses of context. Other corpora, such as the British National Corpus (BNC),[2] is not consistently divided into such topically coherent sections.

The other reason for using the TASA corpus is its convenient size. It is large enough to provide a sound statistical foundation, but small enough to allow for the construction of an *unreduced* word space. For large corpora, such as the BNC,

[1] Kindly provided by courtesy of Thomas Landauer.
[2] http://www.natcorp.ox.ac.uk/

it is necessary to use dimensionality reduction for most experimental settings. I do not want to use dimensionality reduction in these experiments, since it might interfere with the comparison of the different uses of context. In one experiment where there is a limited vocabulary, I also include results from the BNC in order to demonstrate how the corpus size can effect the results. Details of the corpora (after lemmatization) are given in Table 8.1.

Corpus	Size	Tokens	Types	Regions
TASA	56 MB	10 802 187	66 586	$\approx$ 150 words $\approx$ 12 words
BNC	527 MB	106 523 151	317 961	$\approx$ 17 words

Table 8.1: Details of the data sets used in these experiments.

In the remainder of this dissertation, I label the context regions that span approximately 150 words *large* context regions, and the ones that span approximately 12–17 words *small* context regions.

8.2 Preprocessing

I use morphological normalization of the data in these experiments. The main reason for doing so is that all the resources (i.e. tests, thesaurus entries, association norms) used in the following experiments only include base forms of the words. Also, we have demonstrated in previous experiments that the performance of the word-space model benefits from using morphologically normalized data (Karlgren & Sahlgren, 2001). The reason for this is that morphological normalization reduces the number of word types in the data, without affecting the number of word tokens. This at least partially remedies problems with sparse data and improves the statistical foundation of the model. To perform morphological normalization, I use software from Connexor, a Finnish language-technology company that provides parsers and taggers for several different languages.[3]

8.3 Frequency thresholding

Since the word-space model is based on statistical evidence of word distributions, it is imperative that the words we examine have a sufficient statistical foundation. Low-frequency words suffer from sparse data and will not have sufficiently reliable

[3]http://www.connexor.com/

statistics to enable distributional analysis. I therefore use frequency thresholding for words with frequency less than 50 in the following experiments. I depart slightly from this prescript in two experiments. In Chapter 9, I use a frequency threshold of 20 instead, and for the synonym test in Chapter 12 that only involves 400 words, and therefore is much more efficient to execute, I experiment with different frequency thresholds for both low-frequency and high-frequency words.

Frequency thresholding can be done either as a preprocessing step by physically removing the words from the data, or by neglecting the words during processing. I opt for the latter approach, and perform frequency thresholding in the following manner: words that occur outside the stipulated frequency range are not assigned an index vector, and if such non-indexed words occur in the context window, their positions will have weight 0. This way of doing frequency thresholding blanks out positions in the window rather than removing them. It is arguable that an even better approach would be to dynamically modify the extension of the context windows so that they maintain a constant number of word tokens, while still excluding non-indexed words. The advantage with the favored approach is that it is much simpler and much more efficient.

For tests with a specified vocabulary (like the synonym test in Chapter 12 and the antonym test in Chapter 13), the frequency threshold does not apply to the context vectors for the test words. That is, if a test word has a frequency outside the allowed frequency range, it is not assigned an index vector (and thus has no impact on its neighbors' context vectors). However, it still receives a context vector, and can thus be included in the test.

8.4 Transformation of frequency counts

Since I want to minimize the risk of distorting the comparison by using an unfavorable instantiation of any type of context, I use a number of different transformations of the frequency counts when computing the context vectors. By doing so, I optimize the performance of the word spaces with regards to the different types of context in each particular task.

I use four different transformations for the syntagmatic uses of context in these experiments:

- BINARY: the elements of the vectors are either 1 or 0, where 1 indicates that the word has occurred in the context, and 0 that is has not.
- DAMPENED, computed as $\log(\text{TF} + 1)$: the elements are the logarithms of the frequency of occurrence of the word in the context.
- TFIDF, computed as $\log(\text{TFIDF} + 1)$: the elements are the logarithms of the TFIDF-value of the word in the context.

- RAW: the elements are the raw frequency of occurrence of the word in the context.

These transformations are motivated by their widespread use in information-access research (see Section 3.4). Taking the logarithm of the elements in the context vectors is done in order to reduce the effect of high values. The idea is that co-occurrence events that only occur once or twice in the data are less important than co-occurrence events that occur several times, but that it is not important whether they occur 500 or 50 000 times. Note that the value is incremented with one before the logarithm is computed. This is done in order to avoid zero values, for which a logarithm is not defined. In the remainder of the dissertation, I use the following notation for the parameters involved in syntagmatic uses of context:

$$[S : c \in \{+, -\}, a \in \{\text{BINARY,DAMPENED,TFIDF,RAW}\}]$$

where c is the context region ($+$ for large and $-$ for small), and a is the transformation. As an example $[S : +, \text{RAW}]$ means a syntagmatic use of a large context region, with raw frequencies.

8.5 Weighting of context windows

I use two different weighting schemes for the slots in the context windows for the paradigmatic uses of context:

- Constant weighting over the window.
- Aggressive distance weighting according to the formula 2^{1-l} where l is the distance to the focus word.

There are a number of other possibilities for the distance weights. For example, one could imagine using linear distance weighting where the weight decreases with some constant (e.g. 0.1) for every step away from the focus word, or distance weighting according to the formula $\frac{1}{l}$ where l is the distance to the focus word. Figure 8.1 shows the effects of these different weights. Since linear weighting and $\frac{1}{l}$-weighting is intermediate between constant weighting and aggressive weighting, I only use the latter weights in the following experiments.

In the remainder of the dissertation, I use the following notation for the parameters involved in paradigmatic uses of context:

$$[P : s \in \{n + n, n - n\}, b \in \{\text{CONST,AGG}\}]$$

where s is the size of the context window (either the size of one window (e.g. 2+2), the size of several windows (e.g. 3+3,4+4), or the range of several context windows

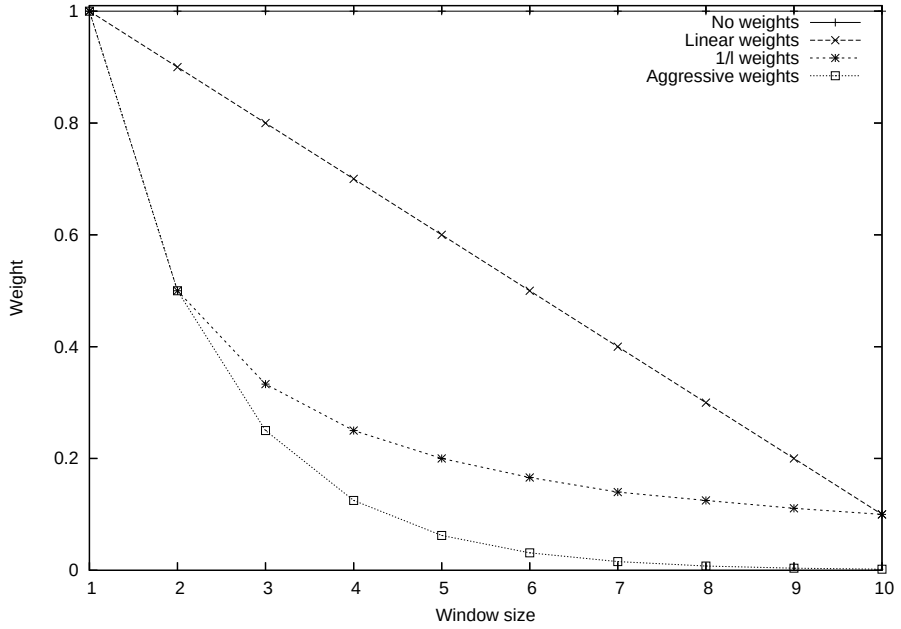

Figure 8.1: Different weighting schemes of the context windows

(e.g. $1-10$, which means that all window sizes between 1+1 to 10+10 are used)), and b is the weighting scheme (CONST for constant weights or AGG for aggressive 2^{1-l} weights). As an example $[P : 2+2, \text{CONST}]$ means a paradigmatic use of a 2+2-sized context window with constant weights.

8.6 Word-space implementation

I do not use any dimensionality reduction of the context vectors in these experiments. The reason, which I already mentioned in Section 8.1 above, is that I do not want any external factors to interfere with the comparison of syntagmatic and paradigmatic uses of context. Thus, I use neither LSA, HAL, nor RI, but the unreduced context vectors from the words-by-documents and words-by-words matrices. For the TASA corpus, the former vectors are 37 619 (for the large context regions) and 784 819-dimensional (for the small context regions), while the latter are 8 217-dimensional (after thresholding low-frequency words). For the BNC corpus, the vectors are 5 794 049 vs. 35 089-dimensional (after frequency thresholding).

8.7 Software

The software used in these experiments is called GSDM (for Guile Sparse Distributed Memory) and was written by Anders Holst at SICS. GSDM is an open source C-library for Guile,[4] designed specifically for the RI methodology. The software provides a number of basic functions for handling texts, vectors, and matrices. All higher-level programming is done in Guile using the basic GSDM functions. GSDM is available under the GNU GPL license.

8.8 Tests

A number of different tests are used in these experiments. Since the point of the experiments is to verify the hypothesis that syntagmatic and paradigmatic uses of context generate different semantic models, I have tried to select tests that measure not only different aspects of linguistic meaning but specifically syntagmatic and paradigmatic relations to different degrees. This has not been an easy task, since syntagmatic and paradigmatic relations are not mutually exclusive. Many words that enter into paradigmatic relations *can* also enter into syntagmatic ones, and conversely. The tests used in these experiments do not provide *exact* measures of these relations.

Most of the following tests use some sort of lexical resource for comparison. One problem with this approach is that freely available lexical resources are scarce, incoherent, and incomplete. It is always an apparent risk when relying on such resources that the resulting evaluation becomes as much an evaluation of the resource itself. We have previously encountered and discussed such problems in Sahlgren & Karlgren (2005a) and Sahlgren (2006). Another problem with lexical resources is that they are typically compiled by humans, who are influenced by extralinguistic factors that are not available to the word-space model. Remember that the word-space model only claims to represent linguistic meaning. In a sense, whatever the word-space model uncovers is the truth, since that is what is *really* there in the data. The linguistic information contained within a corpus such as the TASA does not necessarily have to correspond exactly with general human semantic knowledge. In fact, we would be greatly surprised if it did.

The question how to evaluate models of meaning provides enough material for an entire dissertation in itself. I do not claim to have any definite answer to this question, and I do not intend for these experiments to provide any definite empirical evidence. The following experiments should be viewed more as indications of the viability of the hypothesis. The tests that I use are

[4] http://www.gnu.org/software/guile/guile.html

- direct comparison of the word spaces (Chapter 9);
- thesaurus comparison (Chapter 10);
- association test (Chapter 11);
- synonym test (Chapter 12);
- antonym test (Chapter 13);
- part-of-speech test (Chapter 14).

8.9 Evaluation metrics

The evaluation metrics are somewhat different for the different tests. For the direct comparison, I count overlap between neighborhoods in the word space, as further described in the next chapter. For the synonym test, I report results as percentage of correct answers, as described in Chapter 12.

For each of the four remaining tests, I employ two different evaluation metrics, which I call LAX(h) and STRICT(h) settings. The h defines the number of nearest neighbors extracted from the word space. For example, if $h = 5$, I extract the 5 nearest neighbors from the word space to each word included in the test. The lax setting means looking at whether *any* of the word-space neighbors are listed in the resource (i.e. thesaurus, association norm, antonym list or part-of-speech list) used in these experiments, while the strict setting means looking at whether *all* of the word-space neighbors are listed in the evaluation resource. The results are reported as the percentage of word-space neighbors that are listed in the evaluation resource, defined as:

$$\text{LAX}(h) = 100 \times \frac{\text{\# correct neighbors}}{\text{\# test words}}$$

$$\text{STRICT}(h) = 100 \times \frac{\text{\# correct neighbors}}{\sum h}$$

A couple of examples might help clarify these evaluation metrics. Suppose we use the LAX(1) setting, and that we have 10 test words. This means that we extract 1 nearest neighbor to each of the 10 test words. If 8 of these neighbors are listed in the resource currently used (say, a thesaurus), we will arrive at $100 \times \frac{8}{10} = 80\%$ correct answers. Note that, if we only extract 1 nearest neighbor to each test word, the LAX(1) setting is equivalent to the STRICT(1) setting.

If we instead use LAX(10), we will extract 10 nearest neighbors to each of the 10 test words. However, since we use the lax setting, we are only interested in

whether *any* of these 10 nearest neighbors are listed in the resource. It thus does not matter whether all or 5 or only 1 of the 10 neighbors are listed in the resource. Say that for 9 of the test words, we find that *at least* one of the 10 neighbors (again, we do not care how many) are listed in the resource. We then get $100 \times \frac{9}{10} = 90\%$ correct answers.

Contrary to the LAX(10) setting, if we instead use a STRICT(10) setting, we now count *every* neighbor that is listed in the resource. In the strict setting it *does* matter whether all or 5 or only 1 of the 10 nearest neighbors are listed in the resource. Since we have all together 100 neighbors, our maximum result would be $100 \times \frac{100}{(\sum h)=100} = 100\%$. Say that we only find 26 of all the neighbors in our resource. We then only get $100 \times \frac{26}{(\sum h)=100} = 26\%$ correct answers.

For the thesaurus comparison and the association test, I use LAX(1) and STRICT(h) evaluation, where h is the number of words listed in the association norms and in the thesaurus entries. For the antonym test, which consists of word pairs, I use LAX(10) and STRICT(1). For the part-of-speech test, which does not include a test vocabulary, I use LAX(1) and STRICT(10).

Chapter 9

The overlap between syntagmatic and paradigmatic word spaces

"This is not a clue... or is it?"
(Bart Simpson in "Who shot Mr. Burns?)

The first experiment concerning the difference between word spaces produced with syntagmatic and paradigmatic uses of context is to simply compare the word spaces with each other. However, since mere ocular inspection of word-space neighborhoods is notoriously prone to over-interpretation, it would be useful to define a numeric measurement for the difference between two word spaces.

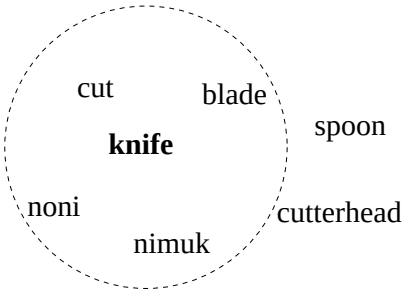

Figure 9.1: Word-space neighborhood produced with $[S : +, \text{TFIDF}]$. The circle indicates what the neighborhood would have looked like if a range around the word "knife" had been used instead of a constant number of neighbors (in this case, 6) to define the neighborhood. "Noni" and "Nimuk" are names occurring in the context of knifes.

One way to arrive at a numeric measurement for the difference (or rather the *commonality*) between two word spaces is to count the overlap, which I define as the number of shared words between neighborhoods in the two spaces. A neighborhood is simply a region in the word space, centered around a given word. Note that such neighborhoods can be defined in two ways: either as a *range* around the word (defined by a similarity threshold), or as a *number* of nearest neighbors. An example of a word-space neighborhood is given in Figure 9.1. In the following experiments, I opt for the latter approach, and define "word-space neighborhood" as a constant number of nearest neighbors to words in the word spaces.

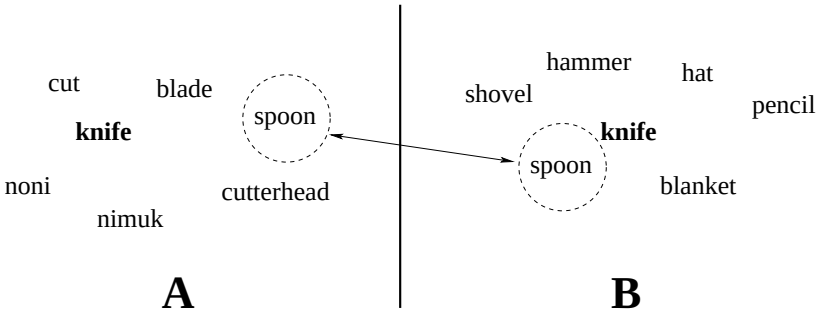

Figure 9.2: Overlap between the neighborhoods of "knife" for $[S : +, \text{TFIDF}]$ (space **A**) and $[P : 2 + 2, \text{CONST}]$ (space **B**).

