Human Information Processing Based Information Retrieval

Erik Kaoru Graf

A thesis submitted for the degree of

Doctor of Philosophy

School of Computing Science
College of Science and Engineering
University of Glasgow

ABSTRACT

This work focused on the investigation of the question how the concept of relevance in Information Retrieval can be validated. The work is motivated by the consistent difficulties of defining the meaning of the concept, and by advances in the field of cognitive science.

Analytical and empirical investigations are carried out with the aim of devising a principled approach to the validation of the concept. The foundation for this work was set by interpreting relevance as a phenomenon occurring within the context of two systems: An IR system and the cognitive processing system of the user. In light of the cognitive interpretation of relevance, an analysis of the learnt lessons in cognitive science with regard to the validation of cognitive phenomena was conducted. It identified that construct validity constitutes the dominant approach to the validation of constructs in cognitive science. Construct validity constitutes a proposal for the conduction of validation in scenarios, where no direct observation of a phenomenon is possible. With regard to the limitations on direct observation of a construct (i.e. a postulated theoretic concept), it bases validation on the evaluation of its relations to other constructs. Based on the interpretation of relevance as a product of cognitive processing it was concluded, that the limitations with regard to direct observation apply to its investigation. The evaluation of its applicability to an IR context, focused on the exploration of the nomological network methodology. A nomological network constitutes an analytically constructed set of constructs and their relations. The construction of such a network forms the basis for establishing construct validity through investigation of the relations between constructs. An analysis focused on contemporary insights to the nomological network methodology identified two important aspects with regard to its application in IR. The first aspect is given by a choice of context and the identification of a pool of candidate constructs for the inclusion in the network. The second consists of identifying criteria for the selection of a set of constructs from the candidate pool. The identification of the pertinent constructs for the network was based on a review of the principles of cognitive exploration, and an analysis of the state of the art in text based discourse processing and reasoning. On that basis, a listing of known sub-processes contributing

to the pertinent cognitive processing was presented. Based on the identification of a large number of potential candidates, the next step consisted of the inference of criteria for the selection of an initial set of constructs for the network. The investigation of these criteria focused on the consideration of pragmatic and meta-theoretical aspects. Based on a survey of experimental means in cognitive science and IR, five pragmatic criteria for the selection of constructs were presented. Consideration of meta-theoretically motivated criteria required to investigate what the specific challenges with regard to the validation of highly abstract constructs are. This question was explored based on the underlying considerations of the Information Processing paradigm and Newell's (1994) cognitive bands. This led to the identification of a set of three meta-theoretical criteria for the selection of constructs. Based on the criteria and the demarcated candidate pool, an IR focused nomological network was defined. The network consists of the constructs of relevance and type and grade of word relatedness.

A necessary prerequisite for making inferences based on a nomological network consists of the availability of validated measurement instruments for the constructs. To that cause, two validation studies targeting the measurement of the type and grade of relations between words were conducted. The clarification of the question of the validity of the measurement instruments enabled the application of the nomological network. A first step of the application consisted of testing if the constructs in the network are related to each other. Based on the alignment of measurements of relevance and the word related constructs it was concluded to be true. The relation between the constructs was characterized by varying the word related constructs over a large parameter space and observing the effect of this variation on relevance. Three hypotheses relating to different aspects of the relations between the word related constructs and relevance. It was concluded, that the conclusive confirmation of the hypotheses requires an extension of the experimental means underlying the study. Based on converging observations from the empirical investigation of the three hypotheses it was concluded, that semantic and associative relations distinctly differ with regard to their impact on relevance estimation.

ACKNOWLEDGEMENTS

This PhD has been quite a journey in many ways. On last count my library of references contained 1272 filed entries. I read through 26 PhD relevant books front to cover, and my code repository consisted of more than half a million lines of code in R, Python, Java, Bash Script, C, and Fortran. I am also occasionally accused of having 'burnt' through four principal supervisors. All lies of course. I was indeed fortunate enough to have been supervised by more than three principal supervisors throughout my PhD, and I count myself lucky to have worked with so many of the great minds in Information Retrieval. Looking back, this PhD has been quite a thing. Without the help and support of many people this would not have been possible. This page aims to acknowledge their part in sharing the fun and in keeping me going when the going got rough.

First of all I want to thank my family who always supported and believed in me. I would like to thank my father Ulli and my mother Kumie who were there for me on the road that led to the PhD and supported me all the way through. A very special thanks goes to my four sisters Imme, Sabrina, Melanie, and Lisa for helping with the proof-reading of this thesis and pushing me all the time to pull through. The biggest thank you goes to Aude for all her patience during the journey and all her support. She celebrated the end of the PhD as much as me.

I also would like to thank all the colleagues and peers I had the pleasure of meeting and working together with. A special thanks goes to Frank Hopfgartner, Ingo Frommholz, and Sachi Arafat for many insightful discussions and the good times we shared. I also would like to thank Hamish Cunningham for the chance to participate in the development of the Mimir project. A heartfelt thank you goes to Mark Girolami for his support in times of need, his encouragement, and his great sense of humour. I am also indebted to Thomas Roelleke for many insightful comments on my work, for being a great host with excellent taste in wine, and for letting me clean his yacht.

I am very grateful for the funding I have received from the Department of Computing Science during my first year of studies. A big thank you goes to the Information

Retrieval Facility for awarding me with a very generous three year 'Open Research' stipend, and for providing me with the chance of being an invited speaker at several Patent related conferences and events.

I also would like to thank my examiners Karen Renaud and Birger Larsen for their insightful comments on my work. A special thanks goes to Karen Renaud for the subsequent meetings in Glasgow and the support and encouragement I received from her. A big thank you goes to Simon Gay for being a terrific and very supportive convener. Further I would like to thank the staff at the University of Glasgow. They are the ones who keep the scientific wheels turning smoothly. A special thank you goes to those four whose work I most frequently interrupted: Douglas McFarlane, Stewart McNeill, Helen McNee and May Gallagher.

Lastly I would like to thank my supervisory team: Keith van Rijsbergen, Mounia Lalmas, Joemon Jose, and Iadh Ounis. I won't praise their work as it speaks for itself. I would like to thank Keith for the great honour of taking me on as his last PhD student. I am immensely grateful for his trust and the level of freedom he granted me for my research. To this day I am amazed at the meetings we had. No matter how much time had passed and how obscure and far-fetched some of my research ideas were, Keith's mind was always right on spot; outlining interesting connections, pointing out flaws in my thinking, and supplying me with a host of additional references and books to read through. He also has a great sense of humour. Mounia I would like to thank for her ability to focus research and her great talent for the dissemination of scientific results. She is a fierce supporter of her PhD students and their research. I also would like to thank Iadh for teaching me the ropes of doing research during my first year of studies. Finally I wholeheartedly thank Joemon for his support during the last part of my journey. I also would like to thank Leif Azzopardi for his many insights, great jokes, and being a terrific second supervisors.

Table of Contents

CHAPTER 1

INTRODUCTION

1.1 Introduction

The last seven decades of Information Retrieval (IR) research have been marked by a continuous stream of advances. Following the pioneering insights gained from uniterm based indexing systems and early work conducted by Luhn (1957, 1958), the history of Information Retrieval has witnessed steady progress. Milestones within this progression are given by the development of the first test collections (Cleverdon, 1967) and an associated evaluation paradigm (Voorhees, 2002), the introduction of term frequency dependent term weighting (Spärck Jones, 1972), the Clustering hypothesis (Jardine and van Rijsbergen, 1971), and the conception of probabilistic retrieval models (Harter, 1975; Robertson, 1977; Rijsbergen, 1979). Based on these advances, Information Retrieval systems have attained an ubiquitous status in daily life – manifested in form of commercially run search engines such as Google[1-1] and Bing[1-2]. A summary of the state of retrieval systems is provided by Spärck Jones (2005) in her analysis of past Text Retrieval Evaluation Conference (TREC[1-3]) results. She remarks (p. 6) that IR systems deliver reasonable retrieval performance on full text sources. This is summarized in the following conclusion (p. 6):

> *In other words TREC appears to endorse, after exhaustive, large experimentation, the modern approach to retrieval, i.e. the approach that is motivated, explicitly or implicitly, by statistical models, that starts from simple natural language terms, that relies on weighting and feedback strategies, and that delivers ranked output.*

[1-1] http://www.google.com
[1-2] http://www.bing.com
[1-3] A yearly held evaluation conference in Information Retrieval, http://trec.nist.gov/

”

However in the same publication Sparck-Jones also emphasises, that recent advances in Information Retrieval consist of 'small' incremental achievements, and that the relevancy of those advances can be questioned. The validity of recent advances is also questioned by Belkin (2008, p.49):

> “ *it is clearly the case that IR as practised is inherently interactive; secondly, it is clearly the case that the new models and associated representation and ranking techniques lead to only incremental (if that) improvement in performance over previous models and techniques, which is generally not statistically significant (e.g. Spärck Jones (2005)); and thirdly, that such improvement, as determined in TREC-style evaluation, rarely, if ever, leads to improved performance by human searchers in interactive IR systems (e.g. Turpin and Hersh (2001); Turpin and Scholer (2006)).* ”

Belkin re-iterates the notion that current approaches to IR research may lead to only incremental advances. Further he emphasises concerns of a mismatch between IR performance measures and the perceived benefit as seen by the users of IR systems. This aspect has received growing attention in the IR community, and apart from the quoted work by Belkin has also been demonstrated by a recent study conducted by Sanderson et al. (2010).

A possible reason for this observed mismatch may be attributed to the prevalent difficulties of defining the concept of 'relevance' in an IR context. As noted by Mizzaro (1997, p. 810) '[r]elevance is not a well understood concept.' The difficulties with regard to the definition of the concept are well illustrated by the large number of presented definitions of 'relevance' in the publications of Schamber (1990) and Mizzaro (1997). Saracevic (1970, p. 121) commented on the large number of 'relevance' definitions via the proposal of an algorithm for the formulation of such definitions.

> “ *Relevance is the (A) gage of relevance of an (B) aspect of relevance existing between an (C) object judged and a (D) frame of reference as judged by an (E) assessor.* ”

The algorithm highlights that through the insertion of appropriate words any number of definitions of 'relevance' can be created. Remarkable with regard to this algorithm, is the large number of intuitively justifiable definitions resulting from its application. The fact, that 'relevance' can be defined as a measure of (utility, usefulness, interestingness, topical distance, ...) existing between a document and a query hints at the underlying complexity of the concept.

These concerns indicate that, despite the undoubted and prevalent successes in IR, there exists a growing perception within the research community of the necessity to introduce

new research paradigms. In particular with regard to the above outlined difficulties of formally defining 'relevance' this work proposes a novel paradigm for its investigation. The proposed 'Correlation to Cognition' paradigm is based on interpreting relevance as a product of cognitive activity. A tenet of the proposed paradigm is given by the notion that the consideration of the mechanics of the mind forms a promising basis for the investigation of relevance. In general the presented approach shares the sentiment expressed by Gardenfors (1999, p. 15):

> *In conclusion, we can expect that in the future, cognitive science will supply man with new tools, electronic or not, that will be better suited to our cognitive needs and that may increase the quality of our lives. In many areas, it is not technology that sets the limits, but rather our lack of understanding of how human cognition works.'*

Against this background, the motive for the proposed paradigm can be expressed in form of two assumptions. The first, resting on the presupposition that core aspects of relevance are still not well understood, assumes that any increase in such understanding would be very beneficial to the IR cause. The second being, that the consideration of internal cognitive processes constitutes a promising and complementary approach to pursuing such investigations.

The next section discusses these assumptions, which form the core of the motivation for this work, in detail. Based on this outline the remaining course of the chapter is as follows. Section 1.3 builds upon the discussion of the motivation by introducing the dissertation's objectives. The objectives provide a high level overview of the attempted contributions of this work. A more concrete definition of the work is provided in Section 1.4, where the problem statement and a set of derived researched questions are presented. Finally, Section 1.5 describes the research methodology.

1.2 Motivation

As indicated in the introduction, the aim of this section lies in exploring the motivation for the correlation to cognition paradigm. This is attempted by looking at the following three aspects:

- The motivation to place the research focus of the dissertation on 'relevance and related concepts'.
- The reasons to base the research of those concepts on a 'cognitive' interpretation.
- The argumentation behind focusing the analytic and empirical parts of the thesis on an information processing interpretation of the mind.

The context for the discussion of the first aspect is set by returning to the earlier mentioned concerns regarding limits in the understanding of central IR concepts. The quoted[1-4] works by Saracevic (2007), Belkin (2008), Mizzaro (1997), and Sanderson et al. (2010) share a critical view regarding the insight to concepts such as 'relevance' and 'user satisfaction'. Cuadra and Katter's (1967) 'Opening the black box of 'relevance'', and Saracevic's (1975) 'Relevance: A review of and a framework for the thinking on the notion in information science' constitute examples from the early days of IR that highlight the difficulties of understanding and defining relevance. Combined with the above quoted observations, they outline that the difficulties of getting a 'firm grasp' on the concept represent a recurring theme in the history of IR. These difficulties of establishing a consensus of the 'meaning' of relevance and related concepts (Schamber, 1990) constitute a strong motivation for placing the focus of this work on this subject. In other words it is believed that the observation of Schamber (1990) that the 'pursuit of a definition of relevance [is] among the most exciting and central challenges of information science' still holds true.

Addressing the second above listed aspect consists of questioning the motivation behind a cognitive approach to the investigation of relevance. Approaching concepts such as 'relevance', 'user satisfaction', and 'aboutness' with a focus on cognitive processing is motivated by recent developments in the field of cognitive science. The first development is given by the increase of available cognitive models relating to language specific cognitive processing. This manifests itself in form of the achieved insights in understanding key components of language processing in the mind (Gaskell and Altmann, 2007). The progress is exemplified by Just et al.'s (2010) work on identifying brain codes underlying the representation of concrete nouns. The work not only demonstrates the contemporary level of insight but also shows, that such insights are driven by a broad array of advances in contributing scientific domains. Progress in the diverse fields of cognitive science combined with the availability of advanced technical instruments and computational modelling constitute a promising base for further advances. The second motivating development is given by the current momentum of Cognitive Science in general. The momentum of cognitive research is demonstrated through recent initiatives such as the 'Human Brain Project' [1-5] (constituting the European Union's largest ever research excellence award) or the U.S. led 'Brain Initiative' [1-6]. The insights represented by the state of the art Cognitive Science as well as its current momentum are interpreted to form a promising basis for the consideration of cognitive mechanics in the investigation of 'relevance'.

Mention of the term 'cognitive mechanics' sheds light onto the motivation behind the third listed aspect: To focus the research approach on an information processing in-

[1-4] see section 1.1
[1-5] http://europa.eu/rapid/press-release_IP-13-54_en.htm
[1-6] http://www.whitehouse.gov/infographics/brain-initiative

terpretation of the mind. Explaining the motivation behind this decision requires to first provide a brief summary of the currently dominant approach to the investigation of the mind. As put by Thagard (2005, p. 10) it is commonly assumed that '[t]he best way to grasp the complexity of human thinking is to use multiple methods'. This is also reflected by the definition of '[c]ognitive science [as] the interdisciplinary study of mind and intelligence, embracing philosophy, psychology, artificial intelligence, neuroscience, linguistics, and anthropology' (Thagard, 2005, p. 3). Historically this is reflected by three main approaches to the investigation of the mind. Early attempts for the investigation of the mind were almost exclusively focused on the observation of the behavior of individuals. A prototypical example for the approach of 'behaviorism' is given by Skinner's (1957) 'Verbal Behavior'. Behaviorism was followed by the Information Processing approach (Massaro and Cowan, 1993) which assumes, that 'knowledge of the internal structure of the organism, [and] the ways in which it processes input information' (Chomsky, 1959, p. 49) is necessary for the investigation of the mind. This focus on internal mental structures is complemented in contemporary cognitive science by the consideration of sociological and cultural aspects (Hutchins, Edwin and Lintern, 1995). As a result of this development, cognitive science advocates that the study of cognitive phenomena requires the consideration of individualistic behavior, internal mental structures, and the interaction between individuals and the outside world. Based on interpreting relevance as a cognitive phenomenon, it seems appropriate to assume that its investigation benefits from following a multi-layered research approach. The focus on internal mental structures that underlies this work is interpreted to constitute a novel and complementing approach to the Cognitive IR tradition and the investigation of relevance.

Based on what was written so far, the motivation for this dissertation can be summarized in the following three points. Firstly, contributing to the understanding of relevance is thought to be one of the central and most important contemporary challenges in Information Retrieval. Secondly, it is believed that the existing knowledge and the momentum of cognitive science provide a promising basis for the interpretation of relevance as a product of cognitive processing. Finally, the focus on internal mental structures is motivated by its conceived novelty and the consensus that the investigation of cognitive phenomena requires a multi-layered approach. This summarizes the motivation behind this work. The following two sections introduce the dissertation's objectives and research questions. Section 1.3 defines the overall aim of this work and outlines the motives underlying the use of a paradigm as a means of structuring the research approach. Section 1.4 concretizes these high level goals through the specification of the disseration's problem statement and the introduction of the research questions.

1.3 Objectives of the Dissertation

The objectives of this dissertation are twofold. Firstly, this work aims at contributing to the understanding of relevance based on the consideration of cognitive information processing. Secondly, it targets at disseminating the underlying research approach through the introduction of a paradigm. Subsequently these aims are concretized by stating the objectives and attempted contributions of this dissertation. The formulation and presentation of these items is based on prior coverage of two points: An exploration of the meaning of the term 'paradigm', and the presentation of an overview of the proposed Correlation to Cognition paradigm. The discussion of these points serves two main purposes. Firstly, it motivates the proposition of a paradigm as a means of advocating a research approach. Secondly, the exploration of the 'paradigm' concept supplies a frame of reference for structuring the research approach underlying this work. To start the discussion, the next paragraph explores the meaning of the term 'paradigm'.

Defining the term 'Paradigm': The term 'paradigm' is central to Thomas S. Kuhn's theories on the progression of scientific development. The following discussion is accordingly based on the various senses of the term in Kuhn's (1962) 'Structure of Scientific Revolutions' (hereinafter referred to as 'Structure'). As outlined by Masterman (1970, p. 61-65), Kuhn used the term paradigm 'in not less than twenty-one different senses' in the initial publication of 'Structure' (Kuhn, 1962) Within these Masterman (1970, p. 65) identified three main types: Metaphysical, sociological, and artefact paradigms. The proposed 'Correlation to Cognition' paradigm falls into the category of the 'artefact' or 'construct paradigm'. At the core of such an artefact or construct paradigm stands 'a concrete picture used analogically; because it has got to be a 'way of seeing" Masterman (1970, p. 76). The statement by Masterman conveys that the purpose of the analogy consists of providing a novel viewpoint; more specifically, a concrete picture of a 'puzzle-solving' (Masterman, 1970, p. 76) situation. This concrete analogy is also referred to as a 'trick' (Masterman, 1970, p. 73), 'the use of which enables the puzzle-solving of normal science to be performed' (Masterman, 1970, p. 73). In short, artefact paradigms are meant to contribute to scientific progress through spurring novel interpretations of scientific challenges. In a metaphorical sense, the analogy at the core of a paradigm acts as a catalysator for the generation of novel problem-solving approaches. The analogical picture and its application in a concrete experimental context together comprise an artefact paradigm. Kuhn referred to this type of paradigm as an 'exemplar'; a 'shared example' Kuhn (1970, p. 187), and called it 'the most novel and least understood aspect' Kuhn (1970, p. 187) of the book. The portrayed role of exemplars in science has also been interpreted on a cognitive level. In the work following 'Structure' Kuhn developed an interpretation of the role of paradigms on the basis of concept theory, and his notions of 'perceptual space' and a domain spe-

cific 'scientific lexicon'(Kuhn, 1990, p. 7-11). An interpretation of Kuhn's work based on cognitive theories on the representation and learning of conceptual structures is provided by Nersessian (2003). An interpretation based on prototype theory is described by Barker et al. (2003). These interpretations relate closely to Masterman's earlier quoted insight that artefact paradigms 'got to be a 'way of seeing" (Masterman, 1970, p. 76). Against this background, the term 'paradigm' within the scope of this work is defined in the sense of an 'exemplar'. At its core stands a concrete analogical picture provided in form of a literal model, picture, or analogy-drawing sequence of word-uses in natural language (Masterman, 1970, p. 79). The exemplar is complemented by a concrete application of this picture that exhibits its use in a puzzle-solving situation. The intended function of the exemplar consisting in the contribution of novel problem-solving approaches, and the extension of a scientific lexicon. In this form, the concept of an exemplar paradigm represents a blueprint for the conception of novel research approaches. The next paragraph provides an interpretation of the proposed 'Correlation to Cognition' paradigm, against the background the so far led discussion.

Correlation to Cognition Paradigm The 'Correlation to Cognition' paradigm is marked by its focus on cognitive processing. Conceived as an artefact paradigm, it is aimed at demonstrating the investigation of Information Retrieval concepts through the consideration of the 'mechanics' of human information processing. To that cause it is based on an information processing interpretation of cognition that focuses on internal mental structures. Based on this view, the proposed research paradigm defines Information Retrieval as a task aimed at maximizing input-output concordance of two systems: The user's mind, and the IR system. This enables the definition of central IR concepts as a state of concordance between those systems. The purpose of this setup is to facilitate reasoning of IR concepts on grounds of the observables and mechanics of cognitive processing. With reference to the above led discussion this concordance based 'picture' forms the central analogy of the paradigm. The aim of this setup consists of paving the way to the investigation of concepts such as 'Relevance' and 'Aboutness' within the context of a cognitive processing level. In short, by embedding the mind of the machine metaphor into an IR context, the paradigm targets the development of novel problem-solving approaches to these central IR concepts. The central element of the paradigm is given by the interpretation of IR in terms of cognitive processing. Based on this overview of the paradigm, the next paragraph specifies the planned contributions of this work.

Attempted Contributions On grounds of the clarification of the meaning of the 'paradigm' the main objective and the attempted contributions of the dissertation can now be specified. As outlined above, the purpose of construct or artefact paradigms consists of enabling the solving of scientific puzzles by introducing novel analogical

interpretations of puzzle solving situations. Consequently this also constitutes the main objective of the dissertation: The introduction of a novel analogical interpretation for the solving of an IR centered scientific puzzle. To achieve this aim, Masterman outlined that a paradigm 'must be a construct, an artefact, a system, a tool; together with the manual of instructions for using it successfully and a method of interpretation of what it does' (Masterman, 1970, p. 70). Based on this statement and the earlier analysis, the attempted contributions are summarized as follows:

Analogy To provide an analogical picture of a puzzle solving situation in form of the proposed 'Correlation to Cognition' paradigm.

Instructions To provide instructions with regard to its application through the development of an accompanying methodology.

Application To provide an example of the application of the paradigm.

Interpretatation Finally, to provide an example of the of 'what it does' (Masterman, 1970, p. 70) through an analysis of the empirical results derived from the paradigm's application.

This listing outlines, that it is the objective of this work to contribute to the investigation of relevance in two ways. Firstly, in form of insights gained through analytical and empirical work, and secondly through the provision of a paradigm aimed at inspiring novel problem solving approaches to the subject. This summarizes the objectives of the dissertation and provides a frame for its interpretation on basis of the artefact paradigm concept. The following section concretizes the so far led discussion by presenting the dissertation's problem statement and research questions.

1.4 Problem Statement and Research Questions

The prior discussions can be summarized as follows. Section 1.2 presented the motivation for the proposal of a cognitive research approach to the investigation of relevance that focuses on human information processing. In Section 1.3 the objectives of the dissertation were defined. This was based on outlining the reasoning behind structuring the proposal in form of a research paradigm. It outlined that the proposed paradigm is devised as an exemplar with the purpose of demonstrating a novel approach to a scientific 'puzzle'. The next step consists of concretizing the focused scientific puzzle and the chosen analytic and empirical approach. In the following this is attempted by briefly analyzing the focused 'problem', the investigation of relevance, in terms of the concept of validity. The results of this analysis are presented in form of the problem statement of

the dissertation. This is followed by the presentation of a set of analytic and empirical research questions.

The context for the discussion of the focused research problem is set by returning to the central concerns regarding limits in the understanding of relevance and its related concepts. The earlier quoted[1-7] works by Saracevic (2007), Belkin (2008), Mizzaro (1997), and Sanderson et al. (2010) expressed a critical view regarding the insight to concepts such as 'relevance' and 'user satisfaction'. Blair (1990, p. 85) illustrated these limitations in the following form:

> *Information retrieval researchers are like automotive engineers who are trying to improve the design of automobiles without being able to measure horsepower or fuel efficiency.*

Subsequently it is attempted to cast this expression of measurement difficulties into more concrete form. This is based on interpreting the identified 'problem' as a question of validation. The following statement by Kelley (1927, p. 14) will act as a peg upon which to hang the discussion:

> *a test is valid if it measures what it purports to measure.*

Contrasting the statements of Blair and Kelly emphasizes that the problem Blair identified consists of an inability to measure certain variables. Blair's examples of 'horsepower' and 'fuel efficiency' are interpreted to be arbitrary and it is assumed that they could have been replaced by others such as 'drag coefficient', 'kilowatt', or 'driving pleasure'. The important point lies in the asserted inability of measuring 'things'. The statement of Kelly allows for the definition of three prerequisites for the ability to measure. The first prerequisite is given by what Kelly refers to as 'a test'. Generally this can be thought of as some form of measurement instrument. The second prequisite is given by the item the test 'purports' to measure; i.e. a concept that constitutes the aim of one's measurement effort. Finally, the third prerequisite is given by the binary attribute of 'validity'. To emphasize this point the statement of Kelly can be rephrased as follows.

> A measurement instrument (a) is valid (c) if it measures the concept (b) it purports to measure.

This allows advancing the discussion by analyzing IR measuring in terms of these prerequisites. Regarding point (a) it can be stated that Information Retrieval seems to have no shortage of target concepts to measure. Mizzaro (1998) outlines the multitude of such concepts through the definition of a 'Relevance pool'. The pool consisting of entries such as 'relevance', 'user satisfaction', 'utility', 'topicality', and 'usefulness'. An even stronger argument for an ample existence of concepts to measure is given by the

[1-7] see section 1.1

earlier mentioned argumentation of Saracevic (1970, p. 121) and his formulation of an algorithm for defining 'relevance' types. Measuring in IR does not seem to suffer from a lack of concepts to measure. Regarding the general availability of measurement instruments (b) it can be stated that IR has developed an array of measurement instruments in form of relevance assessments (Voorhees and Harman, 2005, p. 3-78), user surveys (Ingwersen and Järvelin, 2005), and the analysis of recorded user behavior (Jansen, 2006). This seems to indicate that neither a general lack of (a) concepts or (b) instruments are the primary causes for Blair's attributed inability of taking measurements. This places the spotlight on item (c): Validity. It can be argued the noted measurement difficulties result from a lack of *valid* measurement instruments. In less strict form validity can be defined as the degree to which measurement instruments measure what they purport to measure (Anastasi, 1954, p.29). Based on this definition the result of the so far conducted analysis can be summarized as: Successful measurement of concepts such as 'user happiness', and 'relevance' requires respective instruments that measure the concept with a high degree of certainty. Attributing the concerns regarding IR measurement with the validity of its instruments brings us close to the formulation of the problem statement of the dissertation. The so far quoted definitions of validity were of 'self-referential' nature. Instruments are declared valid if they measure what they ought to measure. The concept of validity that forms the basis for this work and the following problem statement differs from this interpretation. In line with the type of validity that has 'emerged as the central or unifying idea of validity' (Colliver et al., 2012, p. 366) in Cognitive Science it is based on the concept of 'construct validity'. A construct refers to a 'postulated and theoretical concept' (Colliver et al., 2012, p. 367). In contrary to the self-referential definitions of validity, construct validity assumes that validation requires questioning the validity of the applied measurement instrument, as well as the theory of the concept itself. These observations lead to the formulation of the following problem statement for the dissertation:

PS *How can Information Retrieval centric constructs be validated?*

The problem statement represents the result of the effort to cast Blair's noted measurement difficulties and Mizzaro's (1997) observation that '[r]elevance is not a well understood concept' (p. 810) into a more concrete form. The basis for this is set by the application of the term 'construct' to the IR context. In an IR context candidate constructs are given by concepts such as 'relevance', 'aboutness', and 'user satisfaction'. Interpreting IR concepts as constructs emphasizes that investigations of their validity requires questioning the measurement instrument as well as the postulated concept themselves (Cronbach and Meehl, 1955). This means, that questioning the validity of relevance measurements, inherently requires investigating the validity of the concept of relevance. Clarifying if one is measuring relevance, requires clarification of the validity of the postulated theory of the concept of relevance itself. The problem statement aims at addressing Blair's (1990) stated inability of measuring in IR through a valida-

tion of constructs in IR. To address this problem statement, we derive the following research questions:

RQ 1 What constitutes a principled approach to construct validation in IR?

The first research question addresses the need for following a principled approach to the validation task set by the problem statement. The question is of analytical character and focuses on an exploration of the learnt lessons in Cognitive Science with regard to the validation of constructs. As a consequence, a large part of the conducted analysis revolves around one of the core concepts of construct validity: The nomological network. A nomological network constitutes an empirically and theoretically constructed postulated network of concepts and lawful relations that bases validation on an analysis of the relations between constructs. It represents construct validity's answer of how to approach the challenge of validating measurements in face of uncertainty regarding the validity of both, the instrument and the target concept. The exploration of these considerations forms the basis for the work in this dissertation and results in a series of analytical and empirical research questions.

RQ 2 What are potential constructs for the formulation of an IR focused nomological network?

Research question two directly builds up upon the considerations of RQ1 regarding the use of a nomological network for the validation of constructs in IR. A first step in the construction of a nomological network is given by the identification of the pertinent constructs for the network. Addressing this question raises a couple of directly dependent questions. The first of those relates directly to the earlier mentioned 'Correlation to Cognition' paradigm, as the choice of pertinent constructs depends heavily on the underlying 'view'. At the core of the 'Correlation to Cognition' paradigm stands the attempt to create an analogical picture that bridges scientific concepts from the domains of IR and Information Processing focused cognitive science. These two focused scientific domains therefore constitute the primary 'pools' for potential constructs.

RQ 3 What are criteria for the selection of constructs?

Answering the second research question outlined the pool of potential constructs for basing validation in IR on a nomological network. Research question three addresses the question by what criteria constructs should be selected for the inclusion in the devised nomological network. This analytical research question is based on the learnt lessons from Cognitive Science and forms the bridge to the empirical part of the dissertation. The choice of word similarity related constructs and relevance as the empirically investigated constructs results from the analysis of these considerations.

RQ 4 What are valid instruments for the measurement of the grade of relatedness between words?

RQ 5 What are valid instruments for the measurement of the type of relatedness between words?

These both empirically and analytically approached research questions emphasize that the 'triangulation' of a target concept is utterly dependent on the validity of the other concepts included in a nomological network. Answering these questions requires an analysis with regard to pertinent principles of cognitive science, as well as an empirical evaluation aimed at the validation of available instruments. Clarification of the validity of these instruments finally enables the examination of the relation between the constructs of relevance and word relatedness. This is addressed by the last research question.

RQ 6 What are characteristics of the relation of the postulated constructs of relevance and grade and type of word relationships?

The concluding research question is based on the results of the validation studies for word relatedness. It utilizes the resulting validated measurement instruments as means for the investigation of the relation between relevance and word relatedness. The research question aims at empirically gaining insight to the relation of these constructs, and at exploring considerations with regard to the bigger picture of validating the construct of relevance.

Wherever it is applicable and beneficial to the investigation, these questions are broken down into separate and more specic research questions.

1.5 Research Methodology

The research methodology followed in the thesis is fundamentally based on one central premise: The interpretation of 'relevance' as a cognitive phenomenon. It comprises six parts: (1) Reviewing the pertinent literature in IR and Cognitive Science, (2) analyzing the findings, (3) designing an artefact paradigm, (4) designing a methodology for the construction of a nomological network in IR, (5) empirically evaluating and analysing instruments for measuring word relationships, and (6) conducting an empirical analysis of the relation between the constructs of word relationships and relevance.

The initial part of the methodology consists of a literature review. The review aims at providing an overview of the pertinent research paradigms, methodologies, and issues regarding the investigation of cognitive phenomena. In addition, Chapters 5 through 9 each provide reviews of pertinent literature specifically related to the work described in the respective chapters.

Based on these reviews, the second part of the methodology consists of an analysis of the findings. The results of this analysis guide the design of the proposed artefact

paradigm and the methodology for the construction of an IR centered nomological network.

Third, on basis of this fundamental review and discussion, the next step consists of the design of the Correlation to Cognition paradigm. Methodologically this part is based on Masterman's (1970) and Kuhn's (1970) work concerning the establishment of novel research approaches through the use of exemplars.

The fourth element consists of the introduction of a methodology that supports the construction of an IR centered nomological network. The conception of the methodology is based on consideration of two main research tenets from the domain of Cognitive Science: The Information Processing (IP) Paradigm (Massaro and Cowan, 1993) and Newell's (1994) cognitive bands. In particular the design of the methodology is guided by the concept of Identifiability (Massaro and Cowan, 1993) that constitutes the basis for the IP paradigm. The use of a nomological network as a theoretical approach to the investigation of cognitive phenomena is based on the concept of Construct Validity introduced by Cronbach and Meehl (1955).

Step number five consists of the empirical validation of instruments to measure word relationships. Methodologically this step is based on Rubenstein and Goodenough's (1965) word similarity assessments, lexical decision tasks (Meyer et al., 1972), and the use of correlation matrices for validation as proposed by Cronbach and Meehl (1955).

Finally, the last step of methodology concerns the empirical evaluation of the relation between relevance and word relationships. This step is based on measuring relevance by use of pooled assessments (Voorhees, 2002), the use of convergent validation Garner et al. (1956), and the statistical evaluation of retrieval runs Hull (1993).

I

THEORETICAL REALIZATION

Chapter 1 provided an introduction to the dissertation. It defined the relevance as the primary subject of investigation of this work and provided the dissertation's problem statement and research questions. Based on defining the research approach as a task of validation, Section 1.4 listed six research questions. Part I investigates the first three of those research questions.

RQ 1 What constitutes a principled approach to construct validation in IR?

RQ 2 What are potential constructs for the formulation of an IR focused nomological network?

RQ 3 What are criteria for the selection of constructs?

With regard to the aim of structuring the research approach in form of an artefact paradigm, Part I encompasses the following elements.

Analogy To provide an analogical picture of a puzzle solving situation in form of the proposed 'Correlation to Cognition' analogy.

Instructions To provide instructions with regard to its application through the development of an accompanying methodology.

The work in Part I is organized as follows. Chapter 2 introduces two fundamental elements of the dissertation. It first explores the fundamental assumption of this work to interpret relevance as a product of cognitive processing. The result of this investigation is presented in form of the Correlation to Cognition analogy. Based on this discussion, it then provides an analysis regarding principled approaches to the validation of relevance (RQ1). Chapter 3 provides an overview of the principles of the exploration of cognition, and introduces pertinent concepts from Cognitive Science that are referenced throughout this work. Against this background, Chapter 4 addresses the question of identifying potential constructs for an IR focused nomological network (RQ2). Chap-

ter 5 concludes Part I by exploring potential criteria for the selection of constructs for the devised nomological network (RQ3). With respect to the structure of the artefact paradigm, the organization of Part I can be interpreted as follows. Chapter 2 introduces the central analogy of the paradigm, while Chapters 3, 4, and 5 provide instructions regarding its application.

CHAPTER 2

CORRELATING TO COGNITION

In Chapter 1 it was outlined that the primary research subject of this dissertation is the concept of relevance. Two general objectives were defined with respect to that cause. Firstly, to contribute to the understanding of relevance. Secondly, to advocate a novel view for the investigation of relevance by structuring the chosen research approach in form of an artefact paradigm.

Chapter 2 addresses two fundamental steps towards achieving these objectives. Step one consists of pursuing the analogy objective defined in Section 1.4.

> **Analogy** To provide an analogical picture of a puzzle solving situation in form of the proposed 'Correlation to Cognition' paradigm.

Step two consists of the identification of a principled approach to the validation of IR constructs (RQ1).

> **RQ 1** What constitutes a principled approach to construct validation in IR?

The first step addresses the difficulties associated with the choice of focusing on a subject of research with no established consensus of its meaning (Schamber, 1990). As stated by (Masterman, 1970) the purpose of the central analogy of a paradigm consists of establishing novel ways of conceiving scientific problems. In the case of this work this translates to conceiving a novel view on approaching the research of relevance. The task of establishing such a view is pursued throughout Section 2.1. Achieving this goal requires some kind of 'trick' (Masterman, 1970, p. 73), 'the use of which enables the puzzle-solving of normal science to be performed' (Masterman, 1970, p. 73). The 'trick' developed throughout Section 2.1 consists of developing a definition of the system in which the phenomenon of relevance occurs. To emphasize this point, a relation can be drawn to the formulation of the concept of gravity by Newton. Newton's conception of gravity was based on the perception of a system encompassing the earth,

an apple, and the moon (Cohen and Smith, 2002). As described by Cohen and Smith (2002, p. 6) the formulation of Newton's Universal Gravity on his observation of the interaction of the earth and the moon. The objective of Section 2.1 in analogy to this, consists of devising the description of a system that provides a context for the perception of relevance. This constitutes the 'trick' that underlies the analogical picture of the Correlation to Cognition paradigm, and forms the basis for the realization of the dissertation's objectives defined in Section 1.3. Based on these considerations, Section 2.3 addresses the task of the identification of validating relevance and its related constructs (RQ1).

2.1 A Context for Relevance

As stated in the introduction of this chapter, this section aims at providing a context for the description of the phenomenon of relevance in IR. Throughout the following sections this is approached through describing IR based on the formulation of a series of definitions. The objective is pursued by first defining the objective of the Information Retrieval. This is followed by providing definitions of the general purpose, means, and effectiveness of IR systems.

2.1.1 The Objective of Information Retrieval

In order to provide the discussion with a conceptual basis it is necessary to first define the general objective of Information Retrieval. The following definition is based on considering an initial scenario as outlined in Figure 2.1.

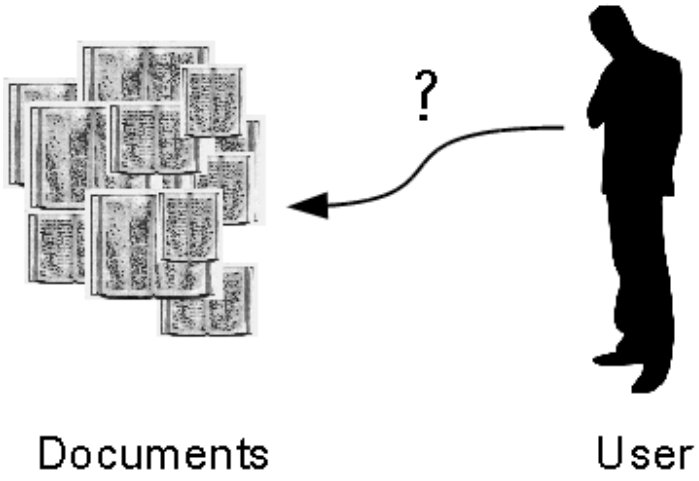


FIGURE 2.1: General Information Retrieval Scenario

The depicted scenario consists of three elements: A document collection, a user, and the task of accessing information in the collection. The term 'document' as part of the IR nomenclature is not restricted to the meaning of 'text documents'. It carries a broader sense that encompasses every potential informational item. Informational items in this regard can therefore refer to pictures, videos, or audio recordings. In the ongoing discussion the term 'informational item' is applied where the aim consists of highlighting this variety. When referring to a 'document' in the common sense, the term 'text document' is used. Information Retrieval as a science can be understood as a reaction to the exponential growth of available informational items, and in this line of thought the objective of Information Retrieval is defined as follows:

> **Definition 1:** The objective of Information Retrieval consists of facilitating access to large repositories of information.

This definition is more general, but closely related to the definition provided by Saracevic (1997, p. 23) on Information Science:

> “ *Information science is trying to organize and make accessible the universe of knowledge records, literature, in a way that 'texts' most likely to be relevant or of value to users are made most accessible intellectually and physically.* ”

This very broad definition serves as a basis throughout the following discussion of elementary considerations of the construction of retrieval devices.

2.1.2 A Prototypical Retrieval Device

The prototypical approach to facilitating access to information consists of the construction of a dedicated system. In the following section the attempt to define principal characteristics of such a system is conducted in order to provide a basis with regard to subsequent discussions in regard of their conception. The non-triviality of the act of constructing such systems is expressed by the following statement by Blair:

'Information retrieval researchers are like automotive engineers who are trying to improve the design of automobiles without being able to measure horsepower or fuel efficiency.' (Blair, 1990, p. 85)

Apart from highlighting the difficulty of the task, the above exercised analogy also provides us with a foothold for the identification of central questions underlying the construction of a retrieval system.

Three basic considerations with regard to the construction of devices can be extracted from it:

1. Purpose: The general aim underlying the construction of a device.
2. Means: The means, the basic underlying mechanical principles, applied in order to achieve that purpose.
3. Effectiveness: The definition of a concept that allows for some form of quantification of the effectiveness of the applied means with regard to the stated purpose.

In the case of an automobile the purpose clearly consists of facilitating transportation. The dominant means to achieve this goal consists of the utilization of the principle of combustion. Measures of the effectiveness with regard to the application of combustion are given by the concepts of horsepower and fuel efficiency. In the case of Information Retrieval the purpose has been defined as facilitating access to information. Subsequently a formal definition of the second and third points is developed based on an examination of the dominant modus operandi of existing retrieval systems.

2.1.3 Principle Function of an IR System

An exploration of the dominant mode of operation will serve as the basis for the definition of the principle mean applied by IR systems in order to facilitate the access to information.

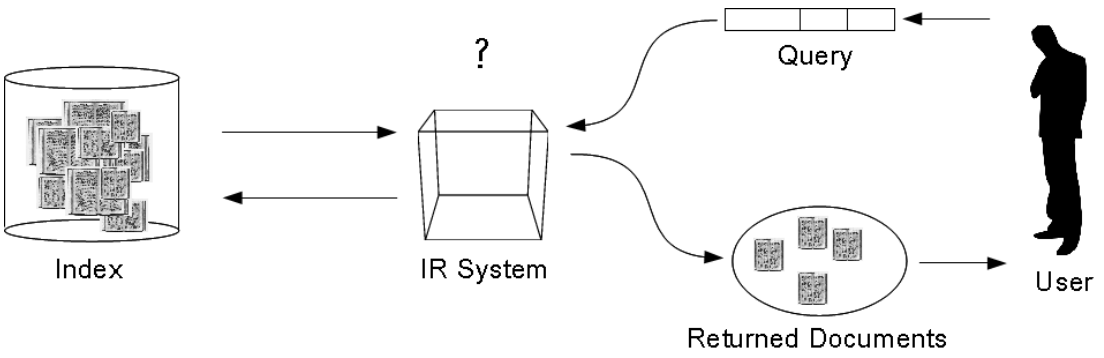


FIGURE 2.2: The dominant mode of operation of an IR system

The dominant mode of operation of an IR system, as outlined in Figure 2.2, can be described as follows. In order to interface with the IR system the user expresses his or her information need in the form of a query consisting of keywords. The query is then submitted to the system. As a result of the query submission the user expects the system to provide him or her with a limited number of suggested documents. The central question with regard to the identification of the principle means then can be stated as: What is the system supposed to do with the query?

Approaching this question from an input-output analysis point of view provides the following scenario. The input consists of an informational item provided by the user as some form of indication with regard to the desired information. The expected output consists of a set of informational items chosen by the system with regard to the provided input. Based on this outline it becomes clear, that the expected function of the system consists of computing an estimate of the relation between the query and every single item or sets of items (Jardine and van Rijsbergen, 1971) contained in its index. The abstract concept of 'aboutnesss' can be utilized to express this functionality with regard to a specific type of such a relation. In this case the purpose of the systems can be defined as retrieving documents that are 'about' the same topic as the query. To do so, the system needs to estimate the topical relation of every document in the corpus with regard to the 'aboutness' of the query. This specific type of relation represents one of many from the user's perspective meaningful relations. The variety with regard to the nature of these relations is indicated by the large amount of reported types of relevance (Schamber, 1990; Mizzaro, 1997), and user goals (Broder, 2002; Rose and Levinson, 2004). With regard to the introduction of our approach we define the principle means of an IR system with regard to its purpose in a general form.

> **Definition 2:** The principle function of an Information Retrieval system consists of providing estimates of relations between informational items.

Formulated in this way, the definition does not specifically refer to any kind of meaningful relation between informational items such as 'relatedness', 'similarity', or 'relevance'. As such, the principal function of a retrieval system is defined in the most general fashion.

2.1.4 Effectiveness

The status of this exploration so far can be summarized as follows. The aim of Information Retrieval consists of facilitating access to repositories of information. To this end an automatic retrieval system is conceived. The system achieves its purpose primarily through providing estimates of the relation of informational items with regard to a token of information provided by the user. The remaining unspecified point with regard to the three items defined to provide a basis for the construction of retrieval systems lies with the question of effectiveness. More precisely, by what means it is possible to gauge to what degree a system achieves the laid out purpose.

The investigation of this question will be based on an exploration of the dominant mode of evaluation.

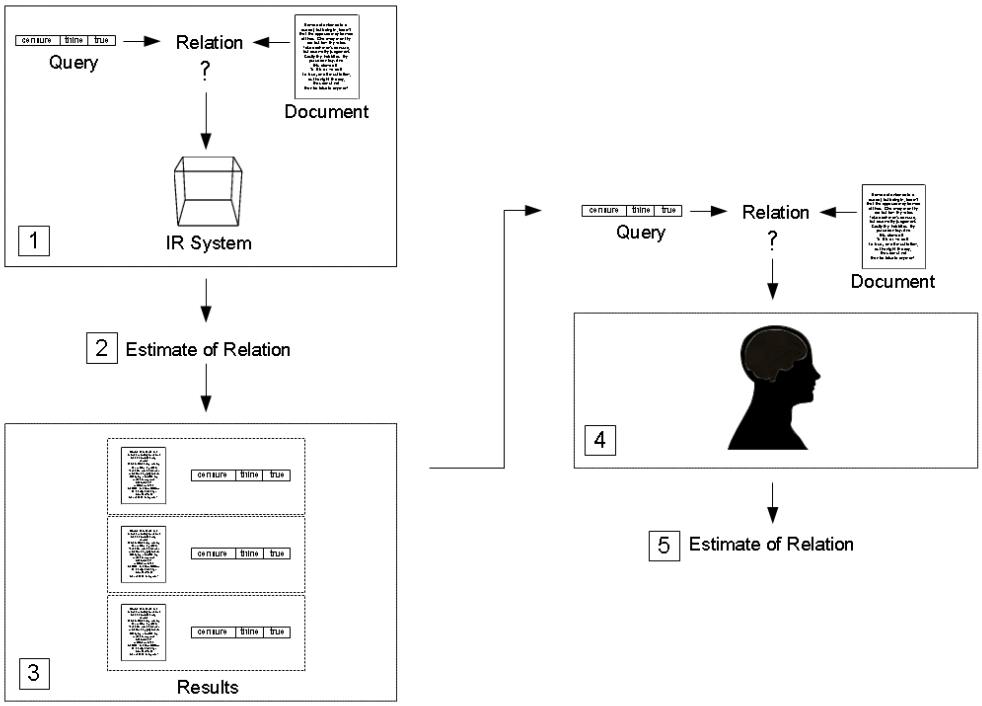


FIGURE 2.3: Evaluation in IR

Figure 2.3 highlights the prototypical evaluation procedure applied to determine the effectiveness of retrieval systems. The process can be summarized in the following way:

1. Given a query the IR system estimates the relation of the query and the items in the corpus.

2. Based on these estimates [2] a selection is made, and a set of items is returned to the user.

3. To evaluate the performance of the system, a human being (often, but not necessarily, the querying person itself) is required to assess the returned results. In order to perform this task the assessor estimates the relation between the original query and each item returned by the system.

4. The verdict of the assessor consists of his or her estimate [5] of the relations between the query and the returned items.

Leaving the specific nature of the relation aside for the moment the following statement can be made. A specific estimate made by the system [2] is perceived as correct if in concordance with the estimate [5] of the human assessor .

> **Definition 3:** Effectiveness of a retrieval system with regard to the estimation of the relation between informational items is defined as a function of the level of concordance occurring between the system's and the querying user's estimates. The higher the concordance the greater the effectiveness of the system.

In essence this statement expresses a non-existence of 'rightness' of the system in the absence of the user. An item returned by the system can only be considered correct if the user of the system perceives it as such. In this regard, an optimal outcome from a system's point of view therefore consists of ranking results in exactly the same way as the user would have done, performing the same task manually. The notion that 'in the end' only the user's verdict is of importance is highlighted in regard of the subsequent section's integration of the user within an interpretation of IR focused on systemic concordance.

2.1.5 A System Concordance Based View of IR

Based on the earlier provided elementary definitions with regard to the purpose of IR, the principle means of an IR system to achieve this purpose, and an interpretation of effectiveness, we will now introduce a paradigmatic view focusing on concordance of systems. The principle function of an IR system consists of providing estimates of the relations of informational items. Following Definition 2, the task of a retrieval system consists of providing an estimate of the relation between two items. The aim of the retrieval system consists of providing an estimate that matches the output of the user's system of judgement as closely as possible. On basis of this view the aim of Information Retrieval can be initially reformulated in the following way.

> **Definition 4:** The aim of Information Retrieval consists of the creation or modification of a system whose estimates exhibit a maximum level of concordance with the output of the user's estimation system.

In a way this can be described as the aim of enabling automation of tasks performed by the human estimation system. In analogy to the automation of physical tasks carried out by human beings the importance resides within input-output concordance and not specificities of the manner in which the task is conducted. To illustrate: An automobile can be interpreted as automation of the human capacity of movement from a point A to a point B via locomotion. The focus of this specific type of automation lies in achieving the result, transition from A to B, and not in exact mimicry of human locomotion. The

above provided definition shares this focus: The aim does not consist of creating a system that processes and operates on information in exactly the same way as a human being. It is focused solely on input-output concordance.

More generally this scenario can be described in the following way: To create an artificial system that, given the same input, will match the output of a natural system. To emphasize the nature of the aim, the optimal outcome of this scenario is depicted in Figure 2.4.

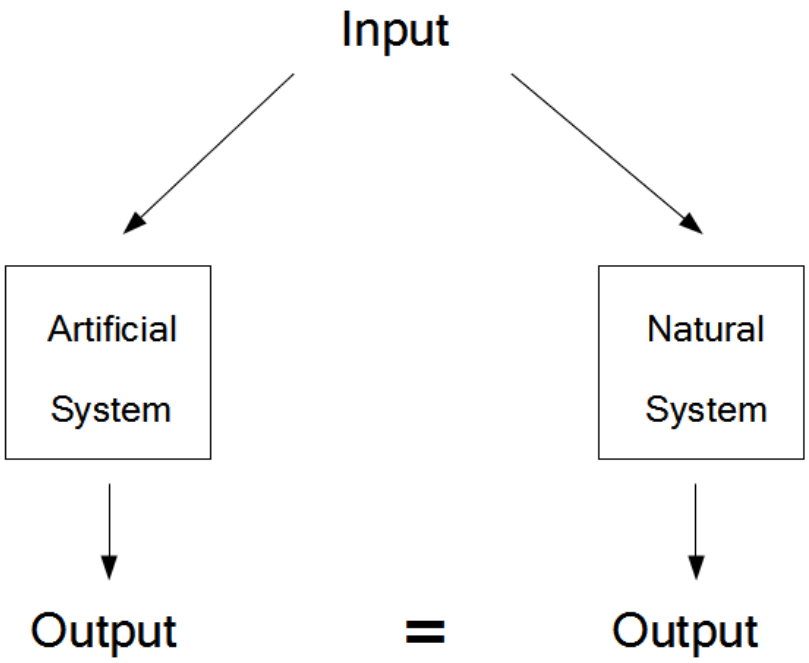


FIGURE 2.4: Concordance of artificial and natural system

As expressed before, the artificial system performs in an optimal manner if its estimates comply entirely with the natural system's estimates. Based on this definition, it is obvious that difficulties arise if the processing details of the natural system are unknown. With regard to the IR perspective this situation applies. The exact specification of the user's estimation system is unknown. In this, the user constitutes a black box, and from this state arise the challenges of Information Retrieval as a science. Based on this outline, the question at hand consists of how to approach the task given in Definition 4: On what grounds can the conception of an artificial system, meant to comply with a natural system be approached, if the latter's mechanics are unknown.

In this form, the listed definitions and the system concordance based view of IR provide the sought after context for the discussion of relevance. The presented context does not suffice to infer the meaning of relevance itself. It represents a systemic description for

the occurrence of relevance. The outline presented in Figure 2.4 in combination with the definitions allows us to make the statement, that relevance is a phenomenon that occurs when the two depicted systems show input-output concordance. This enables to relate to relevance with respect to the properties of the two systems, and provides the basis for the chosen research approach. The next section builds up on the so far led discussion by concretizing the established view and introduces the Correlation to Cognition analogy.

2.2 Correlation to Cognition Analogy

In the following the Correlation to Cognition analogy is introduced. This is pursued on basis of the systemic view presented in Figure 2.4. The figure defined a context for the occurrence of relevance, and enables relating to relevance with respect to the properties of the two systems. The next necessary step towards establishing a novel problem-solving approach to the investigation of relevance consists of substantiating the views of two systems. That is, completing the analogy by concretizing the nature and properties of the two systems.

Defining the nature of the IR system is straightforward. IR systems can be accurately described based on the theory of Information Retrieval and Computing Science. Defining the nature of the natural system, representing a placeholder for the cognition of the user, can be based on various possible interpretations. As mentioned in Section 1.3 cognition can be viewed based on varying levels of abstraction. These abstraction layers encompass the societal and behavioral view, as well as the perspective of the mechanics of the mind. As put by Thagard (2012) the consideration of these various layers is considered necessary with regard to the investigation of complex cognitive phenomena. Within this range of perspectives of the mind, the focus of this dissertation is set on the internal cognitive processes of the 'user system'. As stated in Section 1.2 this is motivated by developments in the field of Cognitive Science and seen as complementing existing efforts to the investigation of relevance in IR. Settling on this perspective represents a fundamental decision underlying the work of the thesis. It is therefore important to outline the argumentation and motivation constituting the basis for the chosen interpretation of the mind. An initial step towards outlining this reasoning, is given by the following two quotations by Chomsky and Hutchins. Chomsky's statement emphasizes the importance of acknowledging cognitive mechanics when researching human behavior.

> *'One would naturally expect that prediction of the behaviour of a complex organism (or machine) would require, in addition to information about external stimulation, knowledge of the internal structure of the organism, the ways in which it processes input information and or-*

> *ganizes its own behaviour. These characteristics of the organism are in general a complicated product of inborn structure, the genetically determined course of maturation, and past experience. Insofar as independent neuro-physiological evidence is not available, it is obvious that inferences concerning the structure of the organism are based on observation of behaviour and outside events. Nevertheless, one's estimate of the relative importance of external factors and internal structure in the determination of behaviour will have an important effect on the direction of research on linguistic (or any other) behaviour, and on the kinds of analogies from animal behaviour studies that will be considered relevant or suggestive. (Chomsky, 1959, p. 27)* ”

As outlined above, Chomsky's main criticism is directed against the attempt to conduct research of complex behaviour without considering the underlying cognitive mechanics. It is important to highlight that the principal distinction does not lie in the mode of experimentation. Every experiment conducted in the Information Processing paradigm as well as the behaviourist approach follows a basic input-output scheme. The two ways of approaching research are demarcated via their experimental focus. This point is concisely expressed by Edwin Hutchins.

> “ *Cognitive science was born in a reaction against behaviorism. Behaviorism had made the claim that internal mental structure was either irrelevant or non-existent - that the study of behavior could be conducted entirely in an objective characterization of behavior itself. Cognitive science's reaction was not simply to argue that the internal mental world was important too; it took as its domain of study the internal mental environment largely separated from the external world. (Hutchins, Edwin and Lintern, 1995, p. 371).* ”

Hutchins' statement emphasizes the importance of considering internal mental structures. It also expresses a necessity for the consideration of the 'external world'. The statements highlight the role of the chosen interpretation of the mind, and motivates the multi-tiered approach to the investigation of the mind that is the field of Cognitive Science. As such, these statements aid in placing the cognitive focus of this work in perspective of the greater picture representing Cognitive Science.

The so far led discussion concludes the definition of the central analogy of the Correlation to Cognition paradigm. The basis for the definition of the analogy was set through the systemic analysis in Section 2.1. The result of the analysis provided a context for relating to relevance. Definition 3 essentially provides a definition of the observance of relevance. The event of concordance between both systems is an observation of relevance. Concordance of the IR system's and the natural system's estimates represents such a manifestation of the concept of relevance. Alike the fall of the apple, and the or-

bit of the moon represent manifestations of the concept of gravity within the Newton's systemic context.

Mere observation however does not imply an understanding of the phenomenon, nor does it necessarily aid in providing a precise definition of the phenomenon. With respect to relevance, answering such questions requires an understanding of the basis for the estimates of the natural system. The interest lies in knowing why informational items are related, and to identify the rules underlying the phenomenon of relevance. Given the focused view of the analogy this translates to acquiring knowledge of the cognitive estimation process that underlies relevance. This observation leads us to the next section, that explores the identification of principled approaches to the validation of cognitive phenomena as a basis for this task.

2.3 Validating Cognitive Phenomena

Sections 2.1 and 2.2 defined a context for researching relevance. Within this scenario, relevance is interpreted as a phenomenon of the interaction between two systems: A natural and an IR system. This scenario was substantiated through the definition of the Correlation to Cognition paradigm introduced in Section 2.2. Against this background, the following section addresses the task of getting a 'grasp' on the concept of relevance. As stated in Section 1.4, this is based on interpreting the problem as a question of the validity of the concept. Throughout the course of the section, this approach is explored as follows. Section 2.3.1 provides an introduction to the concept of validity. Section 2.3.2 complements the initial discussion by exploring the role of construct validity as a means of validating cognitive phenomena. Finally Section 2.3.3 explores the concept of a nomological network and its application within an Information Retrieval context.

2.3.1 Validity

Section 1.4 introduced the concept of validity and based the definition of the problem statement and research questions upon it. This section will expand the so far led discussion of the concept and outline the specific role that validity plays with regard to the observation of cognitive phenomena.

The following excerpt by Lachman et al. (1979, p. 124) serves as an entry point for the discussion.

> *You may wonder how we validate our inferences, and how we confer scientific status on our theoretical statements, which are after all*

> *about unobservables. We do it much as physical scientists do, as for a big part of their subject matter is also unobservable. The temperature of the sun for example, or the activity of molecules, atoms, and electrons cannot be known directly. Nevertheless, scientists have determined that the surface temperature of the sun is about 11,000° and the interior temperature is about 25,000.000°. The key to the problem is that several phenomena that are observable place constraints on the possible interior temperature of the sun. In cognitive psychology we endeavour to make observations that place constraints on the possible workings of the mind. In both cases properties of the system under study give rise to observable data even though these properties are not themselves observable. Interlocking inferences then permit construction of valid factual statements about the unobservable properties. This technique is sometimes called convergent validation (Garner et al. (1956)); when data of several different kinds converge on a conclusion the conclusion is convergently validated.* ”

Lachman's statement expresses that validity in a scientific context represents a tool for the validation of inferences. Historically, validity as a tool was focused on the validation of tests or instruments (Kelley, 1927, p. 14). (Onwuegbuzie et al., 2007, p. 113) describes the validity of a test as the 'extent to which scores generated by an instrument measure the characteristic or variable they are intended to measure [...]'. Under this definition a thermometer constitutes a valid measurement instrument if it accurately measures temperature. This represents probably the most intuitive and widespread application of the concept of validity. Validating gas and water meters for example is a common and everyday task applied to ensuring the validity of instruments aimed at directly measurable observables. A different interpretation of the concept is introduced by Lachman through drawing an analogy that relates validity to measuring the temperature of the sun. This example constitutes a case of validating inferences based on phenomena that are not directly observable ('unobservables'). According to Lachman et al. (1979, p. 124), cognitive phenomena fall into the category of 'unobservables'. This inability of directly measuring cognitive phenomena forms the core challenge of Cognitive Science. This state is described by Chomsky (1959, p. 27) as a lack of 'independent neuro-physiological evidence', and as 'constraints on psychological research' by Massaro and Cowan (1993, p. 388). According to Massaro and Cowan (1993) these constraints are induced by the variability and complexity of behavior, and the issue of identifiability[2-1]. The argumentation outlines the central role of validation within the domain of cognitive science. As Lachman et al. (1979) expressed, the lack of direct observables leads to the adoption of 'indirect' observations that are *assumed* to be

[2-1] A concept from Cognitive Science that describes the inability of determining the correctness of cognitive theories resulting from the lack of direct observation and the theorems of Moore (1956) in formal automata theory. (see Section 3.4)

indicative of specific aspects of cognition. It is precisely the act of forming these assumptions from which stems the need for validation. The act of validation in Cognitive Science can be interpreted as an evaluation of the veracity of these assumptions. This differentiates the interpretation of validity in Cognitive Science from the earlier outlined instrument centered view. These differences are fundamentally rooted in the level of uncertainty inherent to the postulated theories about the concept one aims to measure. Validation of a thermometer does usually not encompass validating the concept of temperature. In contrary to that, the validation of cognitive phenomena is marked by also questioning the validity of the underlying theory. Observation of a phenomena such as 'professionalism' requires validation of the applied tests *and* the theoretic concepts themselves. Validation in Cognitive Science therefore shifts the focus from the instruments towards the concepts of interest.

This can be directly related to the research focus of the dissertation by revisiting the definition of the problem statement.

PS *How can Information Retrieval centric constructs be validated?*

The discussions in this Section offer some preliminary insights towards answering the questions raised by the **PS**. A potential first step to the validation of IR centric constructs is given by their interpretation on a cognitive processing basis. Section 2.1 and Section 2.2 provided the context and perspective for such an interpretation. As described in this Section, the research of cognitively grounded concepts faces inherent challenges regarding the validation of observations and inferences. Given a cognitive interpretation of relevance, Massaro and Cowan's (1993) argumentation provides a possible explanation for the difficulties encountered by past research efforts on relevance in the IR community. As outlined by the discussion based on the quoted excerpt from (Lachman et al., 1979), Cognitive Science utilizes the concept of validation as a means of overcoming those challenges. The development from Kelley (1927) to Colliver et al. (2012) represents close to a century of learnt lessons from past efforts aimed at getting a grasp on cognitive phenomena. Under the premise of viewing relevance as a cognitive phenomenon, these learnt lessons constitute a basis for a principled approach to getting a grasp on the concept. The Section outlined that approaching the validation of cognitive observations shifts the focus of validation from the measurement instruments towards the theoretical concepts themselves. This approach constitutes the core idea behind the concept of construct validity. As a next step towards the validation of IR constructs, the next Section explores the concept of construct validity and relates it to the questions posed by the problem statement and **RQ1**.

2.3.2 Construct Validity

Construct validity has 'emerged as the central or unifying idea of validity' (Colliver et al., 2012, p. 366) in Cognitive Science. The concept has been introduced by Cronbach and Meehl (1955), and represents a unification of earlier conceptions of validity. As described by Colliver et al. (2012), the earlier approaches to determining validity focused on properties of a test or instrument itself. Until the first half of the 20th century the primary approaches to determine validity were based on content and criterion validity. Criterion validity represents the elementary requirement of determining if a test measures what it aims to measure. Content validity focuses on the question, whether an instrument is capable of covering the universe of behaviors it purports to measure. In contrary to these approaches, construct validity encompasses questioning the validity of a test or instrument, as well as the validity of the theoretical postulates one aims to measure. The core idea of construct validity is 'that scientific theory testing [is] seen as part and parcel of test validity, [and] that test validity [is] determined by theory testing' (Colliver et al., 2012, p. 367). To emphasize this, the term construct is defined as a 'postulated and theoretical concept' (Colliver et al., 2012, p. 367). Earlier concepts of validity can be interpreted as attempts of 'fine-tuning' measurement instruments until they measure what they ought to measure with a high degree of certainty. Construct validity places an additional focus on 'fine-tuning' the definition of the postulated concept. Examples for constructs in the cognitive domain are presented by items such as 'burnout', 'stress', and 'empathy'. These examples illustrate the necessity for fine-tuning concept definitions describing cognitive phenomena.

Applying these considerations to IR focused constructs implies that the same considerations should be applied regarding their validation. This means that validating concepts such as 'relevance', 'user satisfaction', and 'user happiness' should be pursued through validating their respective theoretical definitions as well as the applied measurement instruments. This raises the question, how the validation of the theoretical definition of a construct such as 'relevance' can be attempted. Construct validity as defined by Cronbach and Meehl (1955) advocates to pursue this task through an investigation of the relation between targeted constructs and other postulated concepts. It bases the definition of theoretical concepts on an investigation of 'their ties with other theoretical terms and observables' (Colliver et al., 2012, p. 368). Applying construct validity to the concept of relevance therefore translates into the act uncovering its ties with related concepts. To set a methodological basis for this task, Cronbach and Meehl (1955) introduced the concept of a nomological network. A nomological network is defined as a set of constructs and lawful relations between these constructs. Establishing construct validity is realized through the construction of the network. The next Section explores the concept of the nomological network in more detail and investigates its application as a principled means for the validation of IR constructs.

2.3.3 Nomological Network

The so far led discussion in the Chapter can be summarized as follows. Section 2.1 laid the foundation for interpreting relevance as a cognitive phenomena. This view was substantiated through the formulation of the Correlation to Cognition analogy in Section 2.2. Based on interpreting relevance as a product of cognitive processes, Section 2.3.1 outlined the specific challenges underlying the validation of cognitive phenomena. Section 2.3.2 investigated the dominant approach regarding validity in Cognitive Science and outlined how the validation of cognitive observations can be approached through the construction of a nomological network. Against this background the aim of the following Section is as follows. Firstly, the concept of a nomological network is explored in more detail. Secondly, the application of the methodology to the validation of IR centric constructs is investigated.

As stated in the prior Section, a nomological network consists of a set of constructs and definitions of lawful relations between the constructs. This is best illustrated based on an example. Given the task of evaluating the validity of a temperature measurement device (i.e. a thermometer) the respective nomological network consists of the theoretic framework of physics (specifically by the laws of thermodynamics). The validity of a thermometer as a measurement device is determined in reference to these laws. The laws of thermodynamics provide the context for the establishment of validity of temperature measurements. This illustrates that a first step to the establishment of a nomological network requires the definition of its context. It is obvious that a difference exists with regard to the 'quality' or 'sophistication' of the context set by the laws of thermodynamics and that represented by the theoretical basis of cognitive science. Kane (2006, p. 442) addresses this aspect in the following form:

> “ *Initially, the nomological networks were conceived of as formal theories (e.g., Newton's laws), but because such theories are rare to nonexistent in psychology, the requirement was relaxed to include open-ended collections of relationships involving the construct of interest.* ”

Borsboom (2003) expressed this by referring to nomological networks in psychology as 'nomological sketchwork'. This aspect was addressed by Cronbach (1989) through introducing a distinction between weak and strong nomological networks. The term strong nomological network refers to the concept as it was envisioned in line with the then dominant scientific psychology of logical positivism (Colliver et al., 2012). Logical positivism advocates that scientific theory can be established without explicit reference to 'reality'. This view assumes, that the meaning of a construct can be inferred from its role within a tight network of well defined constructs. As expressed in the above quotations, the realization of this vision is hindered by a lack of concretely defined constructs in Cognitive Science. As a consequence, most nomological networks

in Cognitive Science can be categorized as weak networks. In this view, the tight networks featuring lawful relations are replaced by collections of constructs involving any expressions of their relationship. The discussions by Borsboom (2003), Kane (2006) and Colliver et al. (2012) outline the difficulties to establish the meaning of a construct such as 'stress' on basis of weak nomological networks. As a reaction to this, Borsboom et al. (2004, p. 1069) suggests that it is necessary to 'invok[e] realism' with respect to validation. He suggests, that validation requires the measurement of 'attributes' (p. 1069). As described by Colliver et al. (2012), attributes are considered to represent more than just theoretical concepts. Height, weight, blood pressure, and scholastic performance are examples of such attributes. This view shares the sentiment of grounded cognition (Barsalou, 2008), that aims at defining cognition based on modal simulations, bodily states, and situated action. It expresses that validation of a construct is limited when such an attempt is solely based on its investigating its relation to other abstract theoretical constructs. A nomological network *solely* consisting of abstract theoretical constructs is not sufficient to enable convergence on the meaning of constructs. Along this line of argumentation, a tight definition of 'burnout' cannot be inferred solely on basis of its relation to other abstract constructs such as 'frustration' and 'stress'. Such an act requires the consideration of 'attribute' like observations, that could be rooted in bodily state and functions.

At this point the discussion can be related back to RQ1. The question is concerned with the identification of a principled approach for the investigation of the validity of IR constructs. In the beginning of this chapter it was discussed how relevance can be interpreted as a cognitive phenomenon. Following this discussion, it was outlined that the validation of cognitive phenomena is challenged by the constraints of observing cognitive processing and requires a structured approach. Section 2.3.2 showed that the dominant tool for the conduction of such validations is given by construct validity and its nomological network methodology. The Section described how a nomological network functions a means of establishing validity based on creating a net of relations between constructs. The discussion emphasized two aspects that are of importance with regard to the task of establishing such a network. The first is given by the definition of a context for the network. In the case of relevance such a context is provided in form of the Correlation to Cognition analogy. The analogy sets a context for the construction of a nomological network for IR constructs. Based on this context, constructs referring to observations on the side of the IR system are given by concepts such as relevance, aboutness, user happiness, precision and recall. Constructs representing the natural system are given by concepts relating to the information processing interpretation of the mind. The establishment of validity for relevance and its related constructs then consists of the choice of a set of constructs and the investigation of relations between these constructs. Regarding the choice of constructs it has been shown that contemporary insights to construct validity emphasize the need for grounding the validation effort. That is, the selection of concepts should adhere to criteria that consider the realism of

the constructs. In summary, a principled approach to the validation of IR constructs is given by their interpretation as cognitive phenomena and the construction of a nomological network composed of IR and cognitive processing constructs.

2.4 Chapter Conclusions and Answer to RQ 1

The beginning of the chapter laid the foundation for the interpretation of relevance and its related subjects on a cognitive basis. Sections 2.1 and 2.2 resulted in the formulation of the central analogy of the artefact paradigm.

Analogy To provide an analogical picture of a puzzle solving situation in form of the proposed 'Correlation to Cognition' paradigm.

The presented analogy provides a context for the interpretation of relevance as a cognitive phenomenon, and establishes a novel view on IR research by defining its aim as achieving input output concordance with the cognitive processing system of the user.

Section 2.3 outlined considerations regarding the validation of cognitive phenomena. It outlined the research challenges in Cognitive Science stemming from the complexity and variability of behavior and the constraints on direct observation. Against the background of this discussion, it described the central role that validation plays as part of cognitive research.

RQ 1 What constitutes a principled approach to construct validation in IR?

Based on viewing IR constructs as cognitive phenomena, Section 2.3.1 suggested that the task of validating IR constructs should be based on the learnt lessons of validation in cognitive science. Section 2.3.2 showed that the central approach to cognitive validation is given by the concept of construct validity. An approach to validity that requires testing instruments as well as the postulated theory of the constructs. Section 2.3.3 investigated the application of the central methodology of construct validity as a principled approach to construct validation in IR. It outlined how the nomological network methodology can be applied to the validation of IR constructs based on the context defined by the Correlation to Cognition analogy. It further showed, how contemporary insights to the nomological network methodology imply the consideration of two important aspects with regard to its application in IR. The first is given by the choice of context and the set of potential constructs to include in the network. These considerations are represented by the question posed by RQ 2. The second is given by the requirements for 'realism' identified by Kane (2006) and Borsboom et al. (2004). RQ 3 addresses the investigation of these concerns. In summary, it can be stated that the concept of construct validity is seen as a principled approach to the validation of IR

constructs.

CHAPTER 3

PRINCIPLES OF COGNITIVE EXPLORATION

This chapter focuses on an exploration of the fundamental principles underlying the investigation of cognitive mechanics. Chapter 2 defined a context for the phenomenon of relevance. Based on the drawing the analogy that research in Information Retrieval can be conceived as the attempt to maximize correlation of input-output estimates of two systems, a computational system and a cognitive system, it was outlined how relevance can be interpreted as a product of cognitive processing. Section 2.3 was dedicated to the investigation of RQ 1. The result of this investigation consisted of the suggestion to base the validation of IR constructs on the theoretic foundation of construct validity, and the nomological network methdology. In Section 2.3.3 it was emphasized, that the required steps for the construction of the network consist of the choice of a pool of candidate constructs, and the establishment of criteria for the selection of the constructs. These aspects are addressed by RQ 2 and RQ 3.

Addressing these questions requires the consideration of the state of the art of cognitive science. To provide a basis for their exploration, this chapter is concerned with the provision of an overview of principles guiding the investigation of the mind. The structure of the exploration of the relevant subject matter is provided in the subsequent introduction.

3.1 Introduction

The exploration of the principles to the investigation of the mind is structured as follows.

The first part of the exploration, covered in Section 3.2, provides a general outline of the field of cognitive science. With regard to the aim of utilizing knowledge of the mechanics of cognition within an IR specific focus, emphasis is placed on the elaboration of the following points:

1. The main meta-theoretical principles underlying the exploration of the mind in cognitive science.
2. The state of the art of the applied experimental means operating on basis of these principles.
3. An outline of the state of the art of insights to cognition via the delineation of exemplary research.

The relation of these points to the research focus of the dissertation is the following. The exploration of the first two items aims at providing context for the investigating of RQ 2 and 3. The third point serves to establish a knowledge basis for the identification of contributing cognitive processes within the scope of the IR domain. This primarily addresses the question raised by RQ 2. As a reflection of the inter-disciplinary nature of cognitive science the discussions in the Section 3.2 are organized as follows. First a definition of the field of cognitive science is provided. This includes an outline of a selection of its main contributing scientific fields. The description of each field is comprised of a delineation of its state of the art, experimental means, and applied methodologies. These points are outlined on the basis of the presentation of exemplary research.

The second part of the chapter is comprised of Sections 3.3 and 3.4. These sections focus on the exploration of the Information Processing Paradigm. The paradigm constitutes the basis for the mind as a machine view of cognition. In this it forms the basis for the cognitive processing view of cognition that underlies the Correlation to Cognition analogy introduced in Section 2.2. Following this, an exploration of the concept of identifiability is provided. The investigation aims at highlighting the limitations of exploring the mechanics of the cognitive system. These aspects are specifically significant and relevant to the described grounding of IR in cognition discussed in Chapter 4 and the development of a concrete methodology as part of the implementation of the paradigm handled in Chapter 5. The chapter is concluded in Section 3.5.

3.2 Cognitive Science

Rich scientific disciplines defy easy definition. Cognitive Science, with the human mind at the centre of its focus of study, proves no exception. Consequently some deviation can be observed with regard to attaining a concrete definition of the field. A

general definition is provided by Nadel (2005, p. 1) in the 'General Introduction to the Encyclopaedia of Cognitive Science'.

> *For present purposes Cognitive Science can be defined broadly as the scientific study of minds and brains, be they real, artificial, human or animal.*

With regard to its inter-disciplinary nature and the contributing scientific fields Thagard (2005) (also Thagard (2010)) proposes the following selection:

> *'Cognitive science is the interdisciplinary study of mind and intelligence, embracing philosophy, psychology, artificial intelligence, neuroscience, linguistics, and anthropology.'*

In addition to the fields listed by Thagard the literature occasionally also includes the fields of sociology and education. The diversity of the included fields reflects the inherent difficulty of establishing such demarcations. On the basis of its defined subject of study, this applies perhaps more strongly to cognitive science. This large inter-disciplinary span, in combination with the contributing field's complexity, induces the necessity to strategically structure the subsequent exploration. This concretely manifests itself in the following form:

- **Categorization:** A fundamental principle underlying research in cognitive science is presented by its emphasis of the necessity of approaching the research task on multiple levels of abstraction. The principle results in the involvement of a large number of distinct scientific fields and an large body of reported work. As such, it poses a significant challenge with regard to the task of outlining the relevant core aspects in a concise manner. To maximize the utility of the exploration and to enable some form of relative positioning of the knowledge in cognitive science, a weak taxonomic system is introduced. This allows for a categorization of the delineated work.

- **Choice of Portrayed Fields**: Within the exploration, the focus of the discussion is placed on the delineation of a subset of the contributing fields. The choice of fields is made on basis of the categorization scheme. The underlying aim consists of establishing a fundamental basis for the research approach of the dissertation. The scope of the exploration is set to cover multiple levels of abstraction for the investigation of cognition.

Due to their importance these two aspects are subsequently discussed in more detail.

3.2.1 Categorization of Cognitive Science

As noted before, the expanse of cognitive science in terms of the range of topics, different levels of abstraction, and scope is immense. Even on the basis of a selection that is limited to its associated core domains (Miller, 2003; Nadel, 2005; Thagard, 2005), it still includes research fields as diverse as philosophy (Bechtel, 1988), biology, and computer science. In consequence, its scope of investigation ranges from the macroscopic to the microscopic to the virtual world – from the individual to the societal perspective (Zerubavel, 1999) – and its focus entails observations from anthropological history (D'Andrade, 1995) as well as work on artificial intelligence (Neisser, 1967).

This listing, outlining the diversity, inherent contrasts, and depth of the field, highlights cognitive science's intrinsic emphasis to base the investigation of the mind on multiple levels of abstraction. This central tenet of cognitive science and results in the inclusion of a large number of scientific disciplines. This induces the necessity for the introduction of some form of weak taxonomic ordering principle as a means of providing structure to the discussion. A sophisticated effort with regard to the introduction of such an ordering principle was proposed by Newell (1994). The inherent difficulty of the task and a hint with regard to the attainable level of specificity is illustrated by his following statement:

> *'The question for me is how can the human mind occur in the physical universe. We now know that the world is governed by physics. We now understand the way biology nestles comfortably within that. The issue is how will the mind do that as well.'– Lecture Allen Newell, December 4, 1991, Carnegie Mellon University*

The quote illustrates that any form of categorization can only be achieved on the basis of elementary dimensions. The elementary dimension chosen by Newell as a means of positioning cognitive observations is given by 'Time'. Newell utilized this measure to classify the occurrence of human action on the basis of its scale. An illustration of the categorization of human action on basis of a temporal scale is shown in Figure 3.1.

Newell (1994) defined arbitrarily chosen temporal foci of observation. These are utilized as a means of binning actions with regard to their duration. Based on this he distinguishes four distinct cognitive bands. Anderson (2002, p. 87) describes this in the following form:

> *... each successive band captures the human experience at roughly 3 orders of magnitude greater than the previous. Ten millisecond effects are at the upper level of Newell's Biological Band while educational effects of consequence are firmly in his Social Band.*

Time Scale of Human Action			
Scale (sec)	Time Units	System	World (theory)
10^{7} 10^{6} 10^{5}	months weeks days		**Social Band**
10^{4} 10^{3} 10^{2}	hours 10 min minutes	Task Task Task	**Rational Band**
10^{1} 10^{0} 10^{-1}	10 sec 1 sec 100 ms	Unit task Operations Deliberate Act	**Cognitive Band**
10^{-2} 10^{-3} 10^{-4}	10 ms 1 ms 100 µs	Neural circuit Neuron Organelle	**Biological Band**

FIGURE 3.1: Newell's Cognitive Bands. Figure based on Newell (1994)

In his own words Newell (1994, p. 121) remarks that such '[d]ifferent bands are quite different phenomenal worlds as shown in the right-hand column, and are described by different theories.' Through the application of time as a scale it is possible to relatively position these actions to each other. It is of note that such a categorization additionally opens a path to a hierarchical structuring of cognitive actions. Actions falling within the time spans demarked by the cognitive band are composed of a set of actions falling within the biological band, and in turn form part of actions taking place within the range set by the rational band. This allows using Newell's Cognitive Bands as a means of providing structure during the exploration of the different fields of cognitive science. The categorization of cognitive actions on a temporal scale inherently, as described by Newell above, also enables the categorization of associated theories. This notion is expanded to the scope of scientific domains on the basis of the field-specific associated theories. Through this, it is possible to postion the scientific fields on the temporal axis. The categorization of scientific theories on such a basis shares the same general notions as expressed in the statement by Newell (1994, p. 157) below:

> *We don't understand the full variety of mechanisms that contribute to all these system levels. In fact, almost every aspect of the analysis has substantial uncertainty associated with it. That is no reason for not engaging in such speculation. The potential gains are very great,*

namely to get some general guidelines for the cognitive architecture regardless of any specific proposal. ”

To aid with the task of applying structure to the exploration of cognitive science, Newell's bands can be expanded with additional elementary dimensions. The introduction of these dimensions further serves to illustrate the particular challenges and constraints underlying cognitive research. This is also considered a necessary prerequisite for the task of associating the theoretic framework of cognitive science with the domain of Information Retrieval, and as a means of establishing an argumentative basis for the grounding of IR. Figure 3.2 shows these additional dimensions in relation to Newell's Bands. As stated above, the dimensions are primarily chosen with the aim of position-

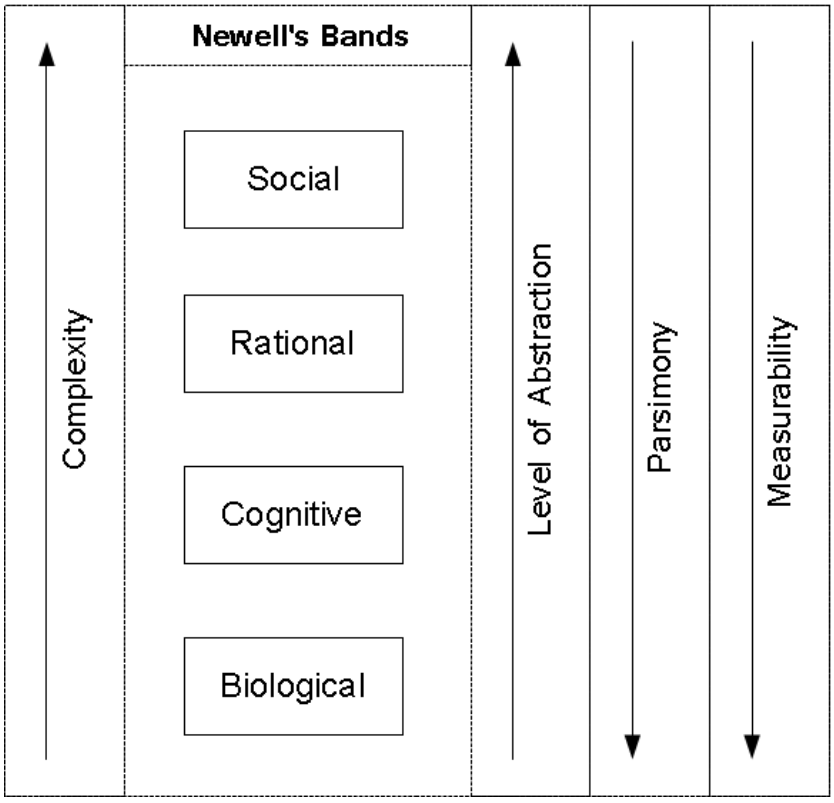


FIGURE 3.2: Interpretation of Newell's Cognitive Bands in terms of complexity, level of abstraction, parsimony, and measurability

ing theories stemming from the various scientific fields in cognitive science. It is of note that the association of such theories with the dimensions rather reflect tendencies than fixed assignments. Statements on grounds of such a basis can only be of relative character. Nevertheless they are considered immensely useful within the following discussion. The justification as well as the underlying motivation for each of the introduced dimensions is provided below.

- Complexity: Complexity is a notion associated with the bands by Newell himself, described in the form of 'Compositional Complexity'. The complexity axis

illustrates that 'higher level bands are hierarchically composed from the rational, cognitive, and biological bands' (Fehling, 1993, p. 58).

- Level of Abstraction: On the basis of such a statement it is possible to make inferences about the implicit level of abstraction of scientific theories. As Rosenbluth remarked: 'No substantial part of the universe is so simple that it can be grasped and controlled without abstraction. Abstraction consists in replacing the part of the universe under consideration by a model of similar but simpler structure.' (Rosenblueth and Wiener, 1945, p. 316). This dimension is to be interpreted in the sense, that a relative assignment of fields in terms of complexity enables their relative assignment in terms of the level of abstraction. The underlying premise being, that higher levels of complexity imply higher levels of abstraction.

- Parsimony: The introduction of parsimony as a concept in the philosophy of science serves the aim of projecting statements with regard to the relative positioning of fields and theories in terms of complexity and abstraction into the context of philosophy of science. This of course can only be done in a general fashion. If parsimony is defined as 'simplicity', bar the connotations of philosophy of science, then its relation to complexity and abstraction is intuitive. On basis of such an interpretation of parsimony, biological aspects are deemed to exhibit more parsimony relative to sociological aspects. The underlying aim of such a definition consists of emphasizing the relation between parsimony in the sense of 'simplicity' and parsimony in the sense of, for example, 'quantitative parsimony' as defined by Nolan (1997). Quantitative parsimony being defined by Nolan as the theoretic virtue of trying to minimize the number of newly postulated individual entities[3-1].

- Measurability: Measurability is an expression of the certainty with regard to *what* one is measuring. It is intuitive that the measurement of highly abstract concepts such as 'health' or 'happiness', attributed to the sociological level, bear less certainty than measurements of the firing rate of a synapsis residing on the biological level. These differences with regard to the quality of measurements are themselves hard to qualify. However, the introduction of the concept allows to illustrate the relationship between complexity, level of abstraction and the ability to measure with certainty. Generally this can be expressed as the notion, that with descending level of bands there is an increase of certainty with regard to *what* one is actually measuring.

On the basis of this basic layout the choice of delineated knowledge is outlined in the

[3-1] Nolan (1997) exemplary illustrates this concept on basis of the discovery of the neutrino. Prior to its discovery physicists were puzzled by the mismatch between the total observed spin of particles before and after decay. The lack of spin of the observed emitted particles lead to the postulation of a new, then unobservable particle as the source of the spin. The most quantitatively parsimonious assumption consists of assuming *one* Neutrino to be responsible for the observed effect.

next section.

3.2.2 Choice of Delineated Fields

In the remainder of the chapter selected parts of the knowledge base of cognitive science are outlined. The purpose of this Section consists of motivating the underlying choice. Re-iterating the agenda of the chapter, the purpose of this delineation is the following:

- A projection of the knowledge in cognitive science towards the domain of IR. The specific motivation herein lies in:
  - Illustrating the relation and significance of such knowledge to IR.
  - To categorize such knowledge on the basis of the dimensions of complexity, level of abstraction, parsimony and measurability that, together, form the introduced weak taxonomic structure.
- Conduct a preliminary analysis with regard to the research approach of the dissertation on basis of a broad survey of the following aspects:
  - Paradigms
  - Methodologies
  - Experimental Means

In view of this, the particular fields are chosen on basis of the following argumentation. The analogy presented in Section 2.2 describes relevance as a phenomenon occuring within the context of two systems: An IR system and the cognitive system of the user. As described in Section 2.3 the analogy provides a general context for the validation of Information Retrieval constructs. On the side of the natural system, the pool of potential constructs for the network is given by the state of the art constructs of cognitive science. In light of this the Section aims at a broad coverage of delineated fields. From another perspective the choice is motivated by the same tenets that provide favourable arguments (Sun et al., 2005) with regard to the existence of the inter-disciplinary field of cognitive science. As expressed in the words of Thagard: 'The best way to grasp the complexity of human thinking is to use multiple methods, especially psychological and neurological experiments and computational models.' (Thagard, 2005, p.10) On this basis the exploration is focusing on the following fields:

- Philosophy
- Cognitive Psychology

- Neuropsychology

Figure 3.3 shows these fields positioned relative to Newell's bands and the additionaly introduced dimensions.

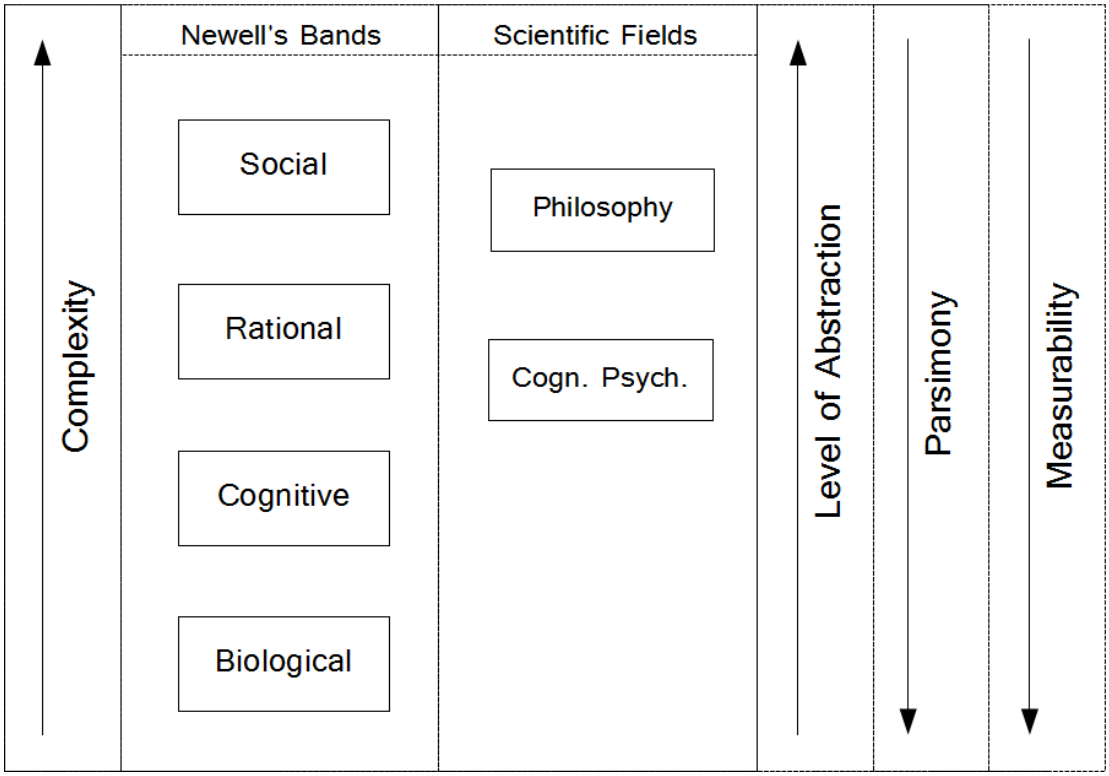


FIGURE 3.3: Coarse Alignment of Fields Within Newell's Cognitive Bands

The argumentation with regard to the positioning of these fields according to Figure 3.3 will be provided subsequently as part of each field's exploration. What can be derived from the figure is the relative position of the fields with respect to the bands and the dimensions. This outlines Cognitive Science's adherence to Thagard's statement regarding the necessity of an inter-disciplinary approach. Further it allows for a qualitative characterization of the fields, and highlights the contrast and the covered span established through the selection. Philosophy as the top-most field in the hierarchy exhibits a tendency to place its research focus on high-level concepts (e.g. Philosophy of Science). This implies a high level of abstraction and complexity of the subjects of research. Relative to the field Neuroscience which focuses on the biological, chemical, and physical aspects of the brain, Philsophy exhibits a measurability that is magnitudes lower. This aspect is discussed in the subsequent section in more detail.

3.2.3 Philosophy

While there is almost no contention with regard to the question whether Philosophy ought to form an integral part of cognitive science, there is large variation with regard to the interpretation of the nature of that role. Subsequently the role of Philosophy is cognitive science is evaluated with regard to this particular question. Further, the field is related to the Cognition to Correlation analogy and the research focus of the dissertation.

Meta-Scientific Considerations

By its intentions, as reflected in its definitions cast in natural language, a possible general statement on Philosophy consists of assigning to it a tendency to cast its focus on the 'edges of knowledge'. The above quoted notion and the resulting role of Philosophy is expressed by Thagard (2009, p. 239) in the following form as part of his discussion of 'Why Cognitive Science Needs Philosophy and Vice Versa'.

> *Whenever science operates at the edge of what is known, it runs into general issues about the nature of knowledge and reality. Mundane science can operate without much concern for methodological and ontological issues, but frontier science cannot avoid them. For example, innovative research in theoretical and experimental physics inevitably encounters fundamental problems about the nature of space and time, as well as methodological questions concerning how scientific investigation should proceed.*

On a general level Thagard (2009) expresses in this work, that some of the most important contributions that Philosophy can make to cognitive science concern questions of the following nature:

- What is the nature of the relations among the fields in cognitive science?
- What constitutes theories and explanations in cognitive science?

Broadly this role can be described as facilitating the interpretation of concepts and enabling reasoning of a relative nature between such concepts. (Dennett, 1984, p. 12) expresses this in the following form: 'Indeed one of philosophy's highest callings is finding ways of helping people see the forest and not just the trees.' In more elaborated form and directly aimed at cognitive science this view is expressed by (Malcolm, 1971, p. 59) in the as follows:

> *'We must reject the doctrine, so powerful in modern philosophy, that we acquire concepts of mental occurrences by observing those occur-*

> *rences taking place in ourselves. In rejecting it we remove the chief source of the temptation to think that a human mind could exist and be provided with concepts, in isolation from a human body and from a community of living human beings.'* ”

Malcolm's statement emphasizes several points. Firstly, it underlines the need for an inter-disciplinary approach to cognition by outlining that cognition is a phenomenon occuring within the context of a human body, a community, and intrinsic observations. The notion is closely related to the discussion surrounding the social, cognitive, and biological bands of Newell. The application of philosophical concepts such as parsimony and measurability in Section 3.2.1 of this chapter can be seen as an example regarding the role of philosophy in cognitive science. As outlined in the next section, another important contribution to cognitive science is given by philosophy's inspiring nature.

Inspirational Aspect

Another important contribution of philosophy is given by its inspirational role. Examples of such a role are provided in the course of Thagard's (2009) discussion of philosophy's role in cognitive science (p. 238):

> “ *Sometimes philosophical ideas can be useful in stimulating scientific investigations, for example, when some of Wittgenstein's ideas about language inspired important 1970s research about the prototypical nature of concepts, and when Daniel Dennett's views of intentional action triggered a flourishing research tradition in developmental psychology concerned with children's judgments about false beliefs.* ”

Shifting the focus back to IR, Philosophy has fulfilled such a role as exemplified by the utilization of work from the sub-fields of Philosophy of Knowledge, Language, and Logic. In particular the work of Dretske (1983), Gardenfors (2004), Barwise and Seligman (1997) and Devlin (1995) is representative with regard to this aspect.

Conclusion

This outline is concluded with a taxonomic interpretation on the grounds of Figure 3.3. As illustrated throughout the discussion in the Section, the main contributions of Philosophy tend to be set on a meta-theoretical level. They concern the analysis of the relationship between fields, or are of inspirational nature. When contrasting the observation with the taxonomic illustration in Figure 3.3 a possible interpretation of this is provided by the following.

Philosophy operates on a high level of complexity and abstraction. Along with this focus comes a higher probability of its output being of significance in terms of the human experience. This can be exemplified by looking at the work of Barwise and Seligman in the sense that their subject of studies, meaning and inference, are of high relevance to the human experience. It should be highlighted again, that all statements with regard to the dimensions are only of relative character. They are not meant to represent universal statements on areas of scientific investigation as diverse as philosophy. However, the discussion aims at establishing guidelines with regard to the methodologies and the nature of knowledge in cognitive science. This is based on the taxonomic system we introduced. The categorization is seen as a first step towards the identification of potential elements for a validation of IR constructs. Further, the discussion in the Section is meant to emphasize the importance of a multi-level perspective on cognition.

3.2.4 Cognitive Psychology

Cognitive psychology is a subdiscipline of psychology. Its primary aim lies in the exploration of internal mental processes. As such, it is highly relevant to the dissertation's research questions concerned with processing models of cognition. The advent of cognitive psychology as an independent field is often seen as being demarked by coinage of the term by Ulric Neisser in the equally named publication of the book 'Cognitive Psychology' (Neisser, 1967, p. 4). The following excerpt taken from this publication serves as a definition of the term 'cognition':

> *The term 'cognition' refers to all processes by which the sensory input is transformed, reduced, elaborated, stored, recovered, and used. It is concerned with these processes even when they operate in the absence of relevant stimulation, as in images and hallucinations... Given such a sweeping definition, it is apparent that cognition is involved in everything a human being might possibly do; that every psychological phenomenon is a cognitive phenomenon. But although cognitive psychology is concerned with all human activity rather than some fraction of it, the concern is from a particular point of view.*

The particular viewpoint expressed in the above situation refers to the focus of cognitive psychology on the internal processes and states of the mind. Lu and Dosher (2007) outline that this specific focus is based on two primary assumptions:

> *Human cognition can, at least in principle, be fully revealed by the scientific method, that is, individual components of mental processes can be identified and understood (1), and (2) Internal mental processes can be described in terms of rules or algorithms in information*

processing models. ”

3.2.5 Conclusion

’The struggle of Psychology has always been to say things of significance to the human experience that have a rigorous scientific foundation.’ [Anderson (2002, p. 85)]

This statement by Anderson outlines that the focus of Cognitive Psychology lies between the philosophical approach and the biological approach. In this role cognitive psychology has integrated experimental means that allow for the measurement of psycho-physical measures such as response time and eye movements, as well as neurological measures on the basis of techniques such as FMRI[3-2] and ERP[3-3]. The application of these means underlies paradigmatic approach that guides their strategic application. The Information Processing paradigm, while not exclusive to cognitive psychology within the cognitive sciences, is utilized in a bridge-building fashion in this field. The next Section explores the concept in more detail.

3.3 Information Processing Paradigm

The Information Processing (IP) Paradigm has been the dominant paradigm for the exploration of the mind within cognitive science in general, and cognitive psychology (see Massaro and Cowan (1993); Lachman et al. (1979); Anderson (2005)) in particular. It holds a central role concerning the execution and posterior interpretation of research stemming from the different disciplines of cognitive science. At the core of the paradigm stands the analogy of interpreting the mind as some kind of apparatus or machine, capable of processing incoming information[3-4]. As outlined in Section 2.2 this analogy is in particular motivated by the aim to acknowledge the importance of the ’internal mental world’ (Hutchins, Edwin and Lintern, 1995) with regard to an exploration of cognition. On the basis of its paramount position with regard to the conduction of such research, the IP paradigm forms a core part with regard to the investigation of the first three research questions of the dissertation. In light of this, subsequently key aspects of the paradigm are outlined.

[3-2]Functional magnetic resonance imaging is a magnetic resonance imaging technique that measures brain activity based on blood flow.

[3-3]ERP refers to electrophysiological measurement of event related potentials in the brain.

[3-4]From a cognitive psychology centered view Massaro and Cowan (1993, p. 384) describes information as: ”Information,’ though difficult to define precisely, refers to representations derived by a person from environmental stimulation or from processing that influences selections among alternative choices for belief or action.’, from a philosophical point of view the difficulties surrounding the concept of ’Information’ are described by Floridi (2004)

Key Assumptions

The enumeration below lists the five key assumptions of the paradigm as defined by Palmer and Kimchi (1984, p. 73).

1. **Informational Description:** 'Mental events can be functionally described as 'informational events,' each of which consists of three parts: the input information (what it starts with), the operation performed on the input (what gets done to the input), and the output information (what it ends up with).'

2. **Recursive Decomposition:** 'Any complex (i.e., nonprimitive) informational event at one level of description can be specified more fully at a lower level by decomposing it into (1) a number of components, each of which is itself an informational event, and (2) the temporal ordering relations among them that specify how the information 'flows' through the system of components. '

3. **Flow Continuity:** 'All input information required to perform each operation must be available in the output of the operations that flow into it.'

4. **Flow Dynamics:** 'No output can be produced by an operation until its input information is available and sufficient additional time has elapsed for it to process that input.'

5. **Physical Embodiment:** 'In the dynamic physical system whose behaviour is being described as an informational event, information is embodied in states of the system (here called representations) and operations that use this information are embodied in changes of state (here called processes)'

The first point reflects the paradigm's view on cognition by rendering the machine analogy into an abstraction consisting of three elements (input, operation, output). In combination with (2) it represents a view of the mind-machine analogy, as well as a strategy for exploration. Rather than being assumptions with regard to the specifics of cognition itself, these first two assumptions represent guidelines for a mode of conducting research. In contrast to this, the remaining three points express assumptions concerning the functioning of the mind. Continuity and dynamics are assumptions referring to the nature of the mechanics. Physical embodiment can be interpreted as an assumption of the nature of the implementation.

Saracevic (1997, p. 17) statement that '... unlike art IR is not there for its own sake, that is, IR systems are researched and built to be used' allows to position the role of Information Processing paradigm with regard to the research approach of the dissertation. The primary goal of the IP paradigm consists of enabling the exploration of the mechanics of the mind on a neurological basis or organisational basis. As shown in Figure 2.4 the defined aim of IR consists of the achievement of input-output convergence.

With respect to the first three research questions of the thesis, the mechanistic view of IP paradigm serves as a source for the identification of observables and constructs for the inclusion in a nomological network for IR. With regard to this, the first 2 key assumptions apply more directly to the research focus of this work.

3.4 Identifiability

The concept of identifiability in cognitive science expresses constraints in terms of direct verification of cognitive effects. This is primarily induced by the limitations of the contemporary experimental means. The inability to directly observe many cognitive phenomena results in the existence of a multitude of plausible explanatory models for the same phenomenon. The lack of means to identify the 'correct' model is referred to as the identifiability issue in cognitive science. Massaro and Cowan (1993, p. 390) refer to this aspect in the following form:

> *'The so-called identifiability issue concerns whether a given model of an experimental result can be identified as the correct one. The issue arises from the theorems of Moore (1956) and from subsequent work in formal automata theory (...). Moore was concerned with the behaviour of sequential machines. Observers of machines or people can record only their inputs and outputs. It is not possible to look, so to speak, inside the black box. The question is: To what extent can the accuracy of one model of the inner workings of a black box be distinguished from that of another model, given only a set of input-output observations? Moore proved that any input-output function can be exactly mimicked by some other such function. No explanatory model of an experimental result can exclude all others.'*

An initial insight on basis of these observations consists of the realization that for every aspect of mental activity there exist various models and theories. Due to the limitations of identifiability it is not possible to judge which of these models is 'right'. This aspect can be interpreted as a limiting factor in regard to the utilization of cognitive models within an nomological network for IR. This can be interpreted as an emphasis of the need for a grounding of the validation effort that was discussed in Section 2.3.3.

3.5 Conclusion

This chapter conducted an exploration of the discipline of cognitive science. The exploration was aimed at establishing a principled foundation for the exploration of the

research questions of the dissertation. Within this exploration the focus was placed on three key aspects underlying the investigation of the mechanics of the mind.

- **Application of multi-layered abstraction**: Based on a survey of exemplary work the necessity for to approach the investigation of the mind on multiple layers of abstraction was illustrated. On grounds of the introduced taxonomic system it has been shown, that higher levels of abstraction are perceived as necessary for establishing significance with regard to human experience on an individual or social level. The utilization of lower levels of abstraction on the other hand is applied on basis of the motivation of establishing 'measurability' through the 'grounding' of higher abstractions in physically and physiologically based measures. Of particular importance with regard to the implementation of the research approach of the thesis is the positioning of the central IR concept of 'relevance' on the upper levels of the abstraction band, and the potential implications with regard to its measurability.

- **Information Processing paradigm**: The IP paradigm as the dominant paradigm underlying research in cognitive science has been outlined in Section 3.3. Regarding the paradigm's focus on the mechanics of the mind, it represents a fundamental basis for the implementation of a nomological network for IR. As noted in Section 3.3, the first two tenets of the IP paradigm (Palmer and Kimchi, 1984) are conceived of being of particular relevance to the instantiation.

- **Identifiability Issue**: The discussion of the identifiability issue highlighted the verification limits of the IP approach. It emphasizes the point, that no specific theory or model within cognitive science can claim to be the 'right' model. This is specifically relevant with regard to the aspect of basing the instantiation of the nomological network on the identification of cognitive processes of relevance to IR.

With regard to the concrete application of these principles within the next chapters, it is important to place these principles in relation to the research focus of the dissertation. This is based on the 'criteria for evaluating theories of mental representations' defined by Thagard (2005). Figure 3.4 shows a listing of these criteria. Without examining each of the listed points in great detail, it is easy to underline the much less ambitious goals underlying the research goals of the thesis. The underlying aim of this work consists of validating IR constructs. As stated in Section 2.4 the validation approach is based on the creation of a nomological network for IR. This requires knowledge of the core constructs relating to the phenomenon of relevance. The exploration of cognition is therefore focused on the extraction of principles of cognitive processing and the identification of contributing constructs. In regard of this, two of the main points of Thagard's listing, 'psychological' and 'neurological plausibility' do not form part of the motivation for the exploration of cognitive aspects in this work. In that regard, the

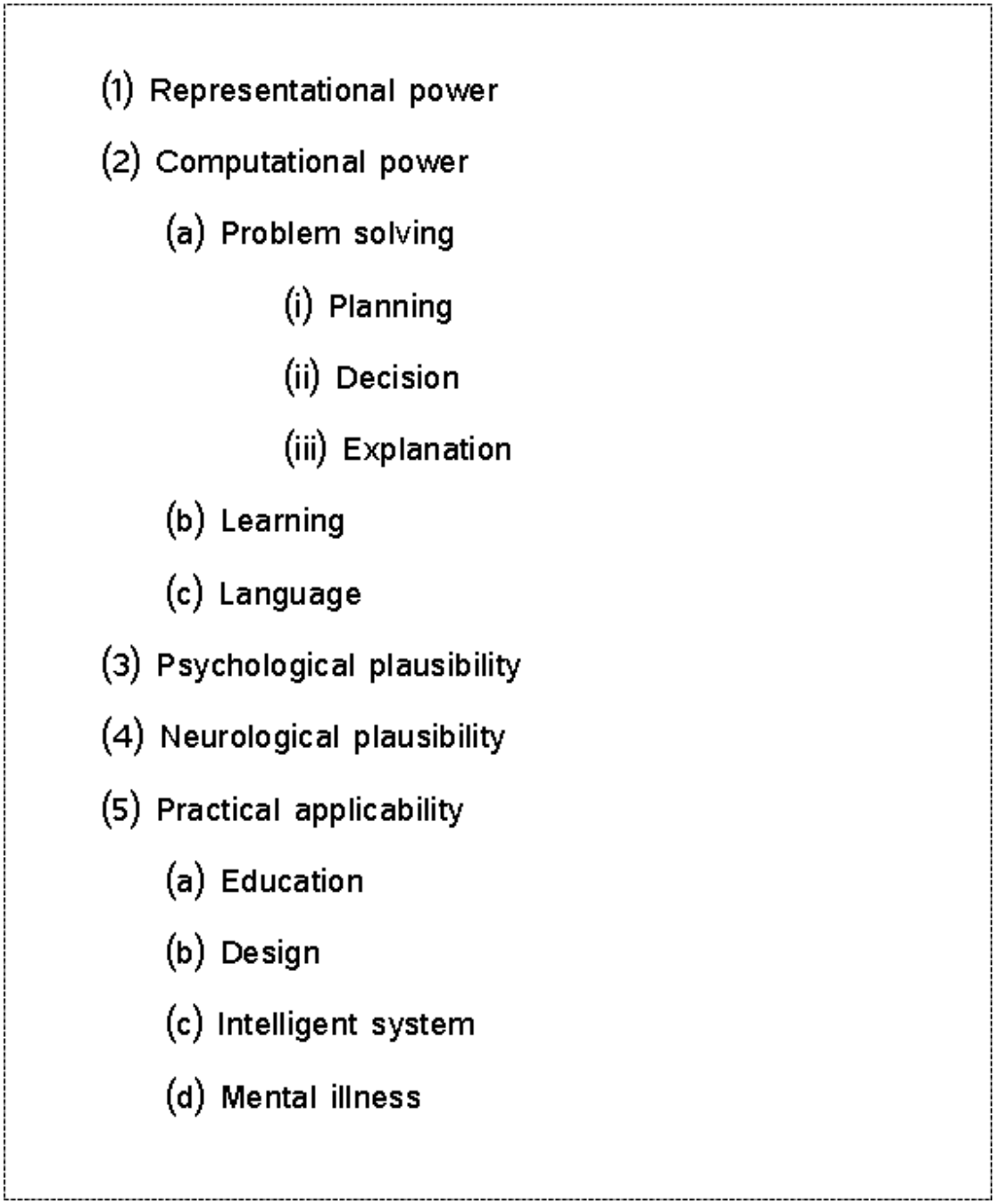


FIGURE 3.4: Criteria for evaluating theories of mental representations based on Thagard (2005)

discussed investigations in the subsequent chapters are of much less ambitious nature relative to the aim of cognitive science. Based on these theoretic foundations, the subsequent chapter explores the identification of cognitive activity that is of relevancy to the context defined by the Correlation to Cognition analogy.

CHAPTER 4

GROUNDING INFORMATION RETRIEVAL IN COGNITION

This chapter explores the identification of cognitive activity pertinent to IR. Chapter 2 outlined the role of construct validity as a principled approach to the validation of IR constructs. Two preliminary steps were identified with regard to the application of the nomological network methodology introduced by Cronbach and Meehl (1955). The first is given by the definition of a context for the network and the choice of the set of constructs to be included in it. These considerations are addressed by RQ 2.

RQ 2 What are potential constructs for the formulation of an IR focused nomological network?

The pool of IR constructs to include in the network is given by the state of the art of IR theory. The pool of constructs referring to the natural system is given by the theories of cognitive processing pertinent to the mental activities conducted by an IR system user. The intrinsic complexity of the task to identify these constructs is illustrated by dissecting the clause 'identification of cognitive activity of relevancy to IR' into the two subclauses 'identification of cognitive activity', and 'of relevancy to IR'. On basis of this division each subclause emphasizes a specific research aspect. The first aspect 'identification of cognitive activity' concerns the development of adequate means for the investigation of cognitive activity. The prerequisites for such an investigation consists strategic research approach and associated experimental methodologies that enable the investigation and 'charting' of the cognitive processing landscape of the mind. The fundamental paradigmatic and strategic research principles of cognitive science represent an implementation of these prerequisites. An overview of these principles and a discussion with regard to their application within the dissertation's research effort was presented in Chapter 3. In summary, these principles are interpreted to provide an argumentative basis for the investigation of RQ 2 in this chapter. The existent body of

knowledge in cognitive science, that resulted from the application of such principles, represents a primary source on which to base the identification of cognitive processes that are pertinent to an IR context.

The exploration of this matter is structured as follows. Section 4.1 provides an outline of cognitive processes pertinent to the IR context. The exploration focuses on the cognition of text based information processing. Limiting the investigation to text based processes narrows the scope of the relevant subject matter for RQ 2. The motivation for this limitation is twofold. Firstly, text based information processing constitutes one of the most well researched and understood phenomena of cognitive processing. Secondly, this state is mirrored on the IR. Text based Information Retrieval represents the dominant and most widely explored and applied use of IR theory. With regard to identification of constructs to include in an IR focused nomological network, it is considered favourably to base the identification effort on a large body of reported work. On basis of these considerations, Section 4.2 reports on an interpretation of 'relevance' on basis of cognitive processing. This 'grounding' of 'relevance' is aimed at the provision of an overview of the cognitive activity underlying the phenomenon of relevance. Specific focus within the outlined discussion, on basis of the introduced mapping, is placed on the elaboration of the following research issues.

- A cognitively grounded discussion with regard to the difficulties to formally define 'relevance' in an IR centric sense.

- A cognitive processing centric analysis with regard to the identification of cognitive processes and constructs contributing to the phenomenon of relevance.

The exploration of these aspects aims at highlighting the benefit of cognitively grounding high-level[4-1] abstractions in cognitive dimensions that are associated with lower levels of abstraction.

On basis of this theoretical background, Section 4.3 adds substance to the analogy defined in Section 2.2. This alignment constitues an inter-disciplinary 'bridging'[4-2] of models from distinct scientific fields of IR and cognitive science. In this form it represents a mapping of cognitive processes and IR retrieval centric computation that enables the strategic selection of cognitive processes with regard to steps of implementing the paradigmatic research approach described in Section 1.3. Further, this cognitive 'grounding' of IR represents a pool for the selection of constructs for an IR focused nomological network.

[4-1] For a clarification of the relative meaning of this expression see Section 3.2.1

[4-2] A detailed elaboration regarding meta-scientific aspects of the 'bridging' of concepts from distinct scientific fields and the associated complexity of the task is provided by Thagard (2006) on base of the inter-disciplinary cognitive science landscape.

4.1 Text Based Information Processing and Reasoning

As a reflection of the later described experimental implementation of the paradigm, this section focuses on an exploration of relevant cognitive knowledge with regard to text based information retrieval as opposed to for example multimedia based information retrieval. In particular the section focuses on outlining the state of the art in the cognitive science of text based information processing and reasoning. Its purpose lies in forming a basis for the subsequently proposed grounding of IR in cognition.

Identifying relevant cognitive processes based on a survey of the existent state of the art relies on a couple of abstract working hypotheses. This aspect can be illustrated through an example. The act of reading forms an integral part of a user's IR interaction. Therefore it is assumed that the cognitive processes attributed to take part in the reading process are of relevance to IR. The abstract hypothesis with respect to the use of the existent body of research from cognitive science is then the following:
The cognitive activity underlying the acts of reading occurring in an Information Retrieval context and a non-IR context do not differ fundamentally.
On basis of this hypothesis the next subsection provides an overview of the state of the art cognitive background of text based discourse processing.

4.1.1 Text Based Discourse Processing

This section provides a brief outline of the cognitive processes thought to play a role in the processing of textual discourse. The majority of the research with regard to such discourse processing is based on the application of the Information Processing paradigm. As a consequence the cognitive activity associated with discourse processing is dissected into a subset of modular processes. In the contemporary literature these processes are attributed to different stages of the reading process. The next subsection provides an overview of such stages.

Stages of Reading

With regard to stages of reading Graesser et al. (1997, p. 167) draw the following summary in his review of 'Discourse Comprehension':

> *Most discourse psychologists adopt van Dijk & Kintsch's (van Dijk and Kintsch, 1983) distinctions among the surface code, the textbase, and the referential situation model. The surface code preserves the exact wording and syntax of clauses. Comprehenders normally retain the surface code of only the most recent clause unless aspects*

> *of this surface code have important repercussions on meaning. The textbase contains explicit text propositions in a stripped-down form that preserves meaning, but not the exact wording and syntax. The textbase also includes a small number of inferences that are needed to establish local text coherence. The situation model is the content or the microworld that the text is about. The situation model for a story refers to the people, spatial setting, actions, and events in the mental microworld. This microworld is constructed inferentially through interactions between the explicit text and background world knowledge.* ”

An overview of these stages in graphical form is presented by figure 4.1. The figure helps to emphasize the complexity of the task. This aspect is described by Graesser et al. (1997, p. 165) in the following form: Such processing ’is intertwined with virtually all cognitive functions and processes, including memory, perception, problem solving, and reasoning’. The complexity is further illustrated through the following statement by Massaro and Cowan (1993, p. 384):

> “ *Recursive decomposition, perhaps better described as hierarchical decomposition, denotes the breaking down of one stage of processing into substages. For example, a memory stage can be broken down into acquisition, retention, and retrieval stages; retrieval can be further broken down into memory search and decision; and memory search can be further broken down into access and comparison stages.* ”

Against the background of this complex set of contributing processes Graesser et al. (1997, p. 168) remark that ’it is a profound understatement to say that these various levels interact with one another in complex ways that are not well understood.’

Nevertheless these stage based models essentially constitute a hierarchical structure of specific processes that contribute to the cognitive activity of reading. As such, they can be utilized as a sort of guiding framework on which to base the implementation of the paradigm. To evaluate this aspect in more detail the three distinct phases are listed below.

- Surface code construction
- Textbase construction
- Situational model construction

Each stage can be understood to represent a different level of abstraction. The construction of the situational model represents the highest level of abstraction. The surface code construction constitutes the lowest level of abstraction. Subsequently some of the important aspects and specific processes of these stages are outlined in more detail.

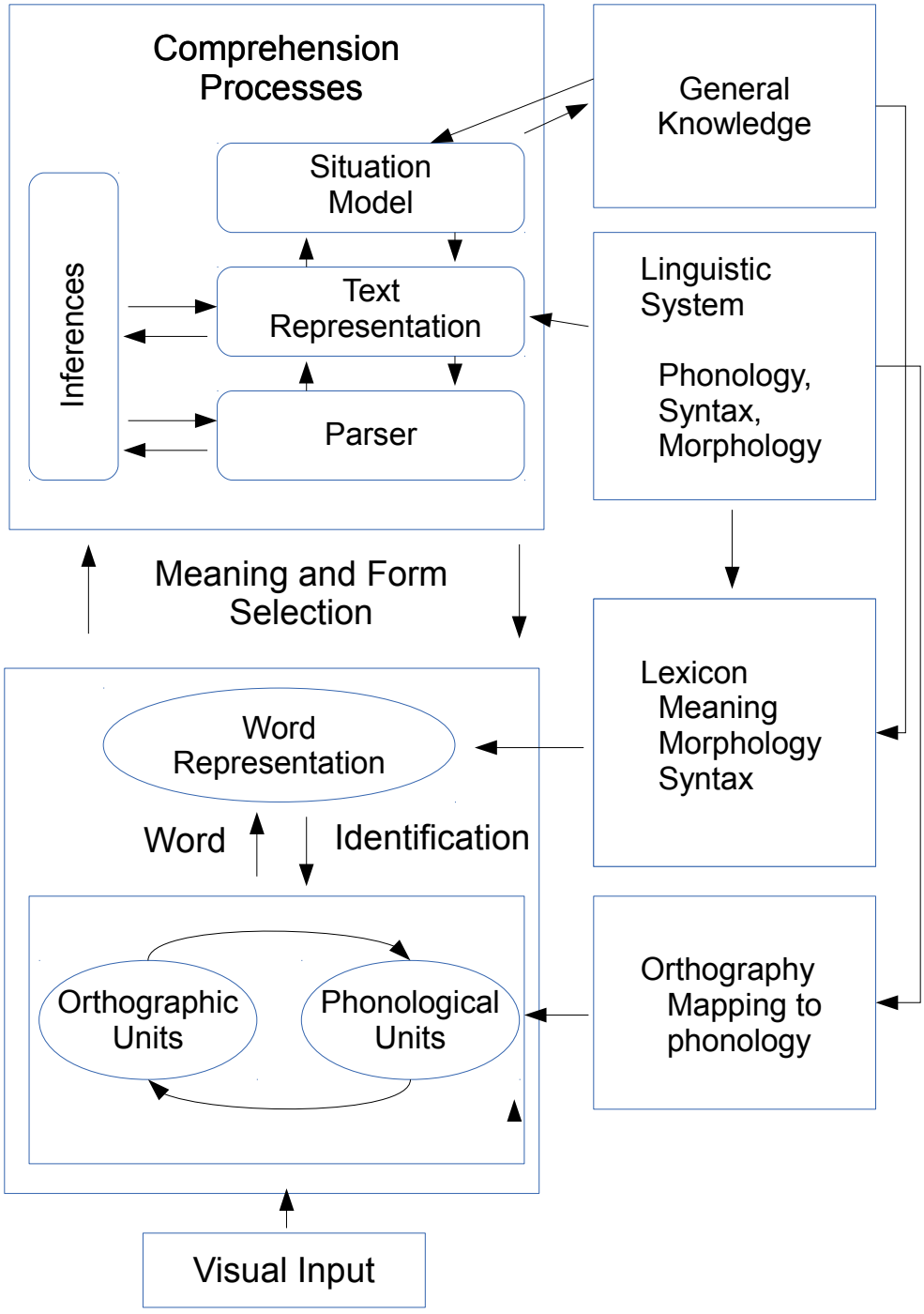


FIGURE 4.1: Perfetti's Model of Discourse Comprehension (Perfetti et al. (2005))

The outlined processes are specifically selected with regard to the aim of identifying constructs and processes pertinent to relevance.

Word Meaning Identification

With reference to figure 4.1 it can be seen that the identification of the meaning of a word is a process that takes a primary role in the construction of the textbase. The below posted sample sentence is well suited to outline this specific process.

> *A number of religions in Ankh-Morpork still practiced [emphasis added] human sacrifice, except that they didn't really need to practice [emphasis added] any more because they had got so good at it. (Pratchett, 1989, p. 309)*

The added emphasis underlines that the meaning of the word *practice* is differing at its occurrences. Inferring the meaning cannot be achieved solely on basis of word recognition alone. It requires some form of contextual analysis. A first step towards identifying the meaning of a word is assumed to consist of the activation of related words (see Swinney (1979)). The underlying mechanism is dependent on the theoretical background. On basis of a connectionist interpretation, such activation processes might be based on a form of spreading activation as described by Anderson (1983). In the above provided example this might result in the activation of words such as 'training, sport, doctor, profession' on encountering the word 'practice'. This sample listing illustrates that the identification of the intended meaning of the word requires more elaborate processes. With regard to such processes Kintsch (1988, p. 170) provides the following overview.

> *As more information about the context becomes available, the sentence and discourse meaning begin to emerge, and more and deeper plausibility checks can be performed as long as there still is time. This is the sense-elaboration phase, in which the meaning of a word is contextually explored and elaborated. ... Thus, word meanings are usually identified long before complex inferences are made in comprehending a discourse. At this point, a 'meaning' has been constructed for the word in this particular context. It consists of the lexical node that has been activated (the contextually inappropriate nodes that had been activated have by now been deactivated through the various context checks), the associative and semantic neighbours of that node, the sentence and discourse context in which the word participated, and some inferences and elaborations that were produced in the course of the various plausibility checks that explored the role of that word in the given context.*

”

The outline illustrates that the process of word meaning identification itself can develop considerable complexity. Of importance with regard to the experimental realization of the paradigm is the process of inferring the relation between words. As noted above a first automated step in word meaning identification consists of an activation of related words. While not implicitly outlined in the description of Kintsch the 'estimation' of such relations of words to each other represents an *elementary* part of any higher level comprehensive processing. The emphasis outlines that this process is interpreted to reside close to the bottom of the hierarchical process chains. This is also illustrated by the fact, that cognitive theories focused on this aspect are commonly proposed in reference to the organization of memory. The process constitutes a process with relatively low level of abstraction. To conclude this brief exploration, the next subsection covers the construction of situational representations; a process operating at a higher level of abstraction.

Construction of a Situational Representation

The construction of a situational representation forms part of all models of discourse comprehension (see Kintsch (1988); McNamara and Kintsch (1996); Zwaan and Radvansky (1998)). Two common elements of such models are presented by the following points:

- The dependency of the construction process on world knowledge.
- The dynamic nature of the process that is implied by this dependency.

'World knowledge' refers to all available memorized information of a specific individual. It is intuitively understandable, that the comprehension of a scientific article for example is highly dependent on the existing knowledge of the reader. Further it is intuitive that this process is dynamic, as the act of reading itself contributes to the pool of memorized information. With regard to Information Retrieval related tasks this aspect is naturally of high importance as it outlines that the interpretation of textual information is dependent on subjective knowledge (Graesser et al., 2002; Hambrick and Engle, 2002).

4.1.2 Text Based Reasoning

The amount of knowledge in the form of descriptive or formal models with regard to reasoning and decision making on basis of text based discourse comprehension is very limited. As noted by Harley et al. (2011) much of the research on language processing has focused on fully or almost fully automatic steps within the process of discourse

comprehension. Automatic processes include visual word recognition, parsing, comprehension, syntactic planning, and lexicalisation. Processing that concerns text based reasoning is referred to by Harley as fully deliberative processes in language.The role of these processes is described by Harley as becoming necessary 'when the going gets tough'. Examples for situations that require deliberative processing are given by the formulation of judgement about linguistic representations, resolving conflict, and engagements in any sort of language planning.

The subsequent quote by Harley et al. (2011, p. 126) illustrates the rudimentary state of knowledge with regard to the deliberative processing of language.

> *They are among the least understood and studied processes in the field of psycholinguistics; indeed, it is not always even recognised that they are involved in mental activity at all. However, these processes control the inputs and the outputs of the language modules, and play a vital role in linguistic behaviour.*

In summary, the quote outlines the limited amount of knowledge concerning the mechanics of high level cognitive reasoning processes. However, while the exact mechanics of such processing is unknown, extensive knowledge exists with regard to contributing sub-processes. With respect to human decision making, an important subprocess is given by the impact of emotions. The next subsection provides a brief overview of the matter.

Emotion

While to our knowledge no research specifically directed at decision making processes that apply to an IR context has been conducted, it is plausible to assume that emotion plays a role within such cognitive processing. The impact of emotion on decision making has long been underestimated. Initially, decision making processes have been interpreted as completely rational. The work of Bechara et al. (2000) showed that emotion plays an important role in such processes. A summary of the role of emotions in decision making processes is given by Bechara and Damasio (2005, p. 368) as follows:

> *Emotions are a major factor in the interaction between environmental conditions and human decision processes, with these emotional systems (underlying somatic state activation) providing valuable implicit or explicit knowledge for making fast and advantageous decisions. Thus the somatic marker view of decision-making is anchored in the emotional side of humans as opposed to the construct of homo economicus. Although the view of maximizing utility of decision-making*

> *is pervasive and has a useful benchmark function, human decision-makers seldom conform to it. The process of deciding advantageously is not just logical but also emotional.* ”

The focus of the above cited study lay on economic decision making. On grounds of emotion impacting such tasks, in general interpreted to be of rational nature, it seems conceivable that emotion also impacts cognitive processing in its occurrence during retrieval tasks. To conclude this short exploration of the relevant state of the art concerning the decision making process, the next section explores the concept of cognitive bias.