Figure 9.2 demonstrates the idea of comparing neighborhoods in two word spaces. The overlap between the spaces is defined as the number of neighbors that occur in both spaces. In the example, only one out of six neighbors ("spoon") occurs in both spaces. If there is a large overlap, we can assume that the spaces are very similar; if there is a small overlap, we can assume that they are not. Such an admittedly simple experiment will not tell us in what sense the spaces differ — if at all — but it *will* tell us whether the hypothesis is worth pursuing or not: I clearly have to admit defeat if the neighborhoods contain exactly the same words, and the overlap therefore is 100%, since that would indicate that the word spaces are identical.

9.1 Computing the overlap

In order to compute the overlap between spaces produced with syntagmatic and paradigmatic uses of context, I first construct word spaces using $[S : +, \text{RAW}]$ (i.e. syntagmatic uses of large context regions with raw frequency counts), $[S : -, \text{RAW}]$ (i.e. syntagmatic uses of small context regions with raw frequency counts) and $[P : 1 - 20, \text{CONST}]$ (i.e. paradigmatic uses of context windows ranging from 1+1 to 20+20 with constant distance weights). I then select all words with frequency above 20 in the TASA corpus, which defines a vocabulary of 13 383 words. For each of these words, I extract the 10 nearest neighbors from the respective spaces. I then count how many words occur in both nearest-neighbor sets for each of the 13 383 words, thus arriving at a measure of the overlap between the different word spaces.

Context window	Overlap			
	$[S : +, \text{RAW}]$		$[S : -, \text{RAW}]$	
	10 NNs	1 NN	10 NNs	1 NN
1+1	1.57	1.18	1.74	1.43
2+2	2.56	1.74	2.93	2.14
3+3	3.78	2.61	4.19	3.29
4+4	4.73	3.29	5.13	4.27
5+5	5.58	4.18	5.95	5.06
6+6	6.34	5.12	6.34	5.12
7+7	6.95	5.88	7.18	7.08
8+8	7.58	6.48	7.74	7.75
9+9	8.05	7.09	8.08	8.32
10+10	8.50	7.73	8.44	8.80
15+15	9.89	9.77	9.31	10.27
20+20	10.52	10.74	9.54	10.62
$[S : -, \text{RAW}]$	37.45	29.71	–	–

Table 9.1: Percentage of nearest neighbors that occur in both syntagmatic and paradigmatic word spaces.

Table 9.1 shows the percentage of mutual nearest neighbors when counting all 10 neighbors, and when only counting the closest neighbor in both spaces. The results clearly show that there is a very small overlap between the different word spaces. The overlap is fairly small even for syntagmatic spaces produced with different sizes of the context region; the overlap between $[S : +, \text{RAW}]$ and $[S : -, \text{RAW}]$ is 37.45% when counting the 10 nearest neighbors, and 29.71% when only counting

the closest one. The overlap between $[S : +, \text{RAW}]$ and $[P : 1 - 10, \text{CONST}]$, and between $[S : -, \text{RAW}]$ and $[P : 1 - 10, \text{CONST}]$ is of course even smaller, and is basically the same when counting the 10 nearest neighbors and when only counting the closest neighbor.

There are three additional noteworthy aspects of these results. The first is that there is a very small difference in the overlap between paradigmatic word spaces and syntagmatic spaces produced with large versus small context regions. The largest overlap in these results are between word spaces produced with a wide context window and a large context region.

The second noteworthy aspect of the results is that the overlap counting the 10 nearest neighbors in most cases is a little bit larger than when only counting the closest neighbor. This might not be surprising considering that there are more possibilities for overlap in larger neighborhoods than in very small ones. What *is* surprising, however, is that this relationship between the overlap counting 10 NNs versus only counting 1 NN is reversed for wider context windows when compared to $[S : -, \text{RAW}]$. Starting with an 8+8-sized context window, the overlap to $[S : -, \text{RAW}]$ is larger when only counting the closest neighbor than when counting all 10 nearest neighbors.

The third noteworthy aspect of these results is that the overlap increases with the size of the context windows. This indicates that wider context windows approximate a syntagmatic use of context. However, consider the difference between the overlap counting the 10 NNs for $[S : +, \text{RAW}]$ and $[S : -, \text{RAW}]$, and the overlap for $[S : +, \text{RAW}]$ and $[P : 20 + 20, \text{CONST}]$: 37.45% vs. 10.52%. The overlap is much larger between the two syntagmatic spaces than between a syntagmatic space and a paradigmatic space produced with a wide context window. In fact, the contexts for $[S : -, \text{RAW}]$ are only ≈ 12 words, while the context windows for $[P : 20 + 20, \text{CONST}]$ span 20+20 words, which is almost double the size of the context. Thus, the difference between syntagmatic and paradigmatic uses of context is not primarily a difference of *size*, but of *type*; syntagmatic and paradigmatic uses of context yield word spaces with different types of relations between words.

One reason why wider context windows seem to approximate a syntagmatic use of context could be that the wider the context is, the wider is the range of *possible* (co-) occurrence events. By contrast, a very narrow context harbors a very limited number of (co-) occurrence possibilities. The chance that *any* two words in the data will share enough of its neighbors to end up close to each other in the word space is therefore much greater in a paradigmatic use of *wide* context windows than in a paradigmatic use of narrow ones. Correspondingly, *any* two words in the data will *co-occur* if the context regions in a syntagmatic use of context are large enough. Thus, the increasing overlap between syntagmatic uses of context and paradigmatic uses of wide context windows might be explained by the increase in general (co-) occurrence events when using wide contexts.

9.2 Computing the density

It seems that a paradigmatic use of wide context windows increases the chance of collecting general occurrence events. This suggests that word spaces assembled from such wide windows should be *denser* than word spaces collected from narrow windows, simply because there are *more* occurrence events in the former spaces. A very simple measure of the density of a word space is to compute the average cosine measure between nearest neighbors in the word space. If the neighbors are very close to each other, we get a high density measure; if they are distant from each other, we get a low score. A high density measure would indicate that the space is very crowded, which could mean one of two things: either that the space rests on a solid statistical foundation, or that it is over-crowded and populated with chance occurrences.

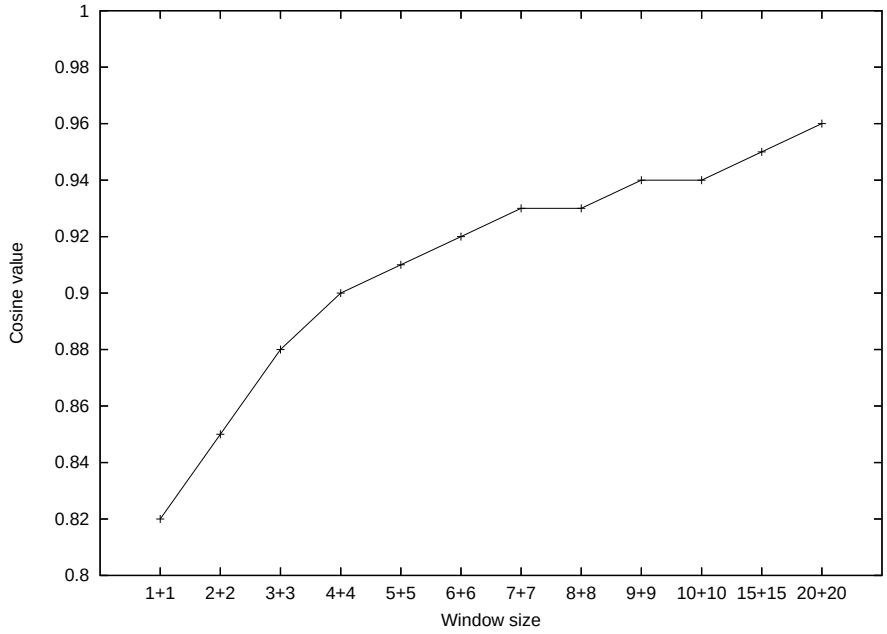

Figure 9.3: Average cosine value between nearest neighbors.

Figure 9.3 shows the results computing such an average density measure for spaces produced with $[P : 1 - 20, \text{CONST}]$. As comparison, the average density measure for spaces produced with $[S : +, \text{RAW}]$ and $[S : -, \text{RAW}]$ is 0.23 and 0.08. These low average cosine scores clearly show the sparse-data effect on the density of the word

spaces; the richer the statistical foundation, the denser the word space. The remarkable difference in density between syntagmatic and paradigmatic word spaces further demonstrates that collecting paradigmatic relations within wide context windows accumulates word spaces that are inherently different from those accumulated by syntagmatic uses of context. We have recently suggested that such density measures can serve as indicators of the topical dispersal of a corpus (Sahlgren & Karlgren, 2005b).

9.3 Conclusion

The comparison presented in this chapter suggests that word spaces collected from syntagmatic and paradigmatic uses of context are inherently different. Only a small percentage of the nearest neighbors occur in both syntagmatic and paradigmatic word spaces, regardless of the size of the contexts. These results refute any suspicion that syntagmatic and paradigmatic word spaces are very similar to each other. On the contrary, they indicate that these two types of word spaces contain inherently different information.

However, it is a legitimate question to ask why there is an overlap at all? If we for a second assume that syntagmatic and paradigmatic word spaces contain completely distinct relations between words, then should we not expect to see an even smaller (possibly even non-existent) overlap? As I noted in the previous chapter, syntagmatic and paradigmatic relations are not mutually exclusive, and these results support this observation. The commonality between syntagmatic and paradigmatic word spaces indicates that some words are *both* syntagmatically and paradigmatically related, and that the difference between syntagmatic and paradigmatic information is not clear-cut.

I conclude this comparison between syntagmatic and paradigmatic word spaces with the observation that, although a wider context window does lead to a larger overlap with a syntagmatic use of context, it does not seem possible to *reproduce* a syntagmatic word space by a paradigmatic use of wide context windows. This means that the hypothesis — that syntagmatic and paradigmatic word spaces contain different semantic information — is not only still valid, but supported by the comparison.

Chapter 10

Thesaurus comparison

"It's like living in a dictionary!"
(Marge Simpson in "I'm Spelling as Fast as I Can")

It would seem that an obvious way to evaluate word spaces would be to use some kind of repository of semantically related words to compare the word space with. One of the first things that comes to mind for a linguist when talking about semantic repositories is *thesauri*, in which words with similar meanings are listed. Curiously enough, thesauri have not (as far as I am aware) been used to evaluate word spaces directly, even though some of the probabilistic approaches that I mentioned briefly in Section 3.2 have used thesauri for evaluation (Grefenstette, 1993).

An example of a thesaurus entry for the word "demon" from the thesaurus used in these experiments (the Moby thesaurus, see below) is displayed in Table 10.1. There are a couple of potential problems with this kind of semantic repository. The first is that thesauri tend to be very wide and include a large number of words. The problem with such lenient repositories of semantic similarity is that they will contain a fair amount of more or less arbitrary terms. Everyone can probably find at least a couple of unexpected, or even a few far-fetched, terms in the example with "demon" in Table 10.1. It will therefore be *very* difficult to reproduce a thesaurus entry using word-space neighborhoods even if we have access to vast amounts of topically pertinent data.

The second potential problem is that there are many different *types* of semantic relations represented in this thesaurus entry; among the more obvious ones are (near) synonyms (e.g. "devil," "satan," "ghoul," "monster"), hypernyms (e.g. "evil spirit"), and hyponyms (e.g. "incubus"). Looking at this example, we can find both syntagmatically and paradigmatically related words. Examples of the former can be "demon–killer" and "demon–violent," and examples of the latter can be

demon:	baba yaga, lilith, mafioso, satan, young turk, addict, afreet, ape-man, atua, barghest, beast, beldam, berserk, berserker, bomber, brute, bug, cacodemon, collector, daemon, daeva, damned spirits, demonkind, demons, denizens of hell, devil, devil incarnate, dragon, dybbuk, eager beaver, energumen, enthusiast, evil genius, evil spirit, evil spirits, faddist, fanatic, fiend, fiend from hell, fire-eater, firebrand, freak, fury, genie, genius, ghoul, goon, gorilla, great one for, gunsel, gyre, hardnose, harpy, hell-raiser, hellcat, hellhound, hellion, hellish host, hellkite, hobbyist, holy terror, hood, hoodlum, host of hell, hothead, hotspur, hound, incendiary, incubus, infatuate, inhabitants of pandemonium, intelligence, jinni, jinniyeh, killer, lamia, lost souls, mad dog, madcap, monster, mugger, nut, ogre, ogress, powers of darkness, pursuer, rakshasa, rapist, revolutionary, rhapsodist, satan, savage, she-wolf, shedu, souls in hell, specter, spirit, spitfire, succubus, sucker for, supernatural being, termagant, terror, terrorist, the damned, the lost, the undead, tiger, tigress, tough, tough guy, ugly customer, vampire, violent, virago, visionary, vixen, werewolf, wild beast, witch, wolf, yogini, zealot

Table 10.1: Thesaurus entry for "demon."

"demon–daemon" and "demon–spirit." This means that it will be difficult to draw any conclusions about the nature of the word spaces based on this comparison. What would it mean if one word space has a particularly high correlation with the thesaurus? Would it mean that it contains more syntagmatic or more paradigmatic relations? Probably, it would mean that the word space contains a fair amount of *both* types of relations between words. Thus, I hypothesize that the word space that correlates the most with the thesaurus contains the widest spectrum of both syntagmatic and paradigmatic relations.

10.1 The Moby thesaurus

The Moby thesaurus (second edition) is a freely available lexical resource that contains more than 30 000 index words, listed together with more than 2.5 million synonyms and related terms.[1] There are two main advantages to using the Moby thesaurus. Firstly, it is much more extensive than other freely available thesauri. Secondly, it is distributed in a very simple plain-text format with one entry (i.e. a root word together with all its related words) per line.

[1] http://www.dcs.shef.ac.uk/research/ilash/Moby/

10.2 Syntagmatic uses of context

Table 10.2 shows the correlation between the Moby thesaurus and word spaces produced with a syntagmatic use of context. As can be seen in the table, the correlation using large context regions is 9.84–10.43%, and the correlation using small context regions is 9.17–9.26%. As can be expected, the correlation is larger for the lax evaluation, where the figures are 22.35–30.10% vs. 23.93–24.77%. This difference indicates that using large context regions yields somewhat "wider" semantic representations that correlate better with the lenient nature of the Moby thesaurus than using small context regions. The reason for this might be that large context regions yield word spaces that contain a more diverse blend of both syntagmatic and paradigmatic information, while smaller context regions yield word spaces than contain more specifically syntagmatic relations.

Transformation	$[S:+]$		$[S:-]$	
	Strict	Lax	Strict	Lax
BINARY	10.40	29.47	**9.26**	24.60
DAMPENED	10.35	27.77	9.23	**24.77**
TFIDF	**10.43**	**30.10**	9.17	23.93
RAW	9.84	22.35	9.17	23.93

Table 10.2: Correlation between thesaurus entries and syntagmatic word spaces.

Note the comparably poor results for $[S:+, \text{RAW}]$. This indicates that, even though I use the cosine measure of similarity, which effectively normalizes the vectors, there is an over-emphasis of high-frequency words when not dampening the frequency counts. Smaller context regions do not have this frequency-induced problem, since those co-occurrence events are much more exclusive.