Cognitive Bias

A group of processes that can significantly impact the outcome of decision making processes is given by different kinds of cognitive bias. The observation of such processes is based on observation on the behavioural level. They constitute high level abstractions of cognitive processing. An example of such a bias is given by the self-confirmation bias (Swann and Read, 1981; Russo et al., 1996). This bias, as suggested by its name, focuses on the observation that humans tend to bias processed information in a way that raises conformity with one's own beliefs and prior knowledge. In other words, this describes a tendency to favor information that enforces prior held beliefs and assumptions. Additional forms of bias exist in the form of source-credibility bias (i.e. the application of bias on basis of the perceived authority of the source of the information (Pornpitakpan, 2004)), heuristic information processing bias (i.e. the introduction of bias due to some form of 'short-circuiting' during the information processing act (Chaiken et al., 1989)). These processes can assume a profound role in human decision making processes.

This concludes the exploration of the state of the art concerning text based discourse comprehension and reasoning. Against the background of this investigation, the next section relates the identified processes to the phenomenon of relevance in IR.

4.2 A Human Information Processing Based Interpretation of Relevance

Section 4.1 provided an overview of the state of the art of text based discourse processing and reasoning. The analogy drawn in Section 2.2 presented relevance as a phenomenon occurring in the context of two systems: The IR system and the cognitive system of the user. As outlined in Section 2.3.3, this analogy provides the context for

the implementation of an IR focused nomological network. A first step towards the implementation of the network consists of the identification of a set of potential constructs. This is addressed by RQ 2 of the dissertation. The following section investigates which cognitive constructs constitute potential candidates for this network. This is based on relating the models of text processing introduced in Section 4.1 with the phenomenon of relevance. Based on the argumentation provided in Chapter 2, relevance is interpreted as cognitive and situational phenomenon. In other words, the question whether a query and a document are considered relevant is entirely based on subjective human assessment.

Against this scenario the concept of relevance is defined as the decision, the output resulting from one or more cognitive processes underlying an act of human judgement. This view in itself is not novel. It is in line with Cuadra et al. (1967, p. 23) expressing that '[r]elevance is not likely to be a very useful concept, so long as it is construed and used only as a relation between strings of written words independent of the judging process'. In relation to this Schamber (1990, p. 759) has stated that: 'While information scientists in general would tend to agree with this statement, they would also be likely to point out the difficulty of understanding a phenomenon as complex and variable as the human relevance judgement process.' Figure 4.2 illustrates the act of relevance judgment by outlining the underlying processes based on the review of reported research from cognitive psychology concerned with human information processing and decision making. The decision process is depicted as an interaction of several distinct cognitive processes. Based on the input of information, relevance is interpreted as the end product of a series of interactions of cognitive processes. While no complete descriptive model has been published that describes the exact mechanism of interaction, the individual role that contributing processes play is documented. Cognitive processes such as visual perception, discourse comprehension (Kintsch and Wharton, 1991; Walker and Kintsch, 1985) , bias (source credibility bias (Pornpitakpan, 2004), self-confirmation bias (Swann and Read, 1981; Russo et al., 1996), heuristic information processing bias (Chaiken et al., 1989)), and emotion (Bechara and Damasio, 2005) are known contributors to the decision process. The presented view is not aiming at completeness or at representing a functional model of the process interaction. The listing of involved cognitive processes is meant to define a scope for potential processes and constructs. Further, it outlines the complexity of the overall process.

The listing of contributing processes allows us to draw relations to observations on the phenomenon of relevance stemming from IR theory. As mentioned by Schamber (1990, p. 757) the problem of the validity of different kinds of relevance 'has most often been expressed in terms of the objectivity (measurability) versus the subjectivity (unmeasurability)'. Concerning this, Fairthorne (1963, pp. 111-112) expressed his criticism of a user oriented definition of relevance in the following form:

> *The only relevance that can be tested or measured is that based on cor-*

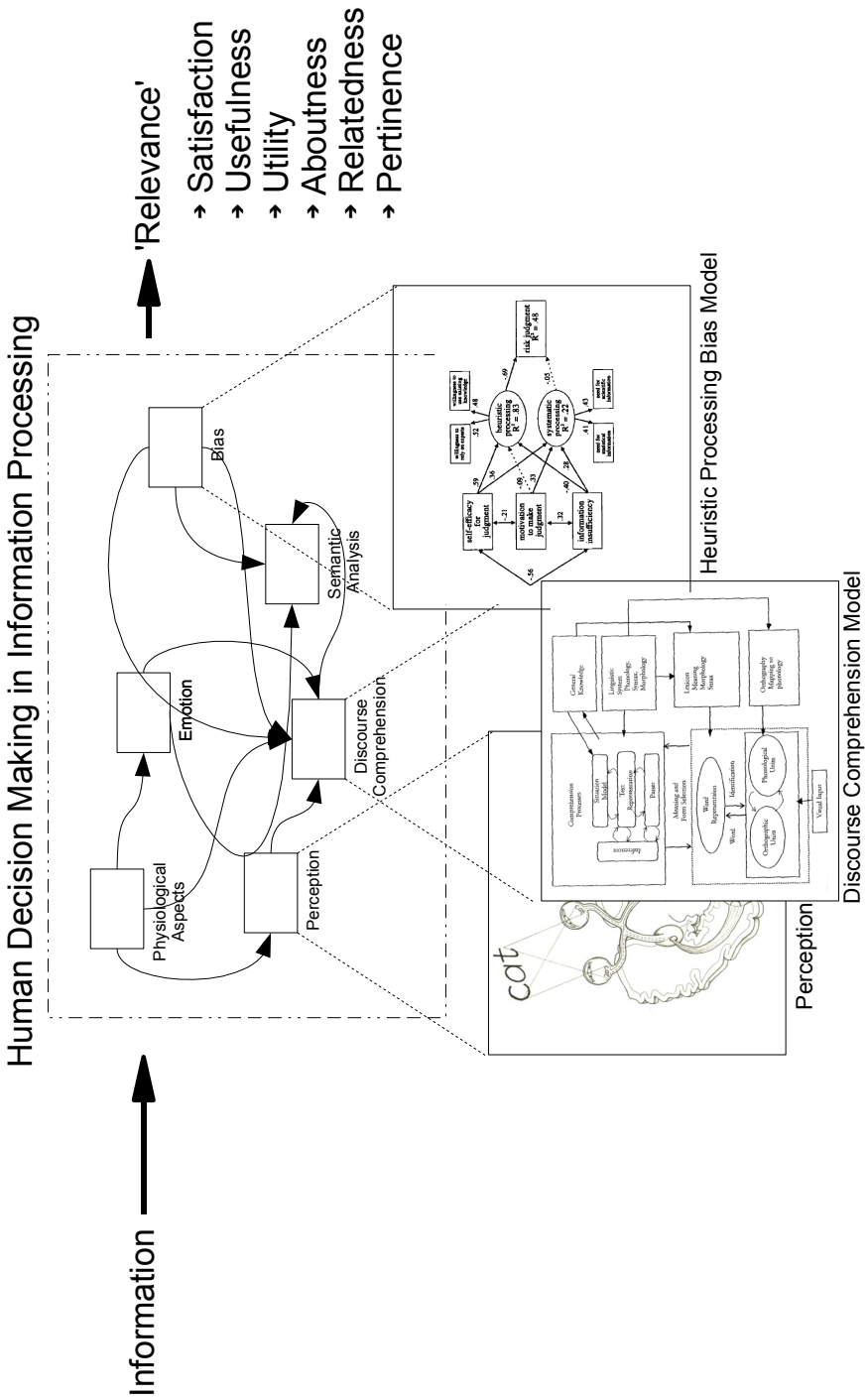


FIGURE 4.2: Relevance as a Product of Cognitive Process Interaction

> *respondences [...] between the words of the request and the words in the collection. [...] If we allow individual inference and understanding when deciding relevance, then any text is relevant to any request from some point of view.* ”

Such an objective and subjective nature of relevance can easily be motivated on ground of the listed processes. A subjective view of relevance for example is well supported based on the reported research on the influence of emotions (Sinclair, 1988; Bechara and Damasio, 2005), individual bias (Swann and Read, 1981), and physiological state (De Vries-Griever and Meijman, 1987; Sanbonmatsu and Kardes, 1988) on decision processes. Whereas evidence for the objective nature of relevance can be deducted from observations that have been reported to be applicable on a general basis. Examples of 'objective' processing is given by the source credibility bias (Tormala et al., 2006), and the mechanics of the discourse comprehension process (Kintsch and Wharton, 1991; Walker and Kintsch, 1985; Wolf and Gibson, 2005). An objective definition of relevance as advocated by Fairthorne (1963) specifically applies to the described processes of surface code construction (Section 4.1.1) and word meaning identification (Section 4.1.1). A second observation that can be made at this stage, consists of the overlap between reported characteristics concerning the nature of relevance and those describing the listed cognitive processes. Discourse comprehension, the act of interpreting a written or spoken message, shares almost all of the reported characteristics of relevance. As reported by Graesser et al. (1997), and Perfetti et al. (2005) it is an inherently dynamic process and is based on the concept of a situational representation. Where the construction of the situational representation is heavily dependent on the current knowledge of the reader. In analogy to that, Schamber (1990) suggests to interpret relevance as a dynamic and situational concept that is dependent on a user's specific information need. In that regard it seems plausible, that many observations with regard to the characteristics of relevance are induced by the nature of the underlying cognitive processes. This also allows for the relation of specific IR techniques to cognitive processes. The use of links (Kleinberg, 1999; Page et al., 1999) to estimate authority can be rationalized based on the reported role of source credibility bias.

The so far led discussion does not aim at providing a formal definition of 'relevance'. On the contrary, by outlining the complexity of the underlying cognitive processing it aims at placing emphasis on the magnitude of the task. The limited list of presented cognitive processes forms a basis for the observation of Saracevic (1970) that there is a limitless amount of definitions of relevance. Saracevic (1970, Table 1) expressed this notion in form of the following algorithm for such definitions.

> “ *Relevance is the (A) gage of relevance of an (B) aspect of relevance existing between an (C) object judged and a (D) frame of reference as judged by an (E) assessor.*

”

In the presented discussion this aspect is illustrated in Figure 4.2 through the listing of such possible definitions of relevance (e.g. utility, satisfaction, usefulness ...). Saracevic algorithm can also be interpreted as a result of constraints to cognitive investigation induced by the variability and complexity of behavior, and the issue of identifiability (Massaro and Cowan, 1993). With regard to the proposed paradigm, these observations are interpreted as motivating a principled approach to the validation of IR constructs.

The aim of this section consisted of investigating which cognitive constructs constitute potential candidates for an IR focused nomological network. This aim was pursued by relating observations on relevance stemming from IR with cognitive models of discourse comprehension and human decision making. The discussion outlined how observations on the nature of relevance in IR can be related to the characteristics of sub-processes contributing to text based decision making. Further, the listing of contributing sub-processes is a form of demarcating pertinent cognitive processes and their associated constructs for the inclusion in the nomological network. With respect to RQ 2 the listing defines the boundary of cognitively focused constructs for the network. The definition of such a boundary represents a coarse pre-selection of potential cognitive constructs. This coarseness is rooted in the nature of the undertaking to validate relevance, and is seen as an implication of the complexity of the human decision making process, and the fuzziness of existing definitions of relevance. Nevertheless it provides the desired starting point for the selection of cognitive constructs for an IR focused nomological network. Based on placing this listing of pertinent sub-processes in context of the Correlation to Cognition analogy, the next section demarcates pools of candidate constructs.

4.3 Mapping the IR and Cognitive Domains

Section 4.2 provided a listing of cognitive sub-processes contributing to text based reasoning and decision making. Based on this listing, the following section aims at demarcating a set of potential constructs for the construction of an IR focused nomological network.

A first step towards pursuing this aim consists of concretizing the goal. This can be based on prior discussions made in Part I. Section 2.3 outlined that the validation of constructs constitutes an iterative task. The necessity for an iterative approach is implied by the ill-defined nature of the focused cognitive constructs. This is intrinsic to the nature of the task of construct validation. It is intuitive that the ill-defined nature of the central construct of a validation effort also has implications on defining the set of constructs meant to enable the validation of the latter. An initial set focused on an

ill-defined construct is likely to be of a preliminary and fuzzy nature. The process of iteratively converging on the meaning of the central construct then enables refining the set of constructs for the network. The difficulties of establishing a consensus of its meaning (Schamber, 1990) outline that these considerations apply to the construct of relevance. In light of this observation, an initial selection constructs for the validation of relevance can be assumed to represent a coarse and preliminary choice.

This raises the question of what the characteristics of the pool upon which to base this initial selection should be. These considerations are represented by the second research question of the thesis. Answering this question can be based on a look at the mode of operation of a nomological network. A nomological network establishes construct validity through the definition of a set of related constructs, and the inference of relations between these constructs and the construct under validation. This implies, that an initial pool for the selection of constructs should ideally encompass the complete set of potentially related constructs. If the meaning of the central construct is ill-defined, it stands to reason that the insights regarding which set of constructs is best suited to the validation effort is limited as well. In light of this, a reasonable approach to the definition of the candidate pool seems to lie in the adoption of an initial strategy of 'casting a wide net'.

This raises pragmatic considerations. As shown in Section 4.1, the pertinent cognitive processing is represented by a complex hierarchical structure encompassing a large amount of sub-processes. Each sub-process represents a set of potential candidate constructs. The number of potential candidates is somewhat multiplied by the implications of the identifiability issue described in Section 3.4. The existence of a variety of theoretic models for each sub-process implies an increase of potential candidates for the construct pool. In light of this it can be assumed that the cognitive processing taken by itself represents a large amount of potential candidate constructs. Given the magnitude of the pool of candidate constructs, it seems beneficial to apply some form of categorization scheme. This requires the utilization of an underlying theoretic framework, or the definition of a context that allows for the inference of criteria forming the basis for such a categorization.

Against the background of these considerations a first step towards the identification of a pool of candidate constructs consists of defining a context for the pool. The Correlation to Cognition analogy provides us with such a context. It provides a context and a frame of reference for reasoning about relevance and its related concepts. Figure 4.3 shows an updated illustration of the Correlation to Cognition analogy based on the insights from Section 4.1 and 4.2. The figure provides a more detailed interpretation of Definition 4 in Section 2.1.5.

The aim of Information Retrieval consists of the creation or modification of a system whose estimates exhibit a maximum level of concordance with the output of the user's

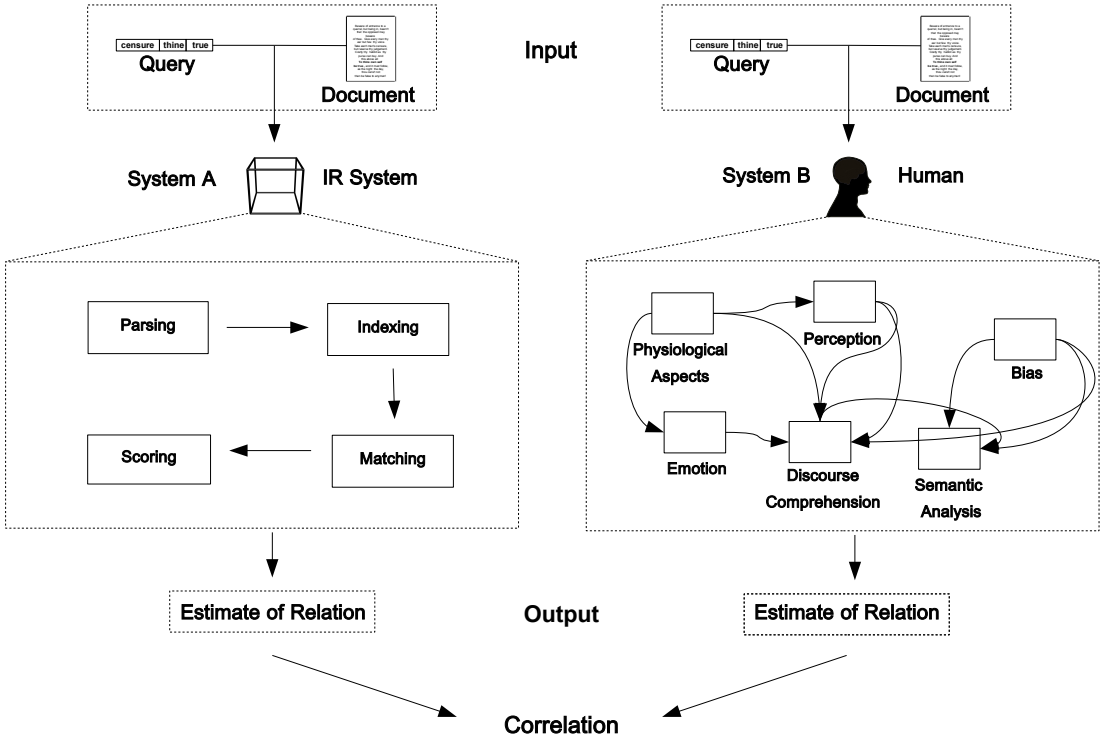


FIGURE 4.3: Interpretation of the Correlation to Cognition paradigm based on sample cognitive processing

estimation system.

In this form, it provides a first demarcation of the pool of candidate constructs. Definition 4 defines relevance as a phenomenon that occurs within the context of the two systems. Potential candidate constructs are given by constructs pertinent to the processes of both systems. On the IR side the pool of candidates is given by the constructs emanating from the theoretic framework of Information Retrieval. On the cognitive side the candidate pool is formed by the constructs associated with the theoretic framework underlying of textual comprehension and decision making. This demarcation results in a large amount of potential candidate constructs for a network.

With respect to this, it was noted that some form of categorization effort is interpreted to be beneficial to the selection of a set of constructs. Figure 4.4 outlines a categorization scheme based on an alignment of cognitive processes and IR processes. The

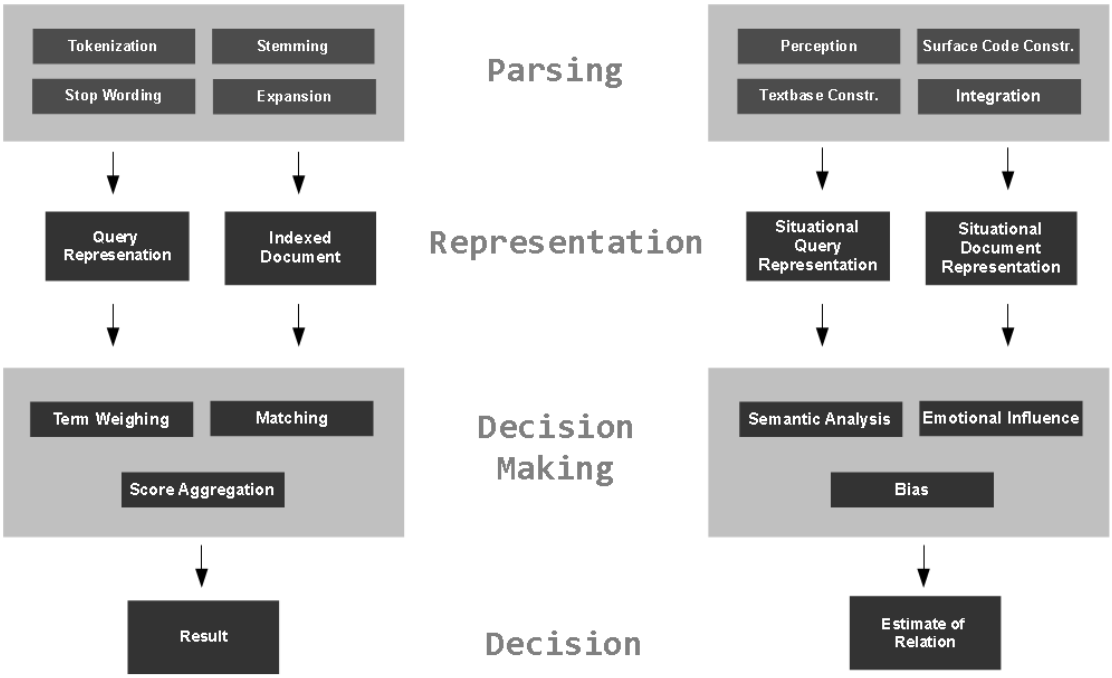


FIGURE 4.4: Mapping of Domains

figure aligns processing on both domains on basis of some elementary phases labelled 'parsing' and 'decision making'. The equivalent on the cognitive side is given by the processes attributed to text based discourse processing described in Section 4.1.1, and text based reasoning covered in Section 4.1.2. The categorization of processing is defined with respect to the level of abstraction. The figure depicts the abstraction level of processing in a top-down order. The processing at the top represents the lowest level of abstraction. This allows for an interpretation of the processes and their associated

constructs with respect to the observations derived from Newell's bands shown in figure 3.2. In this form figure 4.4 provides the sought after initial outline of candidate constructs for the validation of relevance.

4.4 Chapter Conclusions and Answer to RQ 2

Chapter 2 identified construct validity and the nomological network methodology as a principled way to the validation of IR constructs. Two necessary steps towards the establishment of such a network were noted. The identification of potential constructs, and the inference of criteria for their selection. Guided by the second research question, this chapter investigated what constitutes potential pools for the selection of such constructs.

> **RQ 2** What are potential constructs for the formulation of an IR focused nomological network?

The investigation of this question was based on a review of the principles of cognitive exploration in Chapter 3, and the outline of the state of the art in text based discourse processing and reasoning. On that basis, Section 4.2 provided a listing of known subprocesses contributing to text based decision making. This listing demarcates the pool of potential constructs representing the cognitive processing side of the Correlation to Cognition analogy. Section 4.3 placed these observations in context of the IR side of the analogy. The result of this effort is shown in Figure 4.4. The figure outlines the potential pools of constructs and allows for their categorization in terms of the level of abstraction. The definition of these pools, specifically on the cognitive side, is of coarse character. This is interpreted as an implication of both, the complexity of the underlying processing and the fuzziness of the phenomenon of relevance itself. Nevertheless, the outline of the two processing systems provides a frame for reasoning about and selecting constructs from both domains. It is rooted in the nature of validation that these pools should be interpreted as having preliminary character. The next step towards the construction of an IR focused nomological network consists of the development of a methodology for the selection of a set of constructs for the validation of relevance. This is addressed by the third research question of Part I. The next chapter is dedicated to the investigation of RQ 3.

CHAPTER 5

META-THEORETIC CONSIDERATIONS

Chapter 2 explored how construct validity and the nomological network methodology can be applied as a means to the validation of IR constructs. Two necessary steps towards the establishment of such a network were identified. Based on the exploration of cognitive science in Chapter 3, Chapter 4 addressed the first step by investigating the identification of potential constructs for an IR focused nomological network. Section 4.1 described the scope and complexity of the cognitive processing underlying text based human decision making. In Section 4.2 a listing of cognitive processes pertinent to the phenomenon of relevance was presented. Finally, Section 4.3 demarcated a pool of candidate constructs for IR focused nomological networks. Several observations relating to the nature of this demarcation have been made. As outlined in Section 4.4 the demarcation is of coarse and preliminary character. This was seen as an implication of the limited insight to the construct of relevance, and the complexity of the underlying cognitive processing. Further, it was outlined that the pool encompasses a large amount of candidate constructs. Based on the discussions in Section 3.2.1, the notion of the level of abstraction was utilized as a categorization scheme for the pool of candidates. Based on this candidate pool of constructs the next step consists of the selection of a set of constructs for the creation of an IR focused nomological network. This task is based on the third research question of Part I.

RQ 3 What are criteria for the selection of constructs?

This chapter explores this research question based on the development of a methodology for selecting constructs. The motivation for this investigation stems from the large amount of potential constructs, and the observation of Borsboom et al. (2004), Kane (2006) and Colliver et al. (2012) on a need for 'realism' in the construction of nomolog-

ical networks. In light of this, the remainder of this chapter takes the following form. Section 5.1 provides an analysis of the research question and an overview of the investigations in the chapter. Based on this analysis, Section 5.2 discusses the available experimental means in cognitive science and IR. Section 5.3 explores the meta-theoretical considerations of the research question. Finally, Section 5.4 summarizes the answer to RQ 3 and presents conclusions.

5.1 Introduction

RQ 3 focuses on the identification of criteria to guide the selection of constructs for a nomological network. The following section aims at concretizing this question. Accessing the question can be based on a look at the modus operandi of a nomological network. The purpose of the network consists of establishing the validity of a specific construct. To this end the nomological network is defined as a set of pertinent constructs and relations between constructs. Inferences with regard to the validity of the central construct are based on investigating its relation to the pertinent constructs. A technique for such an investigation is represented by convergent validation (Garner et al., 1956). Lachman et al. (1979) describes convergent validation as the idea, that the convergence of several different kinds of data on a conclusion, convergently validates this conclusion. Methodologically Garner et al.'s (1956) approach is based on an analysis of correlations between observations associated with different construct. Inferences with respect to validity are based on the strength and type of correlations.

This outline highlights, that constructs primarily contribute to validation through their associated observations and explanatory power derived from their theoretic basis. This provides a foothold from which to consider criteria for the selection of constructs. To structure this discussion, we introduce a basic categorization. Criteria for the selection of constructs are considered from a pragmatic and a meta-theoretical aspect. Pragmatic considerations can be derived in a straightforward manner. A first pragmatic aspect concerns the experimental means and measurement instruments associated with a construct. The availability of instruments of measurement constitutes a prerequisite for making observations in reference to a construct. It is intuitive that the validity of the measurement instrument is of importance. If the prime tool of inference consists of the correlation of different measured observations, the value of any such inferences depends on the degree of certainty that these measures are valid. Another pragmatic aspect is given by the level of coverage of the variable space associated with the constructs. That is, if the available instruments allow for the measurement of the full spectrum of observables emanating from the construct of interest. Finally, the consideration of the necessary effort for conducting experimentation or taking measurements also represents a pragmatic criterion for the selection of constructs.

The pragmatic aspects represent one side of the potential criteria for the selection of constructs. Exploring the meta-theoretic considerations associated with the call for 'realism' requires a more in depth analysis. The motivation for the call to 'realism' is grounded in the observed difficulties of establishing validity based on nomological networks encompassed solely of highly abstract constructs. A prime example for these difficulties is given by Cronbach's (1989) discussion concerning the validation of the construct of intelligence. The extent of these difficulties is interpreted on different levels. (Borsboom, 2006, p. 431) labelled attempts at defining construct validity using highly abstract constructs a 'black hole from which nothing can escape'. While acknowledging the difficulties, Kane (2006) interprets the statement of Boorsboom as an overstatement of the case. The suggested resolution to resolve these difficulties is a call to 'realism' based on the notion of attributes (Borsboom, 2003; Colliver et al., 2012). Colliver et al. (2012, p. 367) state that attributes are 'more than just theoretical ideas; rather, they are thought to exist independently of their measurement and serve to cause the measurement outcome'. They refer to attributes as 'facts [...] that are thought to really exist out there' (p. 369). Mentioned examples include blood pressure, height, weight, and scholastic performance. Borsboom (2003) and Colliver et al. (2012) express a distinction between attributes and constructs in regard of two points. The first is given by the degree of certainty with respect to the existence of the concept. By their definition, the existence of attributes is considered a certainty, while the one of constructs is principally in doubt. The second distinction concerns the validity of measurements. Variation in attributes is expected to cause variation in their measurement instruments. This causal relationship is interpreted as a given fact.

Attributes are somewhat 'thought to exist apart from theory' (Colliver et al., 2012, p. 367). However, following Rosenblueth and Wiener (1945, p. 316), that '[n]o substantial part of the universe is so simple that it can be grasped and controlled without abstraction'; it seems plausible to deduct that no substantial part of the universe can, from the human perspective, exist outside of theory. This contrary view emphasizes that a demarcation between attributes and constructs is not a straightforward thing. Deriving applicable criteria for the selection of constructs based on these observations requires the derivation of criteria for distinguishing between attributes and constructs. This relates the discussion back to the presented argumentation in Chapter 3. As outlined in Chapter 3, a central role in dealing with the complexity of the mind and the limitations of current experimental instruments is taken by the Information Processing paradigm. One of the fundamental tenets of the paradigm consists of the use of hierarchical decomposition as a means of mitigating research limitations. Taken on its own this 'suggestion' can be understood as merely re-stating the common knowledge that a viable strategy regarding the exploration of 'too' complex problems consists of splitting the task into several sub-components. With regard to such interpretations it is important to highlight the motivation for decomposition in terms of levels of completeness of informational description. Decomposition enables experimentation on basis of the interpretation of

the *levels of the hierarchy* essentially constituting *levels of abstraction*. Experimentation can be understood to be rendered viable on basis of the fundamental relationship between abstraction and completeness of informational description (i.e. lower levels of abstraction are associated with higher levels of completeness of the informational description). This can be paraphrased as the observation, that experimentation focused on lower levels of abstraction is enabled by a more complete knowledge of the set of variables associated with the system under study. These observations substantiate the discussion concerning the 'realism' of constructs and attributes. The notion of 'realism' can be associated with the level of abstraction, the completeness of the informational description and the implications of these criteria on measurability. This allows to interpret attributes as elements with high measurability, and serves as a first indication of how meta-theoretically based criteria for the selection of constructs can be derived. This perspective circumvents the binary nature of the question of the existence outside of theory and provides a preliminary set of criteria for the evaluation of levels of 'realism'.

The so far led discussion can be tied back to the structuring of the chapter. Section 5.2 explores the experimental means in cognitive science and IR. With regard to the discussion, this serves to substantiate considerations of pragmatic and measurability related criteria for the choice of constructs. The association between 'realism' and measurability is investigated in Section 5.3. The focus of the analysis lies in deriving meta-theoretically based criteria for the selection of constructs.

5.2 Experimental Means in Cognitive Science and IR

The prior section outlined the relation of pragmatic aspects and limitations of measurability with respect to the identification of criteria for the selection of constructs. To substantiate these considerations, this section presents a survey of the available experimental means in cognitive science an IR.

5.2.1 Experimental Approaches in IR and Cognitive Science

The purpose of this exploration consists of outlining the main approaches of conducting research in both domains. Figure 5.1 provides an overview of research approaches in the two domains. With regard to the considerations of measurability, the domain specific approaches are interpreted in terms of the underlying level of abstraction. As indicated in Figure 5.1, experimentation in information retrieval can be interpreted to generally focus on a high level of abstraction (henceforth referred to as LOA). As illustrated on the left side of the figure, a high LOA implies a low level of completeness of the informational description. The depicted subset of cognitive science is concerned with

the exploration of specific cognitive processes. Guided by the tenets of the Information Processing paradigm, it is focused on the exploration of lower LOAs that warrant higher levels of completeness of the informational description.

Figure 5.1 categorizes research endeavours in both domains into two distinct branches. The categorization discriminates between research efforts based on direct observations, and those based on observations derived from correlational analysis with computational models. The label of direct observation is applied to all experimental efforts that draw inferences based on data gathered through direct observation of human subjects. Correlation on basis of computational models is defined as a mode of experimentation where inferences are drawn based on correlating the output of a computational system with the output of one or more human subjects.

The significance of this categorization with regard to the selection of constructs is the following. Faced with the task of selecting appropriate constructs for a practical implementation, the introduction of the dichotomy is sought to enable consideration of the viability of experimental 'tools' on a general and abstract level. In light of this the two categories are subsequently explored with the purpose of establishing a basis for the inference of selection criteria. The exploration consists of the provision of brief descriptions of exemplary research from both categories, and their analysis with regard to pragmatic and measurability related aspects.

5.2.2 Direct Observations

Experimental approaches attributed to this category are those whose inferences are based on direct observations of a phenomenon.

Information Retrieval

Within the domain of IR the research that can be attributed to this category is usually focused on the observation of user behaviour. This direct observation of user behaviour in Information Retrieval can be roughly categorized into two main approaches:

- User Surveys
- Recording of User Behavior

As noted by Ingwersen and Järvelin (2005) user surveys constituted the major experimental mean of the empirical information seeking studies in the time span from 1960 - 1985. Predominantly such surveys are of a quantitative nature and based on structured questionnaires and interviews. Within contemporary IR research such surveys

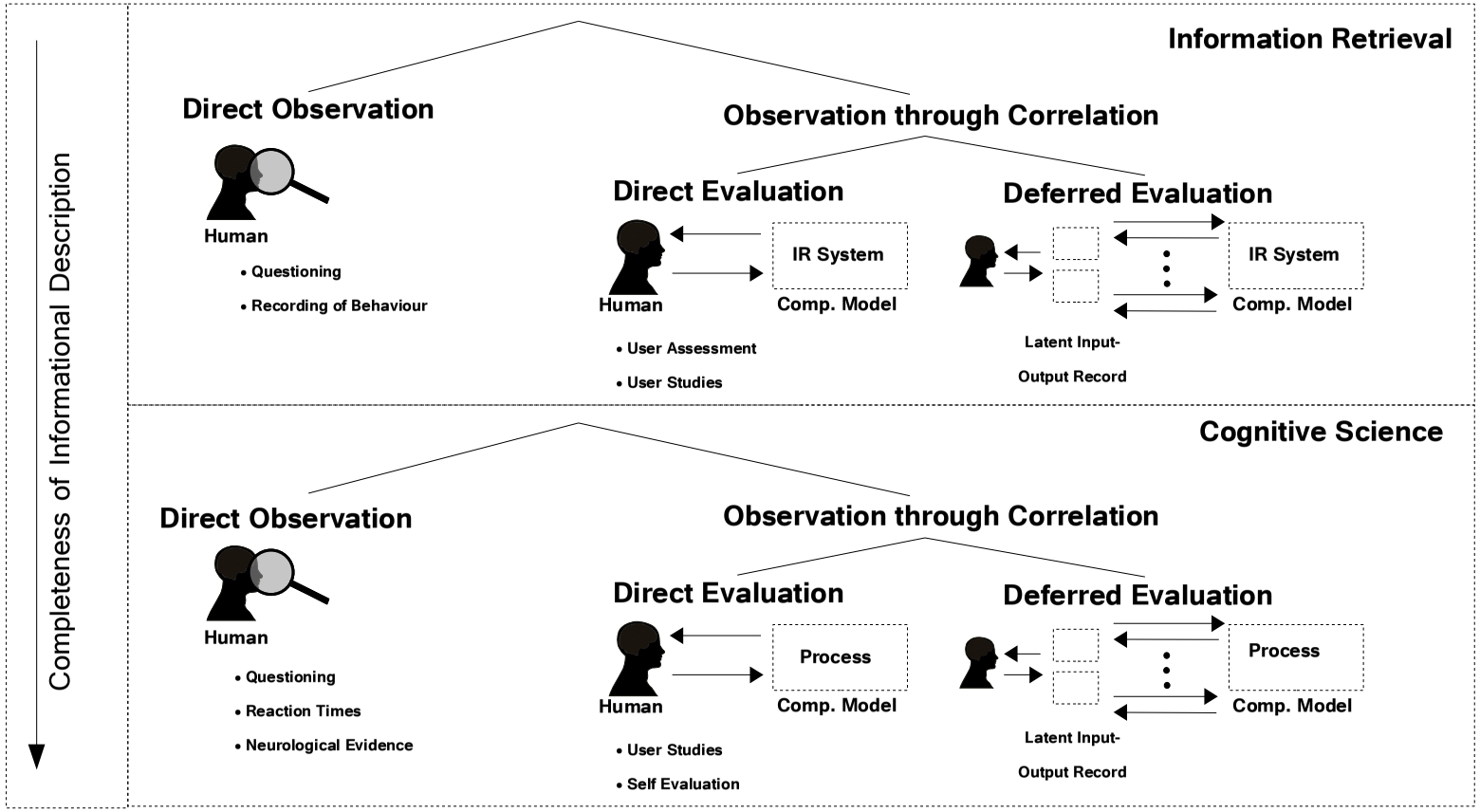


FIGURE 5.1: Principle Experimental Approaches In Cognitive Science and Information Retrieval

76

play an important role in the subfields of Multimedia and Interactive Retrieval. Examples techniques for the recording of user behavior are given by thinking aloud protocols and the capture of a user's interaction with information systems. This form of direct observation is generally focused on a high LOA. In general such studies focus on the observation of external behavior, and are not concerned with the underlying cognitive mechanisms.

Cognitive Science

Research based on direct observation within the domain of cognitive science finds application on many levels of abstraction. The main levels can be identified as the following:

- Neurological level
- Pathological level
- Physiological reaction level
- Speech level

Examples of research on the neurological level consist of the application of FMRI and ERP based analysis. Studies on a pathological level are focused on anatomical observations. Examples of observations on the physiological level are given by eye movements and task specific reaction times. Research based on the speech level is primarily conducted on grounds of verbal or written output of participants. This listing illustrates that cognitive research based on direct observation covers varying levels of abstraction. A general observation with regard to the described research approaches consists of the high level of required effort in terms of time, equipment, and ethical considerations. Neurological as well as pathological research constitute the most 'direct' instruments of observations. The observables targeted by these instruments can be assumed to exhibit high measurability. A limitation of these instruments is given with regard to their ability of enabling inferences concerning the mechanics of cognition (Garrett, 2007).

5.2.3 Observation Through Correlation

As stated before, the label 'observation through correlation' applies to research that is based on drawing inferences by correlating the output of a computational system with the 'output' of human subjects. Subsequently, the application of computational models in both domains is illustrated on basis of the provision of examples. In addition, the motivation for using such models as well as underlying meta-theoretical considerations are explored.

Information Retrieval

The correlation with computational models (i.e. retrieval systems) constitutes possibly the most applied experimental mean within Information Retrieval research.

Motivation for the approach Based on the early work in Information Retrieval (see Bush (1945), Luhn (1957), Luhn (1958)) it seems justified to state that the initial motivation for the use of a computational system within IR is not based on meta-theoretical considerations. Instead it stems from the advantages offered by computational systems in terms of speed of operation and magnitude of storage size.

Principles of the Methodology In a simplifying manner the methodology can be reduced to the following steps:

1. The construction of a functional information retrieval system
2. The application of the system to a specific retrieval task
3. The evaluation of the system on basis of requiring partial (incomplete) or complete judgement of documents and queries with respect to the description of the information need.

Regarding the role of the computational model within these three steps, the following can be said. The first step consists of the construction of a retrieval system. An initial question concerns the underlying basis for the construction of the system. The construction of this system is based on a set of formal and verbal conceptual theories. It constitutes a formalization and integration of a large variety of distinct theories. In the case of IR, examples of such theoretic underpinnings consist of the probability ranking principle, probabilistic term weighting, and relevance feedback. From this perspective, an IR system constitutes a computationally encoded formalization of an integration of various theoretic assumptions. An evaluation of these theories is based on application of the system on a particular IR related task, and subsequent measurement of its performance by correlating the system's output with human assessments. In the original instantiation of the Cranfield experiments (Cleverdon (1967)) the manual assessment comprised the complete set of documents in the collection. In effect the particular retrieval task was therefore conducted by both, the system as well as the human subject. The subsequent evaluation then simply consisted of correlating the judgements of the system and the user. This highlights that the mode of operation consists of correlating observations on a high LOA (i.e. user judgement) with computationally encoded formalizations of various theoretic assumptions. Based on the observations of Newell (1994) it can be assumed, that the measurability of such experimentation is relatively

low. The complexity of the IR system in combination with the focus on high LOA observations places difficulties on the interpretation of measurements.

Cognitive Science

Computational models find wide-spread application within contemporary cognitive research (see Sun (2009) for a general overview).

Motivation for the Approach In contrast to IR, the use of computational models in cognitive science is meta-theoretically motivated. The use of such models can be seen as a reflection of both, the black-box character of the mind (with respect to the limitations of observations), and the limitations with regards to the manipulation of the subject of study. Specifically in regard of the limitations of manipulation, model-based simulation of cognitive processes often constitutes the only viable mean for the experimental verification of theories. The use of computational models within cognitive science is not undisputed. Sun (2009) identifies the following two opposing viewpoints:

- 'In relation to computational cognitive modelling, one possible (and starkly negative) viewpoint is that computational modelling and simulation, including those based on cognitive architectures, should not be taken as theory.' (p. 5)
- 'Computational cognitive models are more than just simulation tools, or programming languages of some sort. They are theoretically pertinent, because they may express theories in a unique and, I believe, indispensable way.' (p. 15)

The second viewpoint expresses the notion, that the model itself can, based on observation of its 'behavior', be interpreted as a theory in its own regard.

Principles of the Methodology With regard to later extensive coverage of these aspects the description of the principles is kept brief. Computational systems in cognitive science are usually applied in form of a coded (programmatically) model of a specific cognitive process. As such, these models can be regarded as a formalization of existing theories of the process. Inferences are drawn based on correlating the model's output with human output. Such modelling targets multiple levels of abstraction. Ranging from models aimed at simulating cognitive architectures (e.g. SOAR Newell (1994)), ACT-R (Anderson et al., 2004)) to models operating on the level of neurons (Hopfield, 1984). The effort and the measurability of such approaches differs based on the focused LOA. It is intuitive that an increase in the complexity of models renders the task of interpreting measurements more difficult.

5.2.4 Conclusion

The section provided an overview of experimental means in cognitive science and IR. This was based on a survey of exemplary research and the distinction between direct observation and observation based on the use of computational models. Based on this survey, some general observations concerning pragmatic aspects and meta-theoretical issues can be made. With regard to pragmatic aspects it can be said, that computational models offer advantages in terms of the manipulation of variables and the simulation of the full spectrum of observables emanating from a phenomenon. Dependent on the LOA of the correlated observations and the complexity of the models, it was outlined that such approaches can be constraint in terms of measurability. Direct observation on lower LOA offers potential advantages with regard to measurability. A general observation with regard to these approaches consists of the high level of required effort in terms of time, equipment, and ethical considerations. Summarizing the approaches in both domains it can be stated, that IR focuses on observations on a high LOA, while cognitive science approaches address a wider spectrum of LOA. These observations on the experimental provide a basis for a listing of pragmatic criteria for the selection of constructs. As a reference for the subsequent investigations in the chapter these aspects are presented below.

- Availability of measurement instruments
- Validity of measurement instruments
- Manipulation of variables
- Coverage of spectrum of observables emanated by phenomenon
- Necessary effort for taking measurements

The list presents a set of criteria for the selection of constructs with respect to practical considerations.

5.3 Recursive Correlation over Levels of Abstraction

The research focus of this chapter lies in identifying criteria for the selection of constructs from the pool of candidates presented in Chapter 4. As part of this investigation, Section 5.2 provided an overview of the experimental means in the domains of IR and cognitive science and listed pragmatic criteria for the selection of constructs. Within this section the research focus is placed on the identification of meta-theoretically based criteria. Approaching this task is embedded into the derivation of a methodology for the construction of IR focused nomological networks. Investigating meta-theoretical

aspects in this way is motivated by the assumption, that highly abstract concepts relating to 'realism', levels of abstraction and measurability are best illustrated in a context outlining their relation to experimental considerations.

5.3.1 Derivation of Methodology

Deriving a methodology for the construction of IR focused nomological networks is based on the introduction of an intuitive analogy. The analogy is chosen with the aim of allowing for an illustration of the considered meta-theoretical aspects. This illustration is based on the presented insights in Chapter 3, that outlined how the consideration of meta-theoretical aspects shapes the principles underlying the investigation of the mind. The development of the subsequently introduced methodology is conducted under consideration of these identified principles.

Recursive Hierarchical Decomposition

A first step towards the development of a methodology consists of outlining the relation between meta-theoretical criteria and the conduction of experimental research. To illustrate this, Figure 5.2 depicts the application of the IP paradigm's tenet of hierarchical decomposition based on a 'weather quality' analogy. The analogy serves to outline the underlying meta-theoretical considerations. It has been chosen on grounds of its intuitiveness and its similarity to the notion of relevance in IR. The focused construct within this analogy is given by the concept of 'weather quality'. Observations of the construct are given in form of a binary categorical variable with the values 'Good' and 'Bad'. Based on this definition, 'weather quality', analogous to 'relevance', can be understood to constitute a highly abstract construct. Additionally, again analogous to 'relevance', it can be interpreted as a product resulting from a highly complex interaction of various cognitive and physiological processes. In light of these shared characteristics the task of exploring the cognitive mechanics underlying human judgements of 'weather quality' closely resembles an exploration of the judgement of 'relevance'. Shifting the focus back to Figure 5.2, it can be seen that the illustration features 3 levels of abstraction portrayed in the form of planes.

The topmost plane is dedicated to *weather quality*; the focused construct. A first step towards the construction of a nomological network focused on 'weather quality' is given by the identification of cognitive tasks that contribute to the targeted decision making process (i.e. judgement of weather quality). This step can be conducted in a similar fashion as described in Chapter 4. The result of such an analysis consists of a hierarchical listing of the cognitive processes that are assumed to contribute to the cognition of 'weather quality' judgement. Dependent on the level of insight to these processes, the

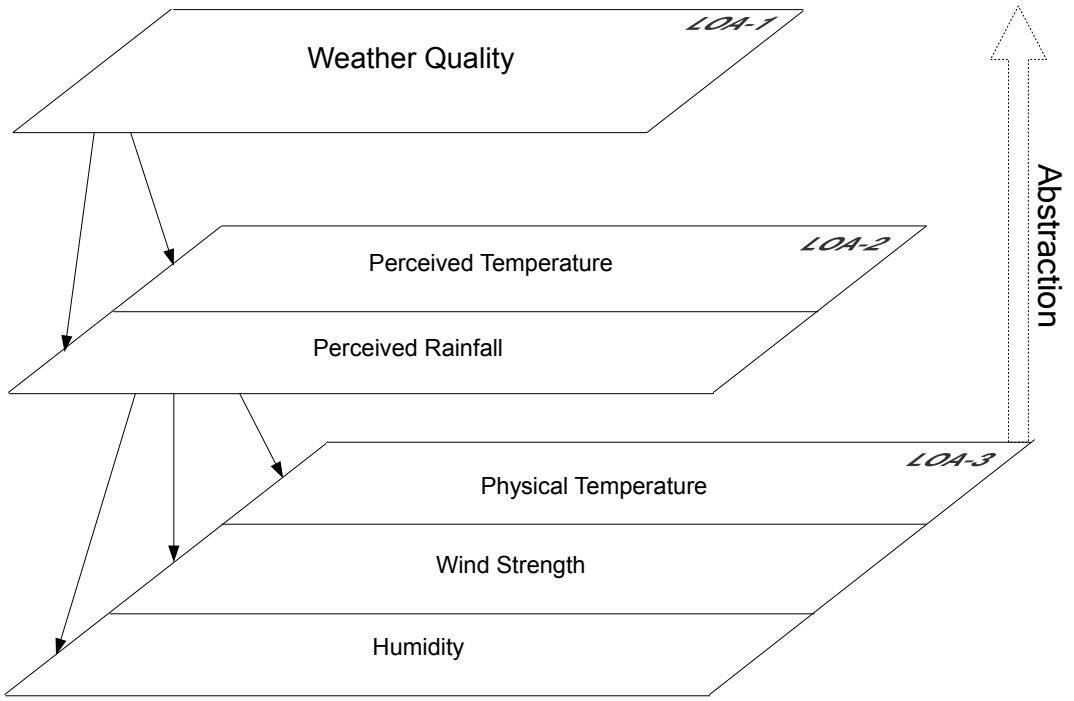


FIGURE 5.2: Recursive Hierarchical Decomposition over Levels of Abstraction (LOA) with Regard to the Measurement of Weather Quality

listing may include statements regarding the relationships between the sub-processes. However, this does not constitute a requirement for the construction of a nomological network. The only requirement with respect to the subsequent use of the processes, consists of the existence of an alignment of these processes with respect to the level of abstraction[5-1]. An example for such an alignment is given by Figure 5.3. Figure

FIGURE 5.3: Sample Table Outlining Hierarchical Decomposition with Regard to Level of Abstraction

5.3 enables reasoning about the processes in terms of their underlying level of abstraction. Figure 5.2 illustrates this via the introduction of a second plane (LOA-2). As indicated by the abstraction-axis in the background, the respective processes belonging to the LOA-2 plane are interpreted to represent lower level abstractions. Based on the argumentation by Newell (1994), this results in higher levels of completeness of their informational description (i.e. a more complete knowledge of the relevant variables of the system under consideration). On the LOA-2 plane the associated constructs are given by 'temperature perception' and 'rainfall perception'. It is intuitively understandable, that these two variables are of less complex and less abstract nature than 'weather quality'. A systematic approach to the identification of additional related constructs consists of a repetition of the steps applied on the highest level of abstraction. Again, this can be primarily based on a survey of prior knowledge and an analysis with regard to contributing processes. In this regard, the methodology can be understood to be of recursive nature. The result obtained by this recursive application is illustrated in Figure 5.2 through the addition of the lowest depicted layer of abstraction (LOA-3). The cognitive process of temperature perception is, in a simplifying manner, composed of the three physical variables temperature, humidity, and wind strength. The so far described steps are analogous to the investigations in Chapter 4 aimed at the identification of a pool of candidate constructs.

Based on these considerations, a first analysis with regard to the notion of 'realism' can be conducted. Section 5.1 outlined that the call to 'realism' is motivated by observed difficulties of establishing construct validity based on highly abstract constructs. To illustrate this, it is supposed that the underlying aim of researching weather quality, consists of the construction of an information system estimating 'weather quality' (in analogy to the aim of IR to construct systems for the estimation of relevance). Based on this setup, evaluating the 'goodness' of the weather quality model could be based on

[5-1] Section 3.2.1 outlined the significance of the concept of levels of abstraction to cognitive research

correlating the computational model's judgements and human judgements of 'weather quality'. The limitations of making inferences based on such correlations, can easily be illustrated. In a more a more realistic example encompassing a more complete list of factors, it is intuitive how aspects such as the age, profession, and prior weather related experiences of the subject might influence the judgement of 'weather quality'. A person accustomed to 'Scottish weather' might rate a weather scenario as 'Good'. However, the same weather scenario might receive a 'Bad' judgement by a subject who has grown up in Spain. In the same manner it is obvious how holding specific professions might bias judgement. A construction worker, pilot, and fisherman might judge the same weather scenario differently based on considerations relating to their profession. If personal planning, hobbies, situational context, personal weather preferences, and the influence of past and forecast weather scenarios are added to the list of influencing factors, it quickly becomes evident how difficult the task of basing inferences on such judgements is rendered. It is, so to speak, hard to infer why a specific scenario received a certain judgement. This raises the question, why it is so hard to make inferences based on judgements of weather quality. Before addressing this question, it is of interest to first investigate how research can be conducted based on such limitations. These considerations are helpful in situating the work of the dissertation in context and to motivate the consideration of meta-theoretical issues.

Starting with the weather analogy, approaching the research problem of building a weather estimation system could be based on two main directives. An intuitive first step to a better understanding of the phenomenon of weather quality could be based on asking users to justify their judgement. A secondary step might consist of creating models and definitions of weather quality that reflect gained insights and observations. Such an approach might result in the identification of more factors influencing judgements. The learnt lessons could result in the definition of different aspects of weather related judgements such as 'weather usefulness', 'weather utility', and 'weather interestingness'. Cognitive models of weather quality could be established, that define weather quality as a relation between a person's idea of good weather and specific weather scenarios. With regard of the task of iteratively improving estimation systems, an obvious step consists of integrating computational modelling of processes assumed to influence weather judgement. Modelling the influence of professions, and past and forecast weather scenarios, could intuitively lead to better estimation systems. This outlines possible strategies for mitigating the *difficulties stemming from the constraints on making inferences*.

The exact same constraints apply to relevance. It is hard to infer why a specific document has been judged relevant. IR is required to develop analogous strategies to mitigate the difficulties associated with the constraints of making inferences. In contrary to the fictional weather quality example, an analysis of such strategies can be based on the historic development of IR. As outlined by Ingwersen and Järvelin (2005), a

major research tool in early IR research consisted of user surveys and questionnaires requiring users to provide insights to the judgement process. Building better relevance estimation systems (i.e. information retrieval systems), is analogously to the weather example, also based on modelling supposed contributing factors. Examples are given by models of authority (Page et al., 1999; Kleinberg, 1999; Ritchie et al., 2006), and term importance (Spärck Jones, 1972). The work of Mizzaro (1998) , Saracevic (2007), and Schamber (1990) outlines, that considerable effort has been dedicated to the establishment of definitions and descriptions of the phenomenon of relevance. The effort has resulted in a categorization scheme of objective and subjective relevance (Swanson, 1986), and the investigation of the relation of relevance to constructs such as satisfaction, utility, and pertinence Saracevic (1975). A framework for the classification of the large number of definitions of relevance was suggested by Mizzaro (1998). Models of the phenomenon of relevance were introduced in form of cognitive models (Belkin et al., 1993; Ingwersen, 1994). Although a large number of relevance related papers has been published (Mizzaro, 1997), no general consensus of the meaning of relevance has been established Schamber (1990).

With regard to construct validity, the strategies applied by the IR community can be interpreted as an act of constructing a nomological network based on highly abstract concepts. Specifically, this applies to the relation of relevance to constructs such as satisfaction, pertinence, aboutness, and utility. In light of this, the lack of a consensus for relevance can be related to Borsboom's (2006) notion of a 'black hole' (p. 431). Defining highly abstract constructs by relating them to other highly abstract constructs is considered a futile attempt by Borsboom (2006). He considers it to be a never-ending process, consisting of the provision of arguments and counter-arguments, from which nothing useful escapes (i.e. a 'black hole'). While acknowledging the limitations of such an approach, Kane (2006) takes a more moderate position and argues, that the expended effort contributes to a better understanding of the phenomenon and to the identification of contributing sub-processes and factors.

These observations provide a foothold for the discussion of meta-theoretical aspects pertinent to the investigation of relevance. This requires to return to the prior deferred question: Why is it hard to make inferences based on judgements of relevance. In the following, it will be argued that this constitutes a problem of measurability. Understanding why a specific person judges extremely strong winds and rain as 'Good' weather, is a question of the validity of the measurements. Validity can be interpreted as the degree of certainty to measure, what one aims to measure. Asking a person to judge the weather constitutes an act of measuring. There exists uncertainty to what degree this constitutes a measurement of the estimation of weather quality, a person's professional background, or his or her origins. With regard to relevance this translates in questioning, to what degree relevance assessments constitute measurements of relevance, a person's interests, prior knowledge, mood, context, assessment related motivation, or general

personal preferences. This outlines the highly abstract nature of the phenomenon of relevance. As outlined by van Rijsbergen (personal communication), 'relevance is a dense measurement'.

This observation can be interpreted on a meta-theoretical level based on Newell's (1994) notions of level of abstraction and measurability. Outlining the relation between level of abstraction and measurability can be based on the depiction of the weather scenario in Figure 5.2. The middle plane of the graphic (LOA-2) is dedicated to the constructs of 'perceived temperature' and 'perceived rainfall'. As depicted in the figure, the construct of 'perceived temperature' is hierarchically decomposed into the constructs of 'physical temperature', 'wind strength', and 'humidity' In light of the so far led discussion, it is obvious that the construction of a 'perceived temperature' estimation system on the LOA-2 level constitutes, relative to LOA-1, a much more feasible endeavour. In the specific example 'temperature perception' can be evaluated on based on correlating model output with human judgements concerning the perceived temperature. This is rendered more feasible through the identified independent variable space consisting of humidity, wind strength, and physical temperature. The decomposition into physical variables indicates a higher measurability of 'temperature perception' relative to 'weather quality'. The validation of measurements is more feasible on the lower levels of abstraction. Construct on lower levels of abstraction offer a higher level of measurability. With respect to the Information Processing paradigm (Palmer and Kimchi, 1984) and Newell's (1994) observations, this can be related to the level of completeness of the informational description on different levels of abstraction. Completeness of the informational description refers to the level of knowledge about the variables contributing to an observed phenomenon. A complete level of informational description suggests, that all contributing variables are accounted for. It is intuitive that level of abstraction, and level of completeness are inversely related. The more complex (and abstract) a construct is, the less likely it is to have knowledge of all contributing variables. Lower levels of abstraction result in higher completeness of the informational description and higher measurability.

Summary: A summary that can be drawn on basis of the portrayed example consists of the following. The application of recursive hierarchical decomposition is interpreted to be a viable proposal for the exploration of complex cognitive decision processes with regard to two main observations.

The first of these being, that lower levels of abstraction exhibit higher levels of informational completeness (more complete knowledge of the relevant variables of the construct under consideration) of experimental scenarios. Secondly, the concepts on lower levels of abstraction constitute constructs of lower complexity. This attributes difficulties concerning the interpretation of variables on the 'relevance' or 'weather quality' level to the large number of contributing factors. As further illustrated, the number of

contributing factors can be interpreted as being inversely proportional to the level of abstraction. Based on these observations, constructs on lower levels of abstraction are attributed with a higher measurability. This observation can be related to the call for 'realism' advocated by Borsboom (2003) and Colliver et al. (2012). Borsboom (2003) and Colliver et al. (2012) express a distinction between attributes and constructs in regard of two points. The first is given by the degree of certainty with respect to the existence of the concept. By their definition, the existence of attributes is considered a certainty, while the one of constructs is principally in doubt. The so far led discussion allows to interpret attributes with regard to the concepts of level of abstraction, completeness of informational description and measurability. The notion of 'attributes' can be interpreted as constructs exhibiting high measurability. According to Colliver et al. (2012), making meaningful inferences based on a nomological network requires the existence of validated constructs and associated validated measurements. The so far led discussion illustrated, that the higher the level of abstraction of a construct is, the less likely one is to have certainty of what one measures. This observation motivates the use of the level of abstraction, the completeness of informational description, and measurability as criteria for the selection of constructs. The focus of the so far led discussion lay on considerations of the feasibility of experimentation with respect to levels of abstractions. The focus of the next subsection is set on exploring how experimentation can be based upon these considerations.

Bottom-Up Correlational Analysis over Levels of Abstraction

The last section outlined how complex tasks can be systematically decomposed with regard to the consideration of levels of abstractions, and introduced meta-theoretically motivated criteria for the selection of constructs. As stated before, the decomposition of a process is initially based on an assumption with regard to contributing sub-processes. The focus of this section can be demarcated by formulating the following two questions.

- Do the assumed sub-processes impact cognitive decisions on the highest level?
- In what ways do the assumed sub-processes impact the high level process?

In the course of this section it will be argued that a test of hypotheses derived from these two questions can be based on aligning measurements associated with different constructs. This aspect is outlined based on a figure. Figure 5.4 shows a slightly expanded depiction of the 'weather quality' example.

As can be seen, the same three levels of abstraction are depicted in the figure. In contrary to earlier illustrations, each level of abstraction relates to measurements of the focused construct. In the case of LOA-1 these measurements refer to 'weather quality'. On basis of the hierarchical structure, the measurements are defined with respect

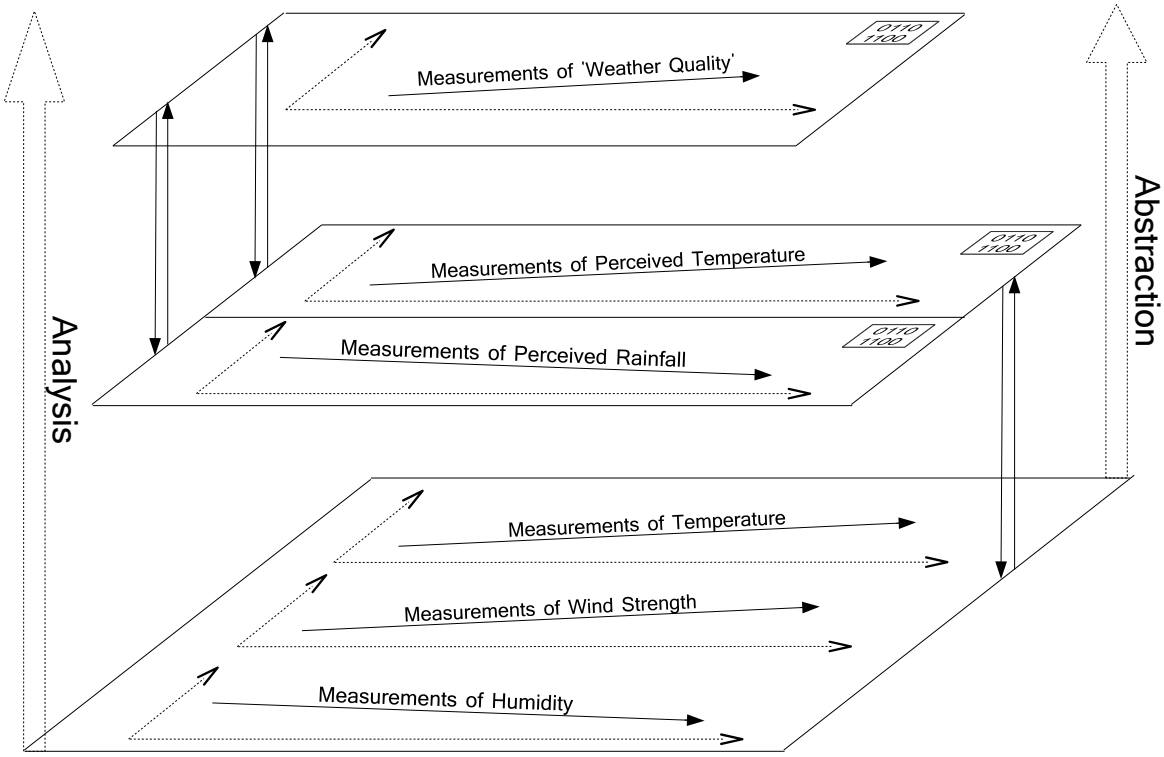


FIGURE 5.4: Measurements

to the variable space spanned by the adjacent lower hierarchical level. Measurements of 'weather quality' on the LOA-1 level are defined with respect to the LOA-2 constructs, namely in the given example measurements of 'temperature perception' and 'rainfall perception'. Measurements of 'temperature perception' themselves are defined on basis of the LOA-3 constructs 'temperature, 'humidity', and 'wind strength'.

Based on such hierarchies, an analysis of construct specific impact can be approached in a bottom-up order. In the figure this is represented through the '↑↓' symbols denoting analysis of correlation ,covariance, or other forms of aligning measurements. With regard to 'temperature perception' an analysis of the impact of the constructs on the adjacent lower level of abstraction could be based on correlating measurements of 'wind strengths', 'temperature', and 'humidity' with measurements (e.g. human judgements) of 'perceived temperature'. As shown in Figure 5.5, the variables on LOA-3 constitute physically grounded constructs. Particularly due to this aspect, an evaluation of their impact on temperature perception can be based on experimentation relying on direct measurements of the constructs.

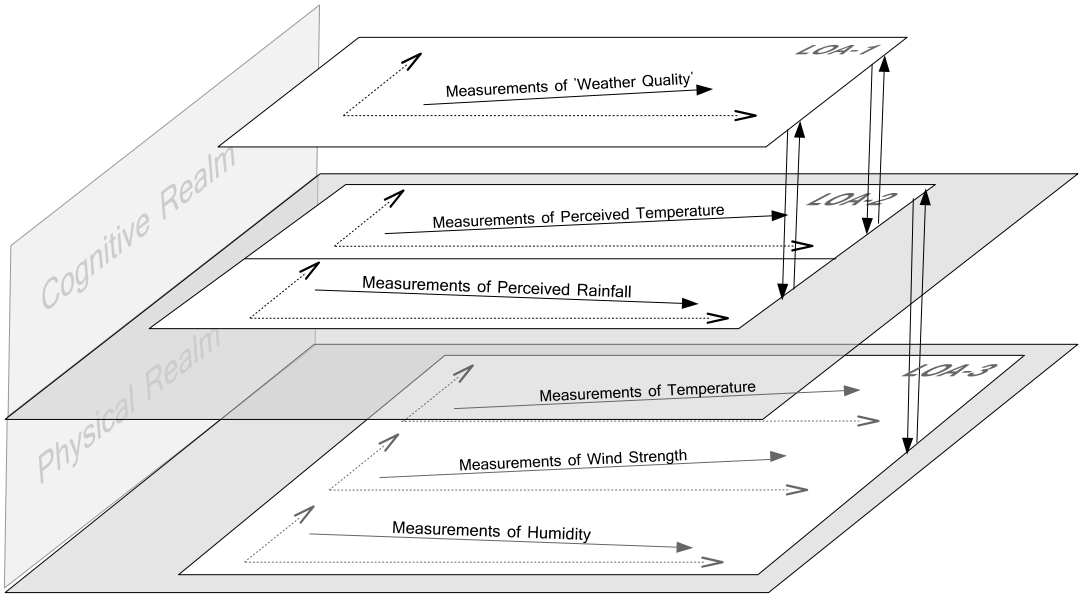


FIGURE 5.5: Levels of abstraction interpretation in terms of cognitive and physical realm

An analysis of the impact of these constructs on 'temperature perception' can be interpreted as being specifically viable due to the ease of measurement of the physically grounded constructs. This observation is important with regard to the cognitive nature of the variables on the LOA-2 and LOA-3 levels. Research on the perception of tem-

perature is more feasible relative research on 'weather quality', due to its adjacency to the physical level of abstraction. In this case, research is rendered more feasible by the high 'measurability' associated with the physical layer and the possibility for manipulation of variables. It is precisely the lack of measurability and limitations with regard to manipulation exhibited by the 'cognitive realm' that induces the necessity for the use computational models. Research of the phenomenon of 'weather quality perception' cannot with the same ease be directly based on manipulation and measurement of the contributing processes with regard to the cognitive nature of these processes.

This observation forms the basic motivation for approaching the analytic task in a bottom-up manner. Specifically, based on the observation that the difficulties with regard to manipulation and measurement of variables are proportional to the level of abstraction. The proposed solution to this observation consists of the construction of computational models of the processes, starting on the lowest considered cognitive level of abstraction. In the given example the lowest cognitive level of abstraction is given by LOA-2. The representation of the processes through a cognitive model is marked in the figure through use of the  symbol. On grounds of the physical nature of the next lower level, it is evident that the evaluation of such a model can be based on aligning measurements of temperature perception with those associated to the three underlying physical constructs. The strategy with regard to enabling research on the higher levels, consists of basing the respective analysis on measurements obtained from validated models on lower levels. In the provided example this consists of the proposal to base research concerning the impact of 'rainfall perception' and 'temperature perception' on 'weather quality' on computational model based measurements of these two constructs. It is intuitive that the use of model based measurements offers large advantages with regard to the measurement and manipulation of the respective constructs. It is further obvious that this benefit is entirely dependent on the validity of the used models. Effectively, as indicated in the figure, this results in a chain of correlations spanning over the levels of abstraction. This constitutes an example of the primary mechanisms of establishing validity based on a nomological network. Inferences with regard to constructs on different levels of abstraction are based on a chain of interlocking inferences. Conclusions with regard to the constructs and their relations are convergently validated (Garner et al., 1956).

Summary: The prior section proposed a methodology aimed at mitigating the difficulties concerning the exploration of high level cognitive tasks via a chain of analytic steps. In light of the relationship between measurability and abstraction it is proposed to conduct these steps in bottom-up order. Further with regard to the general difficulties of researching cognitive phenomena, the use of computational models was suggested. Of high regard is the importance of ensuring the validity of such models.

5.3.2 IR Specific Application of Methodology

The current section outlines the application of the introduced methodology with regard to the construction of an Information Retrieval focused nomological network.

Figure 5.6 depicts an IR focused nomological network consisting of two levels of abstraction.

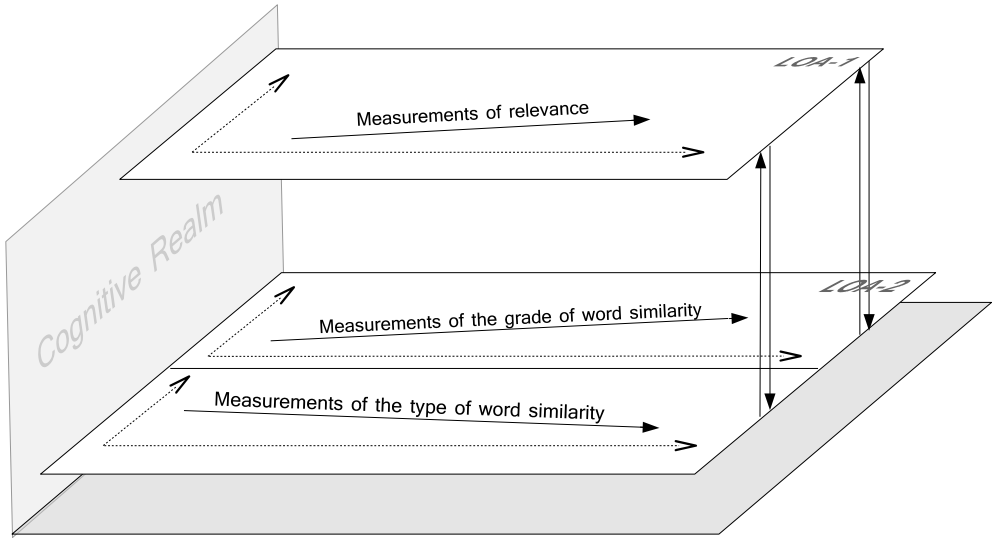


FIGURE 5.6: Correlation of Measurements of Word Similarity with Measurements of Relevance

As can be seen in Figure 5.6 , both levels of abstraction reside in the cognitive realm. The highest level of abstraction is populated by the concept of relevance. As stated before the concept of relevance can be interpreted as a product of complex cognitive processing. The lower level of abstraction is dedicated to the constructs of grade and type of word similarity. On basis of the existent body of knowledge concerning the processing of discourse (see Section 4.1), these processes can be assumed to constitute processes on the lower end of abstraction. The grade of word similarity constitutes a process that can be directly related to the organization of the memory. Further, as will be discussed in more detail in the subsequent chapters, it presents a process that can be directly related to the pathological and neurological level. In this sense the choice of the chosen levels of abstraction is consistent with the meta-theoretical criteria for the selection of constructs. With regard of pragmatic criteria it can be said, that both word

related constructs are associated with a range of measurement instruments. Detailed descriptions of these instruments are provided in the subsequent chapters.

5.4 Chapter Conclusions and Answer to RQ 3

Chapter 2 identied construct validity and the nomological network methodology as a principled way to the validation of IR constructs. Two necessary steps towards the establishment of such a network were noted. Chapter 4 demarcated a pool of potential candidate constructs. Guided by the third research question, this chapter investigated criteria for the selection of constructs.

RQ 3 What are criteria for the selection of constructs?

The investigation of these criteria was approached by considering pragmatic and meta-theoretical aspects. Section 5.2 conducted a survey of experimental means in cognitive science and IR. Based on an analysis of measurement instruments, and experimental approaches in the two domains, it presented five pragmatic criteria for the selection of constructs.

- Availability of measurement instruments
- Validity of measurement instruments
- Manipulation of variables
- Coverage of spectrum of observables emanated by a phenomenon
- Necessary effort for taking measurements

The listed criteria relate to different aspects associated with the task of measuring observables emanating from constructs. The second listed criterion relates to meta-theoretical considerations. Section 5.3 explored these considerations based on the identified principles of cognitive exploration in Chapter 3. The analysis resulted in the identification of three meta-theoretical criteria for the selection of constructs.

- Level of abstraction
- Completeness of informational description
- Measurability

The three criteria allow for an interpretation of constructs with regard to Borsboom's (2003) and Colliver et al.'s (2012) call for 'realism'. It was noted, that Borsboom's (2003) 'attributes' can be interpreted as constructs exhibiting high measurability. In light of this, the meta-theoretical criteria are a reflection of the observed limitations of

nomological networks consisting only of high level constructs (Cronbach, 1989; Borsboom, 2003; Kane, 2006; Colliver et al., 2012).

Based on the identified criteria and the candidate pool demarcated in Chapter 4, the chapter concluded with the introduction of an IR focused nomological network. Section 5.3.2 introduced a nomological network consisting of the constructs of relevance and type and grade of word relatedness. The choice of the word related constructs was motivated by the identified criteria for construct selection. The presentation of the network concluded the theoretically focused part of the thesis. Part II of the dissertation will focus on the investigation of empirical aspects underlying the application of this network.

II

Experimental Realisation

Part I was dedicated to the exploration of theoretical aspects with regard to the validation of IR constructs. Chapter 2 defined a basis for the investigation by interpreting relevance as a phenomenon occurring within the context of two systems: An IR system and the cognitive processing system of a user. Based on this definition, the concept of construct validity was identified as a principled approach to the validation of relevance. An analysis with regard to its application in the context of IR identified two necessary steps. The first step is given by the identification of a pool of candidate constructs for an IR focused nomological network. The second step consists of the inference of criteria for the selection of constructs from this pool. Based on an analysis of the pertinent cognitive research, Chapter 4 demarcated a potential pool of constructs for the network. In Chapter 5, a set of pragmatic and meta-theoretically motivated criteria for construct selection were identified. The consideration of these criteria resulted in the definition of an IR focused nomological network composed of the constructs of relevance and type and grade of word relatedness.

Part II of the dissertation is focused on the investigation of empirical aspects underlying the application of this network. With regard to the aim of structuring the research approach in form of an artefact paradigm, it covers the following element.

Application To provide an example of the application of the paradigm.

A first step with regard to the application of the network is given by the identification of validated measurement instruments for the constructs of type and grade of word relatedness. As outlined in Chapter 5, the availability of validated measurements constitutes a prerequisite for making meaningful inferences based on a nomological network. The identification of validated measurements instruments is addressed by two separate research questions.

RQ 4 What are valid instruments for the measurement of the grade of relatedness between words?

RQ 5 What are valid instruments for the measurement of the type of relatedness between words?

Presupposing the availability of validated measurement instruments, the next step is given by a demonstration of the application of the network. This is addressed by the final research question of the dissertation.

RQ 6 What are characteristics of the relation of the postulated constructs of relevance and grade and type of word relationships?

The investigations in Part II are organized as follows. Chapter 6 provides an overview of the theoretical basis of the constructs of type and grade of word relatedness, and defines the necessary experimental setup for a validation of associated measurement instruments. In Chapter 7 a validation study of measurements instruments of the grade of word relatedness is conducted. A validation study for instruments of the type of word relatedness is presented in Chapter 8. Based on the results of the two validation studies, Chapter 9 is dedicated to an investigation of the relation between the word related constructs and relevance.

CHAPTER 6

MEASURING WORD SIMILARITY PERCEPTION

Part I investigated theoretical considerations with regard to the aim of validating relevance, and resulted in the presentation of an IR focused nomological network. Part II of the dissertation is dedicated to empirical investigations associated with the application of the network. A prerequisite for making inferences based on a nomological network consists of the availability of validated measurement instruments for the word related constructs. This is addressed by RQ 4 and RQ 5 of the dissertation.

RQ 4 What are valid instruments for the measurement of the grade of relatedness between words?

RQ 5 What are valid instruments for the measurement of the type of relatedness between words?

This chapter sets the foundation for the investigation of these research questions by first providing an outline of the theoretic background of the constructs. Based on this outline, a survey of available measurement instruments is presented. Finally, an analysis with regard to the validity and approaches to the validation of these instruments is conducted.

The structure of the chapter takes the following form. Section 6.1 provides a definition of the cognitive process of word similarity perception. An overview of the state of knowledge about the processes is provided. Subsequently, the second section explores general considerations concerning the evaluation of word similarity. To that cause, an overview of the state of the art evaluation approaches is provided. Section 6.3 addresses the question of determining the validity of the outlined evaluation procedures. An analysis of the validity of evaluation procedures is conducted on basis of the existent body

of research. With regard to the identified uncertainty of the validity of these instruments, Section 6.4 formulates a validation strategy. Finally, Section 6.5 presents the experimental setup for the empirical investigation of this validation strategy.

6.1 Cognition of Word Similarity Perception

This section aims at providing an overview of the state of the art of cognitive processing related to word similarity perception. In the words of Vigliocco and Vinson (2007, p. 195) the relevant aspects can be nicely demarcated by the question: 'How are the meanings of different words related to one another?'.

Relations between words have been extensively researched in other domains (e.g. linguistics). With respect to the cognitive focus of this work, the current exploration focuses on reports from cognitive research. The first step in this exploration is focused on outlining the justification for the existence of cognitive processes of word relationships. With regard to this, Vigliocco and Vinson (2007, p. 202) note that 'semantic similarity effects are powerful and well documented in the word recognition literature' (Neely, 1991); word production literature [...] (Glaser and Düngelhoff, 1984; Schriefers et al., 1990) [...] and neuropsychological literature [...] (Ruml et al., 2000; Vinson et al., 2003)'. Empirical evidence with regard to this phenomenon primarily stems from psycholinguistic experiments, specifically reaction time (RT) based experimentation. As noted by Vigliocco and Vinson (2007, p. 203), 'semantic priming refers to the robust finding that speakers typically respond faster to a target word when preceded by a semantically related word than when preceded by an unrelated word (Meyer and Schvaneveldt, 1971)'. The observation, that target word specific reactions occur faster if 'primed' by a similar word, represents the fundamental source of empirical evidence underlying word similarity related cognitive processing.

As noted by McRae and Boisvert (1998) these results also form a principle source for research on the organization of semantic memory. With regard to the focus of this work, this aspect is only of secondary interest. In the centre of interest lie instead the principle characteristics on which basis word relationships can be described. With regard to that (Vigliocco and Vinson, 2007, p. 204) state that 'semantic effects can be graded even for complex semantic categories such as objects, not just for categories that naturally extend along a single dimension (numbers) or only a few dimensions (colours) in both word recognition and picture naming.' Semantic relations between words can therefore be interpreted as being of continuous nature. Further, within the existent cognitive science literature the question of the existence of distinct types of relations is discussed. (Jones et al., 2006, p. 541) noted that there is an overall consensus (see Hutchinson et al. (2003), McNamara (2005a)) that 'there is facilitation for both prime-target pairs that share a purely semantic relationship, and for pairs that share

a purely associative relationship.' A semantic relation in the above context generally refers to words that are related due to overlap of their defining features or their shared category membership. Feature refers to semantic features of a word. Category refers to the concept of a taxonomic semantic category such as e.g. colours, vegetables, and weapons. With regard to these definitions, words are considered semantically related if they overlap in features, belong to the same superordinate category, or exhibit both of these requirements. To illustrate this point, the above described conception is explored using the terms 'Ant' and 'Cockroach'. In Table 6.1 the top-ranked[6-1] features for both words are presented.

Ant	Cockroach
an_insect	an_insect
beh_-_bites	beh_-_flies
beh_-_crawls	found_in_houses
has_6_legs	has_a_shell
has_antennae	has_many_legs
is_black	is_black

TABLE 6.1: Features for the Words 'Ant' and 'Cockroach' Extracted from the Semantic Feature Production Norms Collected by McRae et al. (2005)

In the above case from a featural perspective the two words can be considered semantically related on basis of their shared features such as being black and having a large number of legs. Further with regard to categorical membership the two words are related due to their shared membership of the 'insect' category.

As noted by (Jones et al., 2006, p. 541) 'associated words are produced as responses to one another in free-association[6-2] norms, but are not categorically similar (nor do they have overlapping semantic features)'. Table 6.2 shows associatively related words from the Nelson free association norms (Nelson et al., 2004) fulfilling these requirements.

APART	SAD
APART	WORLDS,
APARTMENT	COMFORT
APARTMENT	COZY

TABLE 6.2: Sample associations for the words 'Apart' and 'Apartment' taken from the Nelson et al. (2004) norms

Based on the so far led discussion, the following summary can be drawn with regard to characteristics of word relationships:

[6-1] with respect to their respective frequency of being named by participants in the underlying study

[6-2] Free association norms are collected by asking participants to freely name related terms for specific target words. (See Nelson et al. (2004))

- Relationships between words are of graded nature.
- Relationships between words can be of semantic or associative type.

On basis of this characterisation of word relationships, the next section explores evaluation approaches with respect to the two above listed aspects.

6.2 Evaluation Procedures

Before delving into the analysis of evaluation approaches, it is important to precisely define the meaning of the term 'evaluation' within the respective context. Evaluation can be loosely described as the act of (b) measuring (a) something in terms of (c) a standard. More detail can be added to this statement by consideration of a simple example. 'Something' usually refers to the variable of focus of an experiment. For the sake of developing an example it is subsequently assumed that (a) is given by 'the fuel consumption of cars per kilometre'. To that cause, it is plausible that an experiment could be conducted on basis of driving a set of cars on an experimental test track. A possible evaluation of the results of this experiment with regard to (a) then consists of (b) calculating the delta of the fuel tank levels at the start and end of the experiment. The utilized (c) standards can be presented by the quantity of three-dimensional space filled by the substance (i.e. volume) expressed in $litres$ or its weight expressed in kg, and the length of the driven distance in km. This illustration highlights the importance of a fourth item with regard to evaluation: (d) An appropriate measurement instrument. In the above example such an instrument might respectively be provided in the form of a scale, a measurement container, or a ruler. Based on this example, evaluation can be defined as:

> **Definition 5:** Evaluation of (a) *a dependent variable of interest* is the act of expressing that variable via (b) *a process of 'measurement'* on grounds of (c) one or more underlying *standard measure* by use of respective (d) *instruments of measurements* of those standards.

On basis of this 'definition' in form of the above sentence it is now possible to outline the significance of this discussion with regard to the point at hand: The evaluation of computational models of word similarity. To illustrate this point another short example is provided in form of the following sentence: I evaluated the (a) *dimensions* of this steel rod by (b) *holding up* this 30 (c) *centimetre* long (d) *ruler*. This simple example, deliberately confined to a context of physical properties, allows for an intuitive application of the above proposed definition of 'evaluation'. However when considering a statement, such as given by the initial sentence of (Voorhees, 2002, p. 355) publica-

tion 'The Philosophy of Information Retrieval Evaluation', things quickly become less straightforward.

> *The evaluation of information retrieval (IR) systems is the process of assessing how well a system meets the information needs of its users.*

Expressed on basis of the introduced schema this translates into:

- Dependent Variable: 'how well a system meets the information needs of its users'
- Process of Measurement: 'assessing'
- Standard Measure: (Precision, Likert scales)
- Measurement Instrument: (Decision making of assessor)

Even without consideration of the points 'Standard' and 'Measurement Instrument', which find no reference in the quoted sentence, it becomes obvious that evaluation in a cognitive context differs from evaluation within a physical context. As illustrated by the first point, a major obstacle seems to be presented by supplying a precise definition of the dependent variable under consideration. This observation naturally also applies equally to the definition of standards. 'Process of measurement' and 'Measurement Instrument' are comparatively easy to define. In essence, this can be understood to resolve to the act of placing a descriptive label on a set of activities (e.g. 'Conducting a user study', and 'Questionnaires'). The real challenge with regard to the black-box nature of cognition consists of defining respective variables, and herein precisely lies the motivation for the so far led discussion. Especially in light of the uncertainty with regard to these definitions it seems advisable to expand considerable effort to the specification of evaluation procedures.

Applying the schema on the evaluation of models of word similarity, allows to define two main classes of evaluation procedures. Table 6.3 provides an overview of these two classes. As can be seen from the table, evaluation procedures pertaining to computational models (CPM) of word similarity can be categorised with respect to their focused dependent variable. With respect to this criterion, the two main classes of evaluation procedures are provided by 'application performance' focused and 'concordance with cognition' focused procedures. The class of application focused procedures comprises all evaluation cases where a word similarity model is evaluated after its integration into an application. Subsequently, as can be seen in the table, the standard measure usually consists of a task and application specific measure. With regard to the role of word similarity models in the paradigm's implementation this type of evaluation is of limited use. The focus on application performance makes it difficult to identify effects directly related to word similarity effects (with reference to this aspect see also Budanitsky and Hirst (2006), and Cramer (2008)). Consequently this procedure is covered only in very brief form within this section.

Concordance with human cognition	
Dependent Variable	Concordance with Human Cognition
Process of Measurement	Calculation of correlation or covariance
Standard Measure	Correlation coefficient
Measurement Instrument	Assessment Data, Priming Exp. Data
Application Performance	
Dependent Variable	Application specific performance variable
Process of Measurement	Analysis on basis of ratings
Standard Measure	Task specific human rating based measure
Measurement Instrument	Assessment Data

TABLE 6.3: Overview over word similarity focused evaluation procedures

The second class of evaluation procedures is focused on measuring the level of concordance between estimates made by the model and estimates made by human beings. The act of 'measuring' consists of statistical analysis on basis of model based output and human based output. The standard measure of the evaluation procedure is often given by a correlation coefficient. The measurement instrument (i.e. the metaphorical 'ruler') consists of data acquired via assessment or priming based experimentation. On grounds of their focus on model output, the members of this class of evaluation procedures are interpreted to represent the most viable 'tool' with respect to the nomological network methodology. Subsequently, the existent procedures of this type are explored. The exploration is structured on basis of the prior identified characteristics of word relations. The next section focuses on evaluation procedures targeting the grade of word relations.

6.2.1 Evaluation Procedures of Graded Word Similarity

This subsection provides an overview of the established methodologies for the evaluation of computational word similarity models with regard to graded similarity effects. The subsection is structured with respect to different measurement instrument types. The first section is focused on evaluation procedures that utilize human assessment as the underlying measurement instrument.

Human Assessment

The evaluation methodology based on human assessment is represented by the procedure subsequently referred to as 'word similarity tests'. Word similarity tests represent an established methodology for evaluating the concordance of model based word similarity estimates with human cognition. The methodology is based on correlating

the scaled similarity ratings of word pairs rated by a group of human assessors with word pair similarity estimates derived from computational models. The data underlying these evaluations most often stems from word similarity ratings obtained on basis of the methodology introduced by Rubenstein and Goodenough (1965).

As noted by Budanitsky and Hirst (2006, p. 18) the Rubenstein word similarity norms represent the result of a cognitively motivated investigation into 'the relationship between similarity of context and similarity of meaning (synonymy) (Rubenstein and Goodenough, 1965)' . To obtain similarity judgements 51 human participants were required to assess the similarity of 65 word pairs. The pairs were chosen with the intention of covering a range of highly synonymous to completely unrelated relations. Subjects were required to initially rank the 65 pairs according to similarity and subsequently rate each word pair's similarity on a scale of 0.0 to 4.0. The assessment methodology is considered reliable on grounds of the verification performed in form of Miller and Charles (1991, p. 13) repeated application of Rubenstein's methodology on a subset of the original word pair set. The reported result of the verification was:

> *That the two sets of ratings are in good correspondence: the Pearson product-moment correlation coefficient is* 0.97*, significant at the 0.01 level. People are not only able to agree reasonably well about the semantic distances between concepts, but their average estimates remain remarkably stable over more than 25 years.*

A similar result is reported by the replication of Resnik (1995) using the Miller and Charles (M&C) subset. The reported correlation between the M&C mean ratings and the mean ratings of his replication was 0.96. The standard correlation measures associated with this type of evaluation are Pearson's product-moment correlation and Spearman's rank correlation. In the described form this evaluation procedure has been widely applied within the course of a large number of studies. Examples of word similarity specific evaluations following this methodology are given by Finkelstein et al. (2002), Budanitsky and Hirst (2006), Durda and Buchanan (2008), and Baroni and Lenci (2010).

Psycholinguistic Measures

A second established methodology consists of basing the evaluation of word similarity models on correlation with psycholinguistic measures. The methodology primarily finds application within the cognitive psychology community.

A prime example of this evaluation procedure, is given by Vigliocco et al.'s (2004) evaluation of the Featural and Unitary Semantic Space (FUSS) model. The accuracy of the estimates made by the FUSS model were evaluated and compared to estimates of

other word similarity models based on the subsequently described approach. The first step consists of defining four distinct word pair sets. Each set relates to one of the below listed categories.

- Very closely related (e.g. dagger-sword).
- Closely related (e.g. dagger-razor).
- Moderately related (e.g. dagger-hammer).
- Non related (e.g. dagger-tongue).

As an attempt to control independent variables the chosen word pairs are 'matched as closely as possible for verbal frequency, number of letters, and minimal orthographic or phonological overlap with the target word' (Vigliocco et al., 2004, p. 462). Following the definition of these sets, priming effects for each of the pairs are measured on basis of the performance of human subjects on a lexical decision tasks. An evaluation of a model's concordance with cognition can then be performed via correlation of pair specific lexical decision response times (RT) and the respective similarity estimates of the model. In this form, the assessment based and priming based evaluation procedures are essentially identical with regard to the underlying approach. Both evaluation procedures are based on performing some form of statistical alignment of cognitive output and model based output. The main distinction is given by the utilized type of cognitive output. The following subsection completes the overview of evaluation procedures by outlining evaluation procedures focused on evaluations with respect to the type of word relations.

6.2.2 Evaluation Procedures of Similarity Type

In contrast to evaluation focused on the grade of similarity, the state of the art of evaluation focused on type of word relations is of less established nature. Analogous to the evaluation centred on grades of similarity, the underlying measurement instruments are presented in form of human assessment and priming experiment based data.

Human Assessment

To date only one study dedicated to the evaluation of computational models with regard to relationship types has been reported. The reported data of this study consists of a set of (word, word, similarity) triples that have been manually annotated in order to 'distinguish between similar and related pairs' (Agirre et al., 2009, p. 24). The reported specification with regard to the terms 'similarity' and 'related' is given by the following excerpt on page 24:

> *We manually split the dataset in two parts, as follows. First, two humans classified all pairs as being synonyms of each other, antonyms, identical, hyperonym-hyponym, hyponym-hyperonym, holonym-meronym, meronym-holonym, and none-of-the-above. [...] This annotation was used to group the pairs in three categories: similar pairs (those classified as synonyms, antonyms, identical, or hyponym-hyperonym), related pairs (those classified as meronym-holonym, and pairs classified as none-of-the-above, with a human average similarity greater than 5), and unrelated pairs (those classified as none-of-the-above that had average similarity less than or equal to 5). We then created two new gold-standard datasets: similarity (the union of similar and unrelated pairs), and relatedness (the union of related and unrelated).*

As becomes evident on basis of this description, the theoretical underpinnings of the relationship classification stem from the linguistic domain. In relation to the interpretation of relationship types in cognitive psychology the following can be noted. The definition of the similarity type in congruent with the concept of a semantic relationship in psychology, in so far as synonymy, antonymy, identical, or hyponymy-hyperonymy are also defined on basis of shared features or category membership. The main difference in contrast to the concept of a semantic relationship consists of the lack of consideration of the 'associative factor'. The word pair 'cat and mouse' for example suffices the definition of a similarity relationship on grounds of hyponymy-hyperonymy but would likely be classified as a semantic-associative relation in psychology on grounds of the words exhibiting a strong associative relation. Concerning associative relations in the psychological sense the diversion of definitions is greater as meronymy-holonymy is regarded as a relationship of semantic type. The dataset therefore can be interpreted as only partly applicable with regard to the aim of evaluating the semantic-associative axis as defined by cognitive psychology.

Psycholinguistic Measures

With regard to the use of psycholinguistic measures, the closest related approach to evaluation on the semantic-associative axis is provided by priming based studies using analytically[6-3] chosen semantic and associative word pair lists. The number of studies following this approach as a means of evaluating computational models of word similarity is very limited. Jones et al. (2006) used a priming data based methodology to evaluate a series of word similarity CPMs with regard to the semantic-associative 'nature' of their estimates. The applied approach can be described as follows. On basis of distinct lists comprised of word pairs that analytically have been assessed to be semantically,

[6-3]The term 'analytical' in this context means that word pairs are chosen manually on basis of the prior listed psychological definitions of semantic and associative relations.

associatively, or semantic-associatively related to each other, similarity estimates of the models were collected for each of the pairs on basis of simulating priming. Assessment of the associative or semantic character of a model was then conducted by comparison of the mean simulated priming values with the reported means of the original priming study. In this form, the evaluation procedure can be interpreted as a form of qualitative assessment of model output with regard to the semantic-associative aspect.

6.2.3 Conclusion

Evaluation of computational models of word similarity can be summarized as follows. Evaluation procedures aimed at such models can be classified into two groups: Procedures focused on indirectly evaluating models on basis of an encompassing application's performance, and those focused on direct evaluation of such models. It has been noted that application-centric evaluation procedures are not viable for this study due to the associated difficulties of isolating word similarity related effects. Concerning evaluation efforts directly focused on word similarity models it has been outlined, that the dependent variable of interest consists of concordance with cognition. That is, the degree to which the model's behaviour mimics the 'behaviour' of the modelled cognitive processes. The aim of these evaluation procedures is congruent with the requirements on measurement in the context of a nomological network. Concerning the outline of these procedures it has been shown, that concordance based evaluations are based on the use of two kinds of metaphorical 'rulers': Assessment data and priming experiment data. Regarding the dichotomy represented by graded similarity and associative-semantic effects it was concluded, that the availability of data is much higher with respect to graded similarity effects.

As a general conclusion it can be stated, that the outlined evaluation procedures in principle constitute viable means for the evaluation of the chosen constructs. With regard to the particular viability of the specific procedures, concerns have been raised that question the adequacy of priming based procedures in terms of sensitivity. With regard to assessment based procedures, the validity of the underlying assessments has been questioned. These noted concerns result in the following situation with respect to the feasibility of taking measurements on the word similarity level of abstraction. Assessment based evaluation constitutes a procedure that, on grounds of the specifics of the underlying assessment approach, is assumed to be sufficient with respect to sensitivity requirements. However, in light of the noted concerns, questions of its validity seem justified. The main concern in regard of priming based approaches is rooted in questions regarding its sensitivity. Against this background, the next step consists of performing a preliminary analysis of the validity of the procedures – specifically the validity of assessment based procedures.

6.3 Validation of Evaluation Procedures

Section 6.1 provided an overview of the cognitive processing underlying word relationship perception. A central statement within this outline consisted of the observation that word relationships can be characterised in the following manner:

- Relationships between words are of graded nature.
- Relationships between words can be of semantic or associative type.

Section 6.2 provided an overview of evaluation procedures for both aspects. Within the following section a preliminary analysis of the validity of such procedures is conducted. To that cause the subsequent section provides a brief overview of the concept of validity.

6.3.1 Preliminary Analysis of the Validity of Evaluation Procedures

The focus of this subsection consists of discussing the validity of the so far introduced evaluation procedures. With respect to the discussion in Section 6.2.3, the analysis is focused on assessment based word similarity tests and priming simulation. Concerns of the validity of application based and experimental priming based evaluation are only briefly explored.

Application Based Evaluation

The evaluation of computational word similarity models within the context of applications provides only very limited possibilities of making inferences aimed at the accuracy of similarity grade and type. As illustrated by Zesch and Gurevych (2006, p. 17) 'a certain measure may work well in one application, but not in another. Application-based evaluation can only state the fact, but give little explanation about the reasons.' The validity of such procedures, purely with respect to a focus on word relationship effects, is therefore assumed to be low.

Assessment Based Evaluation

An investigation of the validity of word similarity assessment based evaluation of computational models can be divided into two sub-investigations:

- Investigating the validity of interpreting the similarity ratings produced by the assessments as measures of word similarity.

- Investigating the validity of interpreting the correlation values between the judgements and estimates by CPMs as a measure of how closely the CPM correlates with the cognitive word similarity processing.

Naturally, the validity described by the second point is dependent on the validity of the first. If the similarity ratings produced in the assessments do not reflect word similarity, then consequently the same is true for values derived from correlation with it. Therefore, validity concerns described by the first point will be initially considered. A first step within these considerations consists of addressing the issue of reliability, as the validity of a measure derived from a test can be ruled out if a test is not reliable. In that regard it can be stated that the reliability of the assessment procedure originally applied by Rubenstein has been shown to be extremely high. This has been evidenced by the high correlation (shown in parentheses) with the assessments of respective subsets of the word pairs as reported by the Resnik (1995) (0.96) and Miller and Charles (1991) (0.97) studies. On basis of the observation, that repeated execution of the procedure leads to highly correlated values, the focus can now be shifted to the question if these values represent reflections of word similarity. This consideration is subsequently conducted based on analysing the reporting of three similarity judgement experiments:

- Rubenstein and Goodenough (1965)
- Finkelstein et al. (2002)
- Cramer (2008)

Since an analysis of the validity of the values generated by these studies is dependent on the declared measured variable the initial discussion is concerned with identifying its declaration within the publications.

Rubenstein and Goodenough: Concerning the question of the declared measured variable, Rubenstein and Goodenough (1965) state that the judgements of the participants are 'Synonymy Judgements'. However, no precise definition of synonymy is provided by them in the publication. In the first sentence of the study it further is stated that 'This study is concerned with the relationship between similarity of context and similarity of meaning (synonymy)' (p. 627). The use of the terms 'similarity of meaning' and 'synonymy' in this form is critical, both with regard to a distinction of semantic relatedness and associative relatedness in the cognitive psychology domain, and the formal definitions of the term 'synonymy' in the linguistic domain. In the linguistic domain a strict definition of synonymity, referred to by Cruse (1997, p. 88) as 'cognitive synonymity', is given by the following excerpt.

> *X is a cognitive synonym of Y if (i) X and Y are syntactically identical, and (ii) any grammatical declarative sentence S containing X has*

> *equivalent truth-conditions to sentence S′, which is identical to S except that X is replaced by Y.* ”

As an example of a word pair sufficing above conditions Cruse (1997) lists ’fiddle’ and ’violin’. It therefore can be stated, that it cannot be verified whether Rubenstein and Goodenough aimed at measuring a specific type of relatedness. This observation is specifically strengthened on basis of the provided instructions to the assessors (p. 628):

> “ *Assign a value from 4.0-0.0 to each pair–the greater the ’similarity of meaning,’ the higher the number. You may assign the same value to more than one pair.* ”

It can be assumed, that the assessors therefore assessed similarity in a ’general fashion’, indifferent of similarity types. However, on basis of the chosen word pairs it can be assumed that Rubenstein and Goodenough aimed at measuring similarity of the type ’semantically related’ in a cognitive psychology sense, and ’synonymous’ in a linguistic sense. Concluding, a statement regarding the declared measured variable can only consist of the following form:

- The applied assessment procedure is meant to measure the similarity of the meaning of words.
- On basis of the chosen word pairs it can be assumed that it is biased towards the measurement of semantic similarity.

Having ’clarified’ the question of what constituted the intended measure, it can now be addressed of how valid the measurement approach is. At this point the question is approached on basis of prior validation efforts of the similarity judgements. To the best of our knowledge no dedicated study focused on the validity of assessment procedures has been conducted. However a variety of studies have been conducted that correlated the similarity judgements with other sources derived on the basis of human word similarity assessments. Initial inferences with regard to the validity of the procedure are therefore based on the reported results of these studies.

Budanitsky and Hirst (2006) conducted a study aimed at ’Evaluating WordNet-based Measures of Lexical Semantic Relatedness’. The focus of the study consisted of the evaluation of WordNet[6-4] based semantic similarity models. Such models use WordNet’s link structure denoting relations between words such as Synonymy[6-5], Hyponymy

[6-4] ’WordNet is an online lexical database designed for use under program control. English nouns, verbs, adjectives, and adverbs are organized into sets of synonyms, each representing a lexicalized concept. Semantic relations link the synonym sets.’(Miller (1995))

[6-5] According to Fellbaum (1998) the underlying definition of synonymy in WordNet is the following: ’The notion of synonymy used in WordNet does not entail interchangeability in all contents; by that criterion, natural languages have few synonyms. The more modest claim is that WordNet synonyms

and Meronymy in order to compute the similarity of word pairs. As noted by Fellbaum (1998), the source of the word and concept links is human assessment. The links stemming from the lexicographers' assessments of the semantic relations between words and between concepts constitute a record of human judgement of semantic relations. In that regard the similarity ratings obtained by Rubenstein and Goodenough and the WordNet based measures of similarity, particularly so with regard of WordNet's lexicographers' assessments forming their basis, can both be interpreted as reflections of human similarity perceptions. They constitute measures derived from independent[6-6] procedures believed to measure the same variable: The perception of word similarity. Table 6.4 presents the Pearson product-moment correlation coefficients between the WordNet based similarity measures and the R&G dataset. The reported mean correlation coefficient of 0.8026 indicates a strong positive association between the two datasets. This can be interpreted as supportive evidence for the hypothesis, that both tests measure the same declared variable. It is a first indication, that word similarity assessment in the form as practised by Rubenstein is valid.

Similarity measure	R&G
Hirst and St-Onge (rel_{HS})	0.744
Leacock and Chodorow(sim_{KC})	0.816
Resnik(sim_R)	0.774
Jiang and Conrath($dist_{jc}$	0.85
Lin(sim_L)	0.829
Mean	0.8026

TABLE 6.4: Correlation between various WordNet based Similarity Models and Rubenstein and Goodenough Word Similarity Judgements as Reported by Budanitsky and Hirst (2006)

Cramer: Cramer (2008) conducted a study titled 'How Well Do Semantic Relatedness Measures Perform? A Meta-Study'. As stated by the title, the studies focus was set on comparative analysis of the reported results of different semantic relatedness measures. In the course of this exploration it also briefly explored the question of the validity of human assessment procedure.

Cramer created seven distinct word pair lists. These lists were subsequently assessed with regard to each pair's similarity and correlated with various assessment based similarity measures. The results of this process are reported in table 6.5.

can be interchanged in some contexts. To be careful, therefore, one should speak of synonymy relative to a context, but in order to facilitate the discussion this qualification will usually be presupposed not asserted.'

[6-6]With regard to the outlined underlying procedures of both sources it is of note that both procedures are quite similar, which is can be considered not optimal with regard to the validation concern.

r	Tree Path	Graph Path	Wu - Palm.	Leac. - Chod.	Hirst-St-O	Resnik	Jiang - Conr.	Lin	Google Norm	Google Quot.	Google PMI
WP1	0.41	0.42	0.36	0.48	0.47	0.44	0.45	0.48	0.27	0.37	0.37
WP2	0.09	0.31	0.33	0.16	0.26	0.37	0.18	0.36	0.24	0.29	0.27
WP3	0.03	0.22	0.24	0.11	0.28	0.19	0.15	0.26	0.46	0.45	0.4
WP4	0.07	0.39	0.11	0.11	0.31	0.11	0.25	0.16	0.34	0.38	0.34
WP5	0.27	0.39	0.26	0.32	0.38	0.31	0.41	0.34	0.19	0.32	0.28
WP6	0.09	0.27	0.15	0.17	0.39	0.24	0.29	0.25	0.26	0.38	0.43
mean	0.16	0.33	0.24	0.23	0.35	0.28	0.29	0.31	0.29	0.36	0.35

TABLE 6.5: Overview of Correlation Results between Word Similarity Judgements and Different Word Similarity Models as Reported by Cramer (2008)

The reported mean correlations uniformly suggest a weak association between the judgments and different word similarity models. The hypotheses drawn by Cramer in regard of these observations are the following (p. 63):

- 'Word nets (and/or corpora) do not cover (all) the types of semantic information required.'
- 'Human judgement experiments are (without clear and standardized specification of the experimental setup) an inappropriate way to evaluate semantic measures.'

These rather broadly formulated hypotheses can be interpreted as a reflection of the following considerations. An analysis of the validity of such similarity assessment procedures can only be addressed in very limited form on basis of the comparison of the results of different studies. This is specifically evident by considering the classification of types of similarity and the level of detail concerning the declaration of the measured variable in the available studies. The following two excerpts from two additional studies illustrate this point:

- Finkelstein et al. (2002, pp. 128-129): 'To this end, we prepared a diverse list of 350 noun pairs representing various degrees of similarity, and employed 16 subjects to estimate the 'relatedness' of the words in pairs on a scale from 0 (totally unrelated words) to 10 (very much related or identical words).'
- (Baroni and Lenci, 2010, p. 20): 'The average rating for each pair is taken as an estimate of the perceived similarity between the two words (e.g., car-automobile:3.9, cord-smile:0.0).'

As outlined by the provided examples, the focused measured variable is often only 'vaguely' defined. Possible reasons for this consist of the lack of a uniformly accepted nomenclature, and the still limited understanding of word relationships in general. On basis of the so far presented data it can be argued, that it is not possible to draw meaningful conclusions simply with regard to the low number of explored experiments. Following that line of thought it can further be stated, that in order to be able to draw more profound conclusions one of the two following items would be necessary. A large num-

ber of studies conducted on basis of differing word pairs, and a precise specification of their main experimental conditions. On basis of such conditions the inference of conclusions could be based on statistical analysis focused on the properties of the word pairs and the experimental conditions.

Such conditions would undoubtedly allow the formulation of more 'narrow' hypothesis with regard to observed differences of the experimental results. However, subsequently it is argued that even if the above listed beneficial conditions would exist[6-7], inferences with regard to the validity of the measurement instrument (i.e. the assessment procedure) are still limited. Different studies do not constitute different 'kinds of data' as referred to by Lachman et al. (1979). To illustrate this point the following example is used. A thermometer constitutes a measurement instrument aimed at measuring temperature. Given two thermometers whose principle of operation rests on the expansion of different liquids, an analogous approach of examining the validity of their measurements, consists of the conduction of measurements with both items under a range of conditions. Presupposing that a very large number of such measurements has been conducted, and a correlational analysis shows very strong positive association, it is argued that only the following can be concluded: Both instruments measure the same '*thing*'. However without the inclusion of *different types of data* it is not possible to assert that the measured '*thing*' is temperature. As historically illustrated[6-8], early liquid expansion based thermometers were also sensitive to atmospheric pressure. A strict limitation of the observations to the values of such thermometers and an examination of the strength of the linear relationship between the vectors of observations therefore obviously is not sufficient with regard to the formulation of a conclusion regarding their validity. Two things are necessary to substantiate such a conclusion.

- **Different kinds of observations**: In the temperature example such observations could be given by solid-liquid-gas state changes of matter.

- **A system of relations**: A nomological network.

Concluding the following summary is drawn. The validity of word pair assessment based evaluation procedures of computational models can be described as a yet to be decided question – particularly so with regard to the consideration of different types of similarities. Subsequently the validation of priming simulation based evaluations is explored.

[6-7] They do not. The number of such studies is very low as a result of the large required effort. Further the reported studies often only provide limited detail of the experimental conditions.

[6-8] 'Instruments of the type used by Galilei were influenced by barometric pressure and are now called barothermoscopes' (Benedict (1984))

Priming Based Evaluation

Inferences on basis of simulating priming on computational models are based on a comparison of the output of the computational model and response times of human subjects. The question of the validity of utilizing this procedure as a measurement instrument for the output of a CPM therefore is first of all a question of the validity of the response time measure. Such considerations of the validity of response time (RT) based priming evaluations are dependent on the interpretation of the RT measure. The most basic interpretation of the RT variable consists of its interpretation as a measure of lexical access time. Concerning this interpretation, the measure can be interpreted as exhibiting a very high validity simply on basis of the convergence of the reported results of a very large body of conducted research. As noted earlier, it has been shown that lexical response time is dependent on factors such as the length of a word (shorter words exhibit faster response times), verbal frequency, number of letters, orthographic or phonological overlap, or contextual diversity (Adelman et al., 2006). Interpretation of response times as an indicator of word similarity, on basis of controlling the above mentioned factors, can be considered a highly valid measure. The validity of these procedures is further confirmed by its reported correlation with neuro-physiologically based procedures such as Electroencephalography (See Bentin et al. (1985), and McNamara (2005b) for a detailed overview of neuro-scientific studies of word priming). In comparison with the noted issues concerning the state of assessment based procedures, priming based evaluations are interpreted as exhibiting higher validity on basis of the following two points:

- The precision of the underlying experimental specifications. (E.g. control of independent variables)
- The performed correlation of RT based results with other kinds of observations such as neuro-physiological data.

6.3.2 Conclusion

The results of the above outlined discussion can be briefly summarized as follows. On basis of the existing state of the art it is *not clear* if assessment based evaluation presents a valid procedure. This aspect specifically applies, when different types of relations between words are taken into consideration. It should be marked, that the above conducted 'analysis' certainly is not considered an in all aspects satisfying and decisive exploration of this aspect. It is not possible to form a conclusive answer concerning the validity of assessment procedures. Contrary to this, the validity of priming based evaluation is interpreted to be high on basis of the nature of the reported state of the art in cognitive psychology.

In combination with the conclusions drawn in Section 6.2.3, this results in the following situation with regard to the experimental application of the nomological network. The basis for making inferences about the relation of the constructs of word similarity perception and relevance is given by aligning measurements associated with both constructs. In light of this, it is imperative that the applied evaluation procedures exhibit the following characteristics:

- **Validity:** Naturally the drawn inferences are subject to the validity of the measures on both levels of abstraction.
- **Sensitivity:** With regard to the characteristics of the task on the higher level of abstraction (i.e. relevance estimation) it is assumed, that it is necessary that the evaluation procedure allows for fine grained measurements concerning the grade of word similarity.

Regarding the first point it has been asserted, that priming data based evaluations are interpreted to suffice the condition of validity. In case of assessment based evaluations it is not feasible to derive an answer on basis of the reported state of the art. With regard to the sensitivity aspect the situation is reversed considering the noted concerns of sensitivity limits of priming based experimentation. Assessment based evaluation is interpreted as a measurement instrument of high sensitivity and potential validity. This induces the necessity for a validation of assessment based procedures. A strategy for such a validation is outlined in the next section.

6.4 Evaluation and Validation Strategy

With respect to the defined nomological network in Section 5.3.2, the formulation of the evaluation and validation strategy focuses on the following two items.

- Grade of relations between words.
- Type of relations between words.

The validation strategy rests on the concept of convergent validation (Garner et al., 1956). With regard to that, the following items are identified as necessary prerequisites for the conduction of such a validation procedure.

- A nomological network.
- The availability of different kinds of measurement data.
- A sufficiently large independent variable space.

- A sufficiently large underlying condition space on which grounds measurements can be made.

Subsequently the validation strategy is introduced on basis of outlining these points in detail. As noted by Lachman et al. (1979), the concept of convergent validation describes the idea, that the convergence of several different kinds of data on a conclusion, convergently validates this conclusion. Lachman et al. (1979) outlined, that the application of convergent validation is based on the utilization of 'several different kinds' of data. In the absence of direct observable phenomena as a verification source, experimental assumptions are validated by theoretically based relation of experimental data and their underlying assumptions. This notion is also referred to as criterion-related validity (i.e., the extent to which measures of a measurement procedure converge with measures of an independent procedure believed to measure the same variable). Mathematical methods applied to verify such convergence are given by correlation matrices and factor analysis. As noted by Cronbach and Meehl (1955, p. 288) the underlying conception is, that 'if two tests are presumed to measure the same construct, a correlation between them is predicted.' As further noted by Cronbach concerning the limitations of the method: 'if the obtained correlation departs from the expectation, however, there is no way to know whether the fault lies in test *A*, test *B*, or the formulation of the construct.' (p.288).

This applies to the task of validating evaluation procedures of word relatedness in the following way. The evaluation procedures discussed in the prior section constitute instruments of measurement. The intended measured variable is given by the degree of concordance of computational word similarity models with human cognition. More specifically, each of the earlier described procedure's purpose can be described as measuring to which degree the output of a computational model is 'the same' as the output of the modelled cognitive process. The use of the term '*output*' in this context provides a convenient peg on which to hang the discussion. It can be noted, that each of the above described procedure's operating principle consists of correlating the *output* of the computational model with *output* of the 'human system'. To avoid confusion, this specific process is henceforth referred to as 'evaluation' in the subsequent discussion. The result of such evaluations consists of a measure representative of the degree of correlation of these *outputs*. Convergent validation of these evaluation procedures can then be conducted on two principle levels: The cognitive output level and the computational output level. Validation on the cognitive output level can be conducted via correlation of the different available types of cognition based data (e.g. the correlation of reaction time based data with verbal assessment based data or neurological evidence). Analogous to this, validation on the model output level can also be based on correlating different types of model output. The validity of such correlations with respect to cognition can then be argued on basis of two distinct interpretations. The first interpretation assigns validity to model output correlations on basis of the acceptance of the validity of the underlying

cognitive data used within the evaluation procedures. If the cognitive data is considered valid, the application of correlation 'propagates' this validity. In the absence of valid underlying data, the interpretation of correlated model outputs is possible on basis of the theoretical context forming the basis of the models. In particular such interpretations can be rendered meaningful through the consideration of additional types of validation such as discriminant validity and divergent validity (Onwuegbuzie et al., 2007). Such an interpretation is closely related to the expressed opinion of Sun (2009) that computational models are generally not restricted to validation on basis of empirical data, and are not solely limited to be interpreted as an abstraction of cognitive architectures, but rather also constitute independent theoretic entities by themselves[6-9].

Based on these consideration, the next subsection explores the available kinds of measurements.

6.4.1 Kinds of Measurements

As noted before, a requirement for the evaluation of validity consists of the availability of different 'kinds' of measurements. To clarify this aspect a short detour is taken by briefly re-iterating 'what is to be validated' within this study. Within the scope of this work two computational models of word similarity, Latent Semantic Analysis (LSA), and Hyperspace Analogue to Language (HAL) are utilized. LSA and HAL can be interpreted as instruments constructed to measure word similarity. This view can be based on the following excerpts:

- **LSA**:'Let us now construe the semantic similarity between two words in terms of distance in semantic space: The smaller the distance, the greater the similarity.' (Landauer and Dumais, 1997, p. 215)
- **HAL**: 'Once the matrices are constructed, similarity measurements can be applied to word vectors; this, we hoped, would yield a measure of semantic similarity between any desired pair of words.' (Lund and Burgess, 1996, p. 204)

Specifically in the case of LSA other interpretations of what LSA constitutes (e.g. a theory for learning) have been proposed. Within the course of this discussion, however, LSA and HAL are firmly and solely interpreted as measurement instruments of word similarity, and considerations of cognitive plausibility and associated aspects are deliberately left out. Thus LSA and HAL are interpreted simply as candidate instruments for the measurement of word similarity, that with regard to their specific role in our study are in need of validation. In the most simplifying manner, validation can be described as making inferences on basis of the act of comparing measurements made by

[6-9] As noted by Sun, this is closely connect to the concept of 'argumentative patterns' discussed by Kitcher (1981)

one instrument with the measurements made by other available instruments designed to measure the same thing. As stated before such validation requires the use of 'different kinds' of data. A more detailed interpretation of this statement can be made in the following way. Validation requires the availability of (a) a set of instruments that are (b) devised to measure the *same* phenomenon on basis of (c) *different* observable, and therefore measurable, manifestations of the phenomenon. The requirements specified by (a) and (b) are completely intuitive. The requirement of (c) is less obvious. Intuitively its requirement can be justified on grounds of a gain of certainty stemming from a consideration of distinct assumptions underlying the attribution of manifestations to a certain phenomenon. On that ground the necessity of its requirement can be interpreted to be proportional to the certainty that is held in the attribution of a manifestation to a phenomenon. Concluding that thermometer *A* is valid on basis of comparing its measurements to a valid thermometer *B* can be seen as a sufficient test of the validity of *A* on grounds of the commonly held certainty that the expansion of a liquid is a manifestation of the phenomenon of temperature. On that note it can be marked that the lower certainty concerning the attribution of manifestations to specific aspects of the mind forms the motivation for requirement (c).

With regard to this, the current point of the discussion is well suited to take look at the available instruments of measuring word similarity. Table 6.6 provides an overview of these.

Measurem. Instr.	Measurem. Mean	Manifest. of Phenomenon	Measured Variable
LSA	Angle betw. 2 word co-occ. vectors	Written Text	Word Similarity
HAL	Dist. betw. 2 word co-occ. vectors	Written Text	Word Similarity
Priming Exp.	Reaction time w.r.t. 2 words	Physiological Reaction	Word Similarity
Assessments	Numerical judgement w.r.t. 2 words	Written Number	Word Similarity

TABLE 6.6: Measurement Instruments of Word Similarity

Four measurement instruments are listed and described in terms of three aspects. As is evident from the last column, all instruments aim at measuring the same variable: Word Similarity. The second column provides an overview of the underlying mean of measurement. The description, while representing an arbitrary simplification of the processes, emphasizes the differences of the 4 approaches. LSA and HAL based measurements stem from co-occurrence vector based calculations. Priming based measurements are calculated on basis of reaction times. Assessment based measurements finally result from verbal or written judgement provided by experimental participants.

To further highlight those differences, the third column lists the manifestations (i.e. the observables) on which grounds the respective measurement instruments operate. Priming based experiments for example are essentially dependent on the physiological reaction exhibited by a finger exerting pressure on a button. Assessments obtain measurements from the formulation of written or verbalized output by participants. An interesting case is given by questioning the manifestation underlying LSA and HAL

based measurements. As can be seen from the table, the proposed manifestation on which basis HAL and LSA derive their measures is 'text'. On first thought this might seem an odd notion, but in comparison with the so far listed manifestations it seems intuitive that (a) text is clearly an output - a product of the mind and thus a manifestation of cognitive processing, and that (b) it is this aspect that all listed manifestations have in common. In that regard the following conclusion is drawn. While in contemporary literature concerning the evaluation and validation of computational models of word similarity one can gain the impression that implicitly a distinction is made between the models (e.g. HAL, LSA) and validation tools (e.g. assessments), on basis of the so far outlined discussion such a distinction seems not warranted. Therefore during the course of the validation study all described procedures are interpreted simply as different measurement instruments of the same phenomenon.

6.4.2 Independent Variable Space

The principle of convergent validation can also be described in the following way. High correlation within the same shared underlying variable space is interpreted as supportive evidence of validity. In regard of this, the single most important characteristic of the underlying independent variable space can be interpreted to be the following: The considered independent variable space should be chosen in a way that causes the dependent variable to vary considerably.

Regarding the dependent variables under consideration, the relationship between the independent and dependent variables is not obvious. As a consequence the independent variable space is chosen to be the maximum *feasible* parameter space of the applied computational models. The limiting factor with regard to feasibility stems from the computational requirements of the models – specifically in terms of memory consumption. In the case of LSA the parameter specific limitations are bounded by dimensionality[6-10]. In the case of HAL the restrictions concerning memory are bound by the size of the sparse matrix holding the co-occurrence counts. The size of the matrix is roughly proportional to the window size parameter. An extensive coverage of these aspects is provided by Golub and Van Loan (1996). As an additional attempt of varying the dependent variable, a series of different weighting and transformation functions for each computational model is used. An overview over the resulting parameter space is provided in Figure 6.1.

The relation with regard to the validation of word similarity evaluation procedures on basis of the illustrated parameter space is subsequently discussed by introducing the term 'estimator'. An 'estimator' is defined as an instance of a computational model

[6-10]Memory usage is proportional to the number of rows and columns of the original matrix and the number of computed singular vectors

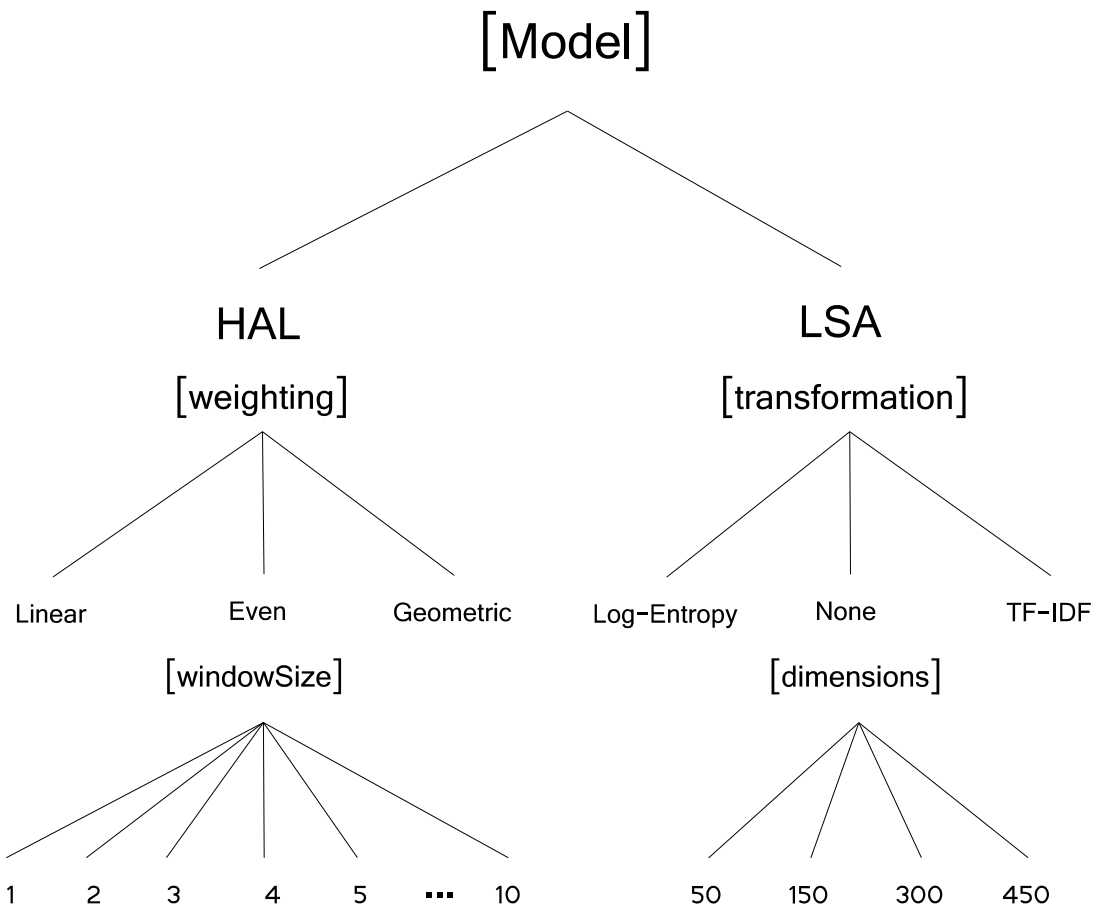


FIGURE 6.1: Overview over Estimator Space

with a specific parameter setting. A HAL based model with a window size of 5 and the 'Linear Term Weighting' function is an example for such an estimator. As such, each of these estimators provides estimates of word similarity that are assumed to correlate to a certain degree to the modelled cognitive process (i.e. word similarity perception). Concerning the determination of the validity of two candidate evaluation procedures A and B and a set of n estimators $E = \{e_1, e_2, ..., e_n\}$, the following statement can be formulated. Upon complete evaluation of E the output of A and B consists of a set of values $a_v = (a_1, ..., a_n)$ and $b_v = (b_1, ..., b_n)$ representing the respective grade of correlation of the estimator based similarity measurements with the similarity measurements of A and B. $a_v = b_v$ then constitutes the case that provides maximum support for the hypothesis that A and B are valid measurement instruments of word similarity perception. In practice, validity can be inferred on basis of calculating the correlation and covariance of a_v, and b_v. With regard to the significance of such calculations it is evident that larger n are favourable. In the case of this study the value of n is 42.

6.4.3 Condition Space

The scope of any inferences of validity is dependent on the size of the underlying condition space. Statements of validity can only be made with certainty in reference to the conditions underlying the measurements. On basis of the utilization of a nomological network these statements can be extended beyond the barriers of the utilized independent variable space. The level of uncertainty of such an extension then of course is dependent on the 'quality' of the nomological network. If the validity of a set of thermometers is evaluated under certain conditions, a statement with regard to their validity is initially limited in scope to exactly those conditions. With regard to the high 'quality' of the nomological network (i.e. physics) these statements can be 'extended'. The respective nomological network concerning word similarity is of much lower 'quality' relative to that of physics. Specifically with regard to this aspect, the underlying set of conditions (i.e. collections of text documents) of this study is chosen with the aim of exhibiting a large diversity with respect to the following distinct properties.

- Topical focus.
- Quality of content.
- Genre of content.
- Size of collection.

On basis of this set of properties four collections are chosen for this study. An interpretation of these collections in regard of these four points is provided in Section 6.5.2. To further vary the condition space a set of four term frequency dependent representations for each collection is created. Concerning the exploration of computational models of

word similarity the set of considered terms is often restricted by the application of a term frequency dependent threshold. Usually in the form of requiring a term to exhibit a minimum document frequency df (i.e. the number of documents in the collection in which the term occurs). Since to our knowledge no dedicated study concerning the effect of such df thresholds has been conducted, the generation of df representations enables an investigation of this aspect, as well as acting as an expansion of the condition space. Details concerning this aspect are provided in Section 6.5.2. Figure 6.2 provides an overview of the resulting condition space.

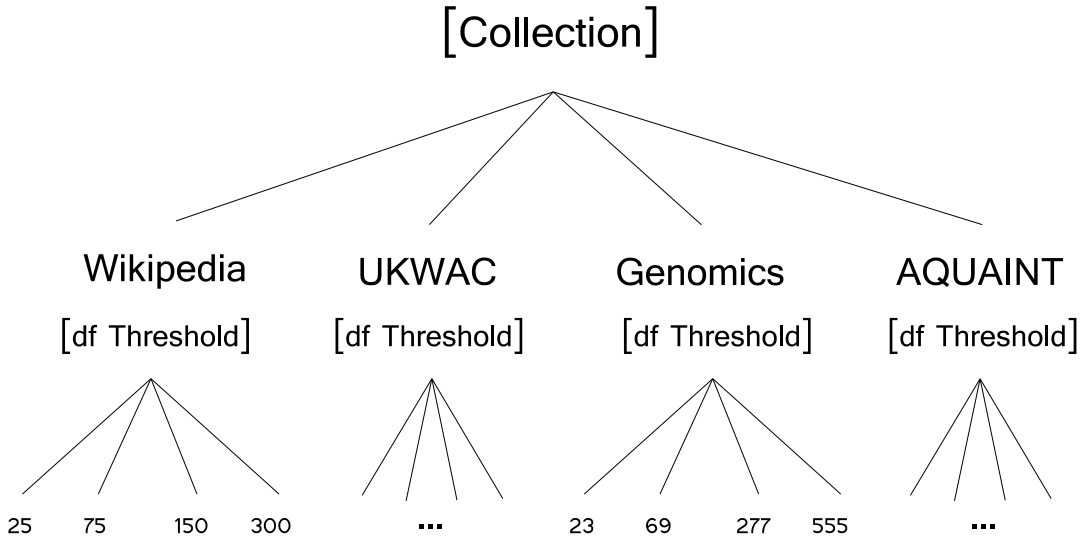


FIGURE 6.2: Overview over Data Space

The total number of utilized representations is 16.