10.3 Paradigmatic uses of context

Figure 10.1 shows that there is a very small variation in the results over the different window sizes for the strict evaluation. The results for the windows with constant weights range between 6.97% for $[P:1+1, \text{CONST}]$ over 7.78% for $[P:3+3, \text{CONST}]$ to 7.10% for $[P:10+10, \text{CONST}]$. The best score for the weighted window is also produced with the 3+3-sized window (7.77%), and the same decreasing trend can be seen here, albeit not as pronounced, when the size of the window increases (7.71% for $[P:10+10, \text{AGG}]$). The lax evaluation follows the exact same pattern: a 3+3-sized window produces the best results for the windows with constant weights (33.81%) and the second best result for the weighted ones (34.20%; $[P:6+6, \text{AGG}]$

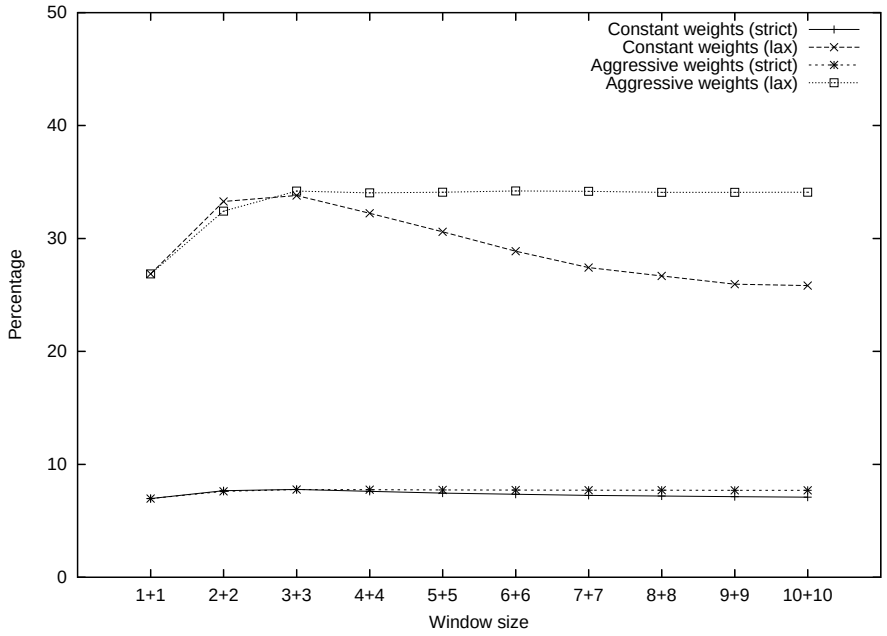

Figure 10.1: Correlation between thesaurus entries and paradigmatic word spaces.

is marginally better with 34.21%), and the results decrease as the window sizes increase. Note that, even though the results for the weighted windows *do* decrease as their size increase, the slope is very flat. This indicates that it is only the few nearest surrounding words that are useful in this particular test. The presence of a slight peak in the results for a window size around 3+3 supports this observation.

10.4 Comparison

We have seen that syntagmatic uses of context yield better results in the strict evaluation (10.43% vs. 7.78%), but that paradigmatic uses of context yield better results in the lax evaluation (34.21% vs. 31.10%). Why would such a discrepancy occur?

I noted above that the Moby thesaurus is very wide, and that it includes a fair amount of both syntagmatically and paradigmatically related words. The fact that syntagmatic uses of context are better than paradigmatic uses in the

strict setting indicates that neighborhoods in the syntagmatic word spaces contain *both* syntagmatically and paradigmatically related words, whereas neighborhoods in the paradigmatic word spaces might be more homogeneous and mainly contain paradigmatic information. If this assumption is true, and syntagmatic uses of context yield more heterogeneous word spaces than paradigmatic uses, there should be a greater chance that semantically unrelated words turn up among the closest neighbors in the syntagmatic word spaces. The fact that paradigmatic uses of context yield better results in the lax evaluation setting supports this observation.

Chapter 11

Association test

"I was just putting words together."
(Homer Simpson in "Old Yeller Belly")

A semantic repository that is very similar to a thesaurus entry is an *association norm*. It too consists of a list of similar words, with the differences that it typically contains fewer words than a thesaurus entry, and that it has been produced by humans that have been subjected to an *association test*. An association test is usually designed so that a human subject is presented with a number of words (usually one at a time), and is asked to provide a number of associatively related words to each presented word. The associations provided by the test subjects are then averaged, and the most frequent associations define the *association norm* for the particular population to which the test subjects belong. An example of an association norm for the word "demon" taken from the University of South Florida Free Association Norms is displayed in Table 11.1 (the numbers indicate the proportion of test subjects that produced the association).

As can be seen in the examples in Table 11.1, these associative relations do not correspond to *one* kind of semantic relation. In the association norm for "demon," we can find (near) synonyms ("devil," "satan," "ghost," "monster"), a meronym ("wings"), a possible antonym ("deacon"), and a number of adjectival modifiers ("evil," "possessed," "bad," "scary"). This means that we will encounter the same problem as the one discussed in relation to thesaurus entries: there are examples of *both* syntagmatically and paradigmatically related words in the association norms. Examples of the former can be "demon–evil" and "demon–scary", and examples of the latter can be "demon–devil" and "demon–satan." However, the association norms contain much fewer words than the thesaurus entries and can therefore be expected to contain fewer arbitrary terms. Furthermore, although the association norms do not correlate with linguistic taxonomy, they *do* represent psychological

demon:	devil	0.553
	evil	0.127
	satan	0.040
	ghost	0.033
	possessed	0.033
	monster	0.027
	bad	0.020
	deacon	0.013
	scary	0.013
	wings	0.013

Table 11.1: Association norm for "demon."

reality — these are, after all, the words that *real* people deem to be intuitively related, and so can be assumed to constitute nearest neighbors in the human associative space.

Another important consideration when using association norms for comparison with word-space neighborhoods is that they are inherently sensitive to a large number of factors; economic and social status, political and religious conviction, cultural and ethnic background, and linguistic proficiency are all examples of factors that influence what kind of associations people make. To take a current example, the word "terrorism" is likely to have entirely different associations for an Islamic fundamentalist than for a conservative American. This presents us with a problem when using association norms to evaluate our word spaces: unless we have produced the word spaces from pertinent data, we will not be able to match the association norm no matter how good a representation of the semantics of the data the word space is. Imagine that we build a word space using newswire text (which is quite common, since many available data collections are based on news articles), and that we compare it to an association norm collected from a population of suburban teens. Chances are we will not see a very large correlation between them. However, that does not mean that the word space is flawed — it only means that it does not reflect the semantics of suburban teens. It may still constitute a truthful reflection of the semantics of the newswire texts.

11.1 The USF association norms

I use the freely available University of South Florida Free Association Norms (Nelson et al., 1998)[1] in the following experiments. The norms were collected from

[1] http://w3.usf.edu/FreeAssociation/

more than 6 000 test subjects who produced responses to 5 019 stimulus words by writing down the first word they could think of that was meaningfully related, or strongly associated, to the stimulus word. Such a procedure is called a *discrete* association task because each participant is asked to produce only a single associate to each stimulus word.

11.2 Syntagmatic uses of context

Table 11.2 demonstrates that the correlation between the association norms and word spaces produced with large text regions is 11.91–13.43%. In contrast to the thesaurus comparison, the small text regions yield better results in this test: 14.93–15.41%. If the analysis in the previous chapter was correct, this should mean that the association norms are semantically more homogeneous than the thesaurus entries. As can be expected, the correlation increases in the lax evaluation setting, where the figures are 27.88–35.73% using the large regions, and 36.45–37.91% using the small ones.

Regarding the frequency transformations, we can note that binary counts produce the best results for the small region but not for the large one, where logarithmic transformation of frequencies seems to be a viable alternative. Note again the inferior results — most pronounced for the large context regions in the lax evaluation setting — using the raw frequency counts.

Transformation	$[S:+]$		$[S:-]$	
	Strict	Lax	Strict	Lax
BINARY	11.91	34.52	**15.41**	**37.91**
DAMPENED	**13.43**	34.52	15.16	37.45
TFIDF	13.07	**35.73**	15.34	37.81
RAW	12.61	27.88	14.93	36.45

Table 11.2: Correlation between association norms and syntagmatic word spaces.

11.3 Paradigmatic uses of context

As can be seen in Figure 11.1, the correlation between word spaces produced with paradigmatic uses of context and human association norms grows with the size of the context windows. The correlation is smallest using $[P:1+1, \text{CONST}]$ (4.17% for the strict evaluation and 12.74% for the lax one), and largest for the strict evaluation using $[P:9+9, \text{CONST}]$ (8.76%) and $[P:10+10, \text{AGG}]$ (6.05%). The

best results for the lax evaluation is produced using $[P : 8+8, const]$ (27.95%) and $[P : 10 + 10, \text{AGG}]$ (18.38%). Regarding weighting of the positions in the context windows, we can see that it has a pronounced negative effect on the overlap. Thus, it seems that wider context windows are preferable to use when reproducing human association norms.

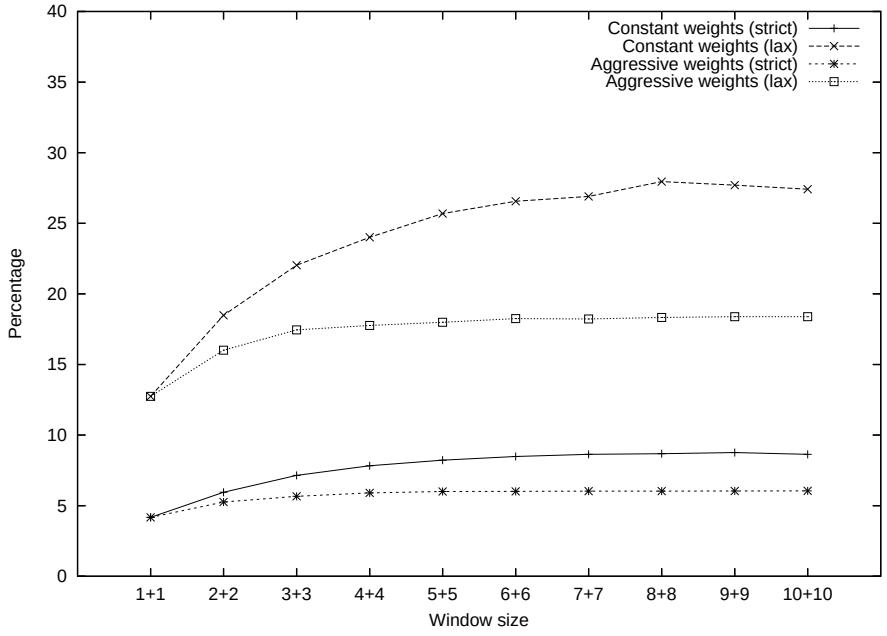

Figure 11.1: Correlation between association norms and paradigmatic word spaces.

11.4 Comparison

The results from these experiments clearly show that syntagmatic uses of context produce word spaces with a much higher degree of correlation to a human association norm than word spaces produced with paradigmatic uses of context. The best result using paradigmatic word spaces (8.76% for the strict evaluation using $[P : 9 + 9, \text{CONST}]$) is far surpassed by the best syntagmatic word space (15.41% for the strict evaluation using $[S : -, \text{BINARY}]$).

This could mean one of two things: either that the association norms contain more syntagmatic than paradigmatic information, or that the word spaces produced with syntagmatic uses of context contain a fair amount of paradigmatic information. As I argued in the previous chapter, the latter hypothesis is likely to be correct to a certain extent. However, note that the small context regions outperform the large ones in this test, while the situation is reversed in the thesaurus comparison. If the assumption in the previous chapter is correct that small context regions yield word spaces that contain more specifically syntagmatic information than the ones produced using large context regions, it is more likely that it is the former hypothesis that is correct: the association norms contain more syntagmatic than paradigmatic information. In that case, these results demonstrate that small context regions yield *more syntagmatic* word spaces than large ones.

Chapter 12

Synonym test

"Inflammable means flammable? What a country!"
(Dr. Nick in "Trilogy of Error")

In contrast to thesauri comparisons and association tests, synonym tests are very common in word-space research, no doubt because a synonym test was used for evaluation in the seminal paper by Landauer & Dumais (1997). In this paper, Landauer & Dumais used 80 test items from the synonym part of the standardized vocabulary test TOEFL (Test Of English as a Foreign Language). The synonym part of this vocabulary test is designed as a multiple-choice test, where the task is to identify the synonym to a given word from four given alternatives. An example of a multiple-choice question from the synonym part of the TOEFL is depicted in Table 12.1.

Word	Alternatives	Synonym
spot	sea	
	location	√
	latitude	
	climate	

Table 12.1: TOEFL synonym test for "spot."

Execution of a synonym test is easily implemented in a word-space framework. Since it is a multiple-choice test, it is straight-forward to calculate the vector similarity between the context vector for the given word and the context vectors for the provided alternatives, and to select the most similar alternative as synonym. This type of test is well-defined and nonparametric (i.e. the test itself involves no

parameters), which makes it reliable and easy to use for comparison of different word-space algorithms, of different parameter settings, and also of different uses of context. In addition to this, most published experiments tend to use the exact same test set — the 80 multiple-choice TOEFL items that were used by Landauer & Dumais (Landauer & Dumais, 1997; Levy et al., 1998; Karlgren & Sahlgren, 2001; Turney, 2001; Rapp, 2003).[1]

In contrast to the thesaurus comparison and association test, the TOEFL synonym test is quite easy. With four alternatives, we can get 25% correct answers by simply guessing at random. Landauer & Dumais report that foreign applicants to American colleges average 64.5%, but they do not present any average results for native speakers of English. It seems reasonable to believe that those figures would be significantly higher. We should therefore expect to see much better results in these experiments than the ones reported in Chapters 10 and 11.

Another difference between the synonym test and the thesaurus and association tests is that it is perfectly feasible to construct a word space that solves a synonym test more or less impeccably while still not being a very good model of word meaning. The point is that we do not know anything about the actual neighborhoods in the word space by merely seeing the results from a synonym test, since the test only utilizes the relative distances between the context vector for the given word and the context vectors for the alternatives. As an example, consider the following word space:

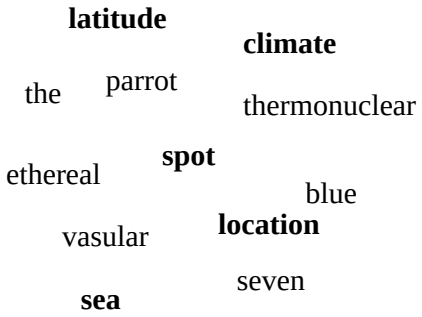

Figure 12.1: Fictive word space.

Now imagine that we use this space to solve the TOEFL item from the example above. Obviously, we would have no problem choosing the correct alternative, but is that fact really significant with regards to determining the quality of this particular word space?

On the other hand, whereas the thesaurus entries and association norms included both syntagmatically and paradigmatically related words, synonyms are

[1]The 80 TOEFL items were kindly provided by Thomas Landauer.

thoroughly paradigmatic: they are the prime example of paradigmatically related words. Synonyms tend to *not* co-occur, except in very specific contexts, such as "w_1 is the same as w_2," or "w_1 means w_2." Rather, synonyms tend to co-occur with similar *other* words. As an example, both "bloke" and "lad" tend to occur in similar contexts, but it is fairly uncommon to see them co-occur. The same applies to "lift" and "elevator," or "neighbor" and "neighbour." Thus, we can expect that paradigmatic uses of context should yield better results on a synonym test than syntagmatic uses.

Amongst previously reported experiments using the TOEFL synonym items, we can note that LSA achieved 64.4% (36% without the use of SVD) on a different corpus (Landauer & Dumais, 1997). The fact that SVD increases the results in TOEFL experiments with LSA, which implements a syntagmatic use of contexts, further supports the hypothesis from Section 7.5 that the SVD approximates a paradigmatic use of contexts. By contrast, RI managed to reach 72% using the lemmatized TASA corpus and a 3+3 sized distance-weighted context window. Turney has reported a score of 73.75% using a probabilistic algorithm called PMI-IR (Turney, 2001), and Levy et al. (1998) reached 76% using the Hellinger distance metric,[2] and a small context window with probabilistic weights for the BNC. Rapp (2003) reports an astounding result of 92.5% correct answers, which was produced by applying SVD to a words-by-words matrix collected using a 2+2-sized window over a lemmatized and stop-word filtered version of the BNC. Rapp identifies three main reasons for his fantastic result: the size of his corpus, the preprocessing (lemmatization and stop-word filtering), and the use of paradigmatic contexts. To this can be added the use of SVD to restructure the co-occurrence data. His best result without SVD is 69% correct answers.

12.1 Syntagmatic uses of context

Since these experiments use a fixed test vocabulary of 400 words, they can be implemented in a comparatively efficient manner (it is not necessary to produce context vectors for *all* words in the data). I therefore use both the TASA and the BNC data in these experiments. As discussed in Chapter 8, the BNC does not directly allow for the use of large context regions, whereas the TASA can be used with both large and small context regions. The results are summarized in Table 12.2 below.

[2]The Hellinger metric is given by:

$$\mathrm{dist}_{\mathrm{H}}(\vec{x}, \vec{y}) = \sum(\sqrt{x} - \sqrt{y})^2$$

Corpus	Transformation	$[S:+]$	$[S:-]$
TASA	BINARY	58.75	51.25
TASA	DAMPENED	56.25	51.25
TASA	TFIDF	**60**	**52.50**
TASA	RAW	53.75	51.25
BNC	BINARY	N.A.	67.50
BNC	DAMPENED	N.A.	67.50
BNC	TFIDF	N.A.	67.50
BNC	RAW	N.A.	63.75

Table 12.2: Percentage of correct answers to 80 items in the TOEFL synonym test for syntagmatic word spaces.

These results follow the same trend as the results in the previous chapters: transformation of the frequency counts leads to better results than using the raw occurrence frequencies. The worst result for both the TASA (using large context regions) and the BNC (using its small regions) is produced using raw frequency counts. Regarding the difference between the context regions, Table 12.2 shows that large context regions produce better results than the small ones for the TASA corpus. It would have been interesting to see what results a corresponding context region could have produced using the BNC data. The difference in the results for the small and large context regions is consistent with the assumption from the previous chapters that a smaller context region is more syntagmatic than a large one, and therefore yields worse result in a thoroughly paradigmatic test.

The best result on the TOEFL synonym test with a syntagmatic use of contexts is 67.50%. This result is reached using the BNC corpus with all tested transformations except the raw frequencies. The best result for the TASA corpus is 60%, using $[S:+, \text{TFIDF}]$.

12.2 Paradigmatic uses of context

The relative efficiency of the synonym test means that I can experiment with different frequency thresholds for the paradigmatic uses of context. I did some initial experiments varying the low frequency threshold from 1 occurrence to 10 with absolutely no effect on the results. By contrast, there is a striking effect when thresholding high frequency words, as can be seen in Figure 12.2. The parameters for these experiments are $[P:2+2, const]$.

The best results for the TASA corpus are produced with an upper threshold of 15 000 or 13 000 occurrences. My version of the lemmatized TASA corpus

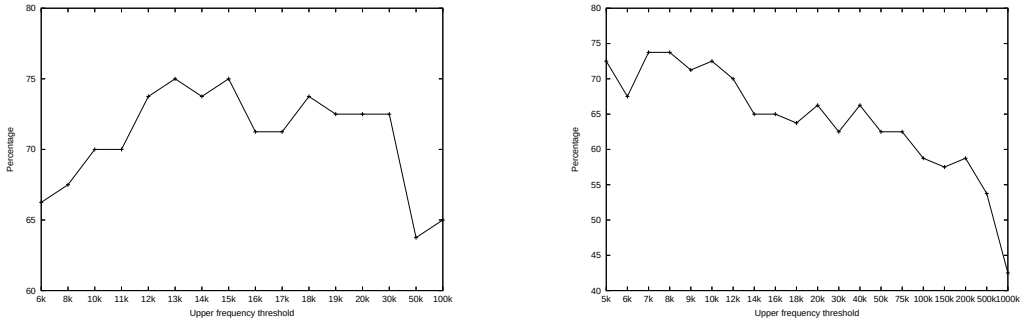

Figure 12.2: Percent correct answers on the TOEFL as a function of upper frequency thresholding for paradigmatic word spaces using the TASA (left graph) and the BNC (right graph).

contains 66 586 word types. Thresholding at 15 000 occurrences affects 83 word types (approximately 0.12% of the words), and leaves 66 503. The best results for the BNC corpus is produced with an upper threshold of 8 000 occurrences. My version of the lemmatized BNC data contains 317 961 unique words. Filtering at 8 000 occurrences removes some 0.4% of the vocabulary (1 346 word types), and leaves 316 615 unique words.

It is somewhat surprising that the larger of the two corpora requires a lower frequency threshold for high-frequency words. One possible reason for this could be that the larger data has a wider genre stratification. Our previous experiments in Sahlgren & Karlgren (2005b), where we applied the density measure introduced in Chapter 9 to (amongst others) the TASA and the BNC corpora, showed a comparatively large difference in word-space density between these two data sets. The BNC received a higher density count than the TASA, which we interpreted as an indication that the BNC is topically more dispersed than the TASA. Such topical dispersal leads to "schizophrenic" distributional behavior that distorts the distributional representations for a larger range of high-frequency words. This points to one of the most serious weaknesses in word-space methodology: it cannot differentiate between several simultaneous distributions. When the data is topically dispersed, word-space representations may become hybrids of different, possibly unrelated, meanings. In the worst-case scenario, the context vector fails to properly represent any one of the meanings involved.

The results using frequency thresholding demonstrate that thresholding distributionally promiscuous words is beneficial for the word space when solving the TOEFL synonym test. Perhaps, then, upper frequency thresholding could improve the results using paradigmatic contexts in the other tests? Unfortunately, optimizing the frequency bounds for the other tests is computationally very expensive,

since they involve all words in the data (and not just a small test vocabulary as in the present experiment). I therefore do not use optimized frequency thresholds in the other experiments. For the TASA corpus, the threshold is set to 15 000 occurrences, and for the BNC to 8 000 occurrences.

12.3 Symmetric windows

In these experiments, I study the use of symmetric context windows that have the same number of window slots on the left and right side of the focus word. The size of the context windows range from 1+1 to 10+10. The results are summarized in Figure 12.3.

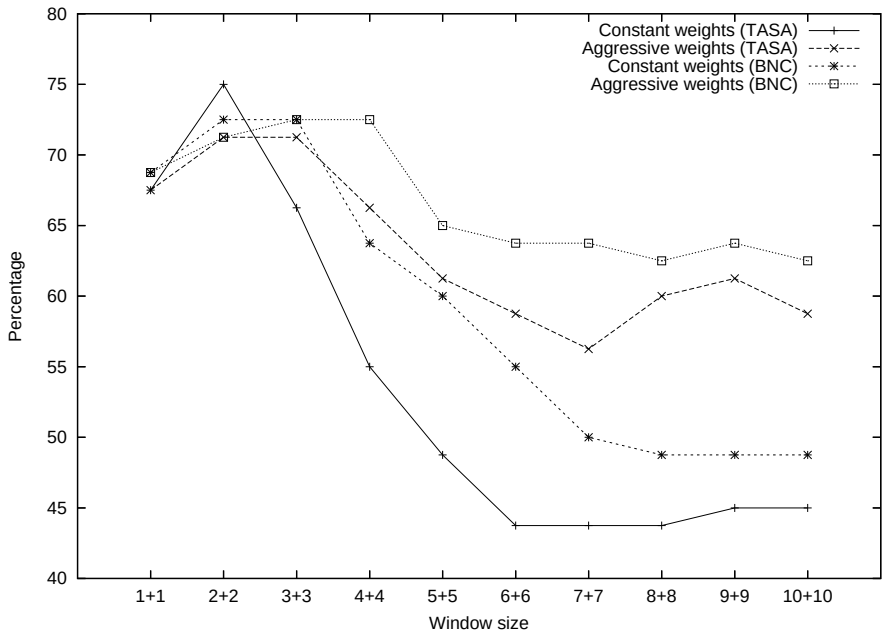

Figure 12.3: Percent correct answers on the TOEFL for paradigmatic word spaces using the TASA and the BNC.

The results clearly show that narrow context windows spanning only a few words on each side of the focus word produce the best results. This concurs in the results reported by Levy et al. (1998). The results decrease for both corpora when the windows grow wider. This indicates that a narrow context window is *more*

paradigmatic than a wide one. The reason for this seems obvious: the further away from the focus word we get, the more possible co-occurrents of the word is there. Recall from Chapter 9 that paradigmatic word spaces grow denser with the size of the context windows. This implies that there might be a correlation between the density of a paradigmatic word space and the amount of "noise" (i.e. chance occurrence events) it contains. When only looking at the immediately surrounding positions, there is a very limited set of possible words that could occupy them, and the resulting word space is consequently less prone to include chance occurrence events. In a sense, we could therefore say that the wider the context window, the more noise we get: the denser a paradigmatic word space is, the more chance occurrences it contains, and consequently the less paradigmatic it is.

A window spanning two words to the left and two words to the right (i.e. a 2+2-sized window) seems to be optimal for the TASA corpus, while a window spanning three preceding and three succeeding words (i.e. a 3+3-sized window) seems to be optimal for the BNC corpus. This (admittedly small) difference might be an artefact of the difference in size between the corpora: the BNC corpus is ten times larger than the TASA corpus, which means that it provides better statistical evidence, and therefore facilitates the use of wider context windows. This suggests that there might be a correlation between data size and the optimal size of the context windows: small data sets require small windows; larger data sets can use larger windows.

It is somewhat surprising that the smaller TASA corpus generates a better top score than the ten times larger BNC data (75% vs. 72.50%). At the same time, the results for the BNC are more stable and do not decrease as drastically as the results for the TASA corpus when the window size increases. This indicates that more data makes wider window sizes more viable, but that the quality of the data is the decisive factor for the performance of the word-space model in this particular test setting.

Regarding the distance-weighting of the context windows, it seems that weighting is only beneficial for the outlier positions in larger windows. As can be seen in Figure 12.3, aggressive weights produce the best results for larger windows. This is not surprising, since aggressive weighting effectively removes the impact of the outlier positions, so that the aggressively weighted windows are comparable to the 2+2-sized or 3+3-sized windows with constant weights. This further supports the assumption that narrower context windows are more paradigmatic than wide ones.

12.4 Asymmetric windows

It is not obvious a priori that a symmetric window is the best option. It might just as well be the case that the left or the right context is more important for

conveying distributional information. Again, since these experiments are efficient to execute, I can examine the effects of using only the left and only the right contexts. As with the symmetric windows, I use two different weighting schemes for the positions within the windows. Figures 12.4 and 12.5 show the results.

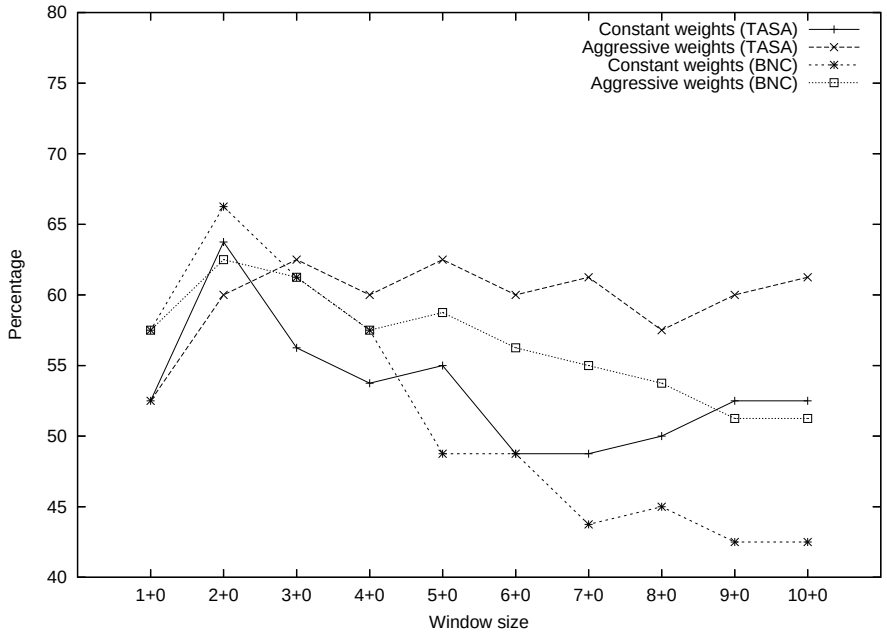

Figure 12.4: Percent correct answers on the TOEFL using only the left context for paradigmatic word spaces using the TASA and the BNC.

It is clear that asymmetric windows do not produce as good results as symmetric ones. The reason for this could be that asymmetric context windows rely on only half as much data as symmetric ones. However, the tendencies in the results are similar: it is the immediately preceding and succeeding words that are the most useful ones. The best result with a left-branching window is a 2+0-sized window (i.e. two words to the left), while the best result with a right-branching window is a 0+3-sized window (i.e. three words to the right). Trying a 2+3-sized window only generates 67.5%, which demonstrates that the optimal window size for the left and right contexts might not correspond to the optimal window size when *both* the left and right context are taken into account. Again, the results with the asymmetric windows concur in the observation that a narrow context window seems to be more paradigmatic than a wide one.

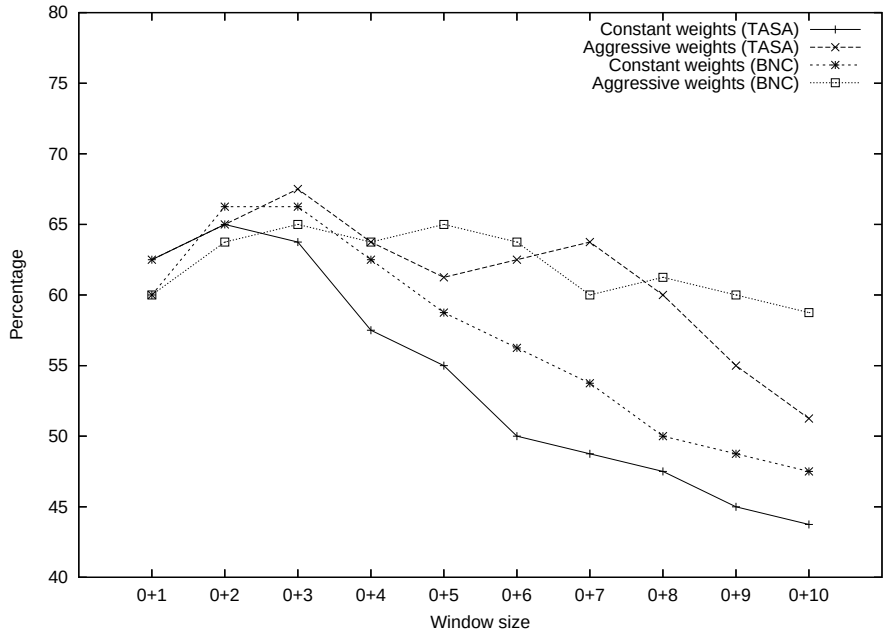

Figure 12.5: Percent correct answers on the TOEFL using only the right context for paradigmatic word spaces using the TASA and the BNC.

12.5 Comparison

If we compare the results produced from syntagmatic and paradigmatic uses of context, it is clear that the paradigmatic word spaces produce better results on the TOEFL synonym test. The best result produced in these experiments is 75% using $[P : 2 + 2, \text{CONST}]$ over the TASA corpus. In comparison, the best result produced from a syntagmatic use of context is 67.5% using the BNC with $[S : -, \text{BINARY}, \text{DAMPENED}, \text{TFIDF}]$. The fact that the smaller data generates a better top score for the paradigmatic word space, while larger data generate a better top score for the syntagmatic word space suggests that *context quality* is more important for paradigmatic uses of context, while *context quantity* might be more important for syntagmatic uses of context. These results alone do not warrant such conclusion, but they do indicate that this phenomenon deserves further study.

Chapter 13

Antonym test

"Ih! I mean: Ah!"
(Groundskeeper Willy in "Lard of the Dance")

Since the last chapter featured a synonym test, it seems appropriate to also include an antonym test in the evaluation of word spaces. Whereas synonyms are words with the *same* meaning, antonyms are words with the *opposite* meaning. Both types of words are substitutable in context, but whereas synonyms retain the meaning of the utterance, antonyms reverse it. As an example, consider how the meaning of the utterance "he looked very alive" is affected when "alive" is substituted with "vital" (a synonym) versus "dead" (an antonym). This means that antonyms are, just as synonyms, primarily paradigmatically related words, which often occur in similar contexts.