6.4.4 Validation Strategy of Graded Perception of Word Similarity

With regard to the so far led discussion a description of the validation strategy can be formulated in a very brief fashion. The applied measurement instruments are, with reference to the arguments presented in Section 6.4.1, given by the set M = {HAL,LSA,Priming Experiments, Assessments}. Necessary variation of the measured variable is aimed to be ensured via the chosen independent variable space described in Section 6.4.2. The

scope and robustness of inferences results from the chosen condition space described in Section 6.4.3.

Validation is based on the *alignment* of different measurement instruments of the same underlying phenomenon. As noted in Section 6.4.1, all four instruments are perceived as such entities. Therefore the maximum amount of information with regard to the validity of the instruments can naturally be derived by alignment of all four instruments with each other. With regard to the scope of this work, specifically from a practical point of view, several restrictions apply. Concerning the large required effort of priming and word similarity assessments it is not feasible to conduct studies of the required scale. Validation on basis of priming and assessment are therefore restricted to the measurements reported by the contemporary literature. These observations are 'bound' to the underlying term pairs. As the pairs in the respective priming and assessment studies are distinct, no direct alignment of the derived measures is possible. As a consequence, the validation is limited to an alignment of computational model results with priming and assessment data. An overview outlining this resulting constellation is provided in Figure 6.3. Consequently the reporting of the respective validation results in Chapter 7

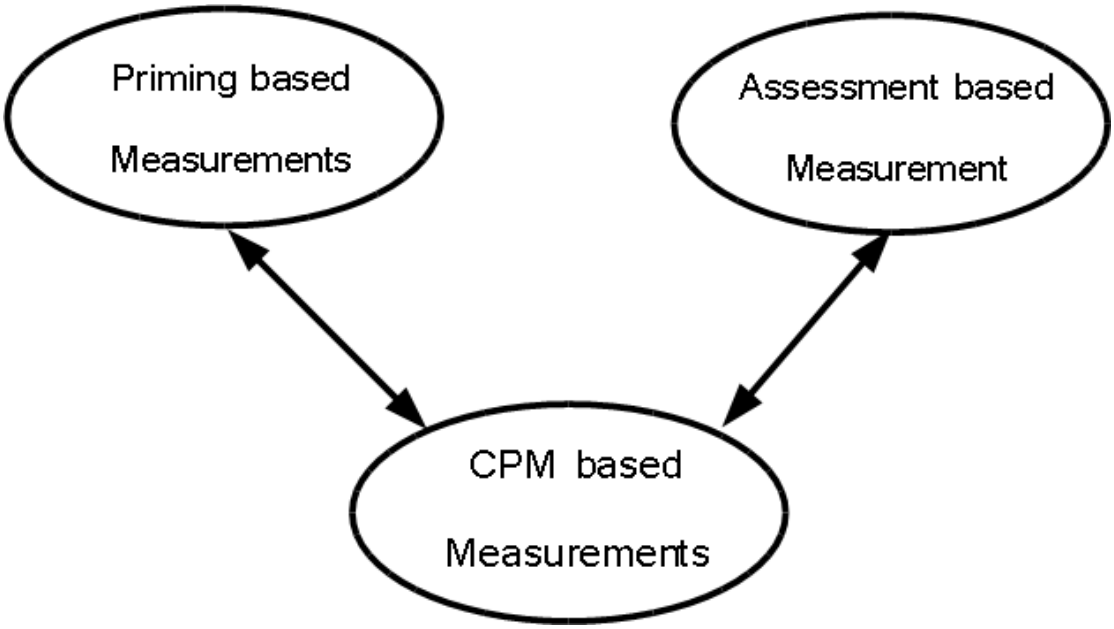


FIGURE 6.3: Possible Validations on Base of Available Measurements

is limited to the illustrated possible alignments of instruments as listed below:

- Computational Model – Assessment Procedure alignment
- Computational Model – Priming Procedure alignment

6.4.5 Validation Strategy of Type of Word Similarity

The validation strategy concerning type of word similarity measurement instruments is in large parts identical to the outlined strategy with regard to graded similarity. The structuring of the reported results in chapter 8 accordingly follows this theme.

In the following section an overview of the shared experimental settings for both respective validation studies is provided.

6.5 Experimental Setup

Subsequently the common elements of the experimental setup of the validation of instruments of graded word similarity and type of similarity measurement are outlined.

6.5.1 Computational Models

There exists a large variety of computational word similarity models. As mentioned before, within the scope of this work the following two models find application:

- Hyperspace Analogue to Language Model (HAL) developed by Lund and Burgess (1996).
- Latent Semantic Analysis Model (LSA) developed by Deerwester et al. (1990).

The motivation for the use of these specific models is based on the following. Both models have been thoroughly established within the research community and integrated in a large array of applications and contexts. As such, they offer a large amount of prior work that can be used to place made observations in context. Secondly the choice for these models was directed by the specifics of the underlying algorithms that are described in the following two paragraphs.

LSA algorithm

Measuring the similarity between words in LSA is based on the following basic algorithm. On basis of a corpus Z consisting of a number of n documents the first step consists of the construction of a co-occurrence matrix. In its initially proposed form this co-occurrence matrix is presented by a $n \times v$ matrix $C = [c_{ij}]$ where v is the number of words, n is the number of documents in the corpus, and c_{ij} is the frequency of the i-th word in the j-th document. Without at this point exploring the specific details it should be noted that the frequency c_{ij} can be weighted on basis of a formula modelling the

global (i.e. corpus-wide) or local (i.e. document-wide) importance. The final preparatory step consists of the application of Singular Value Decomposition in order to reduce the dimensionality of the original co-occurrence matrix. Words are represented in the resulting representation by a $v \times k$ matrix where v is the number of distinct words and k refers to the chosen number of eigenvectors. Without exploring the mathematical aspects of the operation in detail it should be remarked that the effect of this operation is identified to be the following. The resulting word vectors contain components ordered from most to least amount of variation accounted for in the original data. Retaining only the k eigenvectors that account for the most variation of the original data can therefore be seen as a way of reducing noise in the word vectors. In the original publication, LSA was described by Deerwester et al. (1990) as a solution to the 'synonymy problem'.

HAL algorithm

The original HAL algorithm is described by Lund and Burgess (1996, p. 204) in the following form.

> *In this procedure, a 'window' representing a span of words, is passed over the corpus being analyzed. ... By moving this window over the source corpus in one-word increments and recording, at every window movement, the co-occurrence values of the words within it, a co-occurrence matrix can be formed. This matrix has, as axes, the entire vocabulary under consideration, such that each cell of the matrix represents the summed co-occurrence counts for a single word pair.*

In comparison with the description of the LSA algorithm the following observations can be made. Both algorithms are essentially based on the analysis of co-occurrence patterns. A principal difference is given by the focus of the observations. LSA in its original form is based on analyzing those patterns within the scope of a document, while HAL restricts the analysis to the neighbouring terms on basis of using a variable window size.

6.5.2 Collections

As noted in section 6.4.3 the collections were chosen with the aim of ensuring variability with regard to the following attributes:

- **Size of collection**: The size of the collection in terms of the number of documents, the total number of words, and the number of unique words.

- **Quality of content**: Is the contained information of curated high quality, or uncontrolled low quality.
- **Topical focus of content**: I.e. what kind of information is contained in the collection.

The utilized collections and associated attributes are reported in table 6.7. As noted

Collection Name	Document Type	# of items	Reference
Genomics 2004	Scientific Abstracts	4,59m	Hersh et al. (2004)
UKWAC	Web Pages	2,69m	Baroni et al. (2009)
Wikipedia	Encyclopaedia Articles	3,08m	Wikipedia Foundation (2009)
Acquaint	News Paper Articles	1,03m	Voorhees (2006)
TREC 4,5	News Paper Articles	0,53m	Voorhees (2005)

TABLE 6.7: Utilized Test Collections

before, within experimentation of word similarity models it is custom to reduce the amount of considered terms by excluding very frequent and very infrequent terms. In order to enable analysis with regard to the sensitivity of the results concerning Document Term (df) frequency thresholds, as well as additionally introducing variation to the conditions, for each of the above listed collections df dependent subsets were generated. An overview of these representations is given in table 6.8. The df values were arbitrarily chosen for the Wikipedia collection on grounds of preliminary analysis and the reported consensus in prior research. The corresponding values for the other collections are chosen to reflect the Wikipedia values in terms of the df/n ratio. Where n refers to the number of documents in the collection.

In regards of necessary computational effort stop wording and stemming using the Porter stemming algorithm (Porter, 1980) was applied to all test collections.

On basis of this overview of the experimental setup the next chapter reports on the results of the validation study of measurement instruments associated with the grade of word similarity.

Collection Name	df Threshold	# of unique terms
Aquaint	9	119542
Aquaint	27	66855
Aquaint	107	34365
Aquaint	426	17229
Genomics 2004	39	148019
Genomics 2004	118	68045
Genomics 2004	474	26194
Genomics 2004	1894	11352
UKWAC	23	184071
UKWAC	69	91336
UKWAC	277	40499
UKWAC	555	27872
Wikipedia	25	173152
Wikipedia	75	82950
Wikipedia	150	53040
Wikipedia	300	34377

TABLE 6.8: Listing of df dependent collection representations

CHAPTER 7

MEASUREMENT OF GRADED WORD SIMILARITY

Chapter 5 outlined the necessity for validated measurement instruments with regard to the experimental realisation of the paradigm. With reference to this it was further outlined that word similarity assessment based procedures constitute instruments that suffice the assumed requirements of sensitivity. However as illustrated in Section 6.3.1 the question of the validity of these instruments remains unanswered.

In light of this the focus of this chapter consists of reporting on a validation study of such procedures. This is addressed by RQ 4 of the dissertation.

RQ 4 What are valid instruments for the measurement of the grade of relatedness between words?

Section 6.3 provided an overview of the general considerations underlying validation and provided a detailed overview of the underlying strategy developed with regard of the specific context of this work.

Based on these considerations, the structure of the chapter is as follows. On basis of these considerations Section 7.1 outlines the application of these general considerations to the task of evaluating the validity of word similarity test. Following this outline section 7.2 reports on the details of the specific experimental setup. The presentation and discussion of the results of the performed study on basis of the described setup is presented in two distinct sections, each focusing on one of the two utilized computational models. Section 7.3 reports on the validation results obtained on basis of correlating LSA based measurements. The following section analogously reports on results based on HAL measurements. Section 7.5 finally provides an overview and a discussion of the reported results.

7.1 Experimental Outline

This section outlines specific considerations that apply to an evaluation of the validity of word similarity evaluation procedures. As outlined before, the fundamental principle underlying evaluation of a measurement instrument's validity consists of the alignment of its measurement output with the output of different kinds of instruments aimed at measuring the same underlying phenomenon. This principle forms the foundation for the subsequently introduced considerations.

Regarding the task of testing the validity of word similarity assessment procedures the available kinds of measurement instruments are given by the following:

- Computational Models of Word Similarity (i.e. LSA, and HAL)
- Priming Experimentation

A test of the validity of word similarity assessment procedures is limited to the correlation of the output (i.e. measurements) of the above listed two sources with the output of specific implementations of assessment procedures. As noted before, due to the fact that (a) measurements are 'tied' to specific word pairs, and (b) the word pairs of available studies conducted on basis of assessment and priming based procedures are distinct, it is not possible to directly correlate measurements of both sources. A way of mitigating this limitation consists of correlating the measurements of both sources with the measurements of the computational models. To illustrate this aspect Figure 7.1 provides an overview of the resulting situation. On basis of the figure the situation can be described in the following form.

There exist three measurements instruments A, B, and C. The task consists of validating these instruments via alignment of their output. As indicated in the figure it is not possible to directly correlate the measurements of instruments B and C. The proposed approach with regard to that limitation consists of three steps. Step ❶ and ❷ as outlined in the figure consist of correlating the measurements of instrument A with the measurements of instrument B and C respectively. As symbolized in the figure step ❸ then consists of drawing inferences on basis of these correlations. To illustrate the nature of these inferences the following can be considered. Given the case that all three instruments represent perfectly valid instruments of measurement the obtained correlation coefficients between the measurements of $A-B$ and $A-C$ naturally would be 1. In that case it logically follows that, although we cannot directly correlate the measurements of B and C, their measurements must also be correlated perfectly. In reality A does not constitute a perfectly valid measurement instrument. In fact, with regard to the nature of the instrument, the following conclusion can be made concerning the validity of A. Computational models constitute instruments of measurement of word similarity whose validity varies with regard to their underlying parameter settings. If B and C

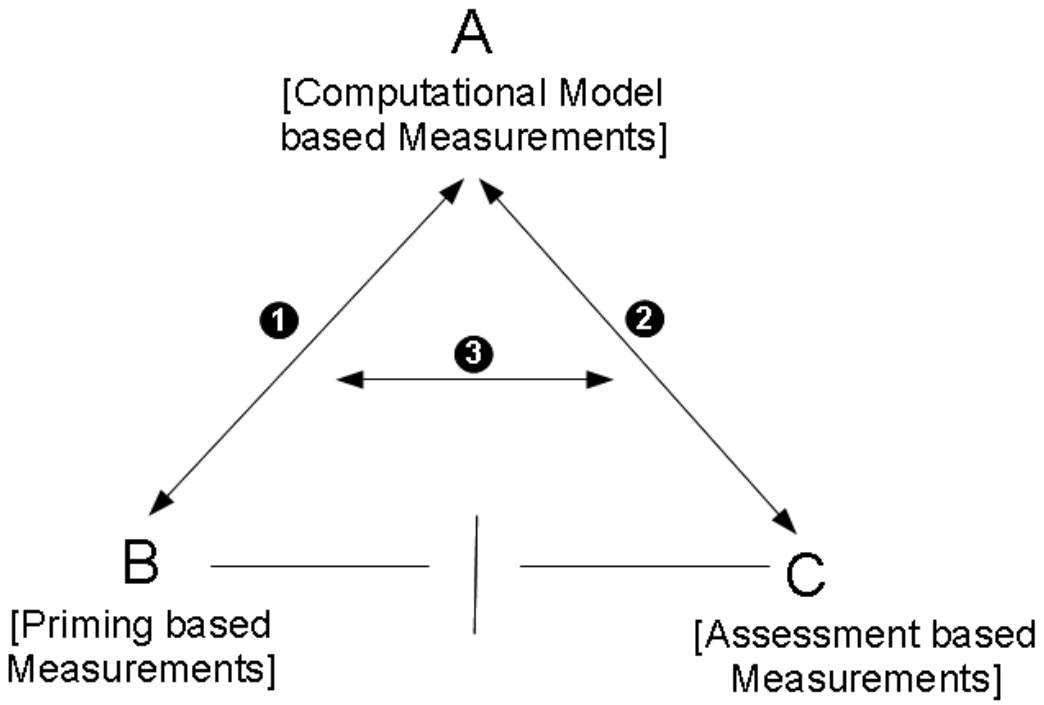


FIGURE 7.1: Validation Strategy - Grades of Word Similarity

both constitute valid instruments this varying degree of validity of A should be equally reflected in the correlations of $A-B$ and $A-C$. Of importance with regard to these considerations is that B on basis of the existent body of evidence (see section 6.3.1) is interpreted to constitute the most valid of the considered instruments. On grounds of this it is therefore concluded that the degree of correlation of $A-B$ and $A-C$ allows to draw inferences with regard to the validity of instrument C representing assessment based procedures of measuring word similarity.

The above described strategy constitutes the basis for the conducted analysis of the validity of assessment based procedures. Consequently this is also reflected within the structure of the chapter. The reporting of the results is provided analogous to the below outlined three steps:

1. Correlation of Priming Based Measurements with Computational Model based measurements.

2. Correlation of Assessment Based Measurements with Computational Model Based Measurements

3. Analysis concerning the validity of Assessment based procedures on grounds of the observations of step 1 and 2.

As noted in the introduction the presentation of the results is provided in form of two distinct sections focused on each of the two computational models.

7.2 Experimental Setup

This section describes the experimental setup concerning items that are specific to the validation of measurement instruments targeted at graded word similarity. With regard to the underlying data and estimator space the setup is consistent with the specifications provided in section 6.5.

7.2.1 Measurements

As stated before the validation effort is based on the below listed four types of measurements.

- HAL based measurements
- LSA based measurements
- Assessment based measurements
- Priming exp. based measurements

The interpretation of HAL and LSA as measurement instruments is based on the discussion provided in Section 6.4.1. Details concerning the respective underlying algorithm have been listed in Section 6.5.1. Subsequently an overview of the utilized assessment and priming based data is provided.

7.2.2 Assessment Based Measurements

Table 7.1 provides an overview of the assessment based data sets that find application in the reported validation.

Name	Abbreviation	# of Pairs	Reference
Finkelstein353	FS353	353	Finkelstein et al. (2002)
Miller & Charles	M&C	30	Miller and Charles (1991)
Rubenst. & Gooden.	R&G	65	Rubenstein and Goodenough (1965)

TABLE 7.1: Assessment based word similarity data sets used in the validation of measurement instruments of graded word similarity

All listed datasets are based on the assessment methodology introduced by Rubenstein and Goodenough (1965) described in detail in Section 6.2.1. The underlying methodology on which basis these measurements of similarity were collected is similar. The distinctive difference regarding the actual procedures underlying these datasets stems

from the sets of chosen word pairs that were used to measure word similarity. With regard to that aspect the three procedures exhibit the following relations. The word pairs of the Miller and Charles (1991) dataset constitute a subset of the Rubenstein and Goodenough (1965) word pairs. The much larger Finkelstein et al. (2002) dataset represents a superset of the Rubenstein & Goodenough word pairs. As shown in Table 7.1 all three datasets focus on a relatively limited set of words. Further the original studies by Rubenstein and Goodenough (1965) and Finkelstein et al. (2002) do not provide details concerning the criteria underlying the selection of the contained word pairs. The small number of word pairs and the lack of control of variables such as word frequency form potential sources of bias. These characteristics constitute motivating arguments for their empirical validation. Regarding the availability of larger data sets an initial study by Snow et al. (2008) showed promising results based on using crowd-sourcing to replicate the task of Miller and Charles (1991).

7.2.3 Priming Experimentation Based Measurements

Table 7.2 lists the priming based datasets used as part of the validation.

Name	Abbreviation	# of Word Pairs	Reference
Vigliocco Object Priming	VigObj	4x32	Vigliocco et al. (2004)
Vigliocco Action Priming	VigAct	4x32	Vigliocco et al. (2004)

TABLE 7.2: Priming based datasets used for the validation of measurement instruments of graded word similarity

As can be seen in Table 7.2, each of the two datasets consists of four sets of 32 word pairs. The pairs in each set have been analytically chosen to reflect one of the below listed grades of relation.

- Very closely related (e.g. dagger-sword).
- Closely related (e.g. dagger-razor).
- Moderately related (e.g. dagger-hammer).
- Non-related (e.g. dagger-tongue).

The psycho-linguistic measures associated with these pairs consist of the measured mean choice-reaction times (CRT) obtained on basis of a lexical decision task experiment. For a detailed description of these sets and the associated experimentation see Section 6.2.1.

7.2.4 Correlational Analysis

Rank Correlation Coefficients: Two rank correlation co-efficients find usage in the correlational analysis.

- Spearman rank correlation coefficient (also referred to as Spearman's ρ)
- Kendall rank correlation coefficient (also referred to as Kendall's τ).

The two tests constitute non-parametric tests for statistical dependence between continuous variables. Since both tests are based on the same assumptions (Kendall, 1938) the choice is guided by the availability of implementations in the used software packages. All computations underlying the correlational analysis are based on the use of the cor and cor.test functions of the stats package in R (2012). RStudio (2012) was used for the development of the R code. Calculation of p-values for reported correlation-coefficients is based on the implementation of Fisher's Z transform in the cor.test package.

Correlation Matrix: The calculation of the correlation matrices used throughout the empirical analysis is based on the pairs package of R (2012). In the correlation matrices the correlation coefficients are shown in the upper panel above the diagonal of the matrix. The calculated p-values are indicated through the use of * symbols referenced in the caption of the plots. The lower panels show scatter-plots of the correlated variables that allow for the visual verification of a monotonic or non-monotonic relationship between the tested variables. Scatter plot smoothing is applied based on the LOESS smoother, a non-parametric locally-weighted polynomial regression method. The purpose of the fitted line consists of supporting the visual analysis of monotonicity and effect size.

7.3 LSA Based Alignment

This section provides an overview of the LSA based analysis concerning the validity of assessment based procedures. The structure of the section follows the above outlined strategy. With regard to considerations of the robustness and scope of these observations the analysis is conducted over the data and condition spaces described in Section 6.4.2 and 6.4.3. As shown in Figure 6.1 the parameter space of LSA therefore resolves to the following:

- **Dimensionality:** $(25, 50, 150, 300, 400, 500)$
- **Transformation Function:** $(\text{LogEntropyTransform}, \text{NoTransform}, \text{TfIdfTransform})$

Inferences concerning the correlation of LSA based measurements with assessment and priming based measurements are based on calculation of correlation coefficients and visual inspection of graphic representations of the underlying data. Subsequently results

obtained from correlating LSA based measures with assessment based data on base of the specified setup are reported.

7.3.1 Word Similarity Assessment based Alignment

To enable the correlation of LSA based measures with human assessments of word similarity the following approach is followed. The underlying human assessment data consists of a set of n triples consisting of $(word_1, word_2, assessedSimilarity)$. To enable the correlation with computational model based measurements, the similarity of $word_1$ and $word_2$ is determined by calculating the *cosine* between the respective LSA based word vectors. The result of this operation consists of n triples of the form $(word_1, word_2, LSASimilarity)$ over the above outlined parameter space.

Interpretation of Correlation coefficients

Figure 7.2 depicts the Spearman correlation coefficients obtained by correlating the similarity ratings of the three word similarity assessment sets with the respective LSA measurements. The underlying purpose of the figure consists of illustrating the implementation of the validation strategy on basis of the LSA parameter space, and of highlighting the relation to the commonly reported form of using assessment based data for model evaluation.

In each of the three depicted graphs the correlation coefficients between the LSA based data and the three assessment based data sets are plotted with respect to the underlying dimensions of the LSA models. Each graph shows the coefficients with respect to one of the three transformation functions. To illustrate this point: In the top left graph the reported correlation co-efficient of 0.84 represents the result of calculating the correlation of the human similarity ratings of the 30 word pairs of the M&C study and the similarity measures for the same word pairs obtained from a LSA model with the parameter settings $(LogEntropy-Transformation, 500)$.

The high positive correlation in this case can be interpreted as an *indication* of validity. The underlying reasoning being, that if the computational model and the assessment procedure measure the same phenomenon, their measurements should be highly correlated. This forms the basis of commonly reported validations (usually referred to as 'evaluations', or 'performance evaluations') of computational models of word similarity. Such evaluations are usually based on the implicit or, as in the case of the following statement by (Budanitsky and Hirst, 2006, p. 17), explicit assumption of the validity of assessment procedures.

> “ *Insofar as human judgements of similarity and relatedness are deemed*

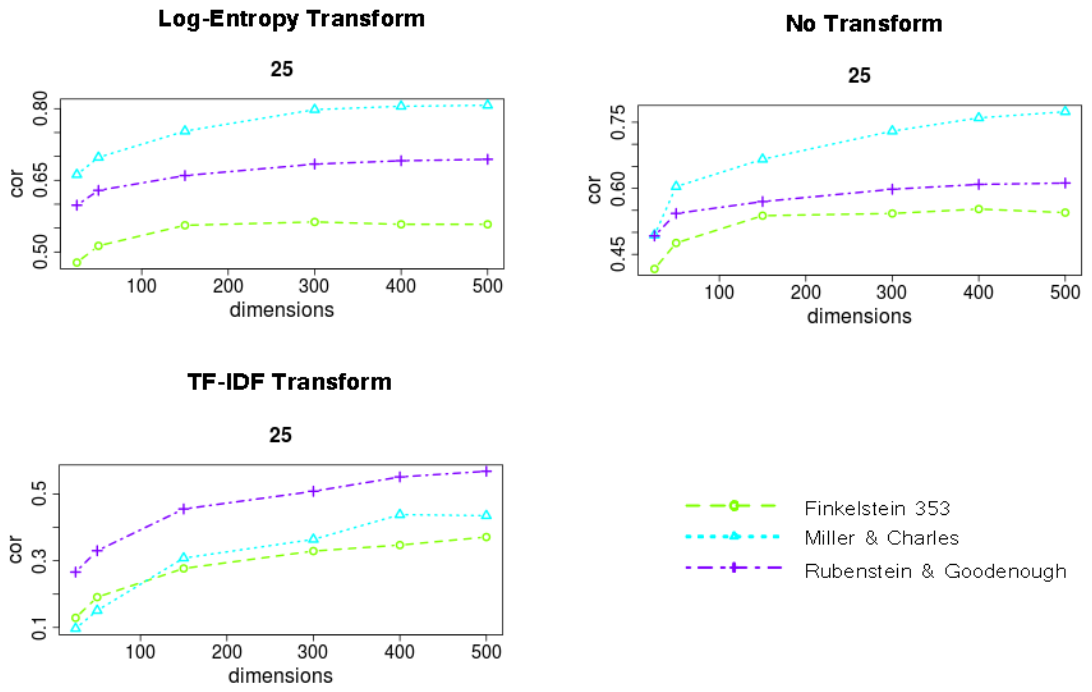


FIGURE 7.2: Correlation Coefficients between FS353, M&C, and R&G and LSA Models Based on Different Transformation Functions on Grounds of the Wikipedia:DF25 collection

> *to be correct by definition, this clearly gives the best assessment of the 'goodness' of a measure.* ”

Examples of such evaluations are given by Cramer (2008), Durda and Buchanan (2008), Gabrilovich and Markovitch (2007), Recchia and Jones (2009), Baroni and Lenci (2010). Disregarding concerns of the truthfulness of this implicit assumption it is obvious that the value of drawn inferences with respect to the validity of a computational model based on the calculation of a single correlation co-efficient is limited. This limitation is primarily induced by limiting the scope of the observation to a single parameter setting of the model in combination with the small number of considered term pairs (this observation specifically applies to the R&G and M&C datasets). The common practice of 'training' a computational model and reporting the '*best*' performance on grounds of correlating the model's output with such datasets therefore appears limited in terms of the ability to draw meaningful inferences. As a consequence, as exhibited in the case of the above cited studies, the evaluation of computational models of word similarity is usually based on the application of a large set of varying tests (see specifically Baroni and Lenci (2010) as an example)

As outlined in Section 7.1, the basic strategy for the evaluation of the validity of assessment based procedures consists of the correlation of their measurements with LSA based measurements and the subsequent alignment of the resulting coefficients with the

respective priming based coefficients. With respect to the figure such a potential alignment can be conducted on basis of one of the dataset specific lines (e.g. the light blue line representing the M&C dataset based coefficients) with its priming based equivalent. With regard to these considerations another aspect that can be remarked on grounds of the graphs in the figure concerns procedure specific analysis. As stated before the three procedures share the same principle methodology but differ in terms of the used word pairs. It is therefore conceivable that even if the underlying procedure should constitute a valid instrument of measuring word similarity, one of the three procedures might not constitute a valid instrument on basis of its chosen set of pairs. Further it is conceivable that through the choice of word-pairs such procedures could exhibit bias with regard to specific types of similarity. With respect to these concerns the observation that the lines representing the three datasets follow very similar trends across all three graphs can be interpreted as support for the validity of the three procedures. If the three datasets all represent valid measurements, the varying degree of validity of LSA with respect to the dimension parameter should be equally reflected in its degree of correlation with the model based measurements. With regard to the figure it can be stated that this *seems* to be the case. Such visual inspection of course only constitutes anecdotal evidence of validity. More profound evidence can be obtained via the actual calculation of the correlation between the data-set specific coefficients.

Estimator Rank Based Correlation

In the prior subsection it was noted that the similar trends exhibited by the plotted correlation coefficients can be interpreted as an indicator of the validity of the three word similarity assessment procedures. A viable way to add substance to this observation consists of calculating the correlation coefficients between the dataset specific coefficients. Figure 7.3 shows a correlogram providing an overview of the linear relationships between the dataset specific correlation coefficients.

The depicted 'correlations of correlations' can be interpreted as follows. FS353, M&C, R&G, and LSA are instruments for measuring word relatedness (as defined in the validation strategy shown in Figure 7.1). Significant positive correlation between LSA and a word similarity test represents supportive evidence for the validity of LSA as a measurement instrument. LSA is an instrument for the measurement of word similarity whose validity varies according to its underlying parameter settings. With regard to the mode of operation of SVD it can be assumed that a reduction to 25 dimensions, on grounds of retaining too few eigenvectors (or more specifically by disregarding too many eigenvectors that represent a large part of the variation in the original data), is likely to result in a model whose measurements are of low validity. Consequently if the procedures underlying the three datasets are valid, the parameter-dependent variation of the validity of LSA should be equally reflected by the correlation coefficients with the

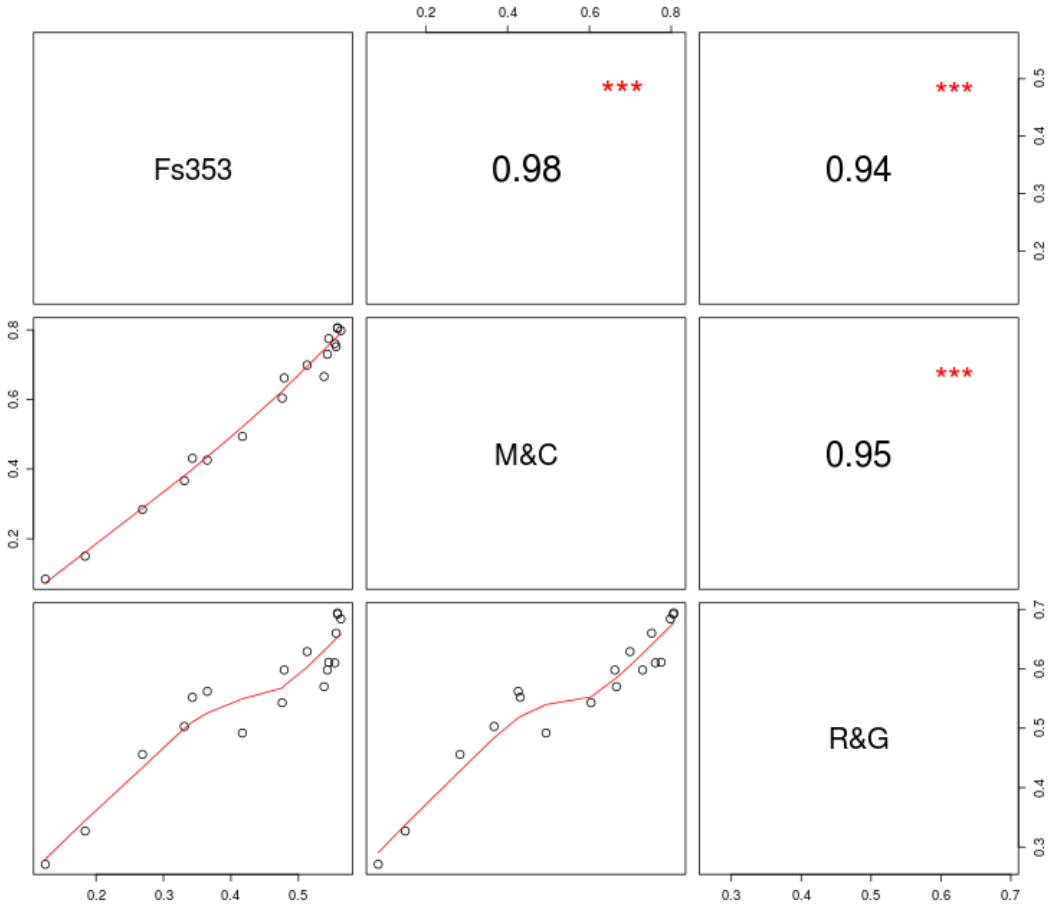


FIGURE 7.3: Correlogram of Kendall's τ between Rubenstein & Goodenough, Miller & Charles and Finkelstein 353 Word Similarity Test Coefficients on the Wikipedia:DF25 collection. $(***; p < 0.0001)$

three word similarity datasets. As can be seen in the correlogram, the reported coefficients and the plots of the correlations indicate that this is the case. This is interpreted as supportive evidence for the validity of the three procedures.

So far the presented results have been limited to LSA models that were based on the Wikipedia corpus with a df threshold of 25. On grounds of the design of the data-space underlying this study it is possible to apply this test of procedure specific validity over different df thresholds.

Correlation over df aspect

Table 7.3 shows the correlation coefficients obtained by correlating the dataset specific coefficients over the available df thresholds of the Wikipedia collection.

	Fs353	M&C	R&G	dfThreshold
Fs353	1	0.991***	0.931***	25
M&C	0.991***	1	0.942***	25
R&G	0.931***	0.942***	1	25
Fs353	1	0.992***	0.931***	75
M&C	0.992***	1	0.943***	75
R&G	0.931***	0.943***	1	75
Fs353	1	0.992***	0.926***	150
M&C	0.992***	1	0.94***	150
R&G	0.926***	0.94***	1	150
Fs353	1	0.993***	0.94***	300
M&C	0.993***	1	0.953***	300
R&G	0.94***	0.953***	1	300

TABLE 7.3: Correlation of data-set specific coefficients over all df thresholds based representations of the Wikipedia collection.($***; p < 0.0001$)

As can be deducted from Table 7.3, the correlation across all df thresholds is uniformly very high and positive. The parameter-dependent varying 'quality' of the measurements of LSA models is therefore almost equally reflected in the data-set specific coefficients resulting from the correlation of their similarity ratings and the LSA measurements.

This observation can be interpreted as a first piece of supportive evidence with regard to the validity of assessments based procedures as instruments of measuring word similarity. The underlying reasoning concerning this statement is as follows. On grounds of the existent body of reported research on LSA, specifically the body of conducted studies evaluating the performance on a variety of different application specific tasks,

it can be inferred that the degree of validity of LSA based models is roughly proportional to the dimension parameter[7-1]. It has been stated before, that application specific evaluations do not constitute ideal instruments of measuring word similarity due to the difficulties of isolating word-similarity effects. However such 'evaluations' of LSA models still offer some implicit evidence concerning the dimension-parameter specific validity of such models. This statement is based on the conclusion that application performance is to a degree dependent on the LSA based measurements and therefore also to a degree on the model's validity as a measurement instrument of word similarity. The observed coefficients for all three data-sets therefore can be interpreted to be 'in line' with the above described general consensus. The supportive evidence therefore in this case stems from the plausibility of the observed results with respect to the underlying nomological network (in this case the reported results of the mentioned studies).

To broaden the scope of these observations and add further substance to them, the next subsection explores the correlation between data-set specific coefficients over the available collections.

Correlation over Collection Aspect

Figure 7.4 provides an overview of the correlation of the dataset specific coefficients across the four different collections.

The four plots in Figure 7.4 show the correlations between the three data-set specific coefficients with respect to the four collections. The correlations are plotted with respect to the collection specific df thresholds (see Table 6.8 for details).

As can be seen in Figure 7.4, the correlations between the datasets are generally very high and positive for the Wikipedia and Acquaint collections. A similar picture is given for the UKWAC collection. Correlation of M&C-R&G and FS353-M&C is very high. Correlation of FS353-R&G is considerably lower but still moderate to high across the df thresholds. An exception to this is provided by the correlations shown for the genomics collection. As can be seen the correlation between the datasets is much lower for the first df threshold level. In contrast to the trends exhibited in the plots for the other three collections the correlation generally degrades with increasing df threshold. A potential explanation for this diversion can be based on the nature and characteristics of the genomics collection. The genomics collection is comprised of abstracts of scientific articles from the area of genomics. As such the documents in the collection are (i) relatively short and (ii) exhibit a mainly technical vocabulary. Point (i) might be interpreted to lead to generally lower and more 'unstable' performance of the underlying LSA models due to LSA's document-wide co-occurrence scope. Aspect (ii) might

[7-1] This statement only applies within the confines of the underlying dimension parameter space applied in this study. It has been shown that higher dimension settings can result in lower validity

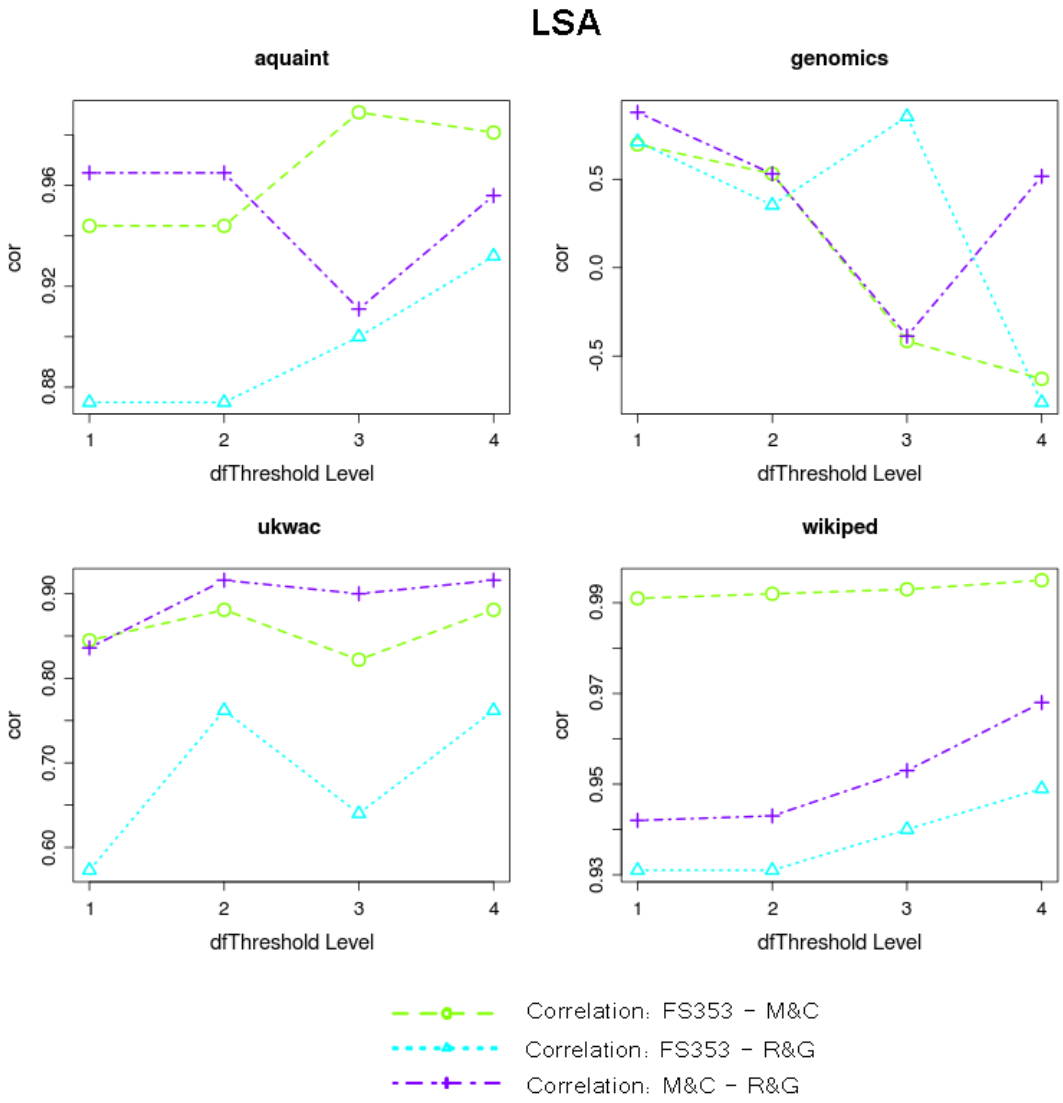


FIGURE 7.4: Correlation of assessment based data-sets over the four utilized collections with respect to collection specific df threshold levels.

result in varying degrees of correlation for each of the specific datasets on basis of the fact that the word pairs in the datasets stem from what can be referred to as 'commonly' used language (e.g. 'sage-wizard', 'food-fruit'). It is intuitive that these commonly used words are less strongly reflected in the technically oriented genomics collection. Further it can be assumed that this lower representation in the collection does not apply uniformly to all the pairs in the three datasets. Therefore it seems plausible to assume that the data-set specific correlation with the LSA measures varies depending on the reflection of the respective word pairs in the genomics collection, and that therefore the correlation between the dataset-specific coefficients is much lower or even negative[7-2].

On basis of the correlations over collections it can stated that the observed high correlation of the three datasets across three collections is interpreted as further supportive evidence with regard to the validity of the specific procedures and of assessment as a measurement instrument in general. The diversion regarding the observed correlations in case of the genomics collection is , on basis of the above presented arguments, interpreted as a constraint of the validity of the LSA computational model. Specifically that the word pair specific validity of LSA as a measurement instrument of word similarity is dependent on sufficient representation of both words in the underlying collection.

The subsequent section explores the alignment of priming based and LSA based measurements.

7.3.2 Priming Based Alignment

As noted in the experimental setup priming based measurements are only available in form of the reported mean reaction times collapsed over the set of word pairs. This limits the possibilities with regard to correlation based analysis. As a consequence the subsequent analysis is mainly based on verification on grounds of visual inspection.

Considerations Regarding Priming Based Measurements

As stated in the experimental setup the utilized priming based data stems from a study reported by Vigliocco et al. (2004).

[7-2]This assumption has not been verified. A viable way to do so consists of analysing the relation between data-set specific correlation with the LSA measures and frequency of occurrence of the dataset specific vocabulary

Interpretation of Priming Results: As part of the introduction of her Featural and Unitary Space (FUSS) model Vigliocco conducted a series of priming experiments aimed at evaluating the correlation of FUSS measurements with human reaction times. Two of those conducted experiments focused on graded relationships between words. The underlying methodology can be summarized in the following form. Four sets of word pairs were created. Each set was comprised of 32 word pairs that exhibited a set-specific grade of relation (e.g. 'very closely related', 'closely related', see Section 7.2.3). To illustrate this, Table 7.4 shows a sample of the used word pairs.

Target	Prime	Prime	Prime	Prime
	Very close	Close	Medium	Far
apple	peach	lemon	bean	raft
camel	zebra	mouse	swan	sofa
shoulder	arm	leg	thumb	bus
fence	gate	wall	roof	bus

TABLE 7.4: Sample of Word Pairs Used in the Vigliocco et al. (2004) Object Based Graded Priming Experiments

Subsequently the mean reaction times for these pairs were recorded by evaluating these sets within a lexical decision type priming experiment (see Section 6.2.1 for a description).

Table 7.5 provides an overview of the obtained results. Consistent with the theories underlying word based priming effects the mean reaction times are inversely proportional to the grade of relation. The reaction times are shorter the closer two words are related to each other. The following section provides details of the applied methodology to

Semantic Distance	Response latencies (ms)	Error rate (%)	Δ (ms)
Very Close	548[8.1]	1.8	na
Close	557[9.3]	1.5	9
Medium	567[9.1]	1.8	10
Far	572[9.8]	1.3	5

TABLE 7.5: Mean Lexical Decision Times for Objects Reported by Vigliocco et al. (2004)

relate LSA based measurements with the above listed results.

Computational Model Based Priming Simulation: In order to allow for a meaningful comparison of the priming based means with LSA measurements it is necessary to simulate the priming task on basis of the LSA model. The applied method is fundamentally similar to those applied by Jones et al. (2006) and Vigliocco et al. (2004). With regard to the fact that LSA models do not contain non-meaningful words such as the non-words used in lexical decision task it is necessary to approximate this condition.

To illustrate this aspect:
In a priming based study the priming effect e in ms is usually calculated as
$e = rt_{n-w}[non-word, target] - rt_p[prime, target]$
where rt represents the measured reaction time.
In the cited results of Vigliocco et al. (2004) the reported means refer to the $rt_p[prime, target]$ response times.

On basis of the informal descriptions provided by Jones et al. (2006) and Vigliocco et al. (2004) it is deducted that LSA based simulated priming effects sp were calculated as:
$sp = \Delta_p[\cos(prime, target)] - \Delta_{rw}[\cos(randomWord, target)]$
Where the use of cos represents the calculation of the cosine between the vectors of the words in parentheses. *randomWord* represents a randomly chosen word from the vocabulary of the underlying collection. Since a larger cosine is interpreted as an indicator of closer relationship between two words, a larger sp can be interpreted as representing a larger priming effect. On basis of the analysis of preliminary results it was decided to apply a slightly altered version of the above priming simulation methodology within this work.

$$sp_n = \Delta_p[\cos(prime, target)] - \frac{\sum_{i=0}^{n} \Delta_{rw}[\cos(randomWord_n, target)]}{n}$$

As can be seen in the formula, instead of calculating sp on basis of subtracting the *cosine* based on a single random word, sp_n is calculated by subtracting the average cosine on basis of n random words. The motivation for this approach is grounded in the relatively small size of the utilized word pair sets. The *randomWord* is supposed to exhibit absolutely no relation to the *target*. In regard to this it is intuitive that the choice of a *randomWord* that is related to the *target* has considerable effect on the derived means if only a small number of word pairs (e.g. 32) is considered. With regard to the assumption that $p(chooseRelated) < p(chooseUnRelated)$ it is therefore concluded that larger n mitigate this potential bias. Within the course of this study n is chosen as 100.

Further specifically with regard to the underlying characteristics of the LSA algorithm a second alteration of the methodology is performed. Due to the underlying properties of the algorithm the mean cosine between vectors is inversely proportional to the number of dimensions. That means, LSA models that retain fewer dimensions generally exhibit larger cosines. An illustration of this is given by the topmost plots in Figure 7.5. The plot shows that the *meanRT* and *meanPT* values with regard to the dimensions parameter of the underlying LSA model. *meanRT* represents the mean of n $\cos(randomWord_n, target)$ values, and *meanPT* the mean of the $\cos(prime, target)$ values . As can be seen in the plot, the values for both means for the dimension parameter of 25 are considerably higher. Of note is that this represents a characteristic of the model and is not, on basis of considering (randomWord,target) pairs, a result of an actual relation of the considered words. With regard to the definition of sp_n provided above this means, that the magnitude of the simulated priming effect is on base of this

overestimated for lower dimensions.

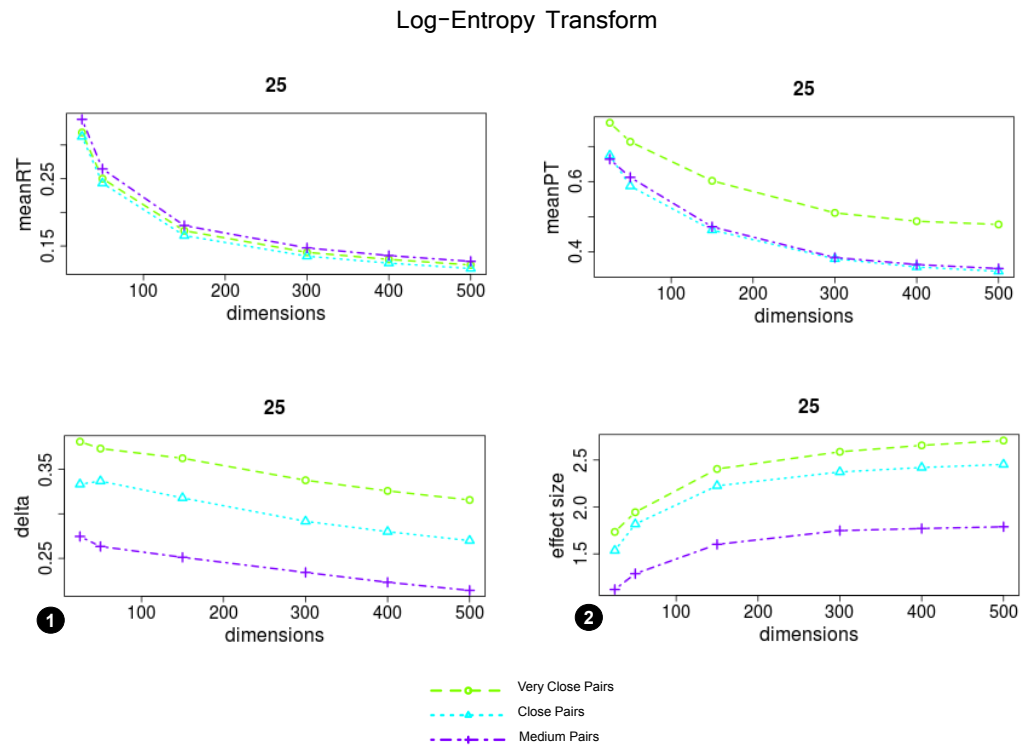


FIGURE 7.5: Application of Cohen's d Illustrated on Basis of Plots Depicting Simulated Priming Values for Vigliocco's Object Pairs on Wikipedia collection

To mitigate this effect it is therefore proposed to apply some form of dimension dependent standardization of the observed effect. A commonly applied technique for the standardization of observed effects consists of computing the effect size. A widely applied method to do so is provided by Cohen's d (Cohen, 1988) defined as

$$d = \frac{\bar{x}_1 - \bar{x}_2}{s}$$

where $\bar{x}_1$ represents the mean of the $(prime, target)$ cosines, $\bar{x}_2$ the means of the $(randomWord, target)$ cosines and s is given by the pooled standard deviation defined as

$$s = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2}}$$

where s_1 and s_2 in the current context represent the standard deviations of the $(prime, target)$ and $(randomWord, target)$ based cosines. The resulting effect size d therefore represents the observed effect in terms of standard deviations.

To illustrate this, plot ❶ in the figure shows the unaltered sp_n values (marked as 'delta'

on the y-axis) over dimensions. As is evident on basis of the plotted lines, this depiction might induce the conclusion that the observed priming effect is largest for the dimensions parameter 25. Plot ❷ shows the picture after calculation of Cohen's d for the respective sp_n values. Within this plot the standardized observed effect is largest for higher dimensions and smallest for the 25 dimensions setting. As can be seen with regard to the curves depicting the priming effects for 'very close' and 'close' pairs the application of this standardization has influence on the distances between the two curves. This however, as is outlined subsequently in detail, is interpreted to be not critical with regard to the interpretation of the priming effects.

On basis of these definitions the next paragraph explores the inferences that can be conducted on basis of relating priming simulation based, priming experiment based, and assessment based measurements.

Priming Based Measurements and Assessment Based Measurements This point is well suited to provide a recap of the so far presented results and conclusions. The so far presented results (see Section 7.3.2) on basis of correlating LSA based measurements with assessment based measurements over the available independent variable and data space are interpreted as supportive evidence for the validity of assessment procedures on grounds of two main aspects. The individual coefficients reported for each of the three available assessment sets represented plausible values with regard to the specifics of the underlying LSA algorithm and the consensus of other reported evaluations of LSA. Further supportive evidence is given by the high correlation of the three sets across the considered experimental spaces.

On basis of this summary the first question concerning the association of priming experiment and priming simulation based measurements, consists of clarifying which possible observations can be interpreted as further supportive evidence of validity. To answer this question it is first necessary to define the basic analytic outline.

The reported priming experimental results in Table 7.5 constitute, on grounds of attesting validity to the priming experiment procedure, the most valid measurements of graded word similarity. A comparison of the observed means of these priming studies with the simulated priming means on basis of an LSA model therefore allows to draw the following conclusion: The closer the means based on LSA models resemble the presented priming experiment based means, the higher the supportive evidence with regard to the validity of the specific LSA model.

With regard of a 'triangulation'[7-3] of such observations with the earlier reported assessment related results the following cases can be identified.

[7-3]For a detailed discussion concerning evaluation of validity on basis of the concept of triangulation see Jick (1979)

- **Concordance of observations**: If the inferences with regard to the validity of specific LSA models on basis of both measurement types are concordant this can be primarily interpreted as supportive evidence of the validity of the assessment procedure as an instrument of measuring grades of word similarity ('primarily' is used on grounds of the already attested validity of the priming procedure).
- **Discordance of observations**: If the drawn inferences on basis of the two measurement types are discordant this can be interpreted as strong opposing evidence with regard to the validity of the assessment procedure.

With regard to the noted limitations of directly correlating priming experiment based and LSA based measurements the basis for inferences with regard to the above outlined reasoning is restricted to visual inspection. Figure 7.6 provides an overview of such a visual evaluation. The figure is divided into three parts, subsequently referred to in top-bottom order as rows $(1,2,3)$. The plots in all three rows are referring to LSA models based on the Wikipedia:DF25 collection. In row 1 the correlation coefficients resulting from correlation of the LSA measurements with the FS353 and R&G datasets are plotted over the dimensions parameter. Each distinct curve represents transformation function specific coefficients. The measurements of LSA models using Log-Entropy transformation show the highest[7-4] coefficients for both datasets.

Row 2 shows plots of the priming simulation based measurements over the exact same parameter and data space for the Vigliocco object word pair sets. The left plot shows the values based on a Log-Entropy transformation model. The observed simulation based priming effects are in principle consistent with the reported experimental means in Table 7.5. The word pairs in the 'Very closely' related set show the highest priming effect, and the observed effects for the 'Close' pairs and 'Medium' pairs are also generally consistent with the experimental observations. To further illustrate this *row*3 presents plots of the deltas between the three sets. These observations are also consistent and plausible with the observations presented in row 1 on grounds of the following reasoning.
As stated above, the assessment based correlational analysis yielded supportive evidence with regard to the validity of both, assessment based procedures and specific LSA models (depending on parameter settings). A strong correlation as observed in the case of the Log-Entropy based models in row 1 therefore indicates that the measurements of the Log-Entropy model are valid. On basis of this, still considered *assumed*, validity the expected observation for the Log-Entropy specific plot in row 2 consists of close resemblance (or correlation) of the valid priming experiment based observations. The observation that the association of the LSA-Assessment measures *and* the LSA-Priming measures *both induce the same conclusion concerning the validity of the respective LSA models*, is therefore interpreted as supportive evidence for the validity of

[7-4]Log-entropy transformation has also shown to result in the highest performance in the IR focused evaluations of Dumais (1991)

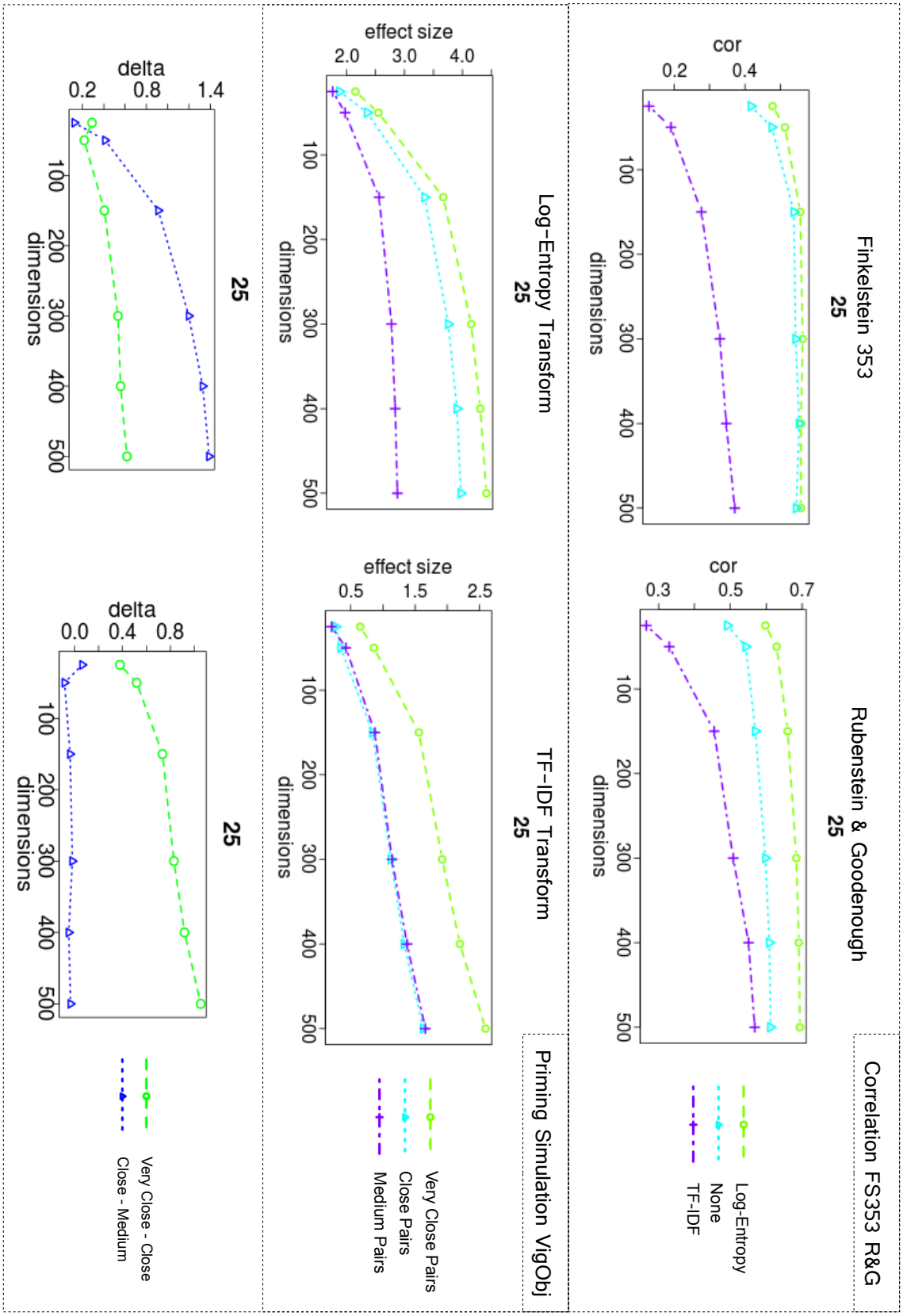


FIGURE 7.6: Priming Simulation Based Analysis: Log-Entropy versus TF-IDF Estimates; Wikipedia:DF25

all three measurement instruments. Further supportive evidence is provided in the form of the right plot in row 2. Here the simulated priming effects on basis of the TF-IDF based LSA models are shown. The observations are not consistent with the experimental priming data due to the missing distinction between the effect size of 'Close' and 'Medium' pairs (evidenced by a delta of close to 0 as shown in *row*3). This observation is plausible with regard to the observations of *row*1 due to the exhibited much lower correlation coefficients of TF-IDF based models. Therefore both, the alignment with priming experiment as wells as the alignment with assessment based measurements, indicate that the TF-IDF measurements are of lower validity.

To add further evidence the correlation between the means reported by Vigliocco et al. (2004) and the priming simulation based means was calculated. The reported coefficients are, on grounds of the few available data points, only of anecdotal character. The results of these calculations are shown in Figure 7.7. As can be seen the picture for the transformation specific curves is consistent with the observations in *row*1 and *row*2. Log-Entropy and No-Transform based means show distinctly higher coefficients than the TF-IDF based means.

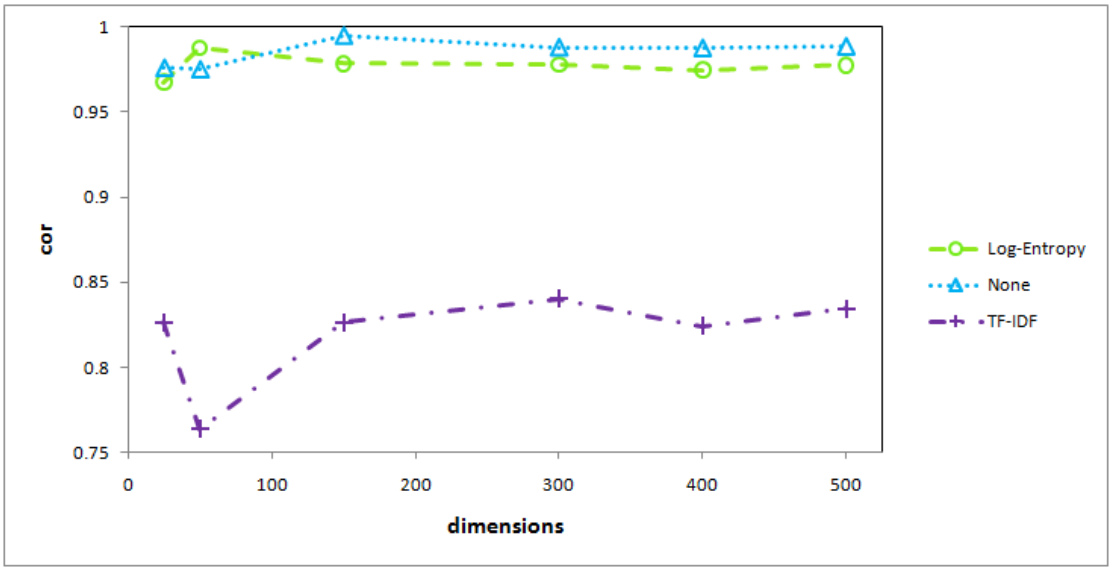


FIGURE 7.7: Correlation of Means Derived from Priming Studies and Means Derived from Simulated Priming on Basis of LSA

7.3.3 Discussion

This section briefly summarizes the so far presented results and conclusions concerning the question of the validity of assessment based procedures.

The presented results in Section 7.2.3 were interpreted as strong supportive evidence for the validity of the procedure as a measurement instrument of word similarity. A

summarization of these results is presented in Table 7.6. As can be seen the mean and median correlation coefficients between the three datasets indicate strong positive correlation across all collections. This induces the conclusion that not only assessment procedures in general, but also each of the three specific procedures constitutes valid instruments. The table also reports the StdDev over all collections and the StdDev excluding the genomics collection. The small size of the latter further emphasizes this observation.

	Mean	Median	StdDev	StdDev w.o. Genomics
FS353-M&C	0.845688	0.9125	0.500654	0.064048041
FS353-R&G	0.797	0.8645	0.423432	0.126585939
M&C-R&G	0.842813	0.916	0.345324	0.038008273

TABLE 7.6: Summary of assessment procedure correlations on basis of LSA models over the collection space

With regard to the large underlying data and condition space it is further concluded that these observations are of a robust nature. Since the conclusion drawn on basis of the alignment with priming based data added further supportive evidence for the validity of assessment based measures, a concluding remark can be stated in the following form.

On grounds of the so far presented evidence a preliminary conclusion that can be drawn is that assessment based procedures constitute valid instruments of measuring word similarity. Consequently they also constitute valid instruments with regard to an evaluation of the validity of LSA models. The subsequent section analogously reports on the application of the so far presented analysis on basis of the HAL computational model.

7.4 HAL Based Alignment

Analogous to the conducted evaluation on basis of LSA models, this section reports on results obtained through alignment of measurement instruments with HAL computational models. Conducting this analysis on basis of HAL is of specific interest since, as outlined in Section 6.5.1, the underlying algorithm substantially differs from LSA.

As was the case with LSA, the reported analysis is conducted over the complete respective parameter space defined in Section 6.4.3. As shown in Figure 6.1, the parameter space of HAL therefore resolves to the following:

- **WindowSize:** $(1,2,3,4,5,6,7,8,9,10)$

- **Weighting Function:** $(EvenWeighting, GeometricWeighting, LinearWeighting)$

The structure of the presentation of the results is slightly altered in comparison to the prior section. While in the LSA section we gradually widened the analytic scope over the data and condition space, the HAL specific exploration starts with the provision of an overview of assessment based correlation coefficients over the available collections.