However, antonyms have a tendency to become syntagmatic in *topical* contexts. As I mentioned in the previous chapter, synonyms can also become syntagmatic, but only in very specific contexts. By contrast, if one of the words in an antonymic word pair occurs in a text on a certain topic, it is not uncommon to encounter the other word as well. As an example, imagine that we talk about the migration patterns of a certain kind of fish, say beluga. If the word "deep" occurs in this context, as in "deep waters," we will probably also encounter the antonym "shallow," as in "shallow waters." We can therefore expect to see good results in this test not only from paradigmatic word spaces, but also from word spaces built from large context regions.

13.1 The Deese antonyms

In order to test the ability of the word spaces to capture antonymy, I follow the procedure described by Grefenstette (1992a), who uses a list of antonym pairs

taken from work done by Deese (1964). Deese's original list included 39 antonym pairs (see Table 13.1), out of which Grefenstette discarded six pairs because they were not correctly tagged in his data. I use 38 of the 39 antonym pairs in the following experiments. The pair "single–married" is discarded because the later word only occurs twice in the TASA.

active–passive	dark–light	easy–hard
bad–good	bottom–top	happy–sad
high–low	long–short	old–young
right–wrong	sour–sweet	alive–dead
big–little	clean–dirty	deep–shallow
empty–full	hard–soft	large–small
narrow–wide	rich–poor	rough–smooth
strong–weak	back–front	black–white
cold–hot	dry–wet	fast–slow
heavy–light	left–right	new–old
pretty–ugly	short–tall	thin–thick
alone–together	far–near	first–last
few–many	single–married	inside–outside

Table 13.1: The 39 Deese antonym pairs.

13.2 Syntagmatic uses of context

Table 13.2 shows the results from syntagmatic uses of context. As can be seen in the table, the best result for the strict evaluation is produced using a large context region (28.95%). At the same time, there is more variation in the results for the large context regions over the different frequency transformations (10.53–28.95%) than for the small context regions (18.42–23.68%). This is true for the lax evaluation as well, where large and small context regions produce the same top result (47.37%), but where the small context region produces this exact same result for all transformations.

Note that the raw frequency counts outperform the other transformations using both large and small context regions, and that the difference to the other transformations is very large for the large regions. This might be explained by the syntagmatic tendency of antonyms in topical contexts. Raw frequency counts in large context regions seems to capture exactly this kind of topical co-occurrence patterns.

Transformation	$[S:+]$		$[S:-]$	
	Strict	Lax	Strict	Lax
BINARY	10.53	15.79	21.05	47.37
DAMPENED	15.79	36.84	21.05	47.37
TFIDF	15.79	21.05	18.42	47.37
RAW	**28.95**	**47.37**	**23.68**	47.37

Table 13.2: Percentage of correct antonyms for syntagmatic word spaces.

13.3 Paradigmatic uses of context

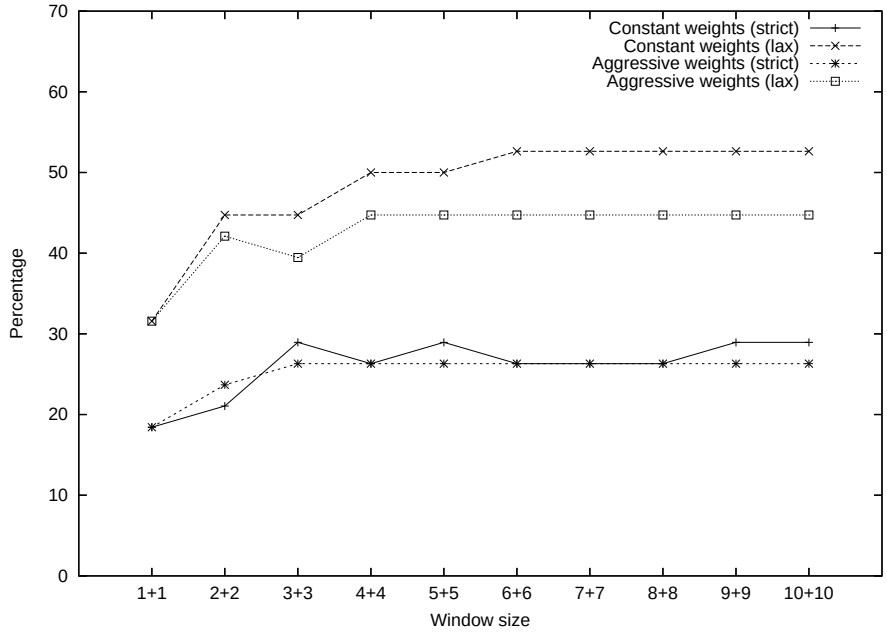

Figure 13.1: Percentage of correct antonyms for paradigmatic word spaces.

Figure 13.1 demonstrates that there is a very small difference between the windows with constant weights and the aggressively weighted ones for the strict evaluation (28.95% using $[P:3+3,5+5,9+9,10+10,\text{CONST}]$ vs. 26.31% using $[P:3-10,\text{AGG}]$). The difference between the weighting schemes is larger for the lax evaluation, where the windows with constant weights produce the best result

(52.63% using $[P : 6-10, \text{CONST}]$ vs. 44.74% using $[P : 4-10, \text{AGG}]$). It seems that wider context windows are preferable in this task. A minimal 1+1-sized window produces inferior results for both weights in both evaluations.

13.4 Comparison

Comparing the results produced from syntagmatic and paradigmatic uses of context, we can note that the top score for the strict evaluation (28.95%) is produced by both syntagmatic and paradigmatic word spaces. For the lax evaluation, the paradigmatic uses of context windows with constant weights yield better results than the syntagmatic uses of context (52.63% vs. 47.37%). These results support the assumption that antonyms become syntagmatic in large contexts (thus the same top score for paradigmatic uses of context and for syntagmatic uses of large context regions in the strict evaluation), but are primarily related through the paradigmatic relation (thus the better results for the paradigmatic word spaces in the lax evaluation setting).

Chapter 14

Part-of-speech test

"Today we're going to talk about predicates and predicate nominatives."
"Boring!"
(Edna Krabappel and Bart Simpson in "The Parent Rap")

Rapp (2002) observes that paradigmatic word pairs typically belong to the same parts of speech, whereas syntagmatic word pairs, although they *can* belong to the same parts of speech, typically do not. This observation implies that it should be possible to test whether a word space contains syntagmatic or paradigmatic relations by looking at how many of the neighbors in the word space have the same parts of speech. The space that has most neighbors with the same parts of speech can be assumed to contain the most paradigmatic relations.

In order to test this idea, I select all words in the TASA with frequency above 50. I tag the words (n.b. *not* the data, but only a list of the words) using the freely available TreeTagger,[1] and use only the most frequent tag for each word type. I restrict the tag set to 9 different parts of speech: nouns, verbs, adjectives, adverbs, conjunctions, determiners, pronouns, interjections, and prepositions. I construct word spaces using untagged data, and extract the 10 nearest neighbors to each word. I then look at how many of the neighbors are listed in the tagged word list with the same part of speech as the word in question. I define a lax evaluation setting as only looking at the closest neighbor, and a strict evaluation setting as looking at all of the 10 nearest neighbors.

Miller & Leacock (1998) argue that the syntactic category of a word can be determined by only looking at the words in its immediate environment. An experimental result that supports this claim is the success of Eric Brill's (1994) simple rule-based technique for recognizing syntactic categories without looking further

[1] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

than three words preceding or succeeding the focus word. This suggests that a narrow context window should lead to better performance in the part-of-speech test than using wide windows or context regions.

14.1 Syntagmatic uses of context

Table 14.1 shows the results from syntagmatic uses of context. As can be seen in the table, the small context regions generally lead to better results the large ones. The only exception is the top score for the lax setting, where large regions yield 57.79%, while the small ones yield 57.56%. The difference is larger in the strict setting where small regions outperform the large ones with 53.37% vs. 52.57%. This superiority of the small regions is a tad surprising, given the assumption in the previous chapters that a small context region is *more syntagmatic* than a large one. Under this assumption, we should expect to see less neighbors with the same part of speech in word spaces produced with small context regions as compared to word spaces produced with large ones. However, I noted above that the part-of-speech test is not very precise, and that syntagmatic neighbors *can* belong to the same part of speech.

	$[S:+]$		$[S:-]$	
Transformation	Strict	Lax	Strict	Lax
BINARY	51.52	**57.79**	**53.37**	**57.56**
DAMPENED	**52.57**	55.91	53.13	57.15
TFIDF	51.72	56.84	53.17	57.39
RAW	52.29	55.98	53.09	56.90

Table 14.1: Percentage of words with the same part of speech for syntagmatic word spaces.

These results follow the same trend as the other ones with regards to the differences between the frequency transformations for the syntagmatic uses of context. Raw frequency counts are generally inferior, with the sole exception of the strict evaluation setting for the large context regions. Binary counts seem to be the best choice in this particular test.

14.2 Paradigmatic uses of context

Figure 14.1 shows the results from paradigmatic uses of context. It is obvious that narrow context windows yield better results than wide ones in this experiment.

Aggressive distance weights clearly outperform the windows with constant weights when the window size grows wider, and there is hardly any difference between a 3+3-sized and a 10+10-sized aggressively weighted window. The fact that narrow windows and aggressive distance weighting outperform wide windows with constant weighting further strengthens the observation made in the previous chapters that a narrow context window is *more paradigmatic* than a wide context window.

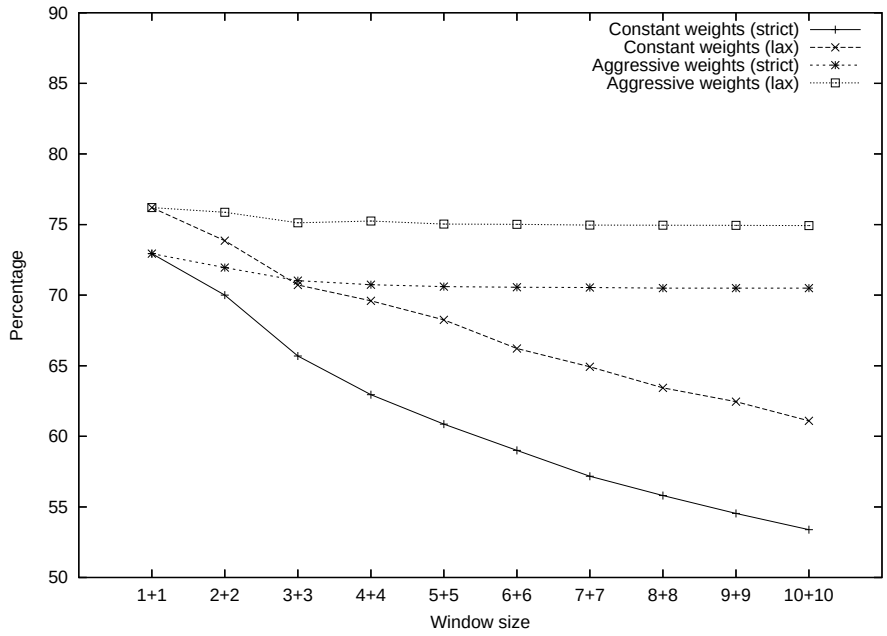

Figure 14.1: Percentage of words with the same part of speech for paradigmatic word spaces.

14.3 Comparison

Comparing the results using syntagmatic and paradigmatic word spaces, it is clear that the paradigmatic space contains more neighbors with the same part of speech than the syntagmatic space. The difference is clearest when using $[P : 1 + 1, \text{CONST}]$ vs. $[S : -, \text{BINARY}]$ (72.94% vs. 53.37% in the strict evaluation, and 76.20% vs. 57.79% in the lax setting). The difference is not as striking when the window sizes grow wider (53.39% vs. 53.37% in the strict setting, and 61.10%

vs. 57.79% in the lax setting using $[P : 10+10, \text{CONST}]$ vs. $[S : -, \text{BINARY}]$). Thus, if Rapp's suggestion is true that paradigmatic word pairs to a larger extent belong to the same part of speech than syntagmatic word pairs, these results clearly support the hypothesis that looking at shared neighbors captures paradigmatic information, while looking at co-occurrences captures syntagmatic information.

Levy et al. (1998) performed a similar experiment where they computed *centroid vectors*[2] for 12 different parts-of-speech by combining the context vectors "from a very large number of different words" belonging to the parts-of-speech. They then extracted the 100 most frequent words from the 12 different parts-of-speech, and looked at whether the words' context vectors were most similar to the centroid vector for the part of speech to which the words belonged. Levy et al. report very good results for their 12 sets of 100 words, and they observe the same trend that narrow context windows yield better performance than wide ones in this kind of "syntactic categorization" task.

[2]A centroid is the average value, or the center point. A centroid vector is simply the vector resulting from combining, normally by standard vector addition, a number of vectors.

Chapter 15

Analysis

"Can somebody tell me what the hell is going on here?"
(Moe Szyslak in "The Itchy and Scratchy and Poochie show")

The experiments reported in the previous chapters demonstrate that syntagmatic and paradigmatic word spaces produce consistently different results on a number of semantic tests. These differences support the hypothesis that syntagmatic and paradigmatic uses of context yield word spaces with inherently different semantic properties. However, it still remains unclear *to what extent* the word spaces contain syntagmatic and paradigmatic information, and *which parameters* influence this distinction?

In this chapter, I will further analyze the results reported in the previous chapters. Note that the actual performance figures are not very interesting in themselves; the fact that a word space has an 8.76% correlation with an association norm does not tell us very much in itself. I am not interested in producing as good results as possible on these tests (and even if I were, it is not obvious how we could know if a result is a good result). Rather, I am interested in what the tests can tell us about the semantic properties of the word spaces. Thus, I will first take a closer look at the effects of the different parameters for syntagmatic and paradigmatic uses of context. Do the parameters influence the *semantic* properties of the word spaces, or do they only have *statistical* effects that optimize a certain representation for a certain test? As should be obvious by now, the parameters *do* influence the results in these tests, but that does not necessarily mean that they influence the syntagmatic and paradigmatic information captured in the word spaces.

After having analyzed the effects of the parameters, I turn to a comparative analysis of the performance of the different word spaces. I then conclude the analysis with a brief review of related research.

15.1 The context region

Starting with syntagmatic uses of context, I have investigated two parameters for the production of syntagmatic word spaces: the size of the context regions, and the transformation of the occurrence counts. Table 15.1 shows the best-performing context region for each test and each evaluation metric. Note that it is often the same region that performs best for a particular test, regardless of which evaluation metric is used. The exceptions are the antonym test and part-of-speech test (**same** in the entry for the lax setting in the antonym test means that large and small regions produce the same top score).

Test	Strict	Lax
Thesaurus	large	large
Association	small	small
Synonym	large	
Antonym	large	**same**
PoS	small	large

Table 15.1: The best context regions for syntagmatic uses of context.

For four out of five tests, large regions produce the best result. This may be explained by the observation that using large regions provides a better statistical foundation than using small ones. Recall from Chapter 9 that the density measure for a word space produced with $[S : +]$ was 0.23, while it was only 0.08 for a word space produced with $[S : -]$. This difference is a clear indication of the statistical contrast between context regions of different sizes.

It is only for the association test and in the strict setting for the part-of-speech test that small context regions outperform the large ones. Considering the statistical incongruity just mentioned, this is an important result. Since the association test is the only test that primarily measures syntagmatic relations, while the other tests are primarily paradigmatic, this result strongly indicates that using small — more linguistically motivated — context regions yields more syntagmatic word spaces than using large ones.

I admit, however, that the discrepancy in the part-of-speech test is baffling. A possible reason is that this test is a thoroughly paradigmatic test, and thus cannot be expected to give precise measures of syntagmatic information.

15.2 Frequency transformations

Table 15.2 summarizes the best frequency transformations for the syntagmatic uses of context. As with the context regions, the best transformations are more or less consistent across the evaluation metrics. The exceptions are small context regions for the thesaurus comparison, large regions for the association test, and large regions for the part-of-speech test. For the thesaurus test with small context regions, binary counts produce the best result for the strict evaluation, but dampened frequency transformation produce the best result for the lax one; for the association test with large context regions, dampened frequency transformation produce the best result for the strict evaluation, but TFIDF-transformation produce the best score for the lax setting; for the part-of-speech test with large context regions, dampened frequency transformation produce the best results for the strict evaluation, while binary counts lead to better results for the lax setting.

Test	$[S:+]$		$[S:-]$	
	Strict	Lax	Strict	Lax
Thesaurus	TFIDF	TFIDF	BINARY	DAMPENED
Association	DAMPENED	TFIDF	BINARY	BINARY
Synonym	TFIDF		TFIDF	
Antonym	RAW	RAW	RAW	**all**
PoS	DAMPENED	BINARY	BINARY	BINARY

Table 15.2: The best frequency transformations for syntagmatic uses of context.

Note that binary counts tend to produce good results for the small regions, whereas TFIDF-transformations produce good results for the large context regions. This might not be all too surprising, considering the statistical differences between the different context regions. The binary counts used in these experiments use a single occurrence as cut-off point. It is arguable that this is sufficient for small context regions, but that one should use a higher cut-off point for large regions. The idea is that a single occurrence of a word in a large context region can be duly treated as a chance occurrence, and thus need not be taken into account (Katz, 1996). Instead, occurrences collected over large context regions can be discriminated by the use of a DF-related transformation. Such transformations have less effects when the context regions are many and small, since most words will occur in very few of these contexts. Finally, the superiority of raw frequency counts in the antonym test was explained in Chapter 13. To this can be added that the raw frequency counts performed substandard in all other tests.[1]

[1]Very few — if any — real-world systems use raw frequency-counting.

15.3 The context window

I have also used two basic parameters for the paradigmatic uses of context: the size of the context windows, and the weighting scheme for the positions in the windows. Table 15.3 demonstrates which window sizes produced the best results for each test and evaluation metric. It is evident that different window sizes are optimal for different tests; narrow windows spanning only a few surrounding words seem optimal for the thesaurus comparison, the synonym test, and the part-of-speech test, while wider windows seem optimal for the association and antonym tests. The optimal window size is consistent over the evaluation metrics.

Test	Strict	Lax
Thesaurus	narrow	narrow
Association	wide	wide
Synonym	narrow	
Antonym	wide	wide
PoS	narrow	narrow

Table 15.3: The best window sizes for paradigmatic uses of context.

It is interesting to note that narrow context windows yield better results than wide windows in both the synonym and part-of-speech tests, which are the most paradigmatic tests in these experiments. In the association test, on the contrary, wide windows prove to be better. This indicates that word spaces produced from a paradigmatic use of contexts become paradigmatically more refined when using narrow context windows. It should be noted that a minimal 1+1-sized window only yields the best result for paradigmatic uses of context in the part-of-speech test. This indicates that the immediate context of a word, which tends to consists of words from closed grammatical classes, is viable for determining the syntactic category of a word, but not as good as the immediately surrounding *content words* for inferring (semantic) paradigmatic relations between words. A 2+2 or 3+3-sized window seems more viable for that purpose.

15.4 Window weights

Table 15.4 demonstrates that weighting of the positions in the context windows is suboptimal for every test except the part-of-speech test, and the lax setting in the thesaurus comparison. In the latter test, constant weighting produces the best result for the strict evaluation (7.78%), but the difference to the top score using aggressive weights (7.77%) is minute. For the part-of-speech test, aggressive

weights clearly outperform the windows with constant weights, demonstrating the superiority of very narrow context windows in this task.

Test	Strict	Lax
Thesaurus	constant	aggressive
Association	constant	constant
Synonym	constant	
Antonym	constant	constant
PoS	aggressive	aggressive

Table 15.4: The best context regions for paradigmatic uses of context.

15.5 Comparison of contexts

This brief summary of the effects of different parameters for the syntagmatic and paradigmatic uses of context demonstrates that there is no single best parameter setting for all tests and all evaluation metrics. This does not necessarily mean that we are forced into the experimentalist fumbling usually favored in word-space research. Using the Saussurian refinement of the distributional hypothesis has made it possible to anticipate the effects of, e.g., the size of the context regions when collecting syntagmatic information, and the size of the context windows when collecting paradigmatic information. Admittedly, it does not provide answers to *all* questions we have about the effects of different parameters for the word-space representations. For example, the frequency transformations do not seem to have any obvious effects on the semantic properties of the word spaces, but they *do* influence the statistical properties of the representations. Such information-theoretic properties of word-space representations is the subject matter for an entire dissertation on its own.

Test	Strict	Lax
Thesaurus	syntagmatic	paradigmatic
Association	syntagmatic	syntagmatic
Synonym	paradigmatic	
Antonym	**same**	paradigmatic
PoS	paradigmatic	paradigmatic

Table 15.5: The best-performing uses of context.

Turning now to the arguably central aspect of the experiments: the comparison between syntagmatic and paradigmatic uses of context. Table 15.5 summarizes the best-performing context type for each test and each evaluation setting. Note that syntagmatic and paradigmatic uses of context are optimal for different tests: syntagmatic uses of context clearly outperform the paradigmatic uses in the association test, but the situation is reversed for the synonym and part-of-speech tests. The difference is not as clear-cut in the thesaurus comparison and the antonym test, where paradigmatic uses of context outperform the syntagmatic uses in the lax evaluation setting, but not in the strict setting (**same** in the table means they produce the same top score). The fact that certain tests show a greater distinction between syntagmatic and paradigmatic word spaces is a good indicator of the extent to which a certain representation contains syntagmatic versus paradigmatic information.

The synonym and part-of-speech tests are clearly paradigmatic tests, while the association test is more syntagmatic. The antonym test is more paradigmatic than syntagmatic, but as I argued in Chapter 13, certain antonyms have a tendency to become syntagmatic in large contexts, which could explain the viability of the syntagmatic use of context in this particular test. The thesaurus comparison is the least precise test in these experiments, and it involves a considerable amount of both syntagmatic and paradigmatic information. This is clearly reflected in the results.

Table 15.6 summarizes the tests used in this dissertation, the semantic relations they primarily measure, the degree to which the relations are essential to the test ($-$ and $+$), and the use of context that yields the best results in the strict evaluation settings.

Test	Relation	Context
Thesaurus comparison	both ($-$)	large region
Association test	syntagmatic ($+$)	small region
Synonym test	paradigmatic ($+$)	narrow window
Antonym test	paradigmatic ($-$)	wide window
Part-of-speech test	paradigmatic ($+$)	narrow window

Table 15.6: The tests used in this dissertation, the semantic relation they primarily measure, and the best-performing context.

I conclude this comparative analysis of the results from the experiments reported in the previous chapters with the observation that distributional information collected from co-occurrences in context regions seems to yield syntagmatic representations that are more refined when the regions are smaller, while distributional

information collected based on the neighboring words seems to yield paradigmatic representations that are more refined the smaller the context windows are.

15.6 Related research

The current investigation is not the first attempt to analyze word spaces in terms of syntagmatic and paradigmatic relations. Schütze & Pedersen (1993) compute syntagmatic and paradigmatic relations between words by using a directional words-by-words matrix similar to that used in HAL, which they transform using SVD. By using a directional matrix and SVD, they produce dimensionality-reduced vectors for both the left and right contexts of a word. The idea is then that paradigmatic similarity can be computed by comparing the left or the right context vectors, and that syntagmatic similarity can be computed by comparing the left *and* the right context vectors. The authors provide some examples of syntagmatic and paradigmatic pairs thus extracted but do not make any systematic evaluation of the approach, or of the differences between the vectors.

Rapp (2002) experiments with techniques to extract paradigmatic and syntagmatic relations between words. He uses narrow context windows for the paradigmatic relations, and a log-likelihood ratio to extract word pairs whose co-occurrence is significantly higher than chance for the syntagmatic relations. Rapp points out that using this kind of probabilistic approach for computing syntagmatic relations is far more efficient than using the word-space model. As evaluation of the extracted relations, he uses the TOEFL synonym test for the paradigmatic relations, and an association test (the Edinburgh Associative Thesaurus) for the syntagmatic ones.[2] However, he does not make any systematic quantitative comparison between the two different techniques.

The only other *experimental* investigation of how different contexts influence the word-space model that I am aware of is Lavelli et al. (2004), who compare what they call *document occurrence representation* (DOR) and *term co-occurrence representation* (TCOR). The former is termed DFITF, and is defined as:

$$\text{DFITF}(t_j, d_i) = \text{DF}(t_j, d_i) \cdot \frac{V}{V_d}$$

where V is the number of unique words in the data, and V_d is the number of unique words in the document. This representation corresponds to the syntagmatic uses of context in this dissertation, but with the difference that the weighting scheme uses something we could call *normalized word type count* instead of the traditional IDF function as second term in the equation.

[2]Rapp concurs in my analysis in Chapter 11 that association tests are not very precise, and that association norms contain *both* syntagmatic and paradigmatic relations.

The other kind of representation used by Lavelli et al. is called TFITF, and is defined as:

$$\text{TFITF}(t_i, t_j) = \text{TF}(t_i, t_j) \cdot \frac{V}{V_t}$$

where $\text{TF}(t_i, t_j)$ is the number of documents in which t_i and t_j co-occur, and V_t is the number of unique words that co-occur with t_i in at least one document. This representation corresponds to the paradigmatic uses of context in this dissertation, but with the difference that the context window is defined as an entire document, and the normalization is again a form of *normalized word type count*.

Lavelli et al. compare these two representations using term categorization and term clustering with WordNetDomains(42) as evaluation resource.[3] The results are very clear: the paradigmatic (TCOR) representations outperform the syntagmatic (DOR) ones in both tests. Lavelli et al. hypothesize that the supremacy of the paradigmatic representation can be explained by its ability to "capture some phenomena related to semantic similarity better" than the syntagmatic ones. They also claim that the paradigmatic representations are more discriminative (as measured using a function based on mutual information) than the syntagmatic ones, and therefore constitute "inherently better features."

Lavelli et al. are not the only ones who argue for the supremacy of paradigmatic uses of context. Stiles (1961) argues that syntagmatic uses of context only generate what he calls *statistical* relationships between terms, and that only a paradigmatic use of context can "project us beyond the purely statistical relationships and into the realm of meaningful associations." Similar claims are made by Rubenstein & Goodenough (1965), Grefenstette (1992b), and Charles (2000).

These claims that paradigmatic uses of context capture semantic similarity better than syntagmatic ones do not seem compatible with the results presented in the current investigation. As can be seen in Table 15.5 above, syntagmatic word spaces outperform paradigmatic ones in several test settings, and most notably in the association test. It is arguable that Lavelli's et al. experiment demonstrates the supremacy of paradigmatic uses of context for the semantic categorization task they defined, but that single experiment does not license the general conclusion that paradigmatic uses of context capture *semantic similarity* better than syntagmatic uses. Paradigmatic uses of context surely *do* capture one kind of semantic similarity better than syntagmatic uses, but the same thing can be said about the syntagmatic uses of context.

The experiments reported in this dissertation have shown that paradigmatic and syntagmatic uses of context generate representations that are viable for different types of semantic tests. Both types of context usage are therefore equally

[3] WordNetDomains(42) is a lexical resource that labels every word in WordNet (version 1.6) according to 42 general semantic categories (Avancini et al., 2003).

well-motivated, both empirically and theoretically. The choice of context usage is a matter of what type of semantic relation one wants to model: the syntagmatic or the paradigmatic type.

15.7 The semantic continuum

It is important to understand that the difference between syntagmatic and paradigmatic word spaces is not discrete. It is not a question whether a word space contains syntagmatic *or* paradigmatic relations, but rather *to what extent* a word space contains these relations. As we have seen in the experiments reported in this dissertation, word spaces produced with syntagmatic uses of context and word spaces produced with paradigmatic uses of context have a small — but still noticeable — overlap. We have also seen that certain contexts (i.e. small context regions) yield more syntagmatic word spaces, while other contexts (i.e. narrow context windows) yield more paradigmatic spaces. The difference between word spaces produced with different types of contexts is more like a semantic continuum, where syntagmatic and paradigmatic relations represent the extremities.

It is an interesting question whether it would be possible to accumulate a *purely* syntagmatic and a *purely* paradigmatic word space. However, even if we could hypothesize a method for producing such refined spaces, we would still be faced with the problem of how to evaluate them. As I noted in Chapter 8, the experiments used in this dissertation do not provide *exact* measures of syntagmatic and paradigmatic relations. It is another interesting question whether it is at all possible to construct tests that measure *only* syntagmatic or paradigmatic information. I leave these as open questions.

Part IV

Curtain call

Chapter 16

Conclusion

"Come on, say something conclusive!"
(Homer Simpson in "Sleeping with the Enemy")

This dissertation has discussed the word-space model. More specifically, it has discussed what kind of meaning is modeled in the word space. The outcome of this discussion has been an identification of two different types of context usage, and an empirical investigation of the different word spaces they produce. The question **what kind of semantic information does the word-space model acquire and represent?** has been answered by *syntagmatic or paradigmatic information, depending on how the context is used.*

This means that we have now reached the end of the line. The theoretical foundations of the word-space model have been charted, and its empirical viability scrutinized. It is time to summarize and conclude this investigation of the semantic properties of the word-space model. First, a quick recapitulation of the highlights from the previous 130 pages.

16.1 Flashbacks

The background-chapters discussed the underlying theoretical assumptions (the geometric metaphor of meaning and the distributional hypothesis), the general word-space methodology (the notion of context vectors), and the most common implementations of word-space models (LSA, HAL and RI). I then discussed the problem how to evaluate word spaces, and concluded that we need a theory of meaning in order to be able to determine how good a model of meaning a given word space is. By excavating the origins of the distributional paradigm (the *Cours de linguistique générale*), I could infer such a theory of meaning in the form of

a structuralist dichotomy of syntagma and paradigm. This in its turn enabled an identification and characterization of two different types of uses of context for producing context vectors: syntagmatic use and paradigmatic use.

The foreground-chapters presented a number of experiments demonstrating the differences between syntagmatic and paradigmatic uses of context. The first experiment investigated the overlap between syntagmatic and paradigmatic word spaces, and showed that only a few percentage of the nearest neighbors occur in both types of word spaces. The rest of the experiments compared syntagmatic and paradigmatic word-spaces to thesaurus entries, association norms, a synonym test, a list of antonym pairs, and a record of part-of-speech assignments. All these tests showed a consistent difference between word spaces produced through syntagmatic uses of context and word spaces produced through paradigmatic uses of context. The last of the foreground-chapters analyzed and scrutinized the results from the experiments, and also provided a brief review of related research.

16.2 Summary of the results

On the theoretical side, I have argued that the word-space model uses the geometric metaphor of meaning as representational basis, and the distributional methodology as discovery procedure. I have argued that the word-space model is a model of semantic similarity, and that it is based on a structuralist meaning theory that posits two fundamental types of semantic similarity relations between words: syntagmatic and paradigmatic relations.

The empirical results presented in the third part of this dissertation have shown that syntagmatic and paradigmatic uses of context yield word spaces with different semantic properties. Syntagmatic uses of context yield more syntagmatic word spaces, while paradigmatic uses of context yield more paradigmatic spaces. The experiments have also shown that using a small context region yields more syntagmatic spaces than using a large one, and that using a narrow context window yields more paradigmatic spaces than using a wide window.

16.3 Answering the questions

I posed a couple of questions in Chapter 1, which I will now answer:

- **Is it at all possible to extract semantic knowledge by merely looking at usage data?** Clearly: yes. Although it is difficult to assess the absolute quality of the results from the experiments reported in this dissertation, they clearly demonstrate that the word-space model is capable of acquiring

semantic information from text data. I also reviewed other evidence of its viability in Chapter 5.

- **Does the word-space model constitute a *complete* model of the *full* spectrum of meaning, or does it only convey *specific aspects* of meaning?** The word-space model constitutes a complete model of meaning, if what we mean by "meaning" is a structuralist dichotomy of syntagma and paradigm. The answer to this question thoroughly depends on our meaning theory; if we believe that meaning is essentially referential, then the answer will be very different.