7.4.1 Word Similarity Assessment Based Alignment

The methodology used to enable the correlation of assessment based measures with HAL based measures is the same as described in the LSA specific section.

Correlation over collection aspect

Figure 7.8 provides an overview of the correlation coefficients between the three assessment based datasets over the 4 collections. As is evident the correlation between the datasets is much lower and exhibits much higher variance relative to the LSA based correlations (see figure 7.4). The observations as such contradict the observations made on basis of LSA models. On basis of LSA models all three datasets exhibited strong positive correlation over the different collections. This was interpreted as strong supportive evidence for the validity of the assessment procedure and the three specific implementations. On grounds of arguing, that if the measurements associated with the three datasets are valid, their CPM based correlation coefficients should be highly correlated. As can be seen in Figure 7.8 this is clearly not the case on basis of an alignment with HAL models. Specifically the correlation between FS353–M&C which was uniformly strong and positive on basis of LSA models stands out in this regard. This fact is further illustrated by examining the means and medians of the correlation coefficients over the collections. Table 7.7 shows the respective values.

	Mean	Median	StdDev	StdDev w.o. Genomics
FS353-M&C	0.3920625	0.356	0.356310299	0.340006417
FS353-R&G	0.6060625	0.7415	0.399648299	0.162926829
M&C-R&G	0.590875	0.5685	0.335040271	0.291822629

TABLE 7.7: Correlation between assessment based procedures on basis of their coefficients from alignment with HAL models.

Based on the presented data several observations can be made. The StdDev exhibited by the coefficients is much larger than the one reported for LSA. Secondly, as noted above, the mean correlations coefficients are generally lower. Thirdly as noted above,

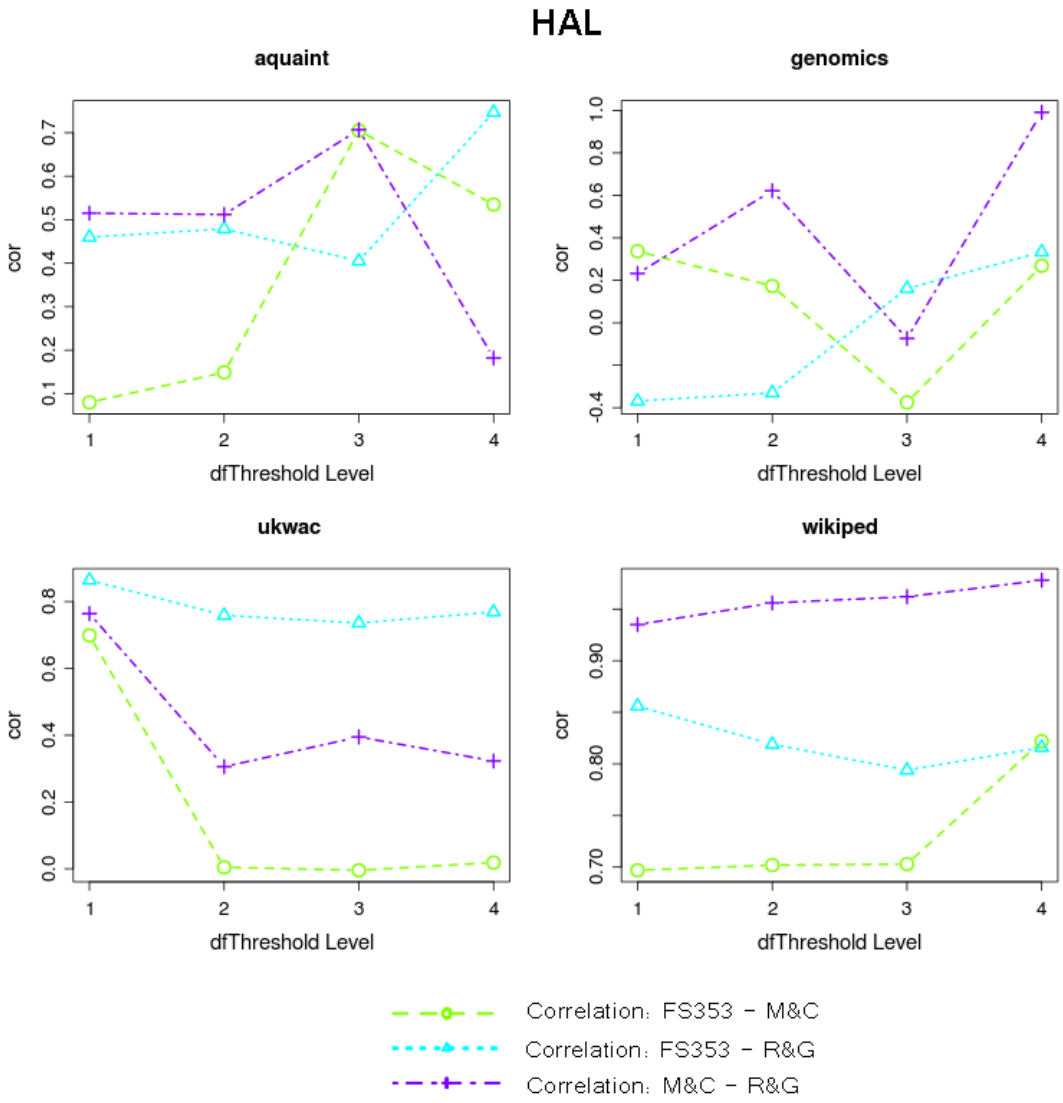


FIGURE 7.8: Correlation of assessment based data-sets over the four utilized collections with respect to collection specific df threshold levels. HAL models.

this is specifically the case for FS353-M&C. A first investigation concerning this aspect consists of an evaluation with regard to the df aspect.

Correlation over df Aspect

Table 7.8 provides an overview of the correlations between the datasets over the different df thresholds of the Wikipedia collection.

	Fs353	M&C	R&G	dfThreshold
Fs353	1	0.697***	0.856***	25
M&C	0.697***	1	0.935***	25
R&G	0.856***	0.935***	1	25
Fs353	1	0.702***	0.819***	75
M&C	0.702***	1	0.956***	75
R&G	0.819***	0.956***	1	75
Fs353	1	0.698***	0.815***	150
M&C	0.698***	1	0.957***	150
R&G	0.815***	0.957***	1	150
Fs353	1	0.703***	0.794***	300
M&C	0.703***	1	0.962***	300
R&G	0.794***	0.962***	1	300

TABLE 7.8: Wikipedia:DF25; Correlation between FS353, R&G, and M&C Word Similarity Tests and HAL models

The correlation coefficients reported in Table 7.8 are consistent with the previous observation of lower correlations for $FS353-M\&C$.

On basis of the so far presented HAL based results the following conclusions can be drawn with regard to the validity of the 3 specific procedures. The generally lower nature and higher variance exhibited by the coefficients of the correlation *between* the datasets outlines that the underlying correlations of the dataset specific measurements with HAL based measurements result in correlations of distinct strength. To illustrate this in more detail: Table 7.8 lists a positive correlation co-efficient of 0.697 between the $FS353-M\&C$ HAL based coefficients. This is significantly lower than the reported value of 0.991 for LSA based coefficients. A value of 0.991 indicates that the 'assessment' of the validity of the LSA models over the parameter space are almost identical. Both datasets, on grounds of their correlation with the LSA measurements, attest very similar degrees of validity to the different LSA models. A value of 0.697 however indicates that the two procedures attest different degrees of validity to the HAL models. Taken by themselves these observations represent contradictory evidence with regard to the validity of each of the two procedures. The need for further investigation of this aspect is further emphasized by the fact, that the other two between-dataset correlations are much stronger $(0.856, 0.935)$.

The next step with regard to such an investigation consists of analysing the impact of the weighting functions with respect to varying window sizes.

Correlation over weighting aspect

Figure 7.9 provides an overview of the dataset specific correlation coefficients with HAL based models with respect to the two underlying parameters of these models.

The figure is divided into three rows that outline the correlations between the datasets over three different collections. The coefficients are plotted with respect to the window size parameter of the underlying HAL model.The subsequent analysis of these plots is focused on the M&C dataset.

As can be seen in the figure the coefficient curves of M&C roughly resemble the curves of the other two datasets in the first row. Notable differences are exhibited by the $TF - IDF$ curves of M&C and R&G and the relatively higher coefficients of FS353 in comparison to both M&C and R&G. In *row*2 and *row*3 however the M&C specific curves of all weightings are distinctly different from those of the other two datasets. In *row*2 this is exhibited by a much higher correlation at window size 1 and a much faster degradation with growing window sizes. In *row*3, while less pronounced, the same stronger degradation, as well as a distinct difference concerning the coefficient of correlation at window size 1 can be observed.

Table 7.9 highlights this M&C specific distinct difference by listing the mean window sizes underlying the highest correlation coefficients for each of the 3 datasets.

Ranking	FS353	M&C	R&G
1	3.67	2.33	3
2	4.67	2.67	4.67
3	5	2.33	3.33

TABLE 7.9: Mean Window Size of Highest Correlation Coefficient with Respect to Assessment Based Sets. Aggregated over Wikipedia, UKWAC, and Aquaint Collection

As can be seen in Table 7.9, FS353 exhibits the largest mean window sizes with regard to the 1^{st}, 2^{nd}, and 3^{rd} highest correlation coefficients. That means the similarity ratings of the FS353 dataset and HAL based ratings are most closely correlated at the mean window size of 3.67. R&G exhibits similar, while slightly lower mean window sizes, compared to those in the FS353 column. A distinct difference to both other sets is given with regard to the M&C column. The listed mean window sizes are considerably lower. The similarity ratings of the M&C datasets are most closely correlated with HAL measurements if the underlying HAL model is based on a relatively small window size. This notable difference is specifically remarkable in regard of the facts that (a) the

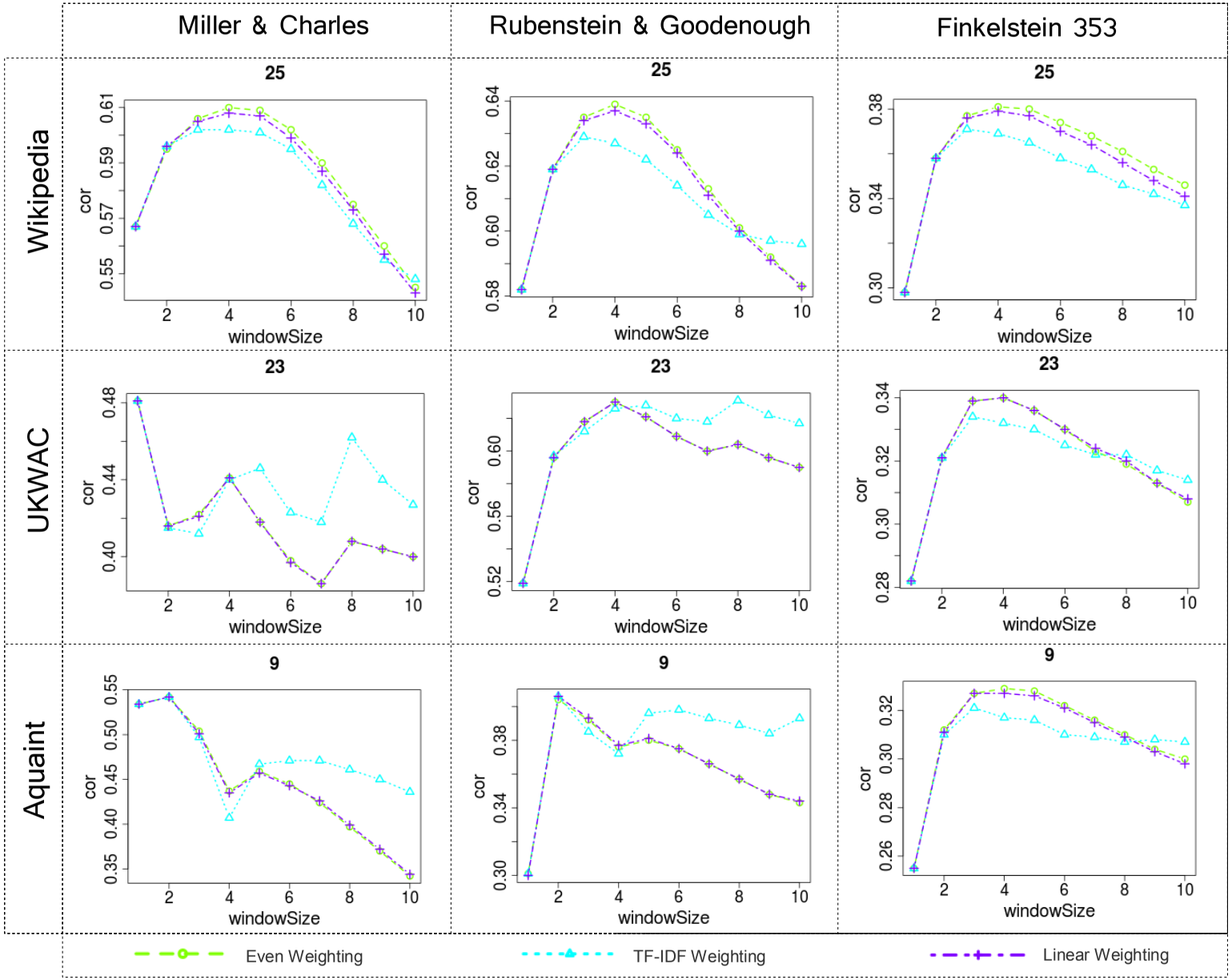


FIGURE 7.9: Comparative Analysis of Word Similarity Data Set coefficients on basis of HAL models. Underlying data consists of the lowest df threshold based representations of the collections.

30 term pairs of the M&S dataset are a subset of the pairs in the R&G dataset, and (b) the correlation of the human ratings for those 30 pairs is reported by Miller and Charles (1991) as 0.97 (Pearson product-moment correlation coefficient). On grounds of this it can therefore be ruled out that the observed diversion of the two datasets stems from the used assessment procedures or the actual similarity ratings as these are both almost identical for R&G and M&C. Consequently the difference must result from the composition of the underlying set of word pairs.

On grounds of the so far presented results and discussions a conclusive statement with regard to the validity of assessment based procedures as instruments of measuring word similarity can be made in the following form. Generally the supportive evidence on basis of the alignment with HAL models is less strong compared to the LSA based analysis. Further this evidence is restricted in scope. As outlined in the presented overview over all collections the correlation of FS353 and R&G is strong and positive in case of the UKWAC and Wikipedia collection. However this does not apply to the same degree to the Acquaint and genomics collection. Further it has been observed that the M&C based correlation coefficients differ distinctly from those of the other two sets. Summarizing it can therefore be stated, that there is supportive evidence for the validity of FS353 and R&G and contradictory evidence for the M&C dataset.

7.4.2 Priming Based Alignment

As was the case concerning the alignment with assessment based measurements, the priming based alignment of HAL models is applied on grounds of the respective methodology described in the LSA counterpart.

Figure 7.10 provides an overview contrasting assessment based and priming based alignments. The respective priming experimental measurements are shown in Tables 7.10 and 7.11. The observations in the tables show that in both cases grade of relatedness of words was inversely proportional to reaction times (i.e. closer related words result in shorter reaction times).

Semantic Distance	Response latencies	Error rate (%)	Δ
Very close	548[8.1]	1.8	na
Close	557[9.3]	1.5	9
Medium	567[9.1]	1.8	10
Far	572[9.8]	1.3	5

TABLE 7.10: Vigliocco Priming for Objects: Lexical Decision Task Response Times Shown.

Regarding the question of the validity of assessment procedures it can be said, that in principle the observations in the plots provide supportive evidence. In *row*1 the (i)

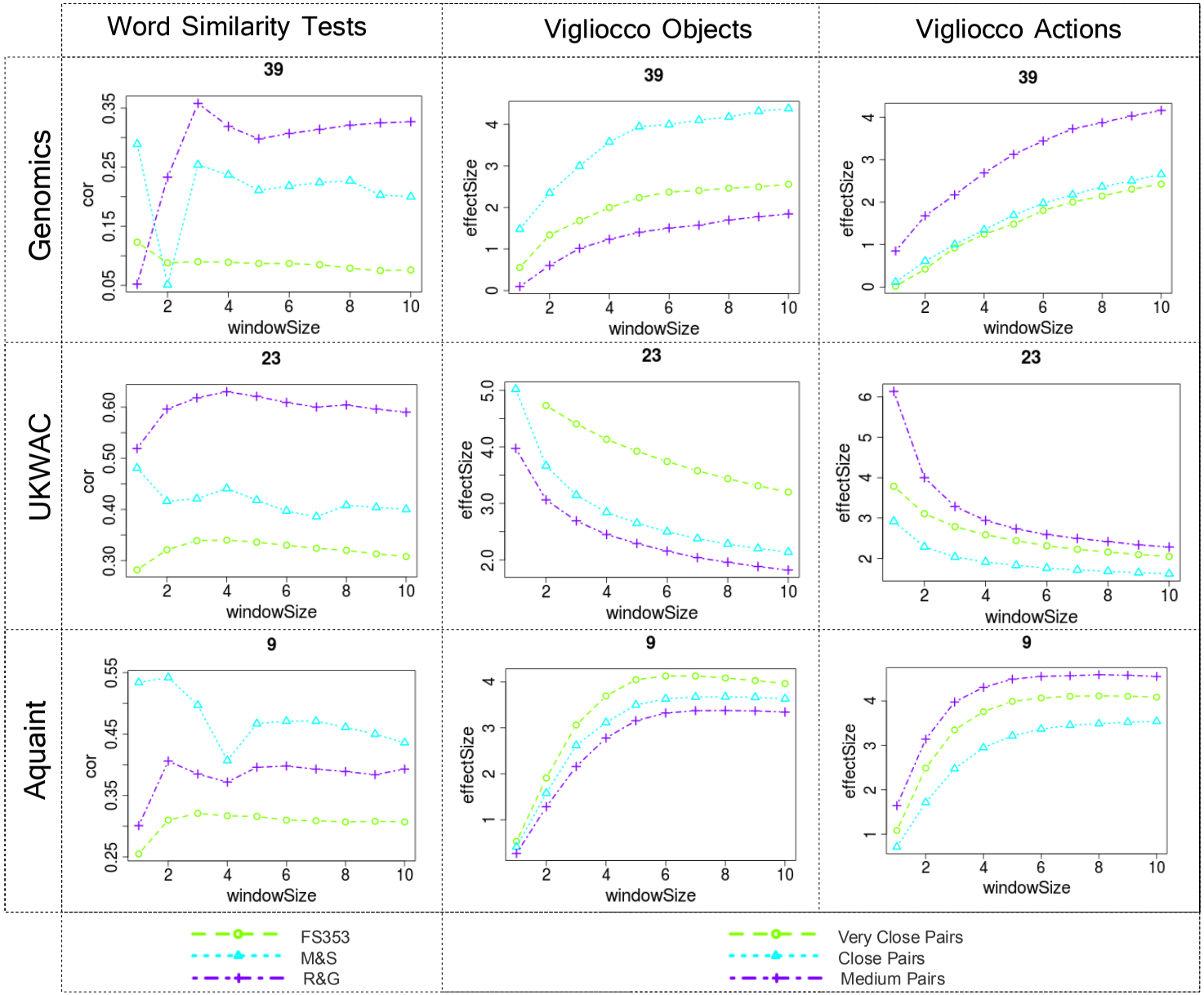


FIGURE 7.10: Priming Simulation Based Analysis on Basis of Alignment of HAL Based Estimates

Semantic Distance	Response latencies	Error rate (%)	Δ
Very close	602[8.6]	3.6	na
Close	613[8.9]	3.1	11
Medium	627[9.4]	4	14
Far	636[10.0]	3.3	9

TABLE 7.11: Vigliocco Priming for Actions: Lexical Decision Task Response Times Shown.

assessment and (ii) priming based alignments are consistent in that (i) attests low validity to the HAL measurements on grounds of the low correlation-coefficients, and (ii) on grounds of the contradiction of the simulated priming effects and the experimental priming effects. This consistency is, at least concerning object priming[7-5], also apparent in *row*2 and *row*3. In both rows the higher (relative to *row*1) correlation coefficients are accompanied by plots showing that the simulated object priming effects more closely match the experimental data. On closer inspection it becomes obvious that this consistency is most robust for the FS353 based correlations. Its exhibited much lower coefficients are confirmed by the priming data in *row*1. The relatively higher coefficients in the other two rows are also paired with consistent observations for the priming based data. In the case of R&G the observations of column 1 and 2 are inconsistent in regard of the exhibited magnitude of the coefficients in plots (r1,c1) and (r1,c3). Both plots show roughly equally sized coefficients, and therefore can be interpreted to suggest equal validity of the observations in the plots (r1,c2) and (r1,c3). This is not the case since plot (r1,c2) shows the highest priming effect for 'Close' instead of 'Very Close' pairs, while plot (r1,c3) features cognitively plausible priming effects. In case of M&C the inconsistency mainly consists with regard to the contradiction caused by the high correlation at window size 1 and the respective observations with regard to the simulated priming effects.

Concluding it can be stated that the reported results from the alignment of HAL measures with priming based observations only provide supportive evidence with regard to the validity of the FS353 procedure. This observation is therefore contradictory to the analogous conclusions drawn on basis of LSA based alignment. The next section provides a summarizing conclusion with regard to the validity of assessment based procedures as instruments of measuring graded word similarities.

[7-5]The observation with regard to the simulated priming effect for actions is consistent over all collections. This might indicate a limitation of HAL based models with regard to the measurement of action relationships. However due to the unavailability of assessment based data focused on actions it was not possible to investigate this in detail.

7.5 Chapter Conclusions and Answer to RQ 4

This section summarizes the results of the validation study concerned with the validity of assessment based procedures with respect to the measurement of grades of word similarity. The validation was conducted in two main steps. Step one consisted of an evaluation on basis of aligning LSA model based, assessment procedure based, and priming procedure based measurements. On grounds of the made observations it was concluded that all three assessment procedures are valid. This conclusion is based on the following two arguments:

- The drawn conclusions on basis of both sources (i.e. priming and assessment) with regard to the validity of specific LSA models are concordant.
- These conclusions are plausible with regard to an analysis of the underlying algorithm.

Step two consisted of an evaluation on basis of aligning HAL model based, assessment procedure based, and priming procedure based measurements. The observations substantially differ from those based on LSA models. Only in the case of the FS353 dataset all observations constitute supportive evidence of validity. In the case of R&G and M&C, specifically on grounds of the alignment with priming based measurements, the observations are to varying degrees not supportive of validity. Based on these results, the following answer can be formulated for RQ 4.

RQ 4 What are valid instruments for the measurement of the grade of relatedness between words?

On basis of the extensive analysis it can be concluded that there is strong supportive evidence that the FS353 measurements are valid measurements of graded word similarity and that consequently the validity of LSA and HAL based models can be evaluated on basis of it. In the case of R&G and M&C the observations are contradictory. With regard to these contradictions their validity remains in question at this point.

The pertinent questions that are raised by these conclusions are the following.

1. Why do the alignments on basis of LSA models suggest validity and those drawn on basis of HAL models in the particular cases do not do so?
2. What are the underlying reasons for the observed differences regarding the correlation of assessment based procedures with HAL models?

If two distinct instruments A and B exhibit varying correlations with a third distinct measurement instrument C, an initial consideration is given by questioning the validity of the measurements of A or B. B might exhibit higher correlation with C due to its measurements being 'more right', and A respectively a lower correlation due to its

measurements being 'more wrong'. With regard to the measurements of *A* constituting the R&G measurements and in case of *B* the M&C measurements this reason can be excluded. The measurements of both datasets are strongly positively correlated (0.97). Therefore the varying correlations of both datasets exhibited on basis of the *same* set of HAL measures must be induced by another aspect. If the measurements of both datasets are 'right', then the only plausible explanation that remains consists of the conclusion that the two datasets represent measurements of different phenomena, and that the occurrence of these phenomena is dependent on the parameter settings underlying the HAL model. On grounds of the existent body of knowledge it can further be assumed that these different phenomena are given by semantic and associative relationships between words.

An evaluation of these assumptions is conducted as part of the validation of measurement instruments of semantic and associative strength of word relationships in the subsequent chapter.

CHAPTER 8

MEASUREMENT OF SEMANTIC-ASSOCIATIVE DEGREE OF WORD RELATIONSHIPS

Chapter 7 presented a validation study of measurement instruments of grade of relatedness between words. This chapter reports on a validation study of procedures aimed at measuring the associative-semantic degree of word relationships. This aspect is addressed by RQ 5 of the dissertation.

RQ 5 What are valid instruments for the measurement of the type of relatedness between words?

The fundamental considerations underlying this exploration are similar to those presented in the previous chapter. A distinctive difference is given by the sophistication of existent measurement procedures. With regard to the measurement of semantic-associative degree only one publicly reported and available assessment-based procedure exists.

The exploration of the validity of the potential procedures is conducted on basis of the following structuring. Section 8.1 reports the details of the specific experimental setup. The presentation and discussion of the results of the performed study is presented in two distinct sections, each focusing on one of the two computational models. Section 8.2 reports on the validation results obtained by aligning HAL based measurements. The following section analogously reports on results based on LSA measurements. Section 8.4 provides an overview and a discussion of the reported results and draws conclusions.

8.1 Experimental Setup

This section describes the experimental setup concerning items that are specific to the validation of instruments targeted at measuring the semantic-associative degree of relationships. With regard to the underlying data and estimator space the setup is consistent with the specifications provided in Section 6.5.

8.1.1 Assessment Based Measurements

Word Similarity Assessments

As noted before, only one dataset aimed at measuring different types of relationships between words has been published. The modelled types of relations in the study are of linguistic nature. Table 8.1 provides an overview of the main characteristics of this dataset. Details concerning the underlying assessment procedure and the definitions of the relationship types have been provided in Section 6.2.2.

Name	Abbreviation	# of Pairs	Reference
FinkelsteinRel	FSRel	252	Agirre et al. (2009)
FinkelsteinSim	FSSim	101	Agirre et al. (2009)

TABLE 8.1: Assessment-Based Procedures Applied in Course of Semantic-Associative Focused Validation

Word Neighbourhood Assessments

With respect to the limited amount of available word similarity assessment data, a second form of assessment is introduced that can be aligned with model based measurements.

The subsequently described assessment procedure is focused on an examination of the neighbourhoods of words generated by computational models. The term 'neighbourhood' refers to the set of terms that are deemed to be most closely related to a specific word. With respect to the focus of the validation the interest lies in the determination of the semantic-ness and associative-ness of a word's neighbourhood. That is, if the n closest related words of a term primarily exhibit a semantic or associative relation. To illustrate this aspect Table 8.2 shows the ten closest related pairs for the term 'ship' on basis of a specific LSA and HAL model. The motivation for integrating this specific type of assessment stems from the following two reasons:

Neighbourhood of word 'ship'		
Rank	Model	
	LSA (LogEntropyTransform,50)	HAL (LinearWeighting,1)
1	vessel	warship
2	sail	somatogyru
3	warship	gunboat
4	salvage	freighter
5	aground	vessel
6	afloat	barge
7	sank	aircraft
8	seaworthy	schooner
9	tug	troop
10	unseaworthy	boat

TABLE 8.2: Neighbourhood of 10 Closest Related Pairs for the Word 'ship' on Basis of Wikipedia:DF25 Collection

1. With regard to the analysis of the relation of 'word similarity' and 'relevance' on basis of computational models it is essential to verify that measurements with regard to the grade and type of similarity are reflected by the semantic space of the model. Assessing the neighbourhood of a word constitutes a mean of directly evaluating model specific output.

2. Contrary to the assessment of the grade of similarity, the assessment of the type of similarity can be based on formally defined criteria and is less likely to result in subjective assessor bias.

On that note the conducted assessment is of the following form:
An initial set of 30 words is chosen. For each of these 30 focused words the 40 closest related words, as calculated on basis of a specific model, are assessed with regard to the exhibited type of the relationship between the neighbour and the focused word. These assessments are conducted with respect to 40 distinct HAL and LSA models. The total number of word pairs that are assessed therefore constitutes 48000.

The assessment of each word-neighbour pair is conducted on basis of the cognitive interpretation of semantic similarity. A relationship between a word and its neighbour is therefore rated *primarily* semantic based on the following criteria:

- share a large number of features.
- share categorical membership.
- exhibit both of the two criteria.

Pairs that are *clearly related* but do *not* exhibit these criteria are labelled associative. Words that exhibit no comprehensible relationship are rated 'false' and not considered

in the calculation of the semantic-associative ratio. Table 8.3 provides a sample of the assessment data with regard to the neighbourhood of the word 'ship'.

Assessed neighbourhood of word 'ship'						
Rank	LSA Model			HAL Model		
#	LogEntropy Transf.,50	cosine	Ass.	Linear Weight.,1	eucl. dist.	Ass.
1	vessel	0.9678	S	warship	0.905636	S
2	sail	0.9656	A	somatogyru	1.01105	X
3	warship	0.958	S	gunboat	1.012974	S
4	salvage	0.954	A	freighter	1.018558	S
5	aground	0.954	A	vessel	1.018776	S
6	afloat	0.9538	A	barge	1.062435	S
7	sank	0.9529	A	aircraft	1.077218	A
8	seaworthy	0.9513	A	schooner	1.079251	S
9	tug	0.9477	A	troop	1.080123	S
10	unseaworthy	0.9476	A	boat	1.081659	S

TABLE 8.3: Assessed Neighbourhood of 10 Closest Related Pairs for the Word 'ship' on Basis of Wikipedia:DF25 Collection. 'S' Marks Semantic Relations. 'A' Marks an Associative Relation. 'X' marks an unrelated term.

The semantic-ness or associative-ness of the neighbourhoods for each word is expressed through the calculation of the ratio of the total number of semantic relations divided by the total number of associative relations, weighted with respect of the rank of each neighbour.

$$\frac{\sum_{1}^{n} wSem_n}{\sum_{1}^{n} wAssoc_n}$$

Where n is the number of considered neighbours.
$\sum_{1}^{n} wSem_n = \sum_{1}^{n} sim_n * \frac{1}{r_n}$ IF relation of term n is semantic
$\sum_{1}^{n} wAssoc_n = \sum_{1}^{n} sim_n * \frac{1}{r_n}$ IF relation of term n is associative
Where r_n is the neighbourhood rank of term n and sim_n the model specific measured similarity of term n (i.e. cosine, or Euclidean distance)

As can be seen, the chosen weighting of $\frac{1}{r_n}$ constitutes a simple linear weighting with respect to the rank of the neighbourhood terms, chosen to reflect this rank in the calculation of the ratio.

8.1.2 Priming Experimentation Based Measurements

Table 8.4 lists the priming based datasets used as part of the validation.

Name	Abbreviation	# of Word Pairs	Reference
Chiarello	Chia	3x48	Chiarello et al. (1990)
Ferrand	Fer	2x44	Ferrand and New (2004)

TABLE 8.4: Priming based procedures applied in course of semantic-associative focused validation

As can be seen in the table two distinct datasets find application. The Ferrand dataset consist of two sets of 44 word pairs that were analytically chosen on basis of exhibiting semantic only or associative only relations. The Chiarello dataset consists of three sets of 48 pairs representing the relations associative, semantic, and semantic-associative. In both cases the initial experiments were aimed at evaluating the magnitude of priming effects with regard to specific relationship types. Details regarding the experimental setup of both studies are provided in the respective references.

Subsequently the results of the validation study are reported. As in the last chapter, the presentation of the results is split into two sections, dedicated each to either the HAL or LSA computational model.

8.2 HAL Based Alignment

Of specific interest with regard to an evaluation of validity that is focused on associative and semantic degree of a relationship is the reported characterisation of the two computational models with regard to this aspect. With regard to this Jones et al. (2006, p. 536) provides a detailed summary of the state of the art interpretation.

> *LSA and HAL consider subtlety different types of information while learning text, and these differences are reflected in the structural representations formed by each model. LSA tends to weight associative relationships more highly than purely semantic relationships. For example, the representation for car is much more similar to the representation to drive ($cos = 0.73$) than it is to members of the same semantic category, such as truck ($cos = 0.49$) or boat ($cos = 0.03$). Further, the verb drive is more similar to car than it is to other action verbs, such as walk ($cos = 0.23$). By contrast, HAL considers distance between intervening words in the moving window; hence, semantic relationships can become more highly weighted in HAL than associative relationships. In HAL, car is more similar to truck ($d = 0.90$) and boat ($d = 0.95$) than it is to drive ($d = 1.12$), and the verb drive is more similar to another action verb like walk ($d = 1.03$) than it is to car ($d = 1.12$). HAL and LSA focus on different sources of information and, thus, make different predictions about the strength of semantic*

and associative relationships in memory. ”

Hutchison (2003, p. 809) further notes that it was demonstrated that ’HAL predicted priming for items that were semantic relatives or semantic-associative relatives, but not for items sharing only an associative relation’. In this form, the above quotations represent a central part of the nomological network underlying the current validation study.

Regarding the order of exploration within this study the following can be remarked. In contrary to the outline underlying the grade of similarity validation, the amount of datasets based on priming experiments is significantly larger than those of assessment-based experiments. As a reaction to this, the exploration of the validity is initially focused on priming based alignment.

8.2.1 Priming Based Alignment

The underlying methodology to simulate priming effects is identical to the description provided in Section 7.3.2. Figure 8.1 shows the simulated priming effects for the Ferrand dataset over the df levels of the Wikipedia collection.

In Figure 8.1 the mean priming effects of the associative and semantic pairs are plotted with respect to the underlying window size of the HAL model. To render the following discussion more intuitive, Table 8.5 provides a small sample of the semantic and associative pairs of the Ferrand and New (2004) dataset. The aim of the original study

Semantic		Associative	
match	LIGHTER	needle	THREAD
beet	RADISH	album	PHOTO
bomb	MISSILE	bulb	LIGHT
bottle	GOURD	anchor	BOAT
thistle	CACTUS	aquarium	FISH
clarinet	FLUTE	spider	WEB
casket	BOX	astronaut	SPACE
zucchini	PUMPKIN	cradle	BABY
dolphin	WHALE	buoy	RESCUE

TABLE 8.5: Sample Associative and Semantic Pairs of the Ferrand and New (2004) Study

consisted of demonstrating that significant priming effects can be measured in terms of reaction times for both relationships. The results are summarized by Ferrand and New (2004, p. 37) as follows:

“ *The present lexical decision experiment demonstrates (1) the existence of automatic semantic similarity priming in the absence of normative*

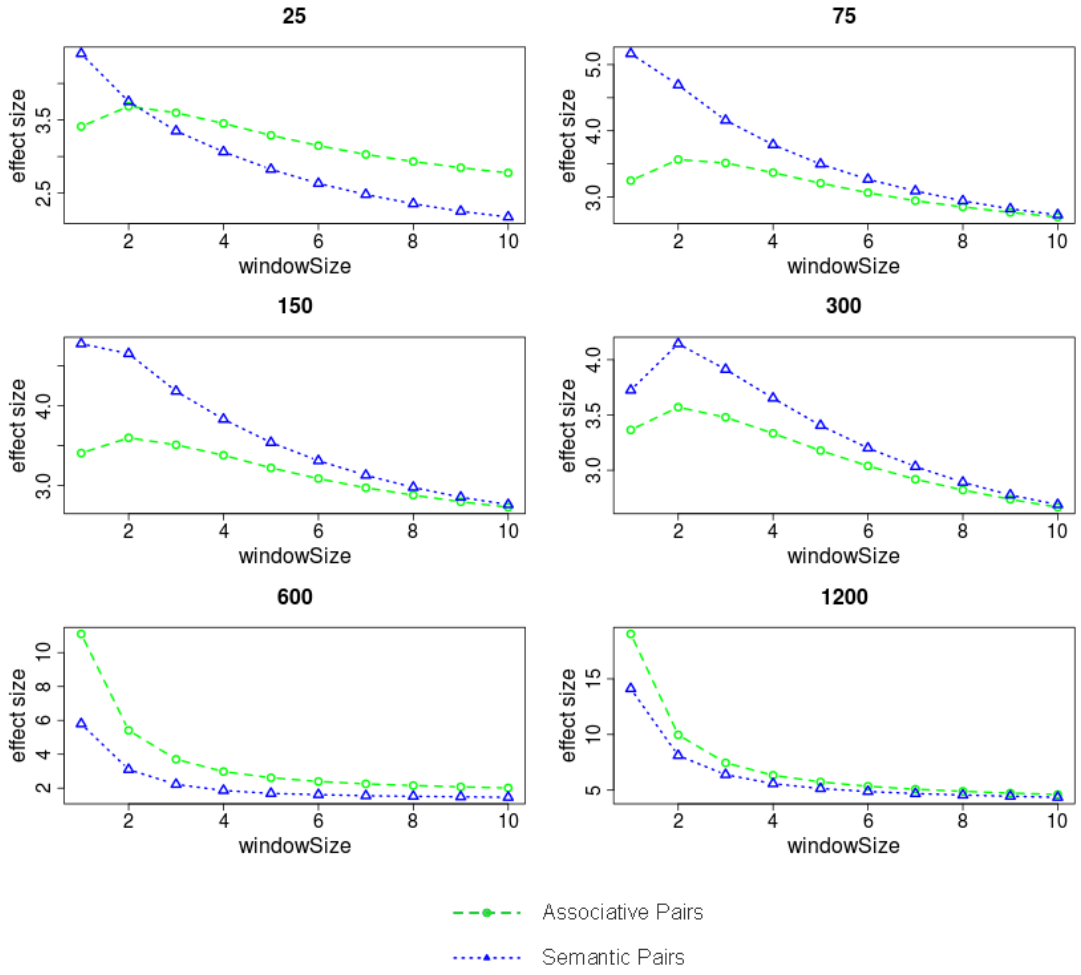


FIGURE 8.1: Simulated Priming Effects for Even Weighting Based HAL Models on Basis of Ferrand and New (2004) Word Pairs; Plotted over Window Size on the Wikipedia collection

> *association (for pairs such as 'dolphin-WHALE'), ... , and (2) the existence of automatic associative priming in the absence of semantic similarity (for pairs such as "spider-WEB") ...* ”

On grounds of this background the observed mean priming effects in the figure can be interpreted in the following manner. Considering the plot of $df25$ it can be seen that the mean effect for semantic pairs is larger for the window sizes 1 and 2. In consideration of the underlying methodology this means, that the distance between the semantic pairs was considerably smaller than the distance of the associative pairs. However as can be seen in the plot with increasing window size the associative priming effect gains in strength and surpasses the semantic effect. Regarding the df levels $75, 150, 300$ it can be observed that the simulated priming effect for the semantic pairs remains superior for larger window sizes. These observations are partly consistent with the observations of Jones et al. (2006) in regard of small to medium window sizes resulting in stronger priming effects for the semantic pairs. However, the top left plot of Figure 8.1 shows that for window sizes larger than 2 the HAL measurements are more associative than semantic. To explore this aspect in more detail the alignment of HAL with a second priming dataset is subsequently explored. Figure 8.2 shows the simulated priming effects on basis of the Chiarello et al. (1990) word pair sets. Analogous to the prior Ferrand dataset, Table 8.6 provides a sample of the underlying pairs of the study.

Semantic		Associative		Both	
Dagger	Rifle	Hockey	Ice	Moth	Fly
Sugar	Salt	Rake	Leaf	Steel	Iron
Floor	Wall	Wave	Ocean	Sword	Knife
Gin	Wine	Book	Page	Army	Navy
Table	Bed	Onion	Tears	Doctor	Nurse

TABLE 8.6: Sample pairs of the Chiarello et al. (1990) study.

As shown in Table 8.6, the study of Chiarello is based on three sets of word pairs analytically chosen to reflect the relationship types associative, semantic, and associative-semantic. Table 8.7 provides an overview of the results of this study that were obtained via the conduction of a lexical decision task.

Central primes	Similar	Δ	Associated	Δ	Similar+Associated	Δ
Related	684	na	646	na	642	na
Neutral	699	15	681.5	35.5	681.5	39.5
Unrelated	715.5	16.5	691	9.5	687	5.5

TABLE 8.7: Chiarello et al. (1990) Lexical Decision Task Response Times. Aggregated over Visual Fields.

As can be seen in Table 8.7, the measured priming effect for the associative pairs is much larger than for the semantic pairs. The largest priming effect is observed for

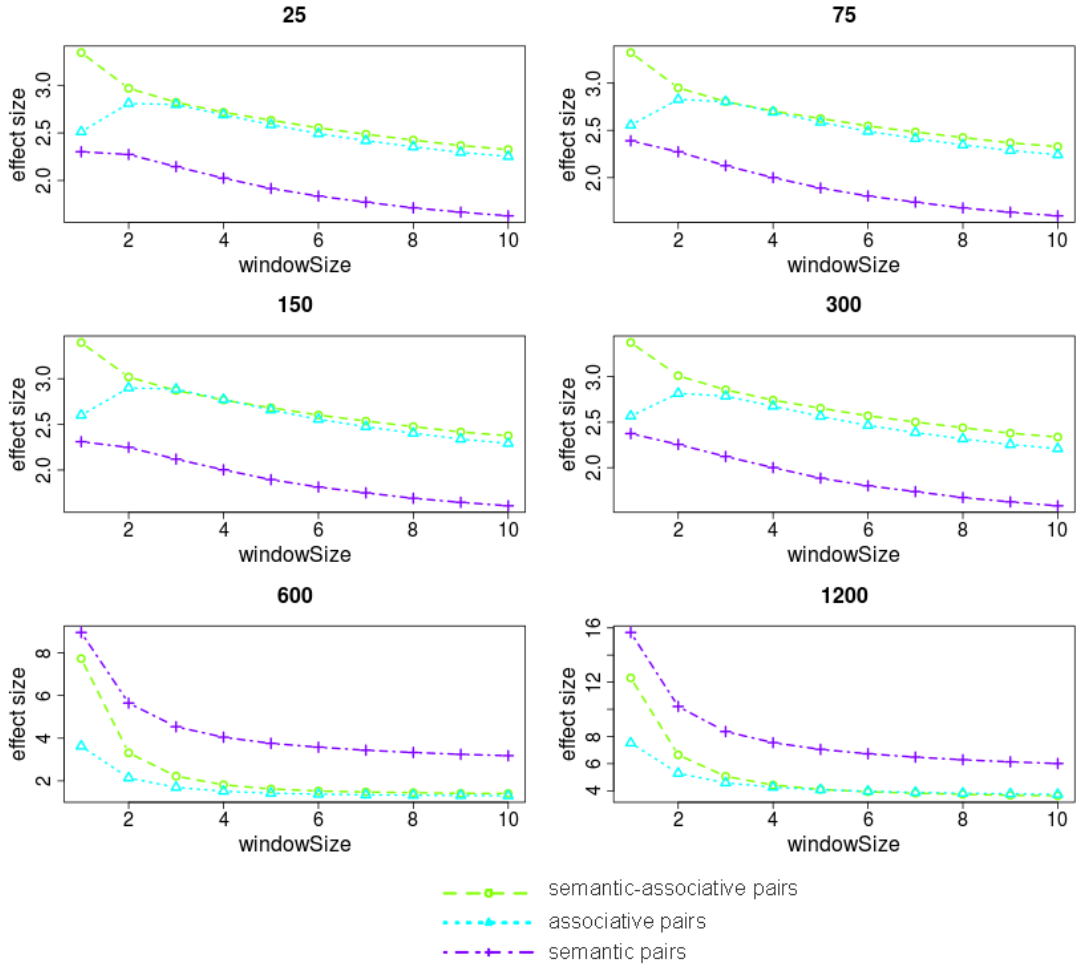


FIGURE 8.2: Simulated priming effects for Even Weighting based HAL models on basis of Chiarello word pairs; plotted over window size on grounds of the Wikipedia collection

the associative-semantic condition. This observation is referred to as the 'associative-semantic boost'.

On grounds of this provided background the observations in figure 8.2 can be interpreted as follows. In the top-left plot the priming effects on basis of the $df25$ representation are plotted. The trend of the *semantic* priming effects is very similar to the Ferrand based simulation. Mean priming effect is highest for window sizes 1 and 2 and diminishing with increasing window size. Further the curve representing *associative* priming effects also resembles the observed behaviour in the Ferrand case. The observed behaviour of the *semantic-associative* pairs is particularly interesting. The HAL based priming effects show a clear associative-semantic boost at the window sizes 1 and 2 that is consistent with the semantic and associative curves at these parameter settings. Concerning the observed magnitude of the priming effects the following considerations are of relevance. In the case of the Ferrand study it could be seen that the simulated priming effect for semantic pairs were larger for some parameter settings. In the Chiarello study the priming effects for associative pairs are uniformly larger than those for the semantic pairs. This can be related to the original experimental results. As can be seen in the respective table the Chiarello study recorded a priming effect for the associative pairs that was more than two times larger than that of the semantic pairs. As noted by Ferrand and New (2004, p. 37) in their study, 'on average, the size of the priming effect was larger for semantic pairs (d = +34.5msec) than for associative pairs (d = +18.5 msec).'[8-1] The magnitude of the priming effects can therefore be interpreted on being highly dependent to the choice of word pairs. As a consequence, this has to be taken into consideration when making inferences with regard to an associative or semantic 'nature' of HAL models. Another distinction concerning the observations on basis of the two studies is given by the steeper decline of the semantic priming effects in the Ferrand case. As Steyvers (2000) remarked the chosen semantic only pairs of Chiarello seem to also exhibit associative relations. The relatively slower decline of the semantic priming effects on basis of the Chiarello pairs seems to confirm Steyvers (2000) observation.

Based on the so far presented observations some preliminary considerations can be made with regard to the validity of the priming effect simulation procedure as an instrument of measuring the semantic-associative degree of HAL based word relationships. First it can be stated, that the observations are plausible with regard to the underlying mechanics of the HAL algorithm. At a symmetric window size of 1 and 2 the words in the HAL model are represented via vectors comprised of words co-occurring in very

[8-1] Ferrand and New (2004, p. 37) explains this observation as follows:

> 'It can easily be explained by the fact that semantic targets had a lower printed frequency ($M = 16.3$ occurrences/million) than associative targets ($M = 98.3$ occurrences/million). Previous studies obtained larger priming effects for low- compared to high-frequency targets'

close vicinity. It is intuitive that the creation of co-occurrence vectors under such parameter settings is more likely to 'pick up' co-occurrence patterns suitable to the distinction of semantically related words. The similarity of a pair such as 'bomb-MISSILE' might, for example, be identified on basis of the common co-occurrence of words such as 'explode, attack, damage'. The similarity of an associative pair such as 'astronaut-SPACE' however might be better represented by vectors originating from larger window sizes. Very short window sizes might, in the case of 'astronaut', result in a sparse vector representation comprised of co-occurrences with words such as 'russian, youngest, works, training'. On grounds of these examples it becomes obvious that the equivalent co-occurrences for 'space' might considerably differ for these small window sizes, and that consequently the measured similarity evidenced by the simulated priming effect is lower. The – on basis of this argumentation – plausible lower priming effect for associatively related pairs at window sizes 1 and 2 is clearly reflected in the simulated priming effects. These considerations constitute a positive indication with regard the validity of the plotted priming effects. That is, of the ability to accurately measure the semantic or associative degree of the word similarity measures of HAL models, with respect to the underlying parameter settings. To place this into a paradigmatic context the following should be considered. The motivation for taking measurements of similarity type stems from the aim of subsequently correlating such measurements with 'relevance' focused measurements. A necessary prerequisite for such an application however consists of determining to which degree a CPM based measure of word similarity is of semantic or associative nature. With regard to that, the observed curves of the simulated priming effects indicate that the underlying simulation procedure might represent such a measurement instrument.

To further add substance to these observations and the claims of validity it is necessary to align the HAL based measurements with additional different kinds of measurements. With regard to this the next section outlines the alignment of HAL based measurements with assessments of the respective model outputs.

8.2.2 Neighbourhood Assessment Based Alignment

As a means of gaining additional supportive evidence for the validity of the priming simulation based measurements an alignment with neighbourhood assessment-based measurements is conducted.

As described in the experimental setup the basis for the subsequently described alignment consists of manual assessment of specific model based neighbourhoods of words. The reasoning is the following. If the measurements of a specific model are primarily representing either a semantic or an associative relation between words, then this should be reflected in the set of terms that are most closely related to a particular word.

Supportive evidence with regard to the validity of the priming simulation methodology would consist of consistency in the observations of both measurement kinds.

Figure 8.3 provides an overview of the measured mean simulated priming effects on basis of the Ferrand word pairs on grounds of HAL models based on top of the UKWAC collection. As can be seen in Figure 8.3, within each df dependent plot, the difference

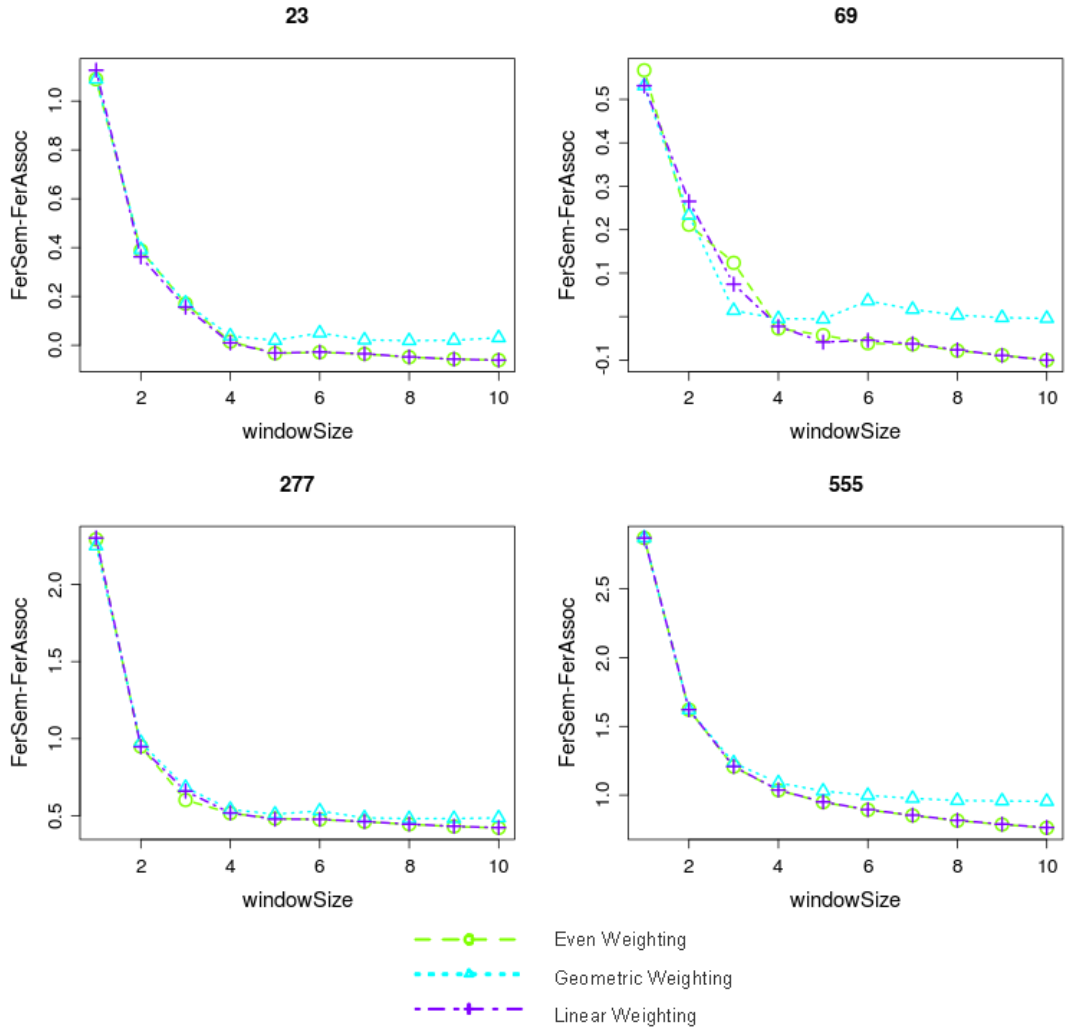


FIGURE 8.3: Plots of Simulated Priming Effects on Basis of Ferrand Pairs and Measurements Produced by HAL Models

between the semantic priming effect and the associative priming effect is plotted over the window size parameter. A positive value reflects a larger semantic mean priming effect. A negative value represents a larger associative mean priming effect. Each curve in the plot represents a specific weighting function. In congruence with the Wikipedia based observations presented so far the simulated mean priming effect for the semantic pairs is considerably larger than the associative effect at window size 1 and 2. If the

underlying measurements of the model are valid, this much larger priming effect for semantic pairs should also be reflected in the neighbourhoods of words. To assess this assumption, Figure 8.4 shows the assessed semantic/associative (sem-assoc) ratio of word neighbourhoods with respect to one of the plotted curves in the top-left plot of figure 8.3.

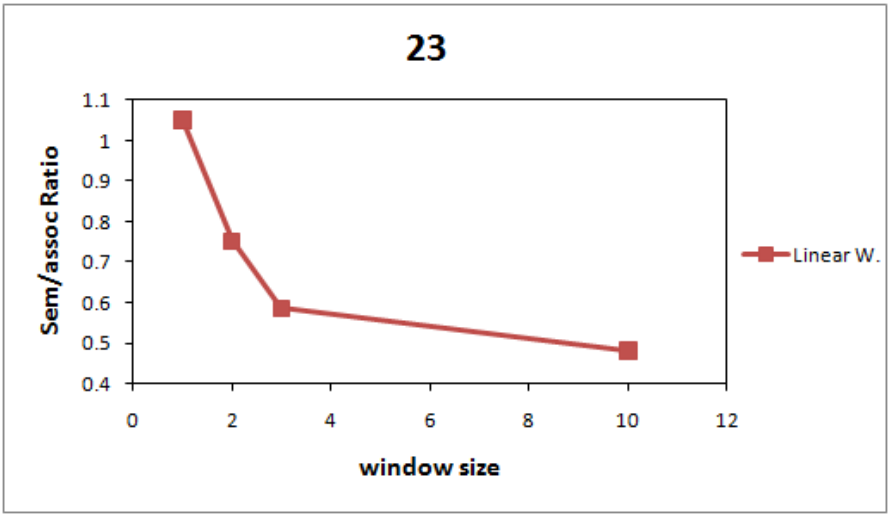


FIGURE 8.4: Semantic/Associative Ratio of Neighbourhood Relations Shown on Basis of UKWAC:df23 Collection

As can be seen the trend of the sem-assoc ratio closely resembles the respective trend of the Linear weighting curve representing the priming simulation effect based sem-assoc ratio. In concordance with the respective priming effects the datapoints at window sizes 1 and 2 indicate, that the neighbourhoods of the HAL models based on these parameter settings are considerably more semantic in nature than those at larger window sizes. Specifically it can be observed that for window size 1 the > 1 sem/assoc priming ratio is also reflected by a > 1 sem/assoc neighbourhood ratio.

To add substance to this observation subsequently an alignment on basis of different data conditions is explored. Figure 8.5 provides an overview of the respective priming effect ratios on basis of the Wikipedia collection. Analogous to the earlier alignment a first investigation consists of a visual comparison with the respective assessment-based ratios. Figure 8.6 shows the curves for Even and Linear weighting with respect of the Wikipedia $df25$ representation. As can be seen the neighbourhood ratio plot again closely resembles the respective plot of the priming effects. However relative to the earlier example the semantic-associative ratio on basis of the Wikipedia collection outlines a much stronger semantic-ness.

A final alignment on basis of neighbourhood assessment data is based on the $df300$ plot of Figure 8.5. As can be seen in the plot the Geometric weighting based HAL models exhibit a distinctly higher semantic-ness compared to the other weighting functions. To evaluate if this observation is also reflected in the word neighbourhoods of

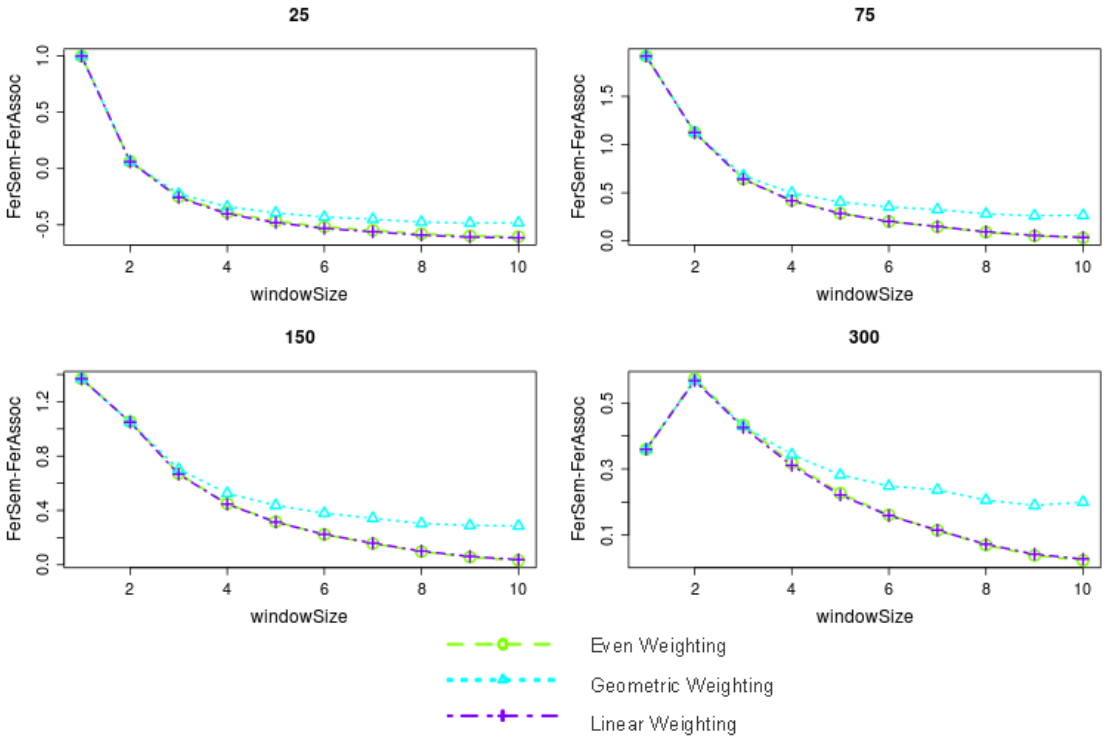


FIGURE 8.5: Ferrand word pair based simulated mean priming effect ratios of HAL models on basis of the Wikipedia collection

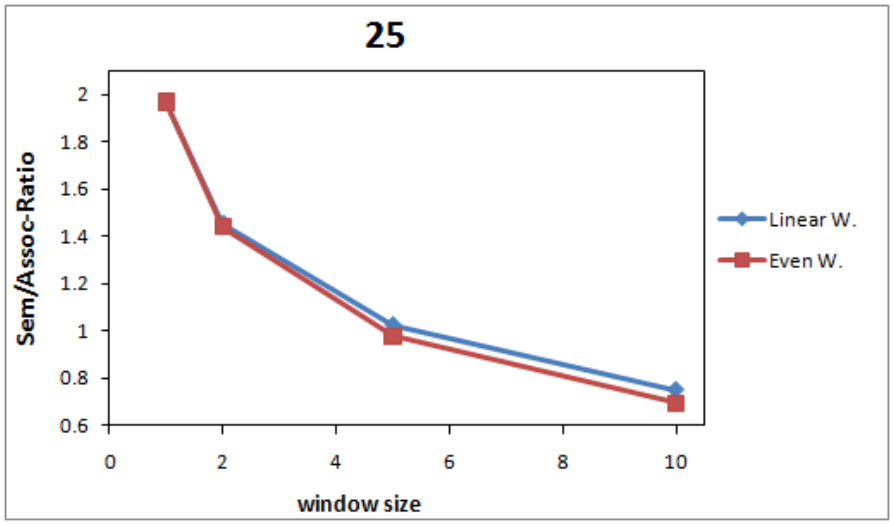


FIGURE 8.6: Semantic/associative ratio of neighbourhood relations shown on basis of wikipedia:df25 collection

the respective models, Figure 8.7 provides a plot of the neighbourhood ratios for the Geometric and Even Weighting based models. As can be clearly seen in the figure the

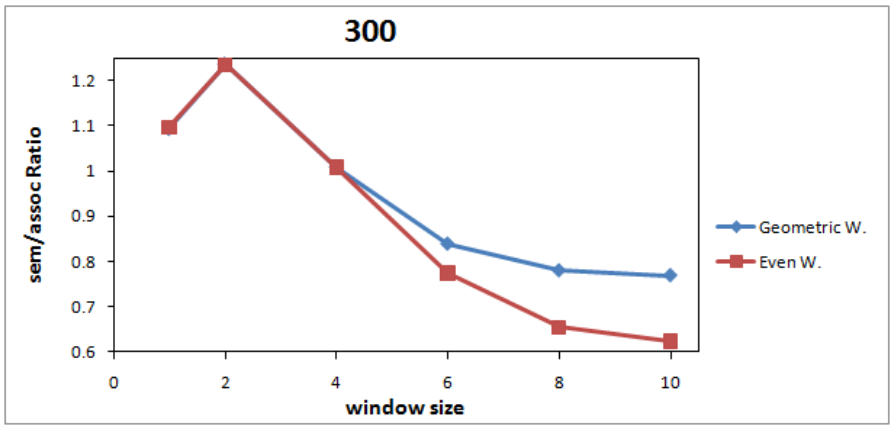


FIGURE 8.7: Semantic/associative ratio of neighbourhood relations shown on basis of the wikipedia:df300 collection

higher sem/assoc ratio exhibited by models based on Geometric weighting for larger window sizes is clearly reflected in the figure. This observation is interpreted as strong supportive evidence with regard to the validity of the simulated priming effect methodology. This is specifically so with regard to the plausibility of this observation in terms of the weighting function formulas.

The Even weighting function simply assigns equal weights to all terms in the window. On basis of this it is plausible that the vectors resulting of a large window size such as 10 will contain relatively more co-occurrences that are beneficial to the identification of associative relations. Further it is intuitive that the recording of such relations can be interpreted as 'noise' with regard to the identification of semantic relations.

The geometric function however geometrically decreases the weight of a co-occurrence with respect to the window size as can be seen in the following formula:

$$w_g = \frac{ws - (pos - 1)}{ws^2}$$

Where w_g is the resulting weight, ws is the window size, and pos is the position of a specific term within the window. It is clear that this weighting formula weighs terms that appear in close vicinity much stronger than those appearing in larger proximity. In the case of a window of size 10 the weight of a co-occurrence adjacent to a term is 10. The weight of a term at position 10 however is 0.01. On basis of this it is obvious that the geometric weighting function can be understood to effectively reduce the window size by applying extremely low weights to terms occurring at larger proximity. Thus the observation that the Geometric weighting based models exhibit significantly larger semantic-ness for larger window sizes is plausible. The fact that this is reflected both

by the assessment and priming simulation based measurements represents strong supportive evidence for their validity. As a final test series with regard to the validity of the priming simulation procedure the next section explores the alignment of the word similarity assessment-based measurements.