16.4 Contributions

The main contributions of this dissertation are the following:

- An identification and discussion of the three main theoretical cornerstones of the word-space model: the geometrical metaphor of meaning, the distributional methodology, and the structuralist meaning theory.
- A critical discussion of word-space evaluations, and the conclusion that we need a theory of meaning in order to devise semantically valid tests.
- The identification of syntagmatic and paradigmatic uses of context.
- **The single most important contribution of this dissertation is the observation that syntagmatic and paradigmatic relations can be modeled by different uses of context when accumulating context vectors.**

16.5 The word-space model revisited

I stated in the Introduction that I wanted to examine how far in our quest for meaning the word-space model can take us. As I have argued in this dissertation, that depends on what we mean by "meaning." Admittedly, such an answer is a bit like ducking the question. I therefore feel obliged to discuss a few remaining concerns before concluding this dissertation on the word-space model.

We have seen that a word space can be more or less syntagmatic or paradigmatic to different degrees, and that the difference between a syntagmatic and a paradigmatic word space is not discrete. On the contrary, word spaces are semantic continua, in which these two types of relations coexist and interact. However, we have (at present) no idea of what the internal structure of the spaces looks like —

maybe word spaces have a common internal structure that can be utilized for different purposes, e.g. to differentiate between different *types* of relations within the word space; or maybe word spaces really *do* have a discoverable "latent" dimensionality, as suggested by proponents of factor analytical dimensionality-reduction techniques. The density measure introduced in Chapter 9 is one attempt at digging a bit deeper into the fundament of word spaces, but we have barely begun the scratch the surface on the internal properties of high-dimensional word spaces.

Furthermore, this dissertation has only been concerned with *word meaning*. However, we would expect of a *complete* model of meaning to be able to handle phrase-, clause-, sentence-, paragraph-, document- and text-meaning too. This brings the question of compositionality to the fore. Now, I am well aware that the subject of compositionality is not without controversy in the philosophy of language. I will not make a stand in this debate here, but merely ask whether the word-space model *can* handle such phenomena? Recall from Chapter 5 that word-space representations have been used for building text representations for a number of tasks, including text categorization and information retrieval, and that this is normally done by simply combining the context vectors of the words in the text. The algebraic nature of the word spaces makes combining context vectors very straight-forward: we can simply use vector addition for producing centroid vectors. However, the fact that we *can*, and normally *do*, use vector addition for producing compositional representations based on context vectors does not mean that it is the *best* possible way to proceed. Combining vectors by vector addition effectively destroys the uniqueness of individual context vectors, and produces an average representation from several, possibly unrelated, distributional profiles. If anything, it is surprising that such a blunt use of vector spaces actually *does* work at times. Much more research is needed on how to best utilize context vectors for producing compositional text representations.

Regardless of the viability of the word-space model for producing compositional representations, it is currently one of the most attractive alternatives within language technology for producing computational models of word meaning. Having said that, I must admit that I have omitted to discuss one of the most unique and attractive properties of the word-space model: its flexibility and ability to continuously evolve over time. It would be greatly interesting to investigate how a word space evolves when subjected to a continuous data flow. I submit this topic as a suggestion for future dissertations.

16.6 Beyond the linguistic frontier

It cannot be stressed enough that the word-space model is a *computational* model of meaning, and *not* a psychologically realistic model of human semantic process-

ing. The only information utilized by the word-space model is linguistic context, which, as we have seen in this dissertation, is a (theoretically and empirically) plausible methodology for acquiring computational representations of structuralist word meaning. However, it is arguable that *human* language users also use *extra*linguistic context when learning, understanding, and using language. It is also arguable — in fact, both Saussure and Harris would insist — that human conceptual meaning involves much more than the structuralist dichotomy of syntagmatic and paradigmatic relations.

The extralinguistic evidence is not available to the word-space model. Viewed with the pretense of creating a cognitively realistic model of human semantic knowledge, this inability to reach beyond the limits of textuality is the single most disqualifying feature of the word-space model. This is perhaps most obvious with regards to the referential aspect of meaning; even though the word-space model might correctly associate "guitar" with "bass" and "lute," it will not be able to reach into the world and pick out the right object.

In order for the word-space model to qualify as a model of human semantic processing, it needs to reach beyond the linguistic frontier into the realms of the extralinguistic world, and to include extralinguistic context in its representation. I do not believe that this is impossible in principle, although I do believe that it would require a radical innovation in how we define and use context for accumulating context vectors. Until that breakthrough, we should be wary about claims for cognitive plausibility.

I conclude this dissertation with a small advise to prospective word-space theorists. Whenever someone claims that their instantiation of the word-space model actually constitutes a viable model of human semantic processing, ask "how many senses does your model have?"

Bibliography

Avancini, H., Lavelli, A., Magnini, B., Sebastiani, F., & Zanoli, R. (2003). Expanding domain-specific lexicons by term categorization. In *Proceedings of the 2003 ACM Symposium on Applied Computing, SAC'03* (pp. 793–797). New York, NY, USA: ACM Press.

Baker, D., & McCallum, A. (1998). Distributional clustering of words for text classification. In *Proceedings of the 21th ACM International Conference on Research and Development in Information Retrieval, SIGIR'98* (pp. 96–103).

Berry, M., Dumais, S., & O'Brien, G. (1995). Using linear algebra for intelligent information retrieval. *SIAM Review*, *37*, 573–595.

Bingham, E. (2003). *Advances in independent component analysis with applications to data mining*. PhD Dissertation, Helsinki University of Technology, Department of Computer Science and Engineering.

Bingham, E., & Mannila, H. (2001). Random projection in dimensionality reduction: applications to image and text data. In *Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD'01* (pp. 245–250).

Black, E. (1988). An experiment in computational discrimination of english word senses. *IBM J. Res. Dev.*, *32*(2), 185–194.

Brill, E. (1994). Some advances in transformation-based part of speech tagging. In *Proceedings of the 12th national Conference on Artificial Intelligence, AAAI'94* (pp. 722–727).

Burgess, C., Livesay, K., & Lund, K. (1998). Explorations in context space: words, sentences, discourse. *Discourse Processes*, *25*, 211–257.

Caid, W., Dumais, S., & Gallant, S. (1995). Learned vector-space models for document retrieval. *Information Processing and Management*, *31*(3), 419–429.

Charles, W. (2000). Contextual correlates of meaning. *Applied Psycholinguistics*, *21*, 505–524.

Chávez, E., & Navarro, G. (2000). *Measuring the dimensionality of general metric spaces* (Tech. Rep. No. TR/DCC-2000-1). Department of Computer Science, University of Chile.

Choueka, Y., & Lusignan, S. (1985). Disambiguation by short contexts. *Computers and the Humanities*, *19*(3), 147–158.

Church, K., & Hanks, P. (1989). Word association norms, mutual information, and lexicography. In *Proceedings of the 27th Conference on Association for Computational Linguistics, ACL'89* (pp. 76–83). Association for Computational Linguistics.

Cottrell, G., & Small, S. (1983). A connectionist scheme for modeling word sense disambiguation. *Cognition and Brain Theory*, *6*, 89–120.

Dagan, I., Marcus, S., & Markovitch, S. (1993). Contextual word similarity and estimation from sparse data. In *Proceedings of the 31st Iinternational Conference on Association for Computational Linguistics, ACL'93* (pp. 164–171). Association for Computational Linguistics.

Damerau, F. (1965). An experiment in automatic indexing. *American Documentation*, *16*, 283–289.

Deerwester, S., Dumais, S., Furnas, G., Landauer, T., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the Society for Information Science*, *41*(6), 391–407.

Deerwester, S., Dumais, S., Landauer, T., Furnas, G., & Beck, L. (1988). Improving information retrieval with latent semantic indexing. In *Proceedings of the 51st Annual Meeting of the American Society for Information Science* (Vol. 25, pp. 36–40). Atlanta, Georgia: Learned Information, Inc.

Deese, J. (1964). The associative structure of some common english adjectives. *Journal of Verbal Learning and Verbal Behavior*, *3*(5), 347–357.

Dumais, S. (1993). Lsi meets trec: A status report. In D. Harman (Ed.), *Proceedings of the first Text REtrieval Conference, TREC1* (pp. 137–152).

Dumais, S. (2004). Latent semantic analysis. *Annual Review of Information Science and Technology (ARIST)*, *38*, 189–230.

Dumais, S., Furnas, G., Landauer, T., Deerwester, S., & Harshman, R. (1988). Using latent semantic analysis to improve access to textual information. In *Proceedings of the Conference on Human Factors in Computing Systems, CHI'88* (pp. 281–285). New York, USA.

Dumais, S., Letsche, T., Littman, M., & Landauer, T. (1997). Automatic cross-language retrieval using latent semantic indexing. In *Proceedings of the AAAI Symposium on Cross-Language Text and Speech Retrieval.*

Gale, W., Church, K., & Yarowsky, D. (1994). Discrimination decisions for 100,000-dimensional spaces. In A. Zampoli, N. Calzolari, & M. Palmer (Eds.), *Current issues in Computational Linguistics: In honour of Don Walker* (pp. 429–450). Kluwer Academic Publishers.

Gallant, S. (1991a). Context vector representations for document retrieval. In *AAAI Natural Language Text Retrieval Workshop.*

Gallant, S. (1991b). A practical approach for representing context and performing

word sense disambiguation using neural networks. *Neural Computation*, *3*(3), 293–309.

Gallant, S.(2000). Context vectors: A step toward a "grand unified representation". In *Hybrid Neural Systems, revised papers from a workshop* (pp. 204–210). London, UK: Springer-Verlag.

Gärdenfors, P.(2000). *Conceptual spaces: The geometry of thought*. Massachusetts: The MIT Press.

Grefenstette, G. (1992a). Finding semantic similarity in raw text: The Deese antonyms. In *Working notes of the AAAI Fall Symposium on Probabilistic Approaches to Natural Language* (pp. 61–65). AAAI Press.

Grefenstette, G. (1992b). Sextant: exploring unexplored contexts for semantic extraction from syntactic analysis. In *Proceedings of the 30th Annual Meeting on Association for Computational Linguistics, ACL'92* (pp. 324–326). Morristown, NJ, USA: Association for Computational Linguistics.

Grefenstette, G.(1993). Evaluation techniques for automatic semantic extraction: Comparing syntactic and window-based approaches. In *Workshop on Acquisition of Lexical Knowledge from Text*. Columbus, OH, USA.

Grefenstette, G. (1994). *Explorations in automatic thesaurus discovery*. Norwell, MA, USA: Kluwer Academic Publishers.

Harris, R.(2001). *Saussure and his interpreters*. New York University Press.

Harris, Z.(1968). *Mathematical structures of language*. Interscience Publishers.

Harris, Z. (1970). Distributional structure. In *Papers in structural and transformational Linguistics* (pp. 775–794).

Harter, S.(1975). A probabilistic approach to automated keyword indexing. *Journal of the American Society for Information Science*, *26*, 280–289.

Hearst, M. (1992). Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th International Conference on Computational Linguistics, COLING'92* (pp. 539–545). Morristown, NJ, USA: Association for Computational Linguistics.

Henley, N.(1969). A psychological study of the semantics of animal terms. *Journal of Verbal Learning and Verbal Behavior*, *8*, 176–184.

Hindle, D. (1990). Noun classification from predicate-argument structures. In *Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics, ACL'90* (pp. 268–275). Morristown, NJ, USA: Association for Computational Linguistics.

Husbands, P., Simon, H., & Ding, C.(2001). On the use of the singular value decomposition for text retrieval. In *Computational information retrieval* (pp. 145–156). Philadelphia, PA, USA: Society for Industrial and Applied Mathematics.

Ide, N., & Veronis, J. (1995). Large neural networks for the resolution of lexi-

cal ambiguity. In P. Saint-Dizier & E. Viegas (Eds.), *Computational lexical semantics* (pp. 251–270). Cambridge University Press.

Isbell, C., & Viola, P. (1998). Restructuring sparse high dimensional data for effective retrieval. In *Proceedings of the 1998 Conference on Advances in Neural Information Processing Systems, NIPS'98* (pp. 480–486). Cambridge, MA, USA: MIT Press.

Jiang, F., & Littman, M. (2000). Approximate dimension equalization in vector-based information retrieval. In *Proceedings of the 17th International Conference on Machine Learning, ICML'00* (pp. 423–430). Morgan Kaufmann, San Francisco, CA.

Johnson, W., & Lindenstrauss, J. (1984). Extensions of lipshitz mapping into hilbert space. *Contemporary Mathematics*, *26*, 189–206.

Kanerva, P. (1988). *Sparse distributed memory*. Cambridge, MA, USA: MIT Press.

Kanerva, P., Kristofersson, J., & Holst, A. (2000). Random indexing of text samples for latent semantic analysis. In *Proceedings of the 22nd Annual Conference of the Cognitive Science Society, CogSci'00* (p. 1036). Erlbaum.

Kaplan, A. (1955). An experimental study of ambiguity and context. *Mechanical Translation*, *2*(2), 39–46.

Karlgren, J. (1999). Stylistic experiments in information retrieval. In T. Strzalkowski (Ed.), *Natural language information retrieval* (pp. 147–166). Kluwer.

Karlgren, J. (2005). Meaningful models for information access systems. In Arppe et al. (Eds.), *A finnish computer linguist: Kimmo Koskenniemi festschrift on the 60th birthday* (pp. 261–268). CSLI Publications.

Karlgren, J., & Sahlgren, M. (2001). From words to understanding. In Y. Uesaka, P. Kanerva, & H. Asoh (Eds.), *Foundations of real-world intelligence* (pp. 294–308). CSLI Publications.

Karlgren, J., Sahlgren, M., & Cöster, R. (2005). Principled query processing. In C. Peters (Ed.), *Working notes of the 6th workshop of the Cross-Language Evaluation Forum, CLEF'05, Vienna, Austria, september 21-23, 2005.*

Karlgren, J., Sahlgren, M., Järvinen, T., & Cöster, R. (2005). Dynamic lexica for query translation. In C. Peters, P. Clough, J. Gonzalo, G. Jones, M. Kluck, & B. Magnini (Eds.), *Multilingual information access for text, speech and images, 5th workshop of the Cross-Language Evaluation Forum, CLEF'04, Bath, UK, september 15-17, 2004, revised selected papers* (pp. 150–155). Springer.

Kaski, S. (1999). Dimensionality reduction by random mapping: Fast similarity computation for clustering. In *Proceedings of the International Joint Conference on Neural Networks, IJCNN'98* (pp. 413–418). IEEE Service Center.

Katz, S. (1996). Distribution of content words and phrases in text and language modelling. *Natural Language Engineering*, *2*(1), 15–60.

Kohonen, T. (1995). *Self-organizing maps*. Berlin, Heidelberg: Springer.

Lakoff, G., & Johnson, M. (1980). *Metaphors we live by.* Chicago: University of Chicago Press.

Lakoff, G., & Johnson, M. (1999). *Philosophy in the flesh: The embodied mind and its challenge to western thought.* New York: Basic Books.

Landauer, T., & Dumais, S. (1997). A solution to plato's problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge. *Psychological Review*, *104* (2), 211–240.

Landauer, T., Foltz, P., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, *25*, 259–284.

Lavelli, A., Sebastiani, F., & Zanoli, R. (2004). Distributional term representations: an experimental comparison. In *Proceedings of the 13th ACM Conference on Information and Knowledge Management, CIKM'04* (pp. 615–624). New York, NY, USA: ACM Press.

Leacock, C., Towell, G., & Voorhees, E. (1996). Towards building contextual representations of word senses using statistical models. In B. Boguraev & J. Pustejovsky (Eds.), *Corpus processing for lexical acquisition* (pp. 97–113). Cambridge, MA, USA: MIT Press.

Lee, L. (1999). Measures of distributional similarity. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics, ACL'99* (pp. 25–32).

Lemaire, B., & Denhiere, G. (2004). Incremental construction of an associative network from a corpus. In *Proceedings of the 26th Annual Meeting of the Cognitive Science Society, CogSci'04* (pp. 825–830).

Levy, J., Bullinaria, J., & Patel, M. (1998). Explorations in the derivation of word co-occurrence statistics. *South Pacific Journal of Psychology*, *10* (1), 99-111.

Lin, D. (1997). Using syntactic dependency as local context to resolve word sense ambiguity. In *Proceedings of the 35th Annual meeting of the Association for Computational Linguistics, ACL'97.*

Lin, D. (1998a). Automatic retrieval and clustering of similar words. In *Proceedings of the 17th International Conference on Computational Linguistics, COLING-ACL'98* (pp. 768–774). Morristown, NJ, USA: Association for Computational Linguistics.