8.2.3 Word Similarity Assessment Based Alignment

As noted before only a single source of word similarity assessment based data with respect to the type of relations exists. Figure 8.8 shows the correlation coefficients on basis of Geometric and Even-weighting HAL models for the Acquaint collection.

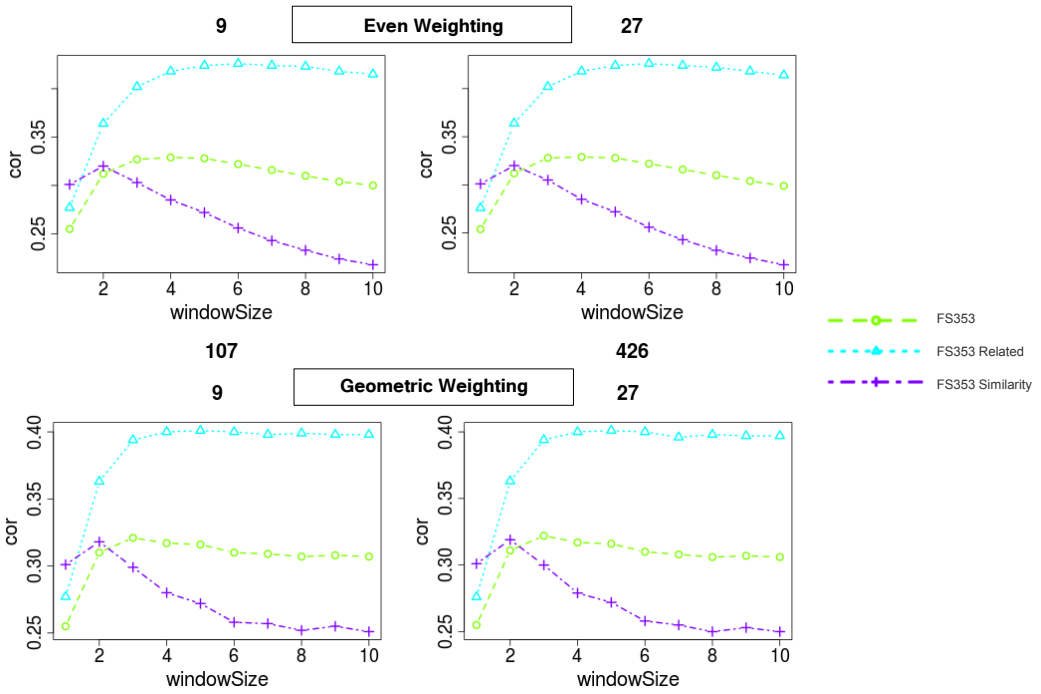


FIGURE 8.8: HAL based correlation coefficients of FS353, FSRel, FSSim on Acquaint collection

As can be seen the df dependent plots each feature the co-efficient curves of the following three assessment based sets:

- **FS353**: The Finkelstein353 set. See Sections 7.2.2 and 6.2.1 for details.
- **FS353 Related**: Subset of FS353 pairs that were assessed to exhibit meronymy-holonymy, and non-classifiable relation types, or unrelatedness. See section 6.2.2 for details.

- **FS353 Similarity**: Subset of FS353 pairs that were assessed to exhibit synonymy, antonymy, hyponymy, unrelatedness. See Section 6.2.2 for details.

On grounds of the above provided short definitions, the observed trends of the curves can be interpreted in the following way. The FS353-Sim curve, which on basis of the underlying definitions is closest to the cognitive definition of semantic relatedness, shows the highest coefficients at very small window sizes. This is consistent with prior observations. As shown on grounds of simulated priming effects and neighbourhood assessment data, these window sizes result in vectors that primarily record co-occurrences that are beneficial to the identification of semantic relations. It is therefore intuitive that the HAL models at these parameter settings show the highest correlation with a dataset that is focused on similarity of a semantic nature. On basis of comparing the trends of the FS353-Sim curve for the specific weightings it can be observed that the Geometric based curve declines more steeply for larger window sizes. This observation indicates that the pairs in the FS353-Sim set also exhibit associative relations. For Even Weighting based models these pairs can be assumed to benefit from larger window sizes. With respect to the prior analysis of the Geometric function this is not the case when applying Geometric weighting and therefore results in a steeper decline. Of note with regard to the 'FS353 Related' curves is its close resemblance of the FS353 curve. Although the term 'Related' is defined as 'Semantically Related' by Agirre et al. (2009), on basis of the so far made observations, the observed trend of the curve indicates that the relations of the pairs in the set mainly seem to be of associative nature.

To investigate the above made interpretations in more detail the next step consist of visually aligning simulated priming effects and word similarity assessments. Figure 8.9 provides an overview of such an alignment on basis of using the Wikipedia and UKWAC collections. The observations of the priming effects and the correlation coefficients can be interpreted as roughly corresponding to each other. As can be seen in plot (r2,c2) and (r3,c2) the coefficients for the FS353-Sim dataset are considerably higher for larger window sizes when based on HAL models with Geometric weighting. This is consistent with the observation that the semantic priming effect in plot (r1,c2) is higher for larger window sizes. Further it can be observed that over all weighting functions the FS353-Sim coefficients generally tend to be higher for smaller window sizes. Specifically so when compared to the coefficients of the other two datasets. On basis of these observations it seems justified to assume that the FS353-Sim dataset is indeed more sensitive to the semantic-ness of the underlying HAL model than the other two datasets.

On basis of this statement a last evaluation with regard to the validity of measurement instruments focused on the semantic-associative degree of word relationships consists of a correlational analysis of the available word similarity assessment sets over different collections. Figure 8.10 shows a correlogram of the correlation coefficients between the FS353, FS353-Rel, FS353-Sim, M&C, and R&G assessment sets.

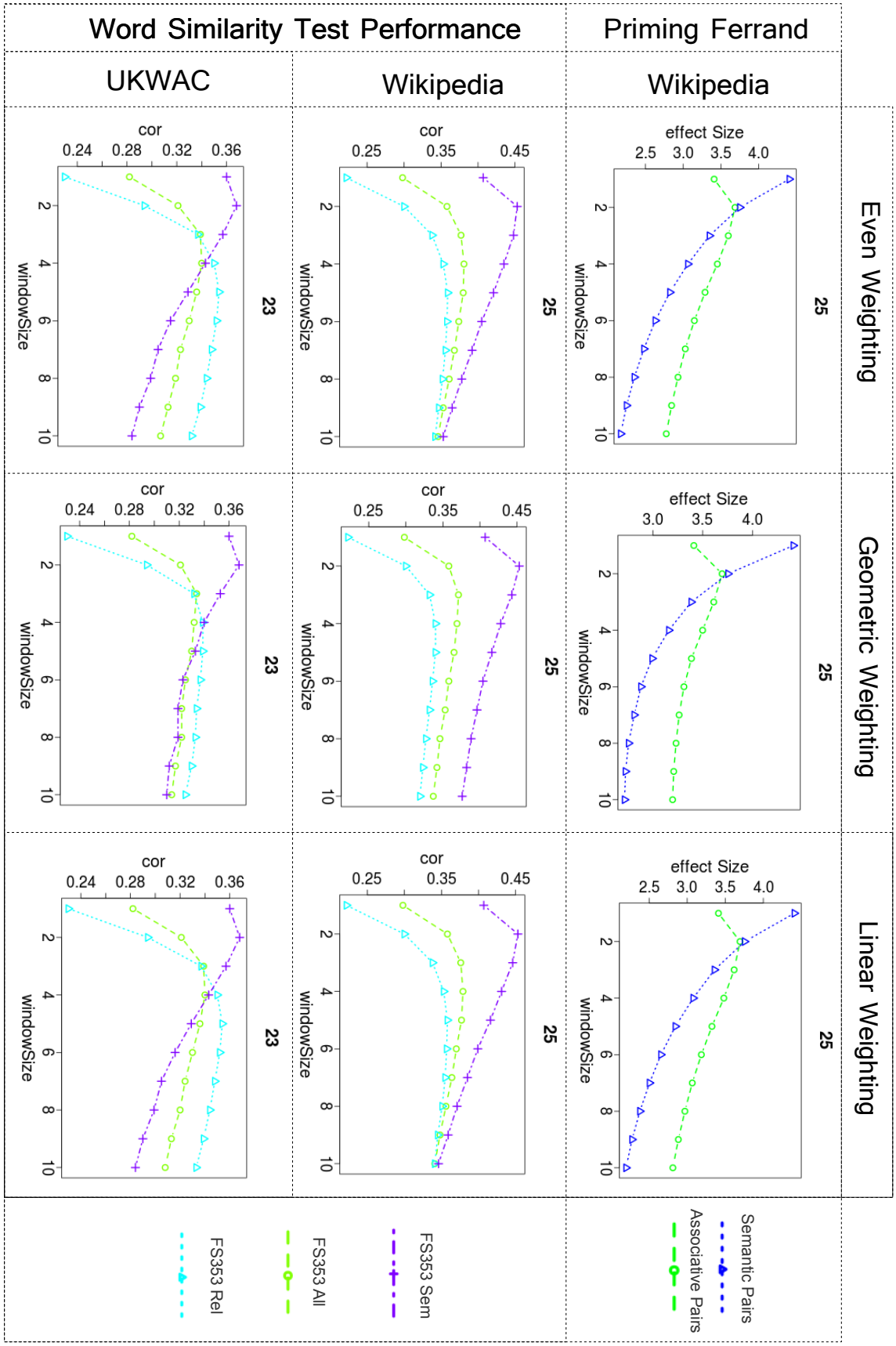


FIGURE 8.9: Word priming based analysis on basis of HAL models

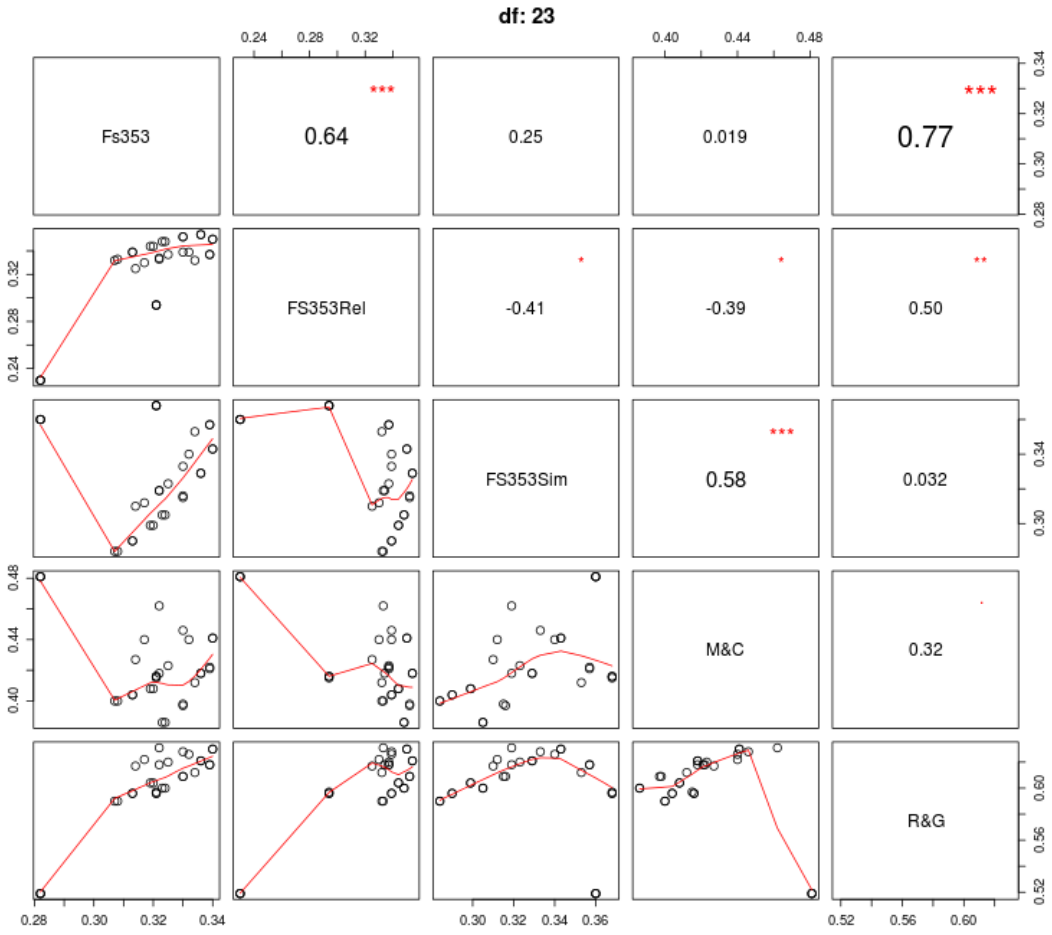


FIGURE 8.10: Correlogram of FS353, FS353-Sim, FS353-Rel, R&G, and M&C coefficients with HAL based models on grounds of UKWAC dataset

As was outlined in Section 7.4.1 the correlations of the FS353, R&G, and M&C on basis of the UKWAC:df23 collection showed considerable diversion. This was interpreted as contradictory evidence with regard to the validity of the M&C dataset. Since the ratings of the M&C dataset are interpreted to be accurate on basis of their very strong positive correlation with the R&G set, and on basis of the so far presented results the hypothesis can be formed, that the diversions of the coefficients of M&C are relationship type dependent. Of specific interest with regard to this hypothesis is therefore the correlation of the semantically focused (i.e. FS353-Sim) and associatively focused (i.e. FS353-Rel) datasets with the non-semantic-associatively classified sets. As can be seen M&C is indeed moderate to strong positively correlated with FS353-Sim and negatively correlated with FS353-Rel. The opposite situation is given for the FS353 and R&G datasets. To add substance to this observation the correlational analysis is extended to the Wikipedia:df25 collection. Figure 8.11

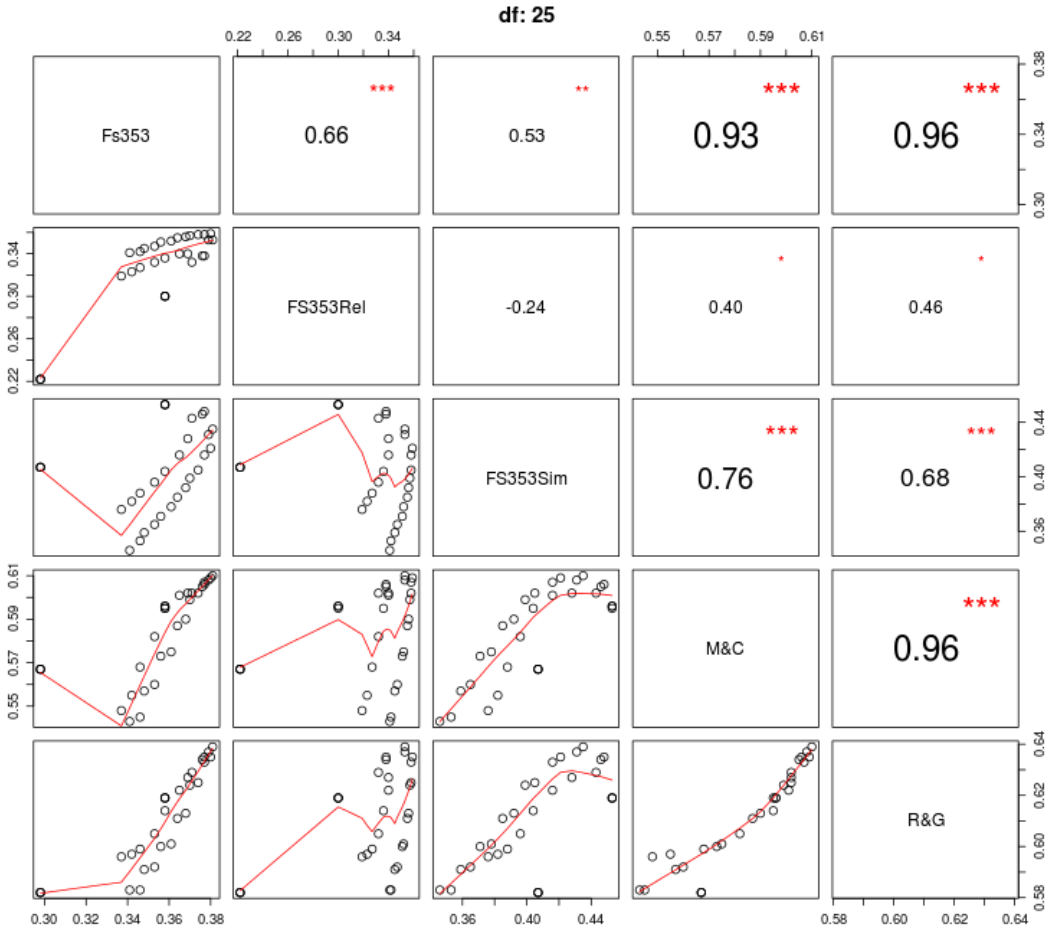


FIGURE 8.11: Correlogram of FS353, FS353-Sim, FS353-Rel, R&G, and M&C coefficients with HAL based models on grounds of Wikipedia dataset

As can be seen in principle the same observations can be made. M&C is more strongly correlated with the FS353-Sim dataset than are R&G and FS353 and vice versa with

regard to the FS353-Rel dataset.

A possible explanation for these observations can be formulated on basis of an inspection of the respective simulated priming coefficients and the assessment based coefficients for both collections. Figure 8.12 shows the respective mean simulated priming effects for both collections.

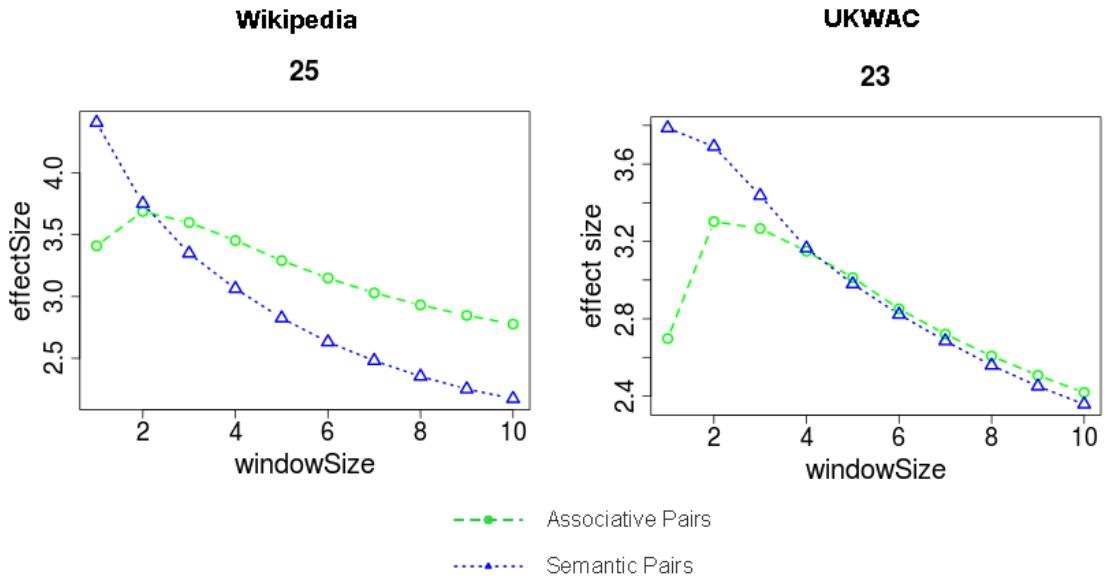


FIGURE 8.12: Simulated Priming effects on basis of Ferrand pairs.

As can be seen the priming effects distinctly differ at window size 1 with regard to the delta of the semantic and associative priming effect. The delta being much larger in case of the UKWAC collection. That means, the measured similarity of the underlying models was much higher for the semantic pairs than for the associative pairs at this window size.

An alignment with the co-efficient curves of the R&G and M&C collection shown in Figure 8.13 shows that this observed difference could potentially be attributed by a higher ratio of semantic pairs in the Miller & Charles dataset.

The reasoning being that if the M&C dataset mainly is comprised of pairs that exhibit a semantic relationship, then it seems plausible, that on basis of the HAL model based measurements at window size 1 being of semantic type, these measurements are more accurate for semantic pairs, and therefore result in a higher correlation co-efficient with the assessed similarities.

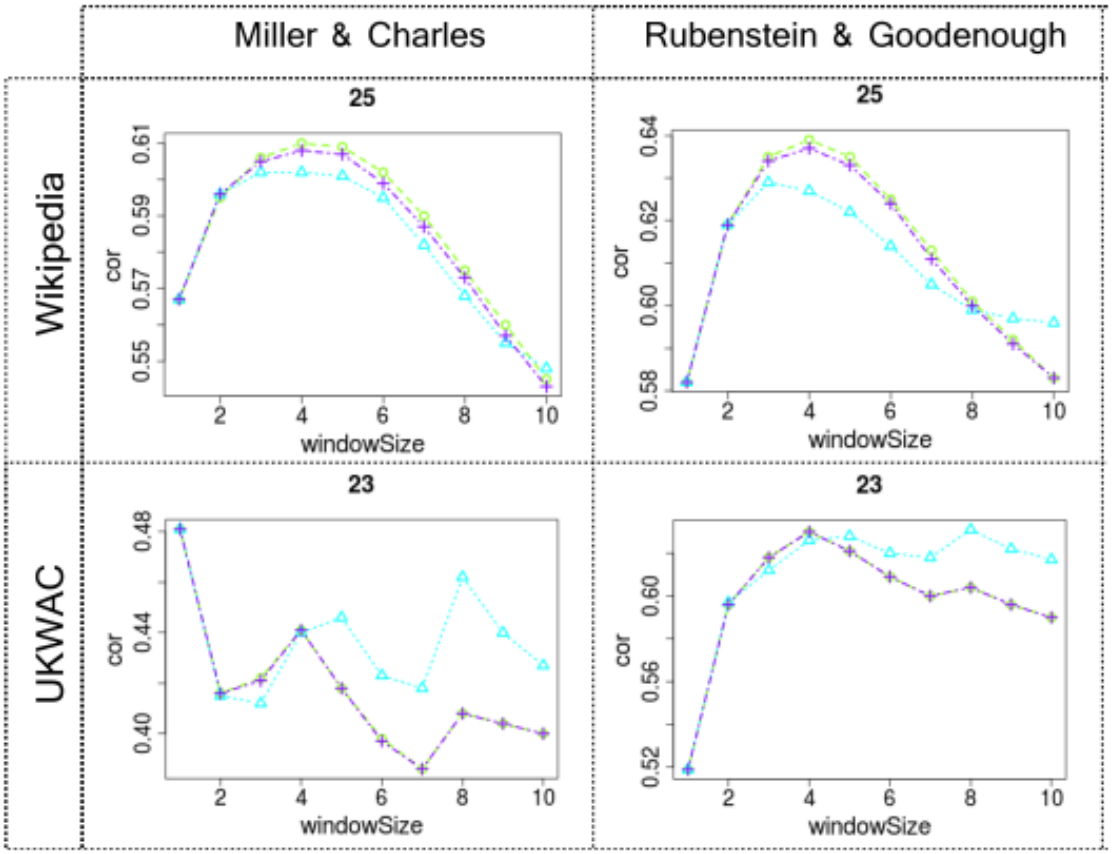


FIGURE 8.13: Plot of correlation coefficients of R&G and M&C collection on basis of HAL models

8.2.4 Discussion

Subsequently a summary with regard to the validity of the word similarity assessment data set is provided. On basis of the presented results it is stated that the FS353 derived semantic and related sets exhibited some degree of validity with respect to the differing definitions of word similarity types applied in this study.

As could be clearly seen the FS353-Sim datasets is sensitive to the higher 'semanticness' of lower window sizes. However, its behaviour is not completely concordant with priming simulation-based and neighbourhood assessment-based measurements. Specifically with regard to larger window sizes the exhibited trends of the curve seem to suggest that the pairs of the set also exhibit associative relations. To investigate this claim in more detail an alteration of the FS353-Sim dataset was conducted. The motivation for this re-assessment consisted of creating a dataset that is more consistent with the cognitive definition of semantic relatedness. To do so the unrelated pairs of the dataset were removed. The reasoning for this removal is , that in the case of unrelated items it is not clear which relationship can be attributed. Further, pairs that were assessed as synonymous but exhibited low similarity ratings (< 3) were removed.

A comparison of the coefficients on basis of the original FS353-Sim and the altered FS353-SimSt dataset is shown in Figure 8.14

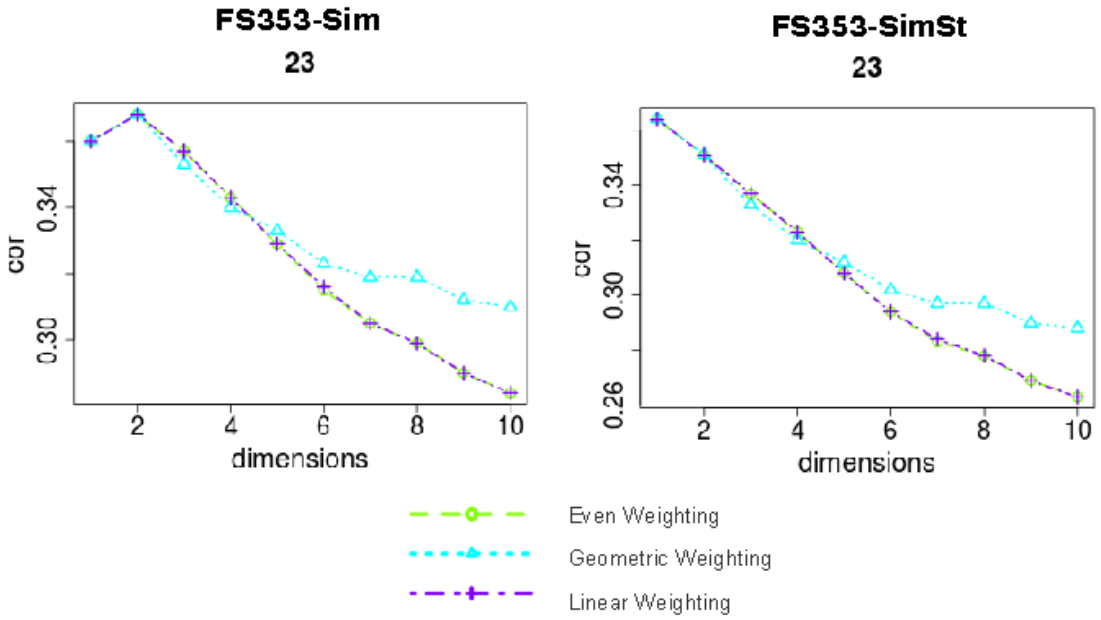


FIGURE 8.14: Comparison of FS353-Sim and FS353-SimSt coefficients on grounds of UKWAC:df23 collection.

As can be seen in the figure the removal of the above listed items results in distinctly different coefficients for the window size 1. The FS353-SimSt dataset, assumed to rep-

resent semantic relations, shows the more plausible trend with regard to the underlying HAL algorithm. Further the observed trend is more consistent with the observation of simulated priming effect based and neighbourhood assessment based observations. A recalculation on basis of dataset correlations on basis of the FS353-SimSt dataset also shows that the prior described observations with regard to the semantic focus of the M&C dataset are emphasized by this alteration of FS353-Sim. The respective results are shown in Figure 8.15.

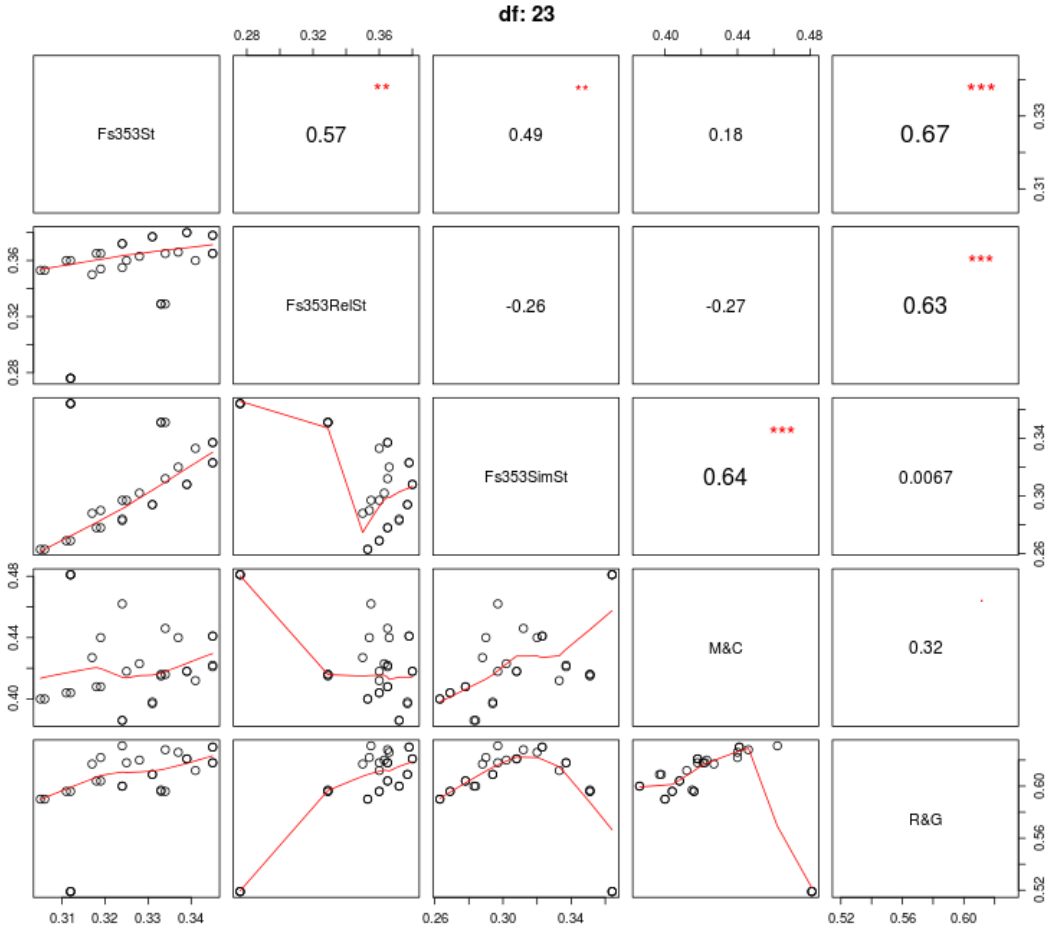


FIGURE 8.15: Correlogram of FS353, FS353-SimSt, FS353-Rel, R&G, and M&C coefficients with HAL based models on grounds of UKWAC dataset

On basis of this, the validity of the FS353-Rel and FS353-Sim datasets as measurement instruments of the semantic or associative degree of a relationship is interpreted as limited. Specifically the inclusion of unrelated pairs in both sets seems to be problematic with regard to this. While as shown on basis of the observations of the FS353-St dataset, a removal of such pairs results in higher concordance with other kinds of measurements, it can be assumed that the creation of a 'cognitively' valid set requires additional steps. The first of these consisting of a re-assessment of all pairs with respect to semantic only and semantic-associative relations as conducted in the Ferrand study. This is not

to be interpreted as a criticism of the applied methodology of the FS353-Sim dataset. It merely results from the fact that control for associative relations does not form part of the linguistic definitions with regard to synonymous similarity.

Conclusion

With regard to the validity of the available measurement instruments the following can be said.

The high concordance of the priming simulation effects on basis of the Ferrand dataset with the neighbourhood assessments constitutes strong supportive evidence for the validity of both procedures. Further supportive evidence as outlined stems from the plausibility of the observed trends with respect to the mechanics of the HAL algorithm.

With regard to the validity of the word similarity assessment based sets the situation is less straightforward. As was outlined on basis of the results, the FS353-Sim procedure exhibits sensitivity with regard to the semantic-associative degree of HAL model based measurements. The observed trends of the respective curves however are only partly concordant with the observations of the neighbourhood assessments based and simulated priming effect based measurements. This represents contradictory evidence with regard to the validity of the dataset as an instrument to measure semantic or associative degree on basis of the cognitive interpretation of these concepts. This does not rule out the potential validity of the underlying assessment procedure itself. As argued before, the differing observations are mainly attributed to differences of the underlying definitions of relationship types.

Concluding, it can therefore be stated, that the priming simulation based procedure constitutes the most appropriate measurement instrument with regard to its applicability to CPMs and the supportive evidence concerning its validity. The next section outlines the application of the so far presented alignments on LSA based models.

8.3 LSA Based Alignment

Subsequently the alignment on basis of LSA computational models is reported. The underlying methodologies and interpretation of the observations is fundamentally identical to those described in course of the presentation of HAL model based alignment.

8.3.1 Priming Based Alignment

A first step with regard to the evaluation of the validity of the priming simulation procedure consists of an analysis of the semantic-associative degree of specific LSA models on basis of the Ferrand and Chiarello word pairs. As noted at the beginning of this chapter LSA is, on grounds of its algorithm, commonly interpreted to be of associative nature.

Figure 8.16 shows plots of the mean simulated priming effects for the Ferrand word pairs. The shown plots are based on two Wikipedia:df representations and plot the priming effects with respect to the underlying transformation function of the models over the dimension parameter.

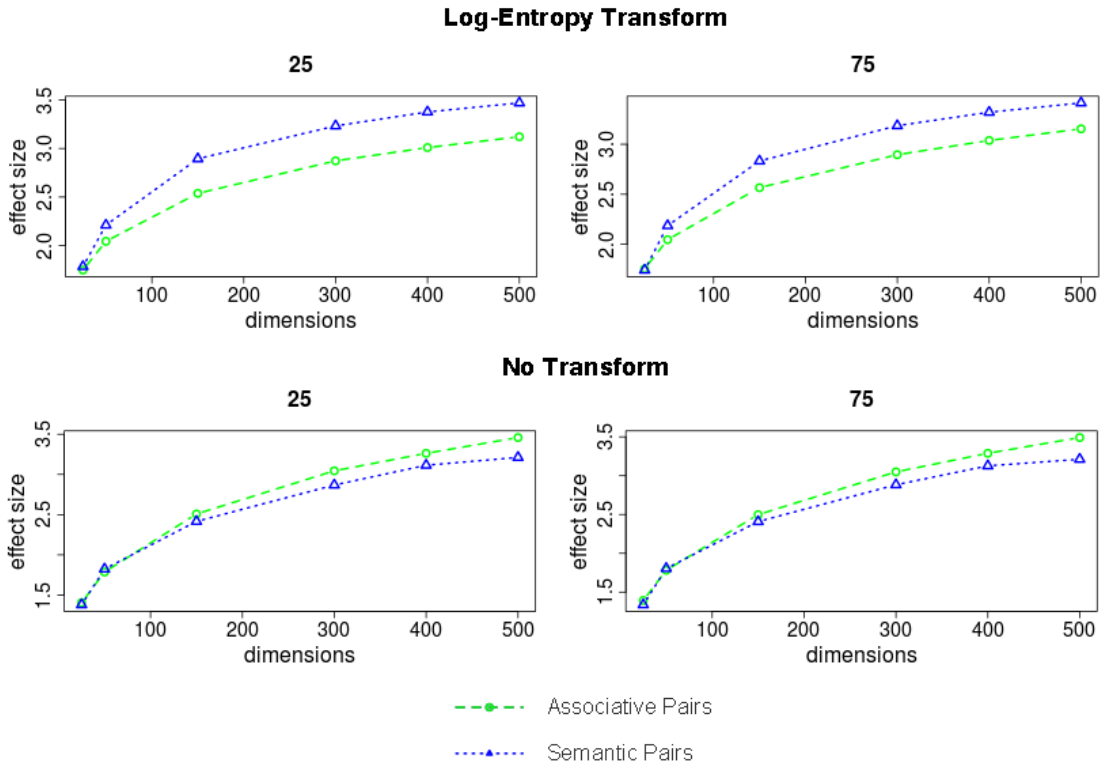


FIGURE 8.16: Mean simulated priming effects, measured on basis of Ferrand word pairs, of LSA models based on Wikipedia collection.

As can be seen on basis of the plots the semantic-associative degree of the model's measurements appears to be primarily dependent on the transformation function. Contrary to the HAL based model's window size parameter, the dimension parameter seems to exhibit less impact on the semantic-associative degree. As stated before, the fact that the semantic curve in the top-left plot is above the associative curve cannot be interpreted directly as Log-Entropy based models being primarily semantic.

This becomes more evident when examining the priming effects based on Chiarello's

word pairs. Figure 8.17 shows the respective plots.

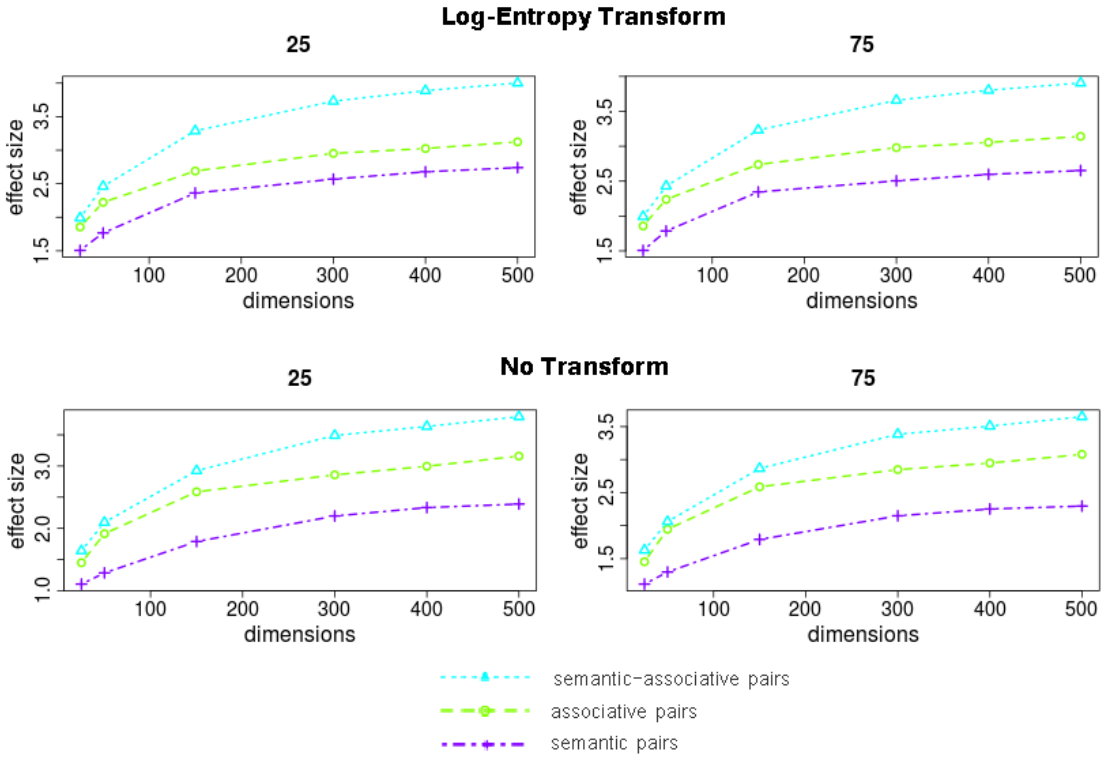


FIGURE 8.17: Mean simulated priming effects, measured on basis of Chiarello word pairs, of LSA models based on Wikipedia collection.

In case of the Chiarello based priming effects it can be seen that the associative curves are generally above the semantic curves. This, as noted before, directly results from the choice of word pairs and is also reflected in the experimental priming results. This confirms that it is not possible to assert that a specific LSA model is associative or semantic in nature solely on basis of such data. However, the priming effects from both figures indicate that in both cases that the Log-Entropy transformation results in *more* semantically focused models. The concordance of this observation for both datasets contributes a first form of supportive evidence for the validity of the priming simulation procedure. Additional supportive evidence in case of the Chiarello based effects consists of the clear visibility of the associative-semantic boost in all plots. The observations are in this regard therefore plausible with regard to the underlying cognitive theory. To investigate the aspect of the higher semantic focus the next section explores the alignment of LSA model output with neighbourhood assessments.

8.3.2 Neighbourhood Assessment Based Alignment

Figure 8.18 shows a plot of the respective neighbourhood assessment based ratios for the Log-Entropy and No-Transform functions.

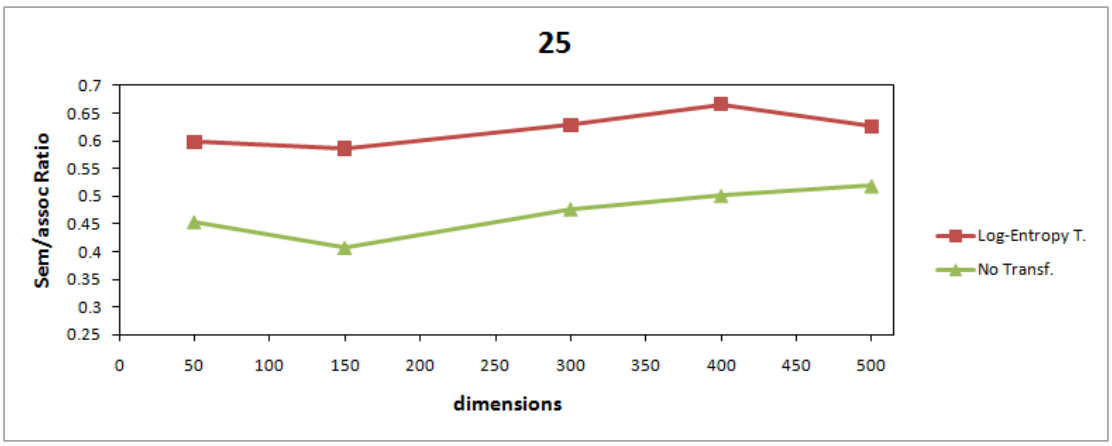


FIGURE 8.18: Neighbourhood assessment based sem/assoc ratios for Log-Entropy and No Transformation based LSA models on Wikipedia:df25 collection.

The observations are consistent with the simulated priming effect based observations in that the Log-Entropy based models are 'more semantic' than those based on no transformation. As can be seen the impact of the dimension parameter on the semantic/associative aspect is small. With regard to a comparison with the HAL based observations it can be remarked that LSA models indeed seem to be of more associative nature than HAL models based on small window sizes. Finally it can be noted, that on basis of the so far presented results, LSA models in general seem to exhibit less variation along the semantic-associative axis compared to the HAL models. This is plausible with regard to the nature of the underlying parameters of each model. As a final evaluative step on basis of LSA models with regard to the validity of the priming simulation method, the next section explores the alignment of word similarity assessment based measurements.

8.3.3 Word Similarity Assessment Based Alignment

Figure 8.19 contrasts priming simulation based measuremens and word similarity assessment based measurements. The plots in the figure are consistent with the prior observations. Larger priming effects for semantic pairs are reflected by the relative performance of the FS353-Sim dataset. This concordance provides additional supportive evidence for the validity of the priming simulation method.

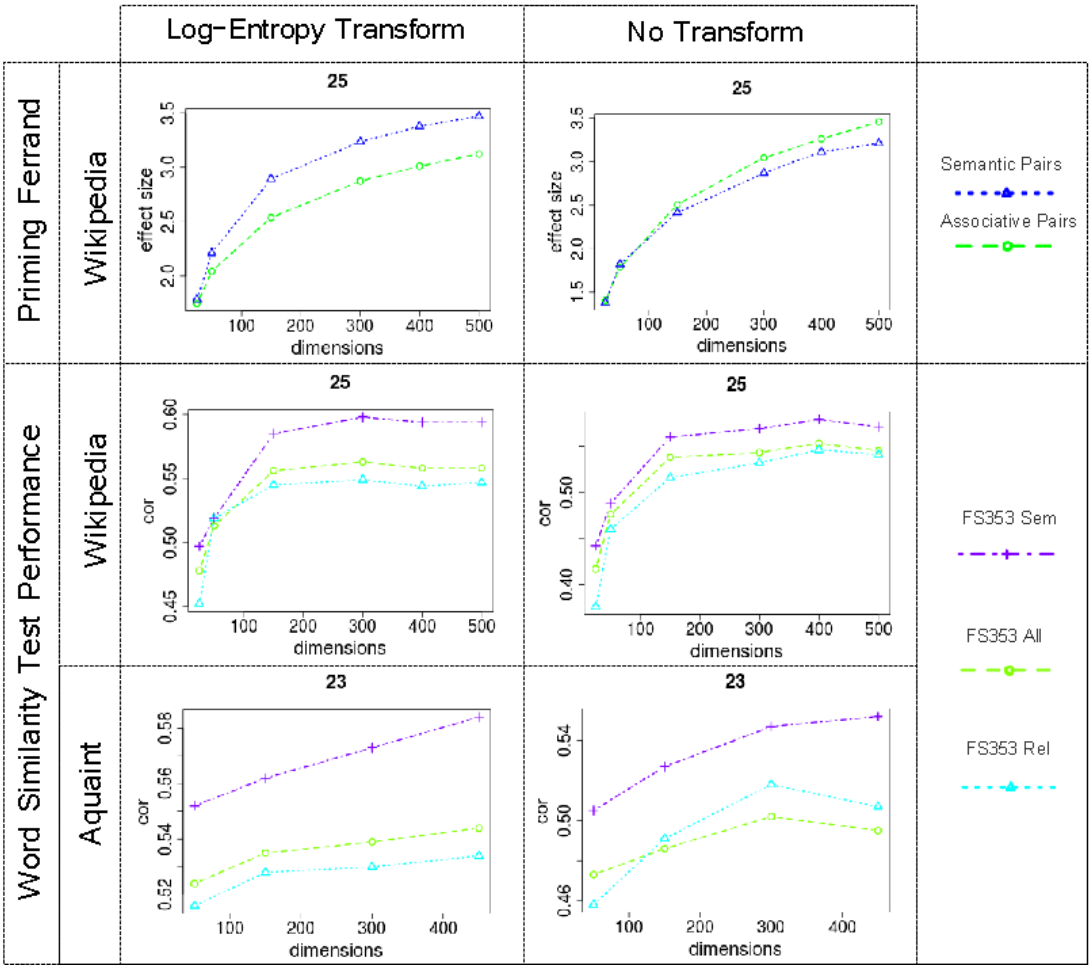


FIGURE 8.19: Simulated priming effects on basis of Ferrand Pairs

8.3.4 Discussion

The presented results of alignments on basis of LSA models can be summarized in the following form.

The shown concordance of the observations on basis of simulated priming effects with neighbourhood assessment based data are interpreted as strong supportive evidence with regard to the validity of both procedures. Secondly the concordance with regard to the observations based on word similarity assessment based measurements additionally is supportive of the validity of the priming simulation method.

8.4 Chapter Conclusions and Answer to RQ 5

On basis of the reported results the subsequent conclusion summarizes the findings of the conducted validation studies with respect to the following two aspects. On basis of the reported results concerning the alignment of different measurement instruments with HAL and LSA based computational models, the following answer to RQ 5 can be formulate.

RQ 5 What are valid instruments for the measurement of the type of relatedness between words?

The supportive evidence for validity is largest for the neighbourhood assessment and priming simulation methods. This observation is grounded on the concordance of the respective measurements and the plausibility of these measurements with the algorithmic characteristics of the models and theoretic observations from cognitive psychology.

With regard to the pragmatic criteria defined in Section 5.2.4, the following can be said. From a practical point of view the priming simulation method presents a more viable means due to the large required effort of assessing word neighbourhoods. A noted limitation concerning measurements based on these instruments concerns the validity of the magnitude of its measurements. As outlined before the magnitude of a measurement can, on grounds of its expression in standard deviations, only be interpreted relative to the underlying computational model. An interpretation of the form, that a specific HAL model's measurements are more semantic than another HAL model's measurements is in this sense interpreted to be a valid inference. However the interpretation that a specific LSA model's measurements are more semantic than a specific HAL model's measurement is not possible on grounds of the magnitude of the measurements.

The presented results can be interpreted to have a distinct impact with regard to the measurement of word similarity in general. The findings presented in this chapter clearly

indicate that the type of word relationships has to be taken into consideration when making inferences with regard to the correlation of a computational model and assessment based similarity ratings. Within the reported results the importance of this aspect is evidenced by the diversion of the correlation coefficients of the M&C and R&G datasets across the HAL parameter space. Inferences on basis of such approaches should, with regard to the findings, not be drawn in disregard of the semantic/associative aspect. To our knowledge this aspect however is only partially or not at all reflected in the contemporary literature[8-2].

Based on the validated measurement instruments identified in the last two chapters, the next step consists of the concrete application of the IR focused nomological network. This aspect is explored in the next chapter.

[8-2] See Krcmár et al. (2011), Recchia and Jones (2009), Jarmasz and Szpakowicz (2003), Gabrilovich and Markovitch (2005), Budiu et al. (2007), Budanitsky and Hirst (2001), Agirre et al. (2009)

CHAPTER 9

RELEVANCE AND WORD SIMILARITY

Part I was dedicated to the investigation of theoretical aspects relating to the problem statement of the dissertation. Chapter 2 addressed RQ 1, and identified construct validity and the nomological network as a principled approach to the validation of relevance. It outlined that two main research questions are implied by the aim of constructing an IR focused nomological network. The first is given by RQ 2 and concerns the identification of the set of candidate constructs for the network. The second question focuses on the identification of criteria for choosing a set of constructs from the pool of candidates, and is addressed by RQ 3. The analysis in Chapter 4 was guided by RQ 2. It identified a potential pool of candidates based on the investigation of the cognitive models of discourse comprehension and reasoning. RQ 3 was investigated in Chapter 5. The analysis resulted in a set of criteria for the choice of constructs was introduced. Based on these criteria, Section 5.3.2 defined an IR focused nomological network consisting of the constructs of relevance, grade of word relatedness, and type of word relatedness.

Part II was dedicated to the empirical investigations relating to the application of the defined network. A prerequisite for the application of the network consisted of a validation of the measurement instruments for grade and type of word relatedness. These research aspects are represented by RQ 4 and RQ 5 of the dissertation. The theoretic background for the validation of these constructs was presented in Chapter 6. Chapter 7 conducted a validation study relating to the grade of relatedness between words. It identified the FS353 word similarity assessment procedure as the instrument with the highest validity for measuring the grade of relatedness between words. In Chapter 8 a validation study focused on the type of relations between words was presented. The identified instrument with the highest validity consists of the priming simulation methodology developed on basis of the Ferrand word pair sets. The listed investigations constituted the necessary steps for the construction of an initial IR focused nomological network.

The concluding chapter of the dissertation is dedicated to an exemplary demonstration of the application of the network as part of the aim to establish construct validity for relevance. This is based on an empirical exploration of the relation between the construct of grade of word relations, the construct of type of word relations, and the construct of relevance.

RQ 6 What are characteristics of the relation of the postulated constructs of relevance and grade and type of word relationships?

Approaching RQ 6 is pursued based on formulating two subquestions.

RQ 6a Does a relation between the postulated constructs of relevance and grade and type of word relationships exist?

RQ 6a is dedicated to the empirical verification of a relation between the three constructs. The construction of the nomological network introduced in Section 5.3.2 is based on the assumption, that the constructs are related. This assumption was formulated based on the analysis of the cognition of discourse comprehension and decision making in Chapter 4. RQ 6a empirically investigates this assumption.

RQ 6b What are characteristics of the relation between the postulated constructs of relevance and grade and type of word relationships?

RQ 6b is dedicated to the identification of characteristics of the relation between the constructs. This constitutes an example of the application of the central tenet of construct validity: to base the validation of constructs on an investigation of the relation between constructs.

The investigations in the chapter are structured in the following way. Section 9.1 investigates the contemporary conceptions of word relatedness in IR. This exploration is based on axiomatic definitions of IR models, and aims to serve as a basis for the analysis of the empirical results of the chapter. A detailed overview of the experimental setup underlying the empirical explorations is presented in Section 9.2. Section 9.3 describes the results, addressing RQ 6a and RQ 6b. We end with a discussion of the investigations of the chapter in Section 9.4.

9.1 Conceptions of Word Relatedness in IR

Subsequently the implicit assumptions underlying state of the art relevance estimation models (i.e. retrieval models) are outlined. This aims at providing a basis for relating the empirical results of the chapter to the formal theoretical background of IR theory. The exploration is based on the axiomatic definitions by Fang and Zhai (2005). As noted by Fang and Zhai (2005), the axiomatic approach aims at formalizing central assumptions underlying IR models.

9.1.1 Relevance Estimation Functions

On basis of their axiomatic approach to IR, Fang and Zhai (2005) formally defined the underlying assumptions of IR models. Their definitions are re-iterated in order to gain a foothold on which to base the discussion.

Following the terminology used by Fang and Zhai (2005) T is defined as the set of all terms. Queries and documents are interpreted as bags of terms; formally expressed as query $Q = \{q_1, ..., q_n\}$ and document $D = \{d_1, ..., d_m\}$, where $q_i, d_i \in T$, and it is possible that $q_i = q_j$ and $d_i = d_j$ even if $i \neq j$. Based on these formal definitions Fang and Zhai (2005, p. 481) take the following approach:

> *Our goal is to define a scoring function $S(Q,D) \in \mathbb{R}$. To help us search through this function space efficiently and define meaningful constraints on the retrieval functions, we propose to define a retrieval function inductively. We start with the base case, when both the document and query contain only one term.*

The purpose of function S can, with regard to the Correlation to Cognition analogy, be interpreted as the attempt to achieve maximum correlation with user estimates of the relation between query Q and a document D. Base Case: assume $Q = \{q\}$ and $D = \{d\}$.

$$S(Q,D) = \begin{cases} weight(q) = weight(d) & \text{when } q = d \\ penalty(q,d) & \text{when } q \neq d \end{cases}$$

A primitive weighting function f can then be defined on basis of function S as described by Fang and Zhai (2005, p. 481) as follows:

> *Function f gives the score of a one-term document and a one-term query and will be referred to as the Primitive weighting function. It rewards the document with a score of $weight(q)$ when d matches q and gives it a penalty score of $penalty(q,d)$ otherwise. We will reasonably assume that $\forall t \in T$, $weight(t) > 0$ and $\forall q, \forall d \neq q, penalty(q,d) < weight(q)$.*

As implied by the primitive weighting function, the case $q = d$ results in $weight(q)$. This expresses the idea of assigning varying importance to the occurrence of this event. Estimating the importance of this event forms a major part of research in IR. A detailed axiomatic overview of the most prevalent techniques is provided by Fang and Zhai (2005) in the same publication. With respect to the explorations' focus on word relatedness, the next subsection aims to outline the implicit assumptions underlying the case $q = d$.

9.1.2 Similarity Based on Graphemically Identical Encoding

A principle assumption underlying the definition of the axiomatic relevance estimation function f is expressed through the statement:

$q = d$

While a necessary condition for $q = d$ is not explicitly stated, it is reasonable to assume that it consists of requiring graphemically identical encoding of the two terms q and d.

A: IF q and d are graphemically identical then $q = d$

A commonly applied generalization of this condition consists of the application of stemming algorithms. Stemming a word results in the removal of common morphological and inflexional endings (Porter, 1980).

B: IF the stemmed representations of q and d are graphemically identical then $q = d$

A and *B* can be interpreted as basic presumptions of the mechanics of human text based processing. Within most IR applications, *A* represents a sufficient condition for supposing a relation between a query Q and a document D. This is grounded in the term independence assumption underlying the inductive approach described by Fang and Zhai (2005). *A* and *B* constitute the primary conditions underlying the estimation of relationships between textual entities (i.e. queries and documents). The conditions represent simplified approximations to human information processing. They can lead to false estimates in the case of homographs. Further, the rather stringent condition induces the problem of vocabulary mismatch (Furnas et al., 1987). The next subsection explores the extension of the graphemic equality condition with regard to these limitations.

9.1.3 Word Relationships

An extension of the Graphemic equality condition is described in axiomatic form by Fang and Zhai (2006). To illustrate the point, the same terminology as applied by Fang will be used.

$s(t,u) \in [0,+\infty]$ is defined as a semantic similarity function between two terms t and u. As defined by Fang and Zhai (2006), larger values of s represent higher similarity. Assuming that term t is closer related to term u than to term v, this should be reflected by the similarity function as $s(t,u) > s(t,v)$. Based on this similarity function, Fang and Zhai (2006) define three retrieval function constraints pertaining to word relatedness.

> “ *STMC*1 : *Let* $Q = q$ *be a query with only one term* q.
> *Let* $D_1 = d_1$ *and* $D_2 = d_2$ *be two single-term documents, where* $q \neq d_1$ *and* $q \neq d_2$. *If* $s(q,d_1) > s(q,d_2)$, *then* $S(Q,D_1) > S(Q,D_2)$. ”

*STMC*1 expresses that a retrieval functions should assign a higher score to the document containing the closer related term. Even though neither document D_1 nor D_2 graphemically match the query Q, document D_1 is assigned a higher score because term d_1 is more closely related to query term q than term d_2. *STMC*2 requires that the match of a original query term is always scored higher than the matching of a related term, even if the related term occurs more frequently than the original term. *STMC*3 captures the assumption, that n matches of an original term and a related term should be scored higher than n matches of just an original term.

The defined constraints express fundamental assumptions with regard to the relation between word relatedness and relevance. *STMC*1 indicates that the existence of relatedness between the terms of two information items (i.e. the query and the document) positively influences the estimation of relevance between those two items. *STMC*2 assumes that graphemical concordance has a higher impact on the estimation process. Finally, *STMC*3 is based on the assumption that graphemical concordance and relatedness between the terms of two information items is a stronger indicator of relevance than graphemical concordance by itself. These initial assumptions provide a foothold for the interpretation of the empirical results of the chapter. Fang and Zhai (2006, p. 119) note that the 'remaining challenge is to define s(t1, t2) in STMC1'. Concerning this point they note that 'co-occurrences of terms obtained from the analysis of a document collection usually reflect underlying semantic relationships that exist between terms'. The terminology used to describe relationships between terms is 'semantic relationships' (Fang and Zhai, 2006, p. 119). The authors do not specify the intended meaning of the term 'semantic'. It is unclear if it constitutes an expression of semantic similarity in a cognitive sense (see Section 6.1), or semantic similarity in a lexical sense (see Cruse (1997)). Within the interpretations of this chapter, it is assumed that the use of 'semantic' was intended to generally refer to relatedness between words. Based on this interpretation, the constraints provide a formalization of the characterization of the relation between relevance and word relatedness in IR.

9.2 Experimental Setup

The focus of this chapter consists of demonstrating the application of an IR focused nomological network as a means of validating relevance. The network defined in Section 5.3.2 consists of the constructs of relevance, grade word relatedness, and type of word relatedness. As outlined in Chapter 5, the choice of the two word related constructs is motivated by their higher measurability.

The application of these constructs as part of the validation of relevance is based on an investigation of their relation with relevance. Establishing a 'grip' on the meaning of a construct by investigating its relation to other constructs constitutes the central tenet of construct validity. That is, by examining the relation between word relatedness and relevance we aim at gaining a better understanding of the meaning of relevance. The basic approach to an investigation of the relation between the constructs consists of inducing variations in one construct and measuring the observable variations in the focused construct. Ideally, the investigation of the relation between the constructs would be conducted in vivo. Constraint on observation and manipulation of the mind exclude this possibility. As outlined throughout the discussion in Chapter 5, an alternative approach in cognitive science is given by basing the investigation on computational models of constructs. Based on the Correlation to Cognition analogy defined in Section 2.2, this can be applied to the network in the following way. The analogy defines the aim of IR systems as achieving maximum input-output concordance with the cognitive estimates of relevance. Essentially, this portrays IR systems as computational models of the relevance estimation process. The result of computing the relation between a query and a document constitutes the system's estimate of relevance.

The empirical investigation in this chapter builds upon this interpretation. HAL and LSA constitute computational models of word relatedness. The applied retrieval system represents a model of the relevance estimation process. Modification of the HAL and LSA models is used to induce variations with regard to word relatedness. This is implemented by altering the query representation. Measuring the effect on the relevance estimation process is based on evaluating the effect of query alterations on retrieval runs using standardized test collections. Based on this approach, the experimentation within this chapter aims at investigating how alterations based on relatedness of words affect the estimation of relevance.

The following subsections describe the elements comprising the experimental setup underlying the empirical investigations of this chapter.

9.2.1 Test Collections and Retrieval Tasks

The selection of the test collections in the experiments is motivated by the aim of investigating the relations of the constructs with respect to different domains, queries, and document types. Subsequently summaries of the collection statistics and the associated retrieval tasks are provided.

.GOV

The .GOV collection is a Web based TREC test collection based on crawl of the .Gov domain in the beginning of 2002. It was conceived as an update for the WT10G collection that was based on a subset of a 100GB crawl by the Internet Archive conducted in 1997. In contrary to the WT10G subset, the .GOV crawl constitutes an 'unedited' (Soboroff, 2002) crawl of a portion of the Web and as such is more closely connected than WT10g (Soboroff, 2002). The collection consist of 1,053,372 documents with the mime type text/html. Documents in the collection were truncated at a size of 100kb. The total size of the collection is 18G.

Within the empirical investigations of this chapter, the associated mixed query retrieval task application. The task combines queries of three different types. Descriptions of these tasks are provided in the following paragraphs. Further details regarding these tasks are provided by Craswell and Hawking (2004).

2004 Mixed Query Web Retrieval Task (WT2004MQ) The mixed query task is a combination of the three subsequently summarized tasks. The task was designed to provide a set of queries with differing goals and challenges. The total amount of queries in the mixed query task is 225. The total number of relevant sites for the task is 1763.

2004 Home Page Web Retrieval Task (WT2004HP) In the home page retrieval task a query consists of the name of a site that the user wants to find. Examples for such queries are given by 'Togo embassy' or 'Baltimore'. The retrieval system is expected to deliver the URL of the sought after site at the highest possible rank (i.e. ideally at rank one). Particular to the WT2004HP task is the low amount of relevant documents. The total number of relevant sites was 83. The number of queries is 75.

2004 Named Page Web Retrieval Task (WT2004NP) The named page finding task is similar to the home page finding task. The goal of the task consists of retrieving a specific web site. In contrary to the home page finding task, the sought after web site does not constitute the main home page of a domain. An example query for the task is given by 'Ireland consular information sheet'. The retrieval system is expected to return the sought after site at the highest possible rank. The total number of relevant sites is 80, and the number of queries is 75.

2004 Topic Distillation Web Retrieval Task (WT2004TD) In the topic distillation task, the queries describe general topics. Examples for such queries are given by 'Amer-

ican music', 'substance abuse', and 'money laundering'. The retrieval system is expected to return the home pages of relevant sites (i.e. sites that provide information relevant to the topic of the query). The total number of relevant sites is 1600, and the number of queries totals 75.

AQUAINT

The Aquaint collection consists of 1,033,000 text documents. All contained documents are newswire documents. The collection consists of documents from three sources: Associated Press (AP) newswire from 19982000, the New York Times newswire from 19982000, and the English documents from the Xinhua News Agency dating from 19962000. More detailed information on the collection and the below described Robust task are provided by Voorhees (2006). The Aquaint collection is associated with the Robust retrieval task.