Lin, D. (1998b). An information-theoretic definition of similarity. In *Proceedings of the 15th International Conference on Machine Learning, ICML'98* (pp. 296–304).

Lowe, W. (2000). What is the dimensionality of human semantic space. In *Proceedings of the 6th Neural Computation and Psychology Workshop* (pp. 303–311). Springer Verlag.

Luhn, H. (1958). The automatic creation of literature abstracts. *IBM Journal of Research and Development*, *2* (2), 159–165.

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces

from lexical co-occurrence. *Behavior Research Methods, Instrumentation, and Computers*, *28*, 203–208.

Lund, K., Burgess, C., & Atchley, R. (1995). Semantic and associative priming in high-dimensional semantic space. In *Proceedings of the 17th Annual Conference of the Cognitive Science Society, CogSci'95* (pp. 660–665). Erlbaum.

McDonald, S., & Lowe, W. (1998). Modelling functional priming and the associative boost. In *Proceedings of the 20th Annual Conference of the Cognitive Science Society, CogSci'98* (pp. 675–680).

McDonald, S., & Ramscar, M. (2001). Testing the distributional hypothesis: The influence of context on judgements of semantic similarity. In *Proceedings of the 23rd Annual Conference of the Cognitive Science Society, CogSci'01* (pp. 611–616).

Miller, G., & Charles, W. (1991). Contextual correlates of semantic similarity. *Language and Cognitive Processes*, *6*(1), 1–28.

Miller, G., & Leacock, C. (1998). Lexical representations for sentence processing. In Y. Ravin & C. Leacock (Eds.), *Polysemy: Theoretical and computational approaches* (pp. 152–160). Oxford University Press.

Nakov, P., Popova, A., & Mateev, P. (2001). Weight functions impact on lsa performance. In *Proceedings of the EuroConference Recent Advances in Natural Language Processing, RANLP'01* (pp. 187–193). Tzigov Chark, Bulgaria.

Nelson, D., McEvoy, C., & Schreiber, T. (1998). *The University of South Florida word association, rhyme, and word fragment norms.* (http://w3.usf.edu/FreeAssociation/)

Nevin, B. (1993). A minimalist program for linguistics: The work of Zellig Harris on meaning and information. *Historiographia Linguistica*, *20*(2/3), 355–398.

Niwa, Y., & Nitta, Y. (1994). Co-occurrence vectors from corpora vs. distance vectors from dictionaries. In *Proceedings of the 15th International Conference on Computational Linguistics, COLING'94* (pp. 304–309). Association for Computational Linguistics.

Osgood, C. (1952). The nature and measurement of meaning. *Psychological Bulletin*, *49*, 197–237.

Osgood, C., Suci, G., & Tannenbaum, P. (1957). *The measurement of meaning.* University of Illinois Press.

Padó, S., & Lapata, M. (2003). Constructing semantic space models from parsed corpora. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics, ACL'03* (pp. 128–135).

Papadimitriou, C., Raghavan, P., Tamaki, H., & Vempala, S. (1998). Latent semantic indexing: A probabilistic analysis. In *Proceedings of the 17th ACM Symposium on the Principles of Database Systems* (pp. 159–168). ACM Press.

Pereira, F., Tishby, N., & Lee, L. (1993). Distributional clustering of english words. In *Proceedings of the 30th Annual Meeting of the Association for*

Computational Linguistics, ACL'93 (pp. 183–190).

Picard, J. (1999). Finding content-bearing terms using term similarities. In *Proceedings of the 9th Conference on European chapter of the Association for Computational Linguistics, EACL'99* (pp. 241–244). Morristown, NJ, USA: Association for Computational Linguistics.

Qiu, Y., & Frei, H.-P. (1993). Concept-based query expansion. In *Proceedings of the 16th ACM International Conference on Research and Development in Information Retrieval, SIGIR'93* (pp. 160–169). Pittsburgh, USA.

Rapp, R. (2002). The computation of word associations: comparing syntagmatic and paradigmatic approaches. In *Proceedings of the 19th International Conference on Computational Linguistics, COLING'02* (pp. 1–7). Morristown, NJ, USA: Association for Computational Linguistics.

Rapp, R. (2003). Word sense discovery based on sense descriptor dissimilarity. In *Proceedings of MT Summit IX* (pp. 315–322).

Rehder, B., Schreiner, M., Wolfe, M., Laham, D., Landauer, T., & Kintsch, W. (1998). Using latent semantic analysis to assess knowledge: Some technical considerations. *Discourse Processes*, *25*, 337–354.

Robertson, S., & Spärck Jones, K. (1997). *Simple, proven approaches to text retrieval* (Technical report No. 356). Computer Laboratory, University of Cambridge.

Rubenstein, H., & Goodenough, J. (1965). Contextual correlates of synonymy. *Communications of the ACM*, *8*(10), 627–633.

Ruge, G. (1992). Experiments on linguistically-based term associations. *Information Processing and Management*, *28*(3), 317–332.

Sahlgren, M. (2004). Automatic bilingual lexicon acquisition using random indexing of aligned bilingual data. In *Proceedings of the 4th International Conference on Language Resources and Evaluation, LREC'04* (pp. 1289–1292).

Sahlgren, M. (2005). An introduction to random indexing. In H. Witschel (Ed.), *Methods and Applications of Semantic Indexing Workshop at the 7th International Conference on Terminology and Knowledge Engineering, TKE'05, Copenhagen, Denmark, august 16, 2005* (Vol. 87).

Sahlgren, M. (2006). Towards pertinent evaluation methodologies for word-space models. In *Proceedings of the 5th International Conference on Language Resources and Evaluation, LREC'06.*

Sahlgren, M., & Cöster, R. (2004). Using bag-of-concepts to improve the performance of support vector machines in text categorization. In *Proceedings of the 20th International Conference on Computational Linguistics, COLING'04* (pp. 487–493).

Sahlgren, M., & Karlgren, J. (2002). Vector-based semantic analysis using random indexing for cross-lingual query expansion. In C. Peters, M. Braschler, J. Gonzalo, & M. Kluck (Eds.), *Evaluation of cross-language information*

retrieval systems, 2nd workshop of the Cross-Language Evaluation Forum, CLEF'01, Darmstadt, Germany, september 3-4, 2001, revised papers (pp. 169–176). Springer.

Sahlgren, M., & Karlgren, J. (2005a). Automatic bilingual lexicon acquisition using random indexing of parallel corpora. *Journal of Natural Language Engineering*, *11*(3), 327–341.

Sahlgren, M., & Karlgren, J. (2005b). Counting lumps in word space: Density as a measure of corpus homogeneity. In *Proceedings of the 12th Symposium on String Processing and Information Retrieval, SPIRE'05.*

Sahlgren, M., Karlgren, J., Cöster, R., & Järvinen, T. (2003). SICS at CLEF 2002: Automatic query expansion using random indexing. In C. Peters, M. Braschler, J. Gonzalo, & M. Kluck (Eds.), *Advances in cross-language information retrieval, 3rd workshop of the Cross-Language Evaluation Forum, CLEF'02. Rome, Italy, september 19-20, 2002, revised papers* (pp. 311–320). Springer.

Sahlgren, M., Karlgren, J., & Hansen, P. (2002). English-japanese cross-lingual query expansion using random indexing of aligned bilingual text data. In K. Oyama, E. Ishida, & N. Kando (Eds.), *Proceedings of the 3rd NTCIR workshop on research in information retrieval, automatic text summarization and question answering.* Tokyo, Japan: National Institute of Informatics, NII.

Salton, G., & McGill, M. (1983). *Introduction to modern information retrieval.* McGraw-Hill.

Salton, G., & Yang, C. (1973). On the specification of term values in automatic indexing. *Documentation*, *29*, 351–372.

Saussure, F. (1916/1983). *Course in general Linguistics.* Duckworth. (Translated by Roy Harris)

Schütze, H. (1992). Dimensions of meaning. In *Proceedings of the 1992 ACM/IEEE Conference on Supercomputing, Supercomputing'92* (pp. 787–796). IEEE Computer Society Press.

Schütze, H. (1993). Word space. In *Proceedings of the 1993 Conference on Advances in Neural Information Processing Systems, NIPS'93* (pp. 895–902). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Schütze, H. (1998). Automatic word sense discrimination. *Computational Linguistics*, *24*(1), 97–123.

Schütze, H., & Pedersen, J. (1993). A vector model for syntagmatic and paradigmatic relatedness. In *Making sense of words* (pp. 104–113). Oxford, England: Ninth Annual Conference of the UW Centre for the New OED and Text Research.

Schütze, H., & Pedersen, J. (1995). Information retrieval based on word senses. In *Proceedings of the 4th Annual Symposium on Document Analysis and*

Information Retrieval (pp. 161–175).

Schütze, H., & Pedersen, J. (1997). A cooccurrence-based thesaurus and two applications to information retrieval. *Information Processing and Management*, *33*(3), 307–318.

Singhal, A., Buckley, C., & Mitra, M. (1996). Pivoted document length normalization. In *Proceedings of the 19th Annual International ACM Conference on Research and Development in Information Retrieval, SIGIR'96* (pp. 21–29).

Small, S., Cottrell, G., & Tanenhaus, M. (Eds.). (1988). *Lexical ambiguity resolution: Perspectives from psycholinguistics, neuropsychology, and artificial intelligence*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Smith, E., & Medin, D. (1981). *Categories and concepts*. Cambridge, MA, USA: Harvard University Press.

Spärck Jones, K. (1972). Exhaustivity and specificity. *Journal of Documentation*, *28*, 11–21.

Stiles, H. (1961). The association factor in information retrieval. *Journal of the ACM*, *8*(2), 271–279.

Strzalkowski, T. (1994). Building a lexical domain map from text corpora. In *Proceedings of the 15th International Conference on Computational Linguistics, COLING'94* (pp. 604–610).

Turney, P. (2001). Mining the Web for synonyms: PMI–IR versus LSA on TOEFL. *Lecture Notes in Computer Science*, *2167*, 491–502.

Vapnik, V. (1995). *The nature of statistical learning theory*. Springer.

Waltz, D., & Pollack, J. (1985). Massively parallel parsing: a strongly interactive model of natural language interpretation. *Cognitive Science*, *9*(1), 51–74.

Weeds, J., Weir, D., & McCarthy, D. (2004). Characterising measures of lexical distributional similarity. In *Proceedings of the 20th International Conference of Computational Linguistics, COLING'04* (pp. 1015–1021).

Widdows, D. (2003, June). Unsupervised methods for developing taxonomies by combining syntactic and statistical information. In *Proceedings of HLT/NAACL* (pp. 276–283).

Widdows, D. (2004). *Geometry and meaning*. Stanford, USA: CSLI Publications.

Wiemer-Hastings, P., & Zipitria, I. (2001). Rules for syntax, vectors for semantics. In *Proceedings of the 23rd Annual Conference of the Cognitive Science Society, CogSci'01* (pp. 1112–1117). Mahwah, NJ: Erlbaum.

Wilks, Y., Fass, D., Guo, C.-M., McDonald, J., Plate, T., & Slator, B. (1990). Providing machine tractable dictionary tools. *Machine Translation*, *5*(2), 99–151.

Wittgenstein, L. (1953). *Philosophical investigations*. Blackwell. (Translated by G.E.M. Anscombe)

Wolfe, M., Schreiner, M., Rehder, B., Laham, D., Foltz, P., Kintsch, W., et al.

(1998). Learning from text: Matching readers and text by latent semantic analysis. *Discourse Processes*, *25*, 309–336.

Yarowsky, D.(1992). Word-sense disambiguation using statistical models of roget's categories trained on large corpora. In *Proceedings of the 14th Iinternational Conference on Computational Linguistics, COLING'92* (pp. 454–460). Association for Computational Linguistics.

Zipf, G.(1949). *Human behavior and the principle of least-effort*. Cambridge, MA: Addison-Wesley.

SWEDISH INSTITUTE OF COMPUTER SCIENCE
SICS Dissertation Series

01: Bogumil Hausman
Pruning and Speculative Work in OR-Parallel PROLOG, 1990.
02: Mats Carlsson
Design and Implementation of an OR-Parallel Prolog Engine, 1990.
03: Nabiel A. Elshiewy
Robust Coordinated Reactive Computing in SANDRA, 1990.
04: Dan Sahlin
An Automatic Partial Evaluator for Full Prolog, 1991.
05: Hans A. Hansson
Time and Probability in Formal Design of Distributed Systems, 1991.
06: Peter Sjödin
From LOTOS Specifications to Distributed Implementations, 1991.
07: Roland Karlsson
A High Performance OR-parallel Prolog System, 1992.
08: Erik Hagersten
Toward Scalable Cache Only Memory Architectures, 1992.
09: Lars-Henrik Eriksson
Finitary Partial Inductive Definitions and General Logic, 1993.
10: Mats Björkman
Architectures for High Performance Communication, 1993.
11: Stephen Pink
Measurement, Implementation, and Optimization of Internet Protocols, 1993.
12: Martin Aronsson
GCLA. The Design, Use, and Implementation of a Program Development System, 1993.
13: Christer Samuelsson
Fast Natural-Language Parsing Using Explanation-Based Learning, 1994.
14: Sverker Jansson
AKL — A Multiparadigm Programming Language, 1994.
15: Fredrik Orava
On the Formal Analysis of Telecommunication Protocols, 1994.
16: Torbjörn Keisu
Tree Constraints, 1994.
17: Olof Hagsand
Computer and Communication Support for Interactive Distributed Applications, 1995.
18: Björn Carlsson
Compiling and Executing Finite Domain Constraints, 1995.
19: Per Kreuger
Computational Issues in Calculi of Partial Inductive Definitions, 1995.
20: Annika Waern
Recognising Human Plans: Issues for Plan Recognition in Human-Computer Interaction, 1996.
21: Björn Gambäck
Processing Swedish Sentences: A Unification-Based Grammar and Some Applications. June 1997.
22: Klas Orsvärn
Knowledge Modelling with Libraries of Task Decomposition Methods, 1996.
23: Kia Höök
A Glass Box Approach to Adaptive Hypermedia, 1996.
24: Bengt Ahlgren
Improving Computer Communication Performance by Reducing Memory Bandwidth Consumption, 1997.

SWEDISH INSTITUTE OF COMPUTER SCIENCE
SICS Dissertation Series

25: Johan Montelius
Exploiting Fine-grain Parallelism in Concurrent Constraint Languages, May, 1997.
26: Jussi Karlgren
Stylistic experiments in information retrieval, 2000
27: Ashley Saulsbury
Attacking Latency Bottlenecks in Distributed Shared Memory Systems, 1999.
28: Kristian Simsarian
Toward Human Robot Collaboration, 2000.
29: Lars-åke Fredlund
A Framework for Reasoning about Erlang Code, 2001.
30: Thiemo Voigt
Architectures for Service Differentiation in Overloaded Internet Servers, 2002.
31: Fredrik Espinoza
Individual Service Provisioning, 2003.
32: Lars Rasmusson
Network capacity sharing with QoS as a financial derivative pricing problem: algorithms and network design, 2002.
33: Martin Svensson
Defining, Designing and Evaluating Social Navigation, 2003.
34: Joe Armstrong
Making reliable distributed systems in the presence of software errors, 2003.
35: Emmanuel Frécon
DIVE on the Internet, 2004.
36: Rickard Cöster
Algorithms and Representations for Personalised Information Access, 2005.
37: Per Brand
The Design Philosophy of Distributed Programming Systems: the Mozart Experience, 2005.
38: Sameh El-Ansary
Designs and Analyses in Structured Peer-to-Peer Systems, 2005.
39: Erik Klintskog
Generic Distribution Support for Programming Systems, 2005.
40: Markus Bylund
A Design Rationale for Pervasive Computing — User Experience, Contextual Change, and Technical Requirements, 2005.
41: Åsa Rudström
Co-Construction of hybrid spaces, 2005.
42: Babak Sadighi Firozabadi
Decentralised Privilege Management for Access Control, 2005.
43: Marie Sjölinder
Age-related cognitive decline and navigation in electronic environments, 2006.

ISSN 1101-1335
ISRN SICS-D–44–SE
SICS Dissertation Series 44