Robust 2005 Task (robust05) The Robust retrieval task is an ad hoc retrieval task. It is aimed at mimicking the retrieval of information from a library. Retrieval is based on a static set of text documents and a set of queries. The queries are supplied in form of topics. The topic set for the 2005 Robust task consisted of 50 topics. These topics had been used in ad hoc and robust tracks in previous years. Common to all these topics was a low median average precision scores in earlier TREC tasks. The topics are assumed to represent difficult queries. Example queries of the task are given by 'Hubble Telescope Achievements' and 'Most Dangerous Vehicles'.

TREC Disk 4&5 without Congressional Records (Disk4&5-CR)

The TREC Disks 4&5 collection consists of 528,000 newswire and Foreign Broadcast Information Service articles. It constitutes a combination of two TREC disks. Disk number four includes material from the Financial Times Limited (1991, 1992, 1993, 1994), the Congressional Record of the 103rd Congress (1993), and the Federal Register (1994). Disk number five contains documents from the Foreign Broadcast Information Service (1996) and the Los Angeles Times (1989, 1990). For the task of the empirical evaluation the content of both disks excluding the congressional records constitutes the test collection. This collection is often referred to as 'Disk4&5-CR'

TREC 7 Ad Hoc Task (Trec7) Trec7 is an ad hoc task for the Disk4&5-CR test collection. It consists of 50 topics labelled with the number 351-400. Topics consists of three main elements: a title element, a description, and a narrative. The title element contains a short query. Example queries for the task are given by 'Falkland petroleum

exploration' and 'ocean remote sensing'. The description provides the query terms in form of a sentence. The narrative provides additional information regarding the criteria for relevance.

TREC 8 Ad Hoc Task (Trec8) The Trec 8 ad hoc task is based on the same methodology and characteristics as Trec7 and supplies a new set of 50 topics labelled with the numbers 401-450.

9.2.2 Retrieval Model

The retrieval model applied throughout the investigations is given by a standard BM25 retrieval function. The formula as presented in Fang et al. (2004) is as follows:

$$\sum_{w \in q \cap d} \left(\ln \frac{N - \mathrm{df}(w) + 0.5}{\mathrm{df}(w) + 0.5} \times \frac{(k_1+1) \times c(w,d)}{k_1((1-b) + b\frac{|d|}{avdl}) + c(w,d)} \times \frac{(k_3+1) \times c(w,q)}{k_3 + c(w,q)} \right)$$

Where b , k_1, k_3 constitute parameters of the model. The parameter k_3 is usually set to a constant 1000. Standard parameter values for b and k_1 are given by 0.6 and 1.2 respectively.

BM25 Okapi represents one of the most successful and widely adopted retrieval models. The BM25 model can be considered as the standard retrieval model in the last two decades (Lv and Zhai, 2011). It serves as a commonly applied model for the generation of baselines in IR. The widespread adoption of the model is also evidenced by its implementation by almost all research focused retrieval systems [9-1] and commercial search systems[9-2].

Apart from pragmatic considerations, the motivation for the choice of the BM25 model is based on the considerations of Section 5.3.1. Section 5.3.1 outlined the role of computational models in cognitively focused nomological networks. Within the empirical evaluation of this chapter, the applied retrieval model is interpreted as a computational model forming part of the relevance estimation process. Since the details of the estimation process are unknown, a plausible heuristic for the selection of such a model seems to consist of choosing the most widespread retrieval model.

[9-1] mg4j.dsi.unimi.it, www.lemurproject.org/indri/, terrier.org, www.lemurproject.org

[9-2] lucene.apache.org/solr/,www.elasticsearch.org

9.2.3 Retrieval System and Preprocesssing

The utilized system for all reported experimentation consists of the MG4J retrieval system (Boldi and Vigna, 2005). The preprocessing of all utilized test collections matches the setup applied for the generation of the word spaces described in Section 6.5. That is, stemming using the Porter stemming algorithm (Porter, 1980) and Stop wording based on the list supplied by Rijsbergen (1979) was applied to all test collections.

9.2.4 Query Expansion Based Integration

Investigating the relation between the set of constructs requires to introduce variation in word relatedness. Modification of the HAL and LSA models is used to vary with regard to word relatedness. The integration into the retrieval process is based on altering the query representation. The applied mechanism for such an alteration is based on query expansion. The following approach is implemented.

Given a set of queries $Q = (q_1, .., q_n)$, where each query q_n is comprised of a set of terms $(t_1, ..., t_m)$, each term t_m is expanded with its K closest related terms. The estimation of the relations of t_m with other terms is based on the available computational models (i.e. HAL, LSA). In this form the integration of the computational models is independent of any specific retrieval models. With regard to a cognitively focused interpretation such an expansion of query terms can be interpreted as a form of activation of related words as described by Swinney (1979). With reference to 9.1.3 it is clear that such an approach is very similar to the 'semantic' matching based retrieval as proposed by Fang and Zhai (2006). Within the subsequently reported experiments a deliberately simplistic retrieval function is applied. The score of a document is calculated as

$$\sum_{1}^{n} s(t_n)$$

where $s(t_n)$ is calculated as:

$$w(t_n) + \beta * \sum_{1}^{k} w(t_{nk}) * r(t_{nk})$$

where $w(t_n)$ is the retrieval function specific weighted score of query term t_n, $w(t_{nk})$ the equivalent score for the k^{th} related expansion term of n, $r(t_{nk})$ is the measured relatedness on basis of the computational model, and β a simple weighting parameter that allows to assign a weight to the scores of the expansion terms. The estimates of the computational models are integrated into the retrieval process in two ways: firstly in form of the chosen expansion terms, and secondly through integrating the measured relatedness into the retrieval process. As such the function is agnostic with regard to specific retrieval methods. This is considered favourable with regard to the investigation

of the relation between word similarity effects and relevance. The choice of a simplistic integration is motivated by the considerations brought forward in Section 5.3.1. As outlined in the section, interpreting measurements becomes more difficult with an increase in the complexity of the system under consideration. The focus with regard to the setup of the retrieval system is set on limiting the complexity of the setup.

9.2.5 Evaluation

Retrieval runs are evaluated based on the relevance assessments provided with the task topics. These assessments are available in form of 'qrel' files. The files contain a listing of relevant documents on a per query basis. Based on these files it is possible to automatically calculate retrieval measures. This is achieved based on matching the unique identifiers (also called 'docnos') of the relevant documents listed in the qrels, and the document set retrieved by the system. All performance measures reported in this chapter are calculated with the TREC Eval tool version 8.1[9-3]

A second step of the evaluation consists of evaluating the significance of retrieval run results. In IR such evaluations are usually based on the formulation of a null hypothesis H_0 that assumes that the tested retrieval models are equivalent in performance. As stated by Hull (1993, p. 333), a 'significance test will attempt to disprove this hypothesis by determining a p-value, a measurement of the probability that the observed difference could have occurred by chance'. Prior to the conduction of experimentation a significance level α is chosen. If the determined p-value is smaller than the chosen α, H_0 is rejected and the observed performance difference is considered significant. A variety of different tests have been applied to the purpose of significance testing in IR. An evaluation by Smucker et al. (2007) showed that the Student's t-test, the bootstrap shift method, and a randomization test all produce comparable significance values when used in the context of TREC ad hoc and web retrieval. Based on their conclusion that no practical difference exists in such a setup, the Student's t-test is chosen for all reported significance testing of this chapter. The significance level α is set to a value of 0.05. Throughout the chapter, reported results exhibiting a lower p-value than α are marked by the ‡ symbol. A p-value lower than 0.01 is marked by the † symbol. All calculations were conducted based on use of the R (2012) toolkit, and the RStudio (2012) development environment.

9.2.6 Word Relationship Measurement

The implementation is directly based on the reported results of chapter 7 and 8. With regard to the drawn conclusions in these chapters measurements of the respective vari-

[9-3] Available at http://trec.nist.gov/trec_eval. Last accessed 22.02.2013

ables will be based on the following instruments:

- **Grade of word relatedness**: The FS353 word similarity assessment procedure. This choice is grounded on the observation that, likely due to its larger size, the FS353 dataset represents a collection of word pairs that reflect both associative, semantic, and associative-semantic relations.
- **Semantic/associative degree of word relationship**: The priming simulation methodology developed on basis of the Ferrand word pair sets.

9.2.7 Relevance Measurement

As stated in Section 9.2.5 measuring the construct of relevance is based on the use of assessments. Due to the central role the issues of validity and measuring take within the research focus of this dissertation, the methodology of assessments is subsequently investigated in more detail.

Assessments constitute records of human judgement of the relevance of documents with respect to an information need. The methodology for measuring relevance based on assessments was introduced by Cleverdon (1967) as part of the Cranfield experiments series known as Cranfield I and Cranfield II. The components of the Cranfield experiments consisted of a small test collection of documents, a set of 225 test queries, and a set of relevance judgements. The collection was comprised of 1398 abstracts of aerodynamics journal articles. As part of the experimentation a complete judgement of all (query, document) pairs was conducted. The Cranfield studies are attributed with founding the system-oriented approach to IR (Hildreth, 2002). That is, the view that IR experiments can be conducted in a controlled, laboratory-like setting based on the use of assessors' judgements. The view assumes that relevance is a context-, and task-independent, objective property of documents (Cooper, 1973), and can be measured objectively.

A limitation of the Cranfield assessments consisted of the very small size and topical focus of the associated test collection. A continuation of the effort to make standardized test collections available was started by the National Institute of Standards and Technology[9-4] (NIST). The Text REtrieval Conference series[9-5] (TREC) conference series were dedicated to the creation of larger test collection sets addressing different domains, tasks, and document types. To enable the assessment of larger test collection the TREC initiative relied assessing only subsets of the (query, document) pairs. The applied assessment methodology is based on pooling and categorized as a form of incomplete relevance judgement (Voorhees, 2002). The application of pooling is motivated by large

[9-4] www.nist.gov
[9-5] trec.nist.gov

necessary effort to assess larger test collections, and the identified necessity to conduct multiple assessment of (document, query) pairs to mitigate the effects of assessor bias. Pooled assessment as implemented by TREC limits the set of documents to be judged based on result rankings obtained from retrieval system runs. That is, only (document, query) pairs returned within result listings of systems participating in the TREC initiative are judged. This constitutes the underlying procedure for measuring relevance in the case of test collections listed in Section 9.2.1. In the case of these collections the judgement of the (document, query) pairs was limited to the first 100 documents returned by the participating systems.

An important consideration with regard to the empirical focus of the chapter consists of the observation, that the pooling methodology introduces bias towards the participating systems and retrieval techniques. Although the TREC collections are widely utilized as a means of assessing retrieval techniques, there is a growing amount of evidence that pooling introduces potential bias towards 'new' systems (Buckley et al., 2007; Soboroff, 2007; Baillie et al., 2008; Webber and Park, 2009). Since relevance is only measured for contributing systems, this might lead to under or over-estimation of the performance of 'new' systems, if relevant documents returned by these systems were not included in the original pools. A good example of how specific techniques might introduce bias is given by the integration of link authority into retrieval models. Techniques such as Page et al.'s (1999) PageRank system or Kleinberg's (1999) HITS algorithm strongly favor document's with many inlinks (i.e. hyperlinks pointing to the document). The utilization of such a technique in a pooling scenario might therefore result in the exclusion of potentially relevant documents with no or only few inlinks. The study by Buckley et al. (2007) demonstrated that pooling induced bias might also extend to favoring documents based on their shared terms with topic titles.

With respect to the empirical investigations of the chapter, the described bias introduced by pooling presents a limitation. As outlined in introduction of this section, the basis empirical approach consists of inducing variations in the word related constructs and to measure its impact on measurements of relevance. Based on incomplete relevance judgement, variation in the construct of relevance might not be reflected. A preferential basis for the planned experimentation would consist of a test collection with complete relevance judgement. Complete relevance judgement ensures, that variation in relevance is fully reflected. However, no large scale test collections with complete relevance judgements are publicly available. An alternative to complete relevance judgements is given by direct assessment of retrieval runs based on alteration of the queries. In this scenario, controlled alteration of a query is followed by direct assessment of the retrieved results. While this approach requires less effort than a complete judgement, the need to conduct those assessments over the variable space covered by the experiments still exceeded the available resources of this study. In light of this, the use of standard test collections constituted the only feasible option. Nevertheless, the impact

of the underlying methodology for the measurement of relevance is to be taken into consideration with regard to the evaluation of the experiments.

9.2.8 Outline of Experiments

The following subsections provide an outline of the experimentation in the chapter. The first subsection describes the mode of operation for varying word relatedness. The second subsection explores the approach to the investigation of the effect of these variations on relevance. The last subsection relates the described approaches to the research questions of the chapter.

Variation in Word Relatedness

Variation in word relatedness is based on the parameter space of the HAL and LSA models. The underlying data set for the models is given by the Wikipedia collection and a df threshold of 25. This is based on the observations in Chapter 7, that the training of HAL and LSA models on this data set resulted in the highest correlations. The parameter space of the HAL and LSA models is defined as shown in Figure 6.1. As outlined in Chapters 7 and 8, variation of the parameters results in word spaces representing different grades and types of word relatedness.

Investigating the Effect of Variation in Word Relatedness

The empirical investigation of the effect induced by this variation is based on the expansion of query terms with their K closest related terms. As outlined in Section 9.2.4, the integration of variation in word relatedness is based on a query expansion mechanism. Given a set of queries $Q = (q_1, .., q_n)$, where each query q_n is comprised of a set of terms $(t_1, ..., t_m)$, each term t_m is expanded with its K closest related terms. The survey by Carpineto and Romano (2012) observed, that the addition of 10-30 terms constitutes the most effective range for query expansion. With regard to these observations, K is defined as $K \in \{1, 2, ..., 30\}$. The motivation for the described expansion technique is the following. The expansion mode aims enabling direct observation of the effect of word relatedness. As outlined by Xu and Croft (1996), directly expanding queries has been observed to have positive or negative impact on query performance. To mitigate 'hurting' queries, query expansion in IR often aims at preselecting the set of potential expansion terms. As outlined by Carpineto and Romano (2012), a commonly applied technique is given by pseudo relevance feedback. The aim of this study, however, is to observe the full range of effects of word relatedness on relevance, and to minimize the complexity of the modelled system. Motivated by this aim, the integration mode aims

at implementing the most direct way of integrating related terms. A similar motivation underlies the choice of using the topic titles (also referred to as short queries) as the basis for the expansion. The topic titles, consisting only of the keywords of a query, constitute the most condensed representation of the information need. Specifically with regard to the range of K, it is assumed that expansion of the titles, in contrary to using the long descriptions of the information need, results in a stronger impact of the effect of word relatedness.

Empirical Approach to Research Questions

RQ 6a questions if the constructs of word relatedness and relevance are related. The work of Borsboom (2003) advocates, that the observation of variance in a construct that is induced by variation of an attribute[9-6] constitutes evidence of a relation between the two. The investigation of this question is based on optimizing the query expansion based retrieval performance. Tuning the BM25 and query expansion parameters aims at inducing maximum variation through the described integration of the constructs of word relatedness. The statistical verification of the relation is based on the significance testing described in Section 9.2.5. The null hypothesis H_0 of the significance testing assumes that the tested retrieval models are equivalent in performance. A rejection of the null hypothesis therefore indicates that the observed variation is significant. This is interpreted as a verification of a relation between the constructs of word relatedness and relevance.

The investigation of RQ 6b is primarily based on the utilization of the complete defined parameter space of HAL, LSA, and the query expansion mode. RQ 6b constitutes a deliberately broadly defined research question. Its primary aim consists of demonstrating how a nomological network might contribute to the understanding of the meaning of relevance. As outlined in Chapter 2, establishing construct validity is pursued by gaining insight to the relation to other constructs. Chapter 5.3 emphasized the important role of the inclusion of highly measurable constructs in such networks. The validation studies conducted in Chapters 7 and 8 ensured us that this applies to the word relatedness constructs. Ideally, the investigation of RQ 6b results in the identification of tight lawful relations between the constructs. A more realistic expectation is, that the investigation of RQ 6b generates insights and provides a basis for the postulation of novel hypothesis about the relation.

[9-6] i.e. a construct with high measurability; See Section 5.3.1

9.3 Results and Analysis

The presentation of the results and their analysis is structured as follows. Subsection 9.3.1 lists the results for the optimized retrieval runs. An relation these results with related work is presented in Subsection 9.3.2. In Subsection 9.3.3 the performance results are interpreted with respect to RQ 6a. Subsection 9.3.4 provides an overview over the observed effect of varying word relatedness based on the defined parameter space. In Subsection 9.3.5 these observations are analysed in relation to the task of identifying characteristics of the relationship between relevance and word relatedness. Finally, Subsection 9.3.6 presents a concluding discussion based on the empirical observations.

9.3.1 Results for Optimized Retrieval Runs

Tables 9.1 to 9.5 provide listings of the best observed retrieval performance. The tables list the three highest performing HAL and LSA models with respect to each of the different term weighting schemes. As demonstrated in Chapters 7 and 8, the different weighting schemes impact the type and grade of word relatedness represented by the HAL and LSA models. In light of this, reporting of the results over the different weighting schemes provides an overview over retrieval performance based on different representations of word relatedness. Each table reports the official evaluation measures associated with the respective TREC task. An exception is given in the case of the TREC8 Ad Hoc task, where in addition to reporting Mean Average Precision (MAP), the performance based on Geometric Mean Average Precision (GMAP) is also reported. The first data row of each table reports the respective baseline set by a BM25 model with optimised parameter values.

AQUAINT Robust 05 GMAP

Run	Weighting	WS / Dim.	GMAP	delta
Baseline	-	-	0.1147	-
DF25_GeometricWeighting_8	GeometricWeighting	8	0.1191	3.8361 ‡
DF25_GeometricWeighting_7	GeometricWeighting	7	0.1189	3.6617 ‡
DF25_GeometricWeighting_6	GeometricWeighting	6	0.1171	2.0924 ‡
DF25_EvenWeighting_5	EvenWeighting	5	0.1171	2.0924 ‡
DF25_EvenWeighting_2	EvenWeighting	2	0.1169	1.918 ‡
DF25_EvenWeighting_9	EvenWeighting	9	0.1168	1.8309 ‡
DF25_LinearWeighting_5	LinearWeighting	5	0.1171	2.0924 ‡
DF25_LinearWeighting_2	LinearWeighting	2	0.1169	1.918 ‡
DF25_LinearWeighting_9	LinearWeighting	9	0.1166	1.6565 ‡
DF25_NoTransform_300	NoTransform	300	0.1214	5.8413 ‡
DF25_NoTransform_500	NoTransform	500	0.1206	5.1439 ‡
DF25_NoTransform_150	NoTransform	150	0.1172	2.1796 ‡
DF25_LogEntropyTransform_300	LogEntropyTransform	300	0.1176	2.5283 ‡
DF25_LogEntropyTransform_150	LogEntropyTransform	150	0.1174	2.354 ‡
DF25_LogEntropyTransform_400	LogEntropyTransform	400	0.1166	1.6565 ‡
DF25_TfIdfTransform_500	TfIdfTransform	500	0.115	0.2616 ‡
DF25_TfIdfTransform_150	TfIdfTransform	150	0.1146	-0.087 †
DF25_TfIdfTransform_400	TfIdfTransform	400	0.1145	-0.174 †

TABLE 9.1: GMAP of HAL and LSA Models on AQUAINT Robust 05 Task

Disk4&5 TREC7 Ad Hoc MAP

Run	Weighting	WS / Dim	MAP	delta
Baseline	-	-	0.19	-
DF25_GeometricWeighting_10	GeometricWeighting	7	0.1938	2 ‡
DF25_GeometricWeighting_9	GeometricWeighting	10	0.1937	1.9474 ‡
DF25_GeometricWeighting_8	GeometricWeighting	6	0.1937	1.9474 ‡
DF25_EvenWeighting_6	EvenWeighting	10	0.1944	2.3158 ‡
DF25_EvenWeighting_5	EvenWeighting	9	0.1944	2.3158 ‡
DF25_EvenWeighting_7	EvenWeighting	8	0.1942	2.2105 ‡
DF25_LinearWeighting_5	LinearWeighting	9	0.1945	2.3684 ‡
DF25_LinearWeighting_7	LinearWeighting	10	0.1945	2.3684 ‡
DF25_LinearWeighting_4	LinearWeighting	7	0.1943	2.2632 ‡
DF25_NoTransform_300	NoTransform	50	0.1945	2.3684 ‡
DF25_NoTransform_50	NoTransform	400	0.1942	2.2105 ‡
DF25_NoTransform_500	NoTransform	150	0.194	2.1053 ‡
DF25_LogEntropyTransform_50	LogEntropyTransform	150	0.1928	1.4737 ‡
DF25_LogEntropyTransform_150	LogEntropyTransform	50	0.1926	1.3684 ‡
DF25_LogEntropyTransform_300	LogEntropyTransform	25	0.1921	1.1053 ‡
DF25_TfIdfTransform_500	TfIdfTransform	300	0.1932	1.6842 ‡
DF25_TfIdfTransform_50	TfIdfTransform	500	0.1927	1.4211 ‡
DF25_TfIdfTransform_25	TfIdfTransform	400	0.1927	1.4211 ‡

TABLE 9.2: MAP of HAL and LSA Models on TREC7 Ad Hoc Task

Disk4&5 TREC8 Ad Hoc GMAP

Run	Weighting	WS / Dim	GMAP	delta
Baseline	-	-	0.1266	-
DF25_GeometricWeighting_8	GeometricWeighting	8	0.1565	23.618 ‡
DF25_GeometricWeighting_7	GeometricWeighting	7	0.1558	23.065 ‡
DF25_GeometricWeighting_9	GeometricWeighting	9	0.1555	22.828 ‡
DF25_EvenWeighting_6	EvenWeighting	6	0.1578	24.645 ‡
DF25_EvenWeighting_5	EvenWeighting	5	0.1565	23.618 ‡
DF25_EvenWeighting_7	EvenWeighting	7	0.1559	23.144 ‡
DF25_LinearWeighting_6	LinearWeighting	5	0.1563	23.46 ‡
DF25_LinearWeighting_5	LinearWeighting	10	0.156	23.223 ‡
DF25_LinearWeighting_7	LinearWeighting	7	0.1557	22.986 ‡
DF25_NoTransform_150	NoTransform	150	0.1534	21.169 ‡
DF25_NoTransform_25	NoTransform	25	0.1522	20.221 ‡
DF25_NoTransform_300	NoTransform	300	0.1493	17.93 ‡
DF25_LogEntropyTransform_150	LogEntropyTransform	150	0.1585	25.197 ‡
DF25_LogEntropyTransform_400	LogEntropyTransform	400	0.1579	24.724 ‡
DF25_LogEntropyTransform_300	LogEntropyTransform	300	0.1568	23.855 ‡
DF25_TfIdfTransform_500	TfIdfTransform	500	0.1528	20.695 ‡
DF25_TfIdfTransform_400	TfIdfTransform	400	0.1491	17.773 ‡
DF25_TfIdfTransform_50	TfIdfTransform	50	0.1491	17.773 ‡

TABLE 9.3: GMAP of HAL and LSA Models on TREC8 Ad Hoc Task

Disk4&5 TREC8 Ad Hoc MAP

Run	Weighting	WS / Dim	MAP	delta
Baseline			0.2384	
DF25_GeometricWeighting_10	GeometricWeighting	10	0.2594	8.8087 ‡
DF25_GeometricWeighting_9	GeometricWeighting	9	0.2593	8.7668 ‡
DF25_GeometricWeighting_8	GeometricWeighting	8	0.2583	8.3473 ‡
DF25_EvenWeighting_6	EvenWeighting	6	0.2607	9.354 ‡
DF25_EvenWeighting_5	EvenWeighting	5	0.2601	9.1023 ‡
DF25_EvenWeighting_7	EvenWeighting	7	0.2582	8.3054 ‡
DF25_LinearWeighting_5	LinearWeighting	5	0.2598	8.9765 ‡
DF25_LinearWeighting_7	LinearWeighting	7	0.2585	8.4312 ‡
DF25_LinearWeighting_4	LinearWeighting	4	0.2579	8.1795 ‡
DF25_NoTransform_300	NoTransform	150	0.2561	7.4245 ‡
DF25_NoTransform_50	NoTransform	25	0.2544	6.7114 ‡
DF25_NoTransform_500	NoTransform	300	0.2542	6.6275 ‡
DF25_LogEntropyTransform_50	LogEntropyTransform	150	0.2588	8.557 ‡
DF25_LogEntropyTransform_150	LogEntropyTransform	400	0.2579	8.1795 ‡
DF25_LogEntropyTransform_300	LogEntropyTransform	300	0.2575	8.0117 ‡
DF25_TfIdfTransform_500	TfIdfTransform	500	0.2562	7.4664 ‡
DF25_TfIdfTransform_50	TfIdfTransform	150	0.2557	7.2567 ‡
DF25_TfIdfTransform_25	TfIdfTransform	300	0.2557	7.2567 ‡

TABLE 9.4: MAP of HAL and LSA Models on TREC8 Ad Hoc Task

.GOV Mixed Query Task MAP

Run	Weighting	WS / Dim	MAP	delta
baseline	NA	NA	0.2516	NA
DF25_GeometricWeighting_8	GeometricWeighting	7	0.257	2.1463 ‡
DF25_GeometricWeighting_7	GeometricWeighting	9	0.2569	2.1065 ‡
DF25_GeometricWeighting_6	GeometricWeighting	4	0.2559	1.7091 ‡
DF25_EvenWeighting_9	EvenWeighting	4	0.2565	1.9475 ‡
DF25_EvenWeighting_7	EvenWeighting	1	0.2539	0.9141 ‡
DF25_EvenWeighting_6	EvenWeighting	9	0.2538	0.8744 ‡
DF25_LinearWeighting_9	LinearWeighting	3	0.2561	1.7886 ‡
DF25_LinearWeighting_7	LinearWeighting	6	0.2543	1.0731 ‡
DF25_LinearWeighting_6	LinearWeighting	1	0.2543	1.0731 ‡
DF25_NoTransform_300	NoTransform	150	0.2561	1.7886 ‡
DF25_NoTransform_50	NoTransform	25	0.2544	1.1129 ‡
DF25_NoTransform_500	NoTransform	300	0.2541	0.9936 ‡
DF25_LogEntropyTransform_50	LogEntropyTransform	150	0.2588	2.8617 ‡
DF25_LogEntropyTransform_150	LogEntropyTransform	400	0.2579	2.504 ‡
DF25_LogEntropyTransform_300	LogEntropyTransform	300	0.2575	2.345 ‡
DF25_TfIdfTransform_500	TfIdfTransform	500	0.2562	1.8283 ‡
DF25_TfIdfTransform_50	TfIdfTransform	150	0.2557	1.6296 ‡
DF25_TfIdfTransform_25	TfIdfTransform	300	0.2556	1.5898 ‡

TABLE 9.5: MAP Performance of HAL and LSA Models on .GOV Mixed Query Task

9.3.2 Analysis of Retrieval Performance

Tables 9.1 to 9.5 show that the expansion with related terms leads to performance gains across all considered collections. Some general observations can be made with regard to the observed performance.

With the exception of the performance on the TREC 8 Ad Hoc task, the direct expansion results in a small improvement over the baseline. It seems plausible to assume, that this is implied by the chosen mode query expansion described in 9.2.4. As outlined by the studies of Collins-Thompson (2009) and Xu and Croft (1996), the direct expansion of queries results in positive or negative impact on query performance. To improve query performance, a chosen expansion term has to match the intended 'meaning' of the information need. Homonymy, and situational and contextual mismatch may result in the expansion of a query with 'wrong' terms. Depending on the information need, the expansion of the query 'fix bad banks' with terms relating to either the geographic or institutional interpretation of 'bank' may result in a strong negative impact on query performance. As previously mentioned, a commonly applied technique to mitigate the risk of expanding with mismatching terms consists of contextual analysis. As described by Carpineto and Romano (2012), pseudo relevance feedback is often applied to that cause. The study of Fang and Zhai (2006) and Carpineto and Romano's (2012) survey demonstrate the common application of such approaches, and their potential benefit as a means of minimizing a negative impact on query performance. However, with respect to the empirical aims of the chapter, the positive or negative impact resulting from direct query expansion is interpreted as beneficial to the investigation. The aim of this study differs from performance focused approaches to query expansion through its focus on the investigation of the relation between the selected constructs. As outlined in Section 9.2.4, the aim of *directly* observing the effect of varying word relatedness (positive or negative) motivated the choice of a deliberately simple expansion mode.

Nevertheless, it is of interest to embed the observed performance in the context of other reported results and techniques. The strongest improvements were registered for the TREC 8 Ad Hoc task shown in Tables 9.3 and 9.4. As listed in Table 9.2, the expansion on the TREC 7 Ad Hoc task led only to small improvements. The number of studies reporting on the use of direct expansion mechanisms is limited. The study by Fang (2008) outlines, that the observed performance in Table 9.2, 9.3, and 9.4 is comparable to the performance of direct query expansion based on lexical resources. Specifically, if the potential qualitative advantages of using manually curated resources are taken into consideration. Expansion based on difficult queries, led to modest increases in retrieval performance (Table 9.1). With regard to the Robust 05 task Fang and Zhai (2006) demonstrated the benefit of integrating pseudo relevance feedback. Voorhees (2005) and Liu and Yu (2005) emphasize the benefit of integrating web based resources into the retrieval process. Only slight improvement was achieved on the .GOV related

task (Table 9.5). As described in Section 9.2.7 this might be implied by bias introduced by pooling based measurement. As outlined by Craswell and Hawking (2004), 16 out 18 participating systems used URL, link or anchor text based techniques. The exclusion of these techniques might have resulted in the retrieval of relevant documents not represented in the assessments. The potential performance gains by integrating URL length, link information, document structure, and anchor text into a web based retrieval scenario are described by Craswell and Hawking (2004). The above listed studies demonstrate the possible performance improvements by integrating additional resources, techniques, and models into the retrieval process. With reference to the discussion of Section 5.3, the application of these techniques is deliberately excluded to keep the experimental system complexity at a minimum.

Generally it can be observed that the chosen integration approach resulted in improved query performance across all tasks and applied word spaces. This observation forms the basis for the investigation of RQ 6a in the next subsection.

9.3.3 Analysis with Regard to RQ 6a

A preliminary step towards the analysis of the relation of constructs is given by a test of the hypothesis that the constructs are related. Subsection 9.2.8 outlined that the empirical investigation of this question is based on the observed retrieval performance described in Section 9.3.1.

The investigation of the question is based on a re-interpretation of the null hypothesis of the significance tests. The null hypothesis H_0 that assumes that the tested retrieval models are equivalent in performance. With respect to the integration of word relatedness into the retrieval process H_0 can be interpreted as assuming, that word relatedness does not impact the estimation process of relevance. Formulated in this way, a rejection of H_0 signifies a relation of the constructs of word relatedness and relevance. As shown throughout the Tables 9.1 to 9.5, H_0 is rejected for all task, HAL, and LSA combinations.

The rejection of the hypothesis can be interpreted from two perspectives. From an Information Retrieval perspective the observation of significance is not surprising. The work of Xu and Croft (1996); Collins-Thompson (2009); Carpineto and Romano (2012) documents the considerable effect of query expansion on retrieval performance. Specifically the study by Collins-Thompson (2009), that aims at mitigating the negative effects of expansion emphasizes this observation. The observance of significant variation as a result of the integration of word relatedness into the retrieval process is in concordance with these observations.

From a cognitive processing perspective, the rejection of H_0 is not surprising due to

the strength of the hypothesis that the constructs of word relatedness and relevance are related. The work of Swinney (1979) and Anderson (1983) outlined the mechanisms underlying relations between words. The models of discourse comprehension by van Dijk and Kintsch (1983), Graesser et al. (1997), and Perfetti et al. (2005) emphasize the important role that word relations play with regard to the identification of the meaning of a word, and the inference of the meaning of discourse. The important role of word relations in identifying the meaning of a document, added strength to the hypothesis that word relatedness also impacts the estimation of relevance between two textual items.

The rejection of H_0 is also in concordance with the observations in Section 9.1. Fang and Zhai's (2006) definition of $STMC1$ states, that the estimation of relevance between two information items Q and D is positively influenced if Q contains terms related to the terms in D. This expresses a general assumption of a relation between word relatedness and relevance. The retrieval results in Section 9.3.1 confirm the assumption of a relation of the grade of word relatedness, the type of word relatedness and relevance. The confirmation stems from the observation of the significance of HAL and LSA based retrieval runs over all applied tasks. Based on the confirmation of the relation between the constructs, the next subsection investigates retrieval performance over the full parameter space.

9.3.4 Analysis of Variation over Parameter Space

The prior subsection presented a confirmation for the existence of a relation between the constructs of grade and type of word relatedness and relevance. Guided by research question RQ 6b the following subsection explores the identification of characteristics of the relation between the constructs. In the context of the dissertation, the exploration of RQ 6b constitutes a demonstration of the application of the nomological network described in Section 5.3.2. Investigating the relation of the constructs is based on inducing variance in the word relatedness constructs and the observance of its effect on relevance. Varying word relatedness is based on the defined parameter space. As described in Section 9.2, the parameter space is composed of the parameters of the HAL and LSA models, and the query expansion parameters. The first step of the analysis consists of outlining the relation between the parameters and the constructs of grade and type of word relatedness.

The validation studies in Chapters 7 and 8 demonstrated, that different settings of the weighting, window size, and dimension parameters of the HAL and LSA models result in word spaces representing different types of word relatedness. The validation studies outlined, that this applies specifically HAL models defined on different window sizes. HAL models trained on higher window sizes result in word spaces that favor associative

terms as nearest neighbours (i.e. expressing an association such as 'spider – web'). Whereas models defined with a window size of 1 result in spaces that primarily map semantic relations (i.e. expressing a similarity based on shared features such as 'van – car'). Iterating over the window size parameter space therefore allows to induce variance with regard to the type of word relatedness. Inducing variance with regard to the grade of word relatedness is realized based on the K parameter of the query expansion mode. The expansion mode is based on altering the query through expanding its terms with the K closest related terms (where $K \in \{1, 2, ..., 30\}$). This enables the investigation of the impact of expanding with progressively lesser related terms. In this form, the defined parameter space allows to induce variance with regard to the type and grade of word relatedness.

Based on this outline, the first step of the analysis consists of examining the observed retrieval performance over the full parameter space and retrieval tasks. The Figures 9.1 and 9.2 show the resulting MAP of HAL based expansion. The figures depict the performance of HAL models based on where terms in the included in the sliding window are evenly weighted. The observed retrieval performance over K is shown for the tasks, Aquaint Robust 05, TREC 7 Ad hoc, Trec 8 Ad hoc and .GOV TD. Each row in the figure refers to one of these tasks. Each of the three columns in the figures is dedicated to a range of K values, and shows the arithmetic mean of MAP performance for this range. That is, the first column shows the average MAP measured for expansion with 1-5 related terms. The second column shows the observed performance for expansion with 6-10 terms. Binning of K is motivated by the prior made observations regarding the 'mismatch risk' of query expansion (see discussion in 9.3.2). Aggregating observed performance based on equally sized bins aims to mitigate these effects and to support the identification of underlying trends in the data. To aid the visual interpretation of the data a cubic smoothing spline[9-7] is fitted to the data. This is pictured in form of a continuous line coloured red. Further, if the observed values are in the range of the measured performance of baseline runs, a continuous horizontal line indicates baseline performance.

In this form the figures allow for making some basic observations with respect to the observed retrieval performance over the parameter space. The first of those observations is, that expansion with related terms can be beneficial to retrieval performance. This is indicated by the performance shown in the first two columns of the Robust 05 , TREC 7 Ad Hoc, and TREC 8 Ad hoc task in Figure 9.1. For all three tasks, the mean performance of certain window sizes exceeds baseline performance. The last row in both figures shows, that this does not apply to the .GOV Topic Distillation task. Topic Distillation constitutes the most difficult task of the .GOV 2004 track (Craswell and Hawking, 2004). The utilized direct expansion technique does not result in an improvement over the baseline in this case. A second observation is that retrieval performance

[9-7] Based on the implementation of the smooth.spline package of R (2012)

degrades with continued addition of expansion terms. This becomes evident by looking at the development of mean performance in the respective plots of the bins of $K <= 5$, $K <= 10$, and $K <= 15$ in Figure 9.1, and $K <= 20$, $K <= 25$, and $K <= 30$ in Figure 9.2. In the case of TREC 7 Ad Hoc for example, the top performaning window sizes degrade from a MAP over 0.19 to slightly over 0.18. A third observation that can be made based on the observations of Chapter 8. With the exception of the top left plot, observed performance exceeding the baseline is highest for very small and medium to large window sizes. This becomes evident through a distinct 'V' shape formed by fitted splines. The validation study in Chapter 8 outlined, that these window sizes result in semantically focused or associatively focused word spaces showing the highest correlation with human assessments of word relatedness. Figure 8.14 illustrates that HAL models trained on window sizes of 1 and 2 result in semantically oriented word spaces. Figure 8.9 shows that window sizes of 5 or higher result in HAL models with the highest correlation for the associatively focused 'FS353 Rel' word relationship assessments.

These initial observations indicate the existence of characteristic patterns with regard to the relation of the constructs of word relatedness and relevance. In the subsequent section these characteristics are explored in more detail on basis of the performance over the parameter space.

9.3.5 Analysis with Regard to RQ6b

RQ 6b is aimed at characterizing the the relation between the postulated constructs of relevance and grade and type of word relatedness? Motivated by the research focus of the dissertation, the question is formulated in an open-ended form. An underlying aim of the research question consists of providing a demonstration of the application of the nomological network defined in Section 5.3.2. As outlined in Subsection 9.2.8, the empirical investigation of the question is based on the defined parameter space. The prior subsection provided a first outline of pursuing such an analysis. Based on the observed retrieval behavior it presented three observations indicating potential characteristics of the relation between the constructs. Subsequently, an extension of this initial analysis is presented. To provide structure to the investigation, the results of the analytic investigation are reported based on the definition of three hypotheses pertaining to the relation between the constructs of word relatedness and relevance. The hypotheses are derived based on the initial observations in the prior subsection. In the following three subsections, each hypothesis is explored in relation to the observed retrieval results in Subsection 9.3.1, the observed retrieval performance over the full parameter space, and the observations with respect to grade and type of word relatedness in Chapters 7 and 8.

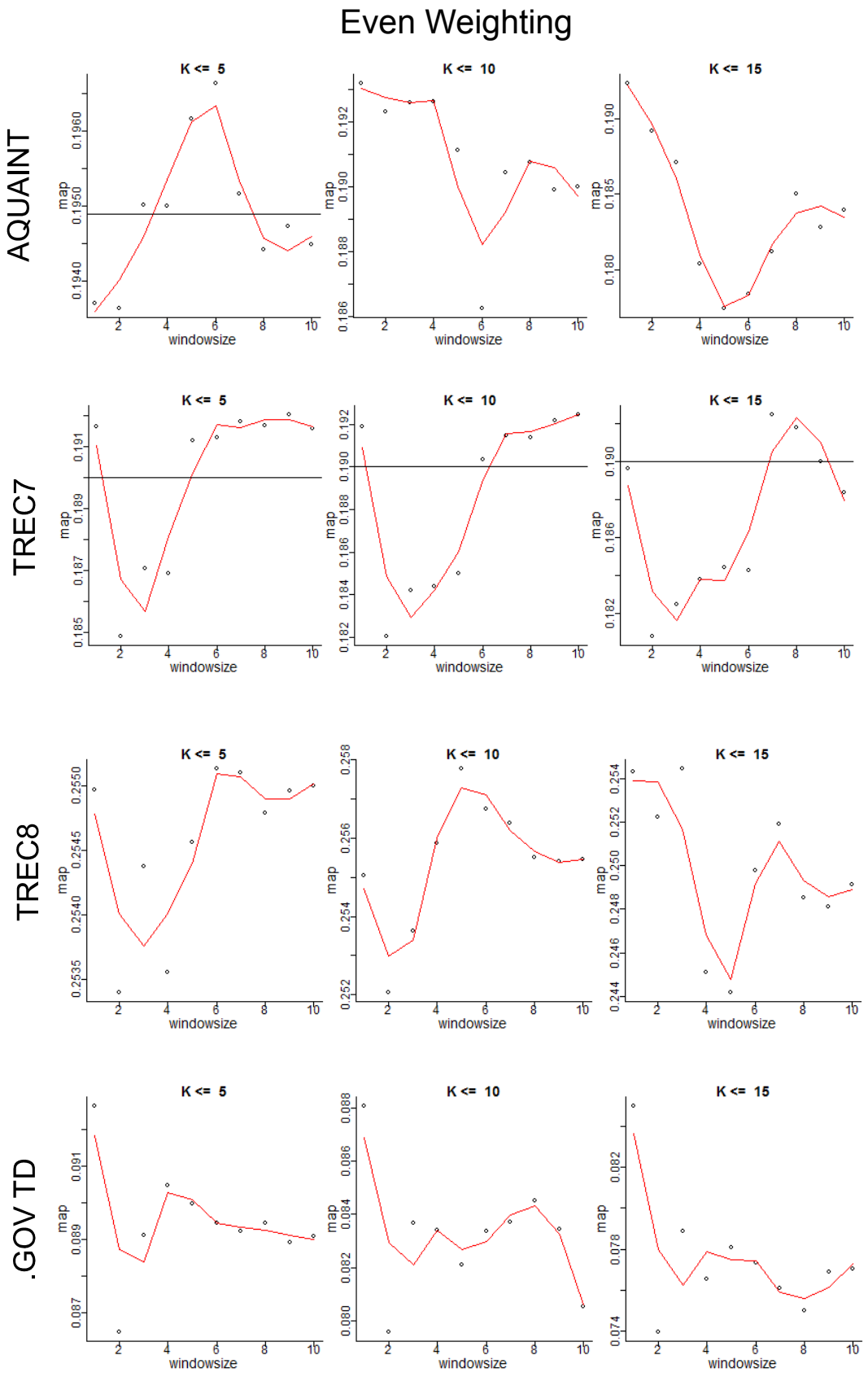


FIGURE 9.1: Arithmetic Mean of MAP Performance for the Expansion with Ranges $K <= 5$, $K <= 10$, and $K <= 15$

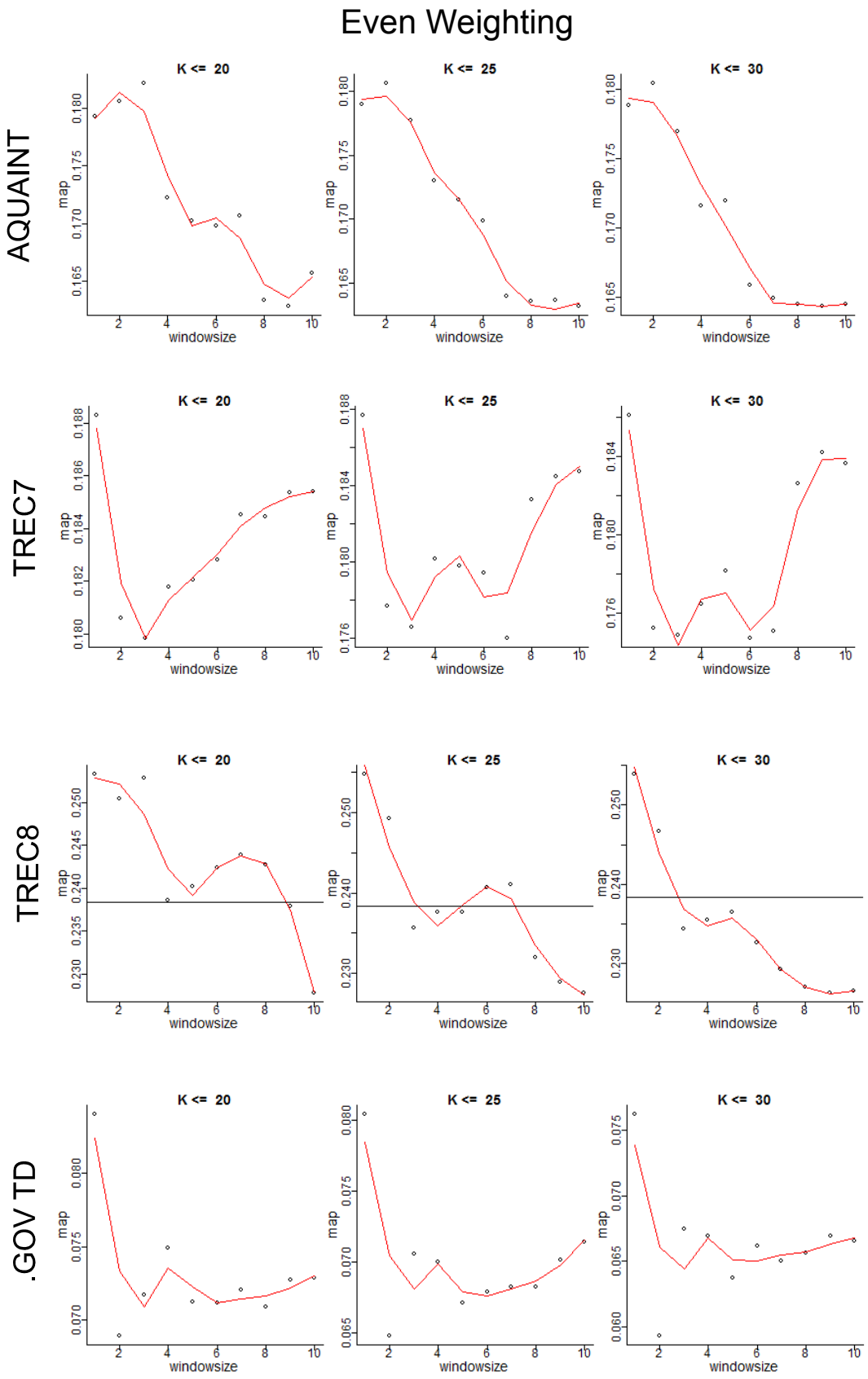


FIGURE 9.2: Arithmetic Mean of MAP Performance for the Expansion with Ranges $K <= 20$, $K <= 25$, and $K <= 30$

Effectiveness of Semantic and Associative Expansion

As outline in the prior section, the exploration of the relation is based on the formulation of three hypotheses. The initially informally (and without rigour) defined hypotheses are aimed at providing structure to the exploration of an open-ended research question. Subsequently these hypotheses are evaluated based on an analysis of the observed retrieval performance over the full parameter space, and the observations with respect to grade and type of word relatedness in Chapters 7 and 8.

The first hypothesis labelled 'Effectiveness of Semantic and Associative Expansion' hypothesis, relates to one of the basic observations made in Subsection 9.3.4.

> **Hypothesis 9.1:** The existence of associative or semantic relations between the terms of two textual items positively impacts the estimation of their relevance.

Based on the axiomatic definitions of IR presented in Subsection 9.1.3, the hypothesis represents an extension of Fang and Zhai's (2006) definition of $STMC1$.

Let $D_1 = d_1$ and $D_2 = d_2$ be two single-term documents, where $q \neq d_1$ and $q \neq d_2$. If $s(q,d_1) > s(q,d_2)$, then $S(Q,D_1) > S(Q,D_2)$.

Function s is the similarity function between two terms. As noted earlier, the definition of $STMC1$ does not provide a specific definition of the type of similarity or relatedness between terms with regard to s. The above defined hypothesis can be interpreted as an extension of s by defining:

$s(t,u) > s(t,v)$ if t and u are associatively related
$s(t,u) > s(t,v)$ if t and u are semantically related

Where t, u, and v represent terms. This allows for an interpretation of the hypothesis from a retrieval performance perspective. With regard to query expansion it can be stated, that query expansion with associatively related as well as semantically related terms is potentially beneficial to retrieval. Subsequently, it is investigated if the hypothesis can be strengthened based on the empirical observations.

Retrieval Performance A first step with regard to an analysis of the hypothesis can be based on relating it to the observed performance of the optimized retrieval runs. In Subsection 9.3.1 it was noted, that expansion based on word space models resulted in significant improvement across all tasks. Relating these observations to the hypothesis requires their interpretation with regard to the type of word relations.

Based on the results in Tables 9.1 to 9.5, supporting evidence for the hypothesis consists of significant performance improvements for both, HAL and LSA based expansion. The investigations in Chapter 8 showed that HAL models with small window sizes primarily represent semantic relations, and that LSA models represent associative relations between terms. The conclusion was based on two core observations. Firstly, the observed difference in the semantic/associative ratio of neighbourhood relations. Figures 8.4 and 8.7 show that HAL models trained on small window sizes result in a primarily semantic neighbourhood, while the opposite applies to LSA models. Secondly, the higher correlations of LSA models with human assessments focused on associative relations supports the conclusion. Figure 8.9 and 8.19 show the higher correlations of LSA models with the 'FS353 Rel' data set. These observations document the associative and semantic nature of the respective LSA and HAL models. The observation of significant improvement for expansion based on both models, represents supportive evidence of the hypothesis that associative and semantic relations positively impacts the estimation of relevance. Further supportive observations can be made based on the performance of Tables 9.1 and 9.5. The listing of the best performing HAL models in both tables consists of HAL models with either small *or* large window sizes. Table 9.6 demonstrates the respective semantic and associative nature of the resulting HAL models. It shows the nearest neighbours of the word 'cancer' from the Robust05 title only query 'radio wave brain cancer'. The table lists the results of the nearest neighbours assessments for

Nearest neighbours for term 'cancer'					
10 Even Weight.			2 Geom. Weight.		
term	eucl. dist.	S/A	term	eucl. dist.	S/A
ovarian	0.692512	A	adenocarcinoma	1.057155	S
chemotherapi	0.696284	A	melanoma	1.074517	S
breast	0.701709	A	mesothelioma	1.075439	S
metastat	0.701984	A	leukaemia	1.086389	S
prostat	0.705444	A	patient	1.095584	A
diagnos	0.710013	A	tuberculosi	1.098431	S
colorect	0.718181	A	diseas	1.105329	S
metastasi	0.749785	A	implant	1.106676	A
metastas	0.765439	A	metastasi	1.114785	A

TABLE 9.6: Nearest neighbours of term 'cancer' shown for 2 specific HAL models based on Wikipedia:df25 collection

an Even weighting based HAL model of window size 10 and a Geometric weighting based HAL model of window size 2. The semantic-associative assessments ('S','A') outline, that the window size 10 based model's nearest neighbours are dominated by terms that exhibit an associative relationship with the word 'cancer'. The window size 2 model exhibits dominantly semantically related terms. This illustrates the type related focus of HAL models with respect to the window size parameter. The observed significant results for both, small and large window sizes represent tentative supportive

evidence the hypothesis. To substantiate this conclusion, the next paragraph explores the observed results over the full parameter space.

Parameter Space Performance The last paragraph outlined how the performance of HAL models differing in window size can be interpreted as supportive evidence for the 'Effectiveness of Semantic and Associative Expansion' hypothesis. Subsequently this observation is explored in more detail by relating the observed effect of window size on retrieval performance over the parameter space, and measurements of the type of word relatedness.

Section 9.3.4 described a distinct pattern in the retrieval results shown in Figures 9.1 and 9.2 . With the exception of the top left plot, observed performance exceeding the baseline is highest for very small and medium to large window sizes. This becomes evident through a distinct 'V' shape formed by fitted splines. Based on the previous discussions, this pattern can be associated with the semantic and associative focus resulting from the respective window sizes. HAL models trained on small window sizes are best suited to capture and map semantic relations between terms. HAL models trained on larger window sizes are best suited to capture and map associative relations. The relation between window size and retrieval performance becomes more clear by concentrating the analysis on the tasks showing the most robust improvements over K. Figure 9.3 shows the performance of HAL models with different weighting schemes for the TREC 8 Ad hoc and TREC 7 Ad hoc tasks. In the figure, the performance of HAL models defined on even, linear, and geometric weighting schemes are shown. Each plot shows the arithmetic mean of the performance in the range of $K \in \{1,2,...,15\}$. Averaging performance over K is applied as a means of mitigating outliers induced by the earlier noted 'risky' nature of query expansion. The first and second plot columns confirm the conclusion that retrieval benefits from expansion with semantic and associative terms. Retrieval performance is strongest for those window sizes that result in the HAL models showing the highest respective performance for semantic and associative term relations. The observed retrieval behavior in case of the geometric weighting scheme adds confirmation to this interpretation. As noted in Subsection 8.2.2, the application of geometric weighting results in applying a weight of 10 to adjacent terms and a weight of 0.01 to terms at distance 10. This effectively reduces the ability of HAL models to capture associative term relations. The resulting effect on retrieval performance is clearly visible in the rightmost plots of Figure 9.3. The discussion can be substantiated by considering the observed behavior for associative and semantic priming effects on HAL models. Figure 9.4 shows plots of semantic and associative priming effects of HAL models.

The left plot shows the difference resulting from subtracting the measured mean associative priming effect from the mean semantic effect. A positive value reflects a larger semantic priming effect. The right plot shows the result of subtracting the measured

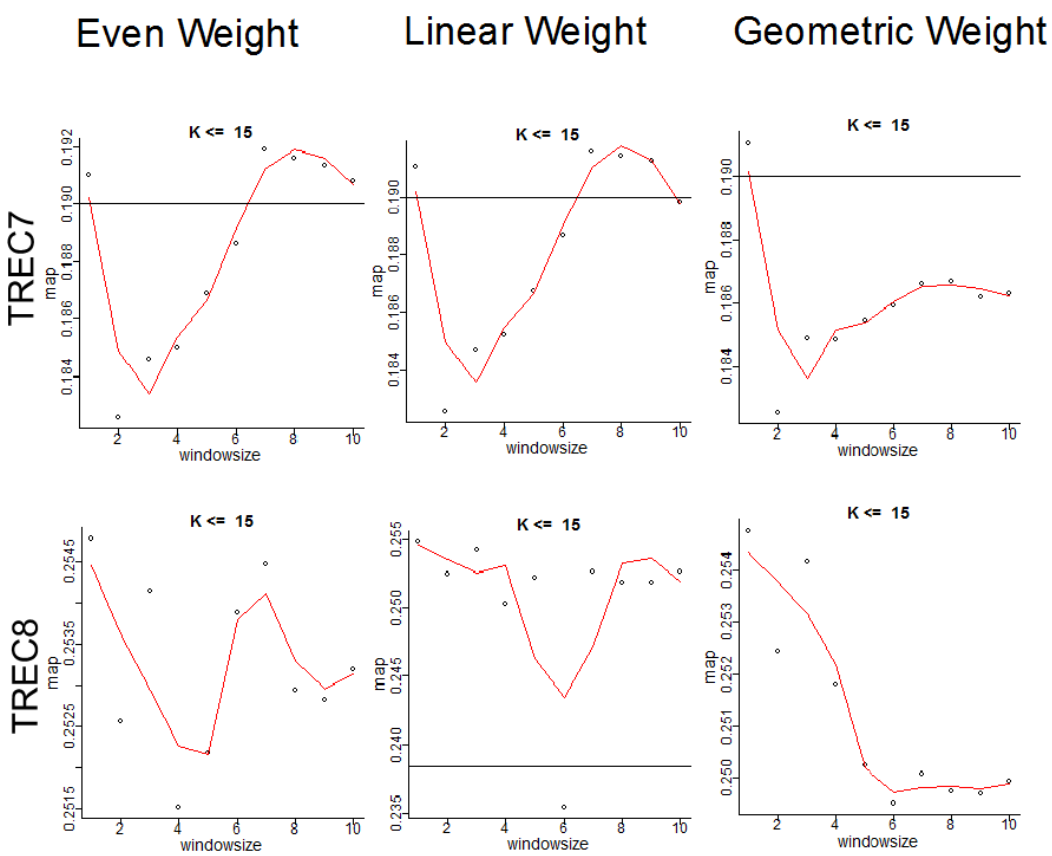


FIGURE 9.3: Mean MAP Performance over Window Size Parameter. TREC 8 Ad Hoc, TREC 7 Ad Hoc

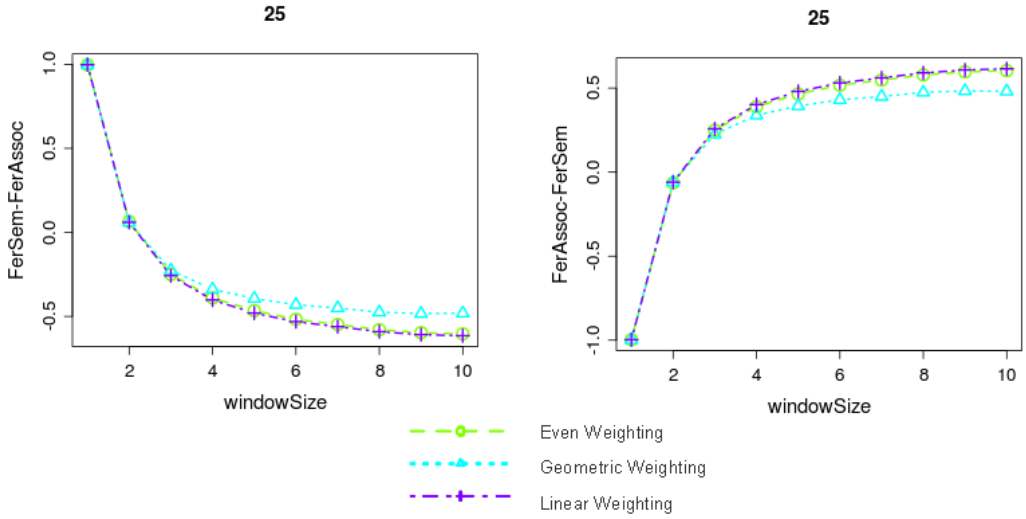


FIGURE 9.4: Semantic and associative priming effects of HAL models on basis of the Wikipedia:df25 collection

mean semantic priming effect from the mean associative effect. A positive value reflects a larger associative priming effect. Relating the observations in Figure 9.4 and 9.3 further confirms the hypothesized impact of associative and semantic term relations on relevance estimation. Mean retrieval performance is highest for HAL models mapping associative *or* semantic relations.

Discussion Subsequently, the so far led discussion is summarized with regard to two aspects: The 'Effectiveness of Semantic and Associative Expansion' hypothesis, and its interpretation with respect to the aim of validating relevance.

The observed results provided supportive evidence with regard to the hypothesis. The convergence of observations from term neighbourhood assessments, priming effect simulations, and retrieval performance support the hypothesized effectiveness of associative *and* semantic expansion. The observations extend prior studies in IR (Carpineto and Romano, 2012), that have provided evidence for the effectiveness of query expansion, by relating query expansion to the construct of type of word relatedness. The observed effectiveness of semantic and associative expansion is intuitive with regard to the listed terms in Table 9.6. The listing of neighbours of the term 'cancer' illustrates, how a query such as 'radio wave brain cancer' might benefit from the expansion with semantic *or* associative terms.

In relation to the aim of validating relevance, the following observations can be made. The chosen approach to the validation of relevance is given by construct validity and the nomological network methodology. A nomological network establishes the validity of a construct based on characterizing its relation to other constructs. The mechanism was outlined in Section 5.3.1. Characterizing the relation is based on aligning measurements of observables associated with the investigated constructs. The discussion in this section constitutes an example of such an alignment. Measurements of the construct of relevance were aligned with measurements of the construct of type of word relatedness. This alignment provided supportive evidence for the hypothesis, that the existence of associative or semantic relations between the terms of two textual items positively impacts the estimation of their relevance. The hypothesis constitutes a characterizing statement about the relation between the two constructs.

Assessing its significance with regard to the validation of relevance requires to interpret this observation in relation to the concept of convergent validation. As noted by Lachman et al. (1979), the concept describes the idea, that the convergence of several different kinds of data on a conclusion, convergently validates this conclusion. A prior example of the application of convergent validation is given by the validation study in Chapter 8. The semantic and associative orientation of HAL models with differing window sizes, was convergently validated based on different kinds of data. A result of the validation effort consisted of a validated measurement instrument for the type of word

relatedness. The validity of this instrument is based on a chain of 'interlocking' converging conclusions. The chain of interlocking interlocking inferences is composed of priming experimentation data (Chiarello et al., 1990; Ferrand and New, 2004), manual assessments of word type, and human assessments of the grade of relatedness (Rubenstein and Goodenough, 1965; Miller and Charles, 1991; Finkelstein et al., 2002). This illustrates the basic mechanism underlying the establishment of validity through the application of convergent validation. It is established based on a chain of converging and interlocking inferences. Confirmation of observances through convergence with other observations, constitutes an elementary mechanism underlying scientific investigation. In light of this, it is important to emphasize, that a distinction with regard to its application in construct validity is given by the consideration of the validity of each observation in such chains. This underlines the importance of ensuring the validity of the utilized measurement instruments. It is intuitive, that the merit of insights derived from a chain of interlocking inferences is entirely dependent on the validity of the inferences it is composed of.

The so far led discussion illustrated how validity can 'arise' from the formation of a chain of convergently validated inferences. It also provided a summary of the validation effort with respect to the construct of word relation types. To relate these considerations to the validation of relevance requires to explore the motivation for the application of convergent validation. The necessity for its application can be derived from an evaluation of the 'construct' concept. The term 'construct' emphasizes the abstract nature of a phenomenon. Type of word relatedness constitutes a construct. It is a theoretic postulated concept. Based on an extensive body of interlocking inferences, it constitutes a well established constructs. Nevertheless, alike relevance, it still constitutes an abstract theoretic concept. Its formulation and interpretation is based on a chain of interlocking observations. In this specific case, the definition of associative and semantic relation types is based on converging observations of priming reaction times and assessments of the grade of relatedness of words (Chiarello et al., 1990; Ferrand and New, 2004; Vigliocco et al., 2004). Considered in isolation, physiological reaction times and human assessments offer limited insights with regard to word relations. Insights derived from an investigation of their relation, formed a basis for the the definition of the concept and an understanding of its meaning. This illustrates the primary role that relations fulfil in a nomological network. They contribute to the understanding of the constructs. The eventual aim of construct validity consists of establishing a precise definition of a construct. This requires a precise understanding of its meaning.

The so far led discussion allows to interpret the empirical results of the current section with regard to the validation of relevance. The empirical investigation indicated a convergence of retrieval performance and the associative and semantic orientation of the underlying HAL models. With respect to convergent validation, this result can be interpreted as an additional link in a chain of converging inferences. It convergently

validates that the window size parameter impacts the associative-semantic orientation of HAL models, and the validity of the priming simulation methodology for measuring this orientation. The measurement instrument itself was convergently validated based on the alignment with priming reaction times, and human assessments. This outlines, that the starting point of the chain of converging validations lies in cognitive psychology experimentation. An examination of the underlying chain of convergent validations constitutes a prerequisite for making inferences based on the observed convergence of relevance and word relation type. The reliability of any inferences with respect to relevance entirely depends on the validity of the construct of word relation type and its associated measurements. As outlined in Section 5.3.2, the choice of the type of word relations construct was based on considerations of its validity. In light of this, it is considered that the alignment of its measurements with relevance, allows for drawing meaningful inferences. The 'Effectiveness of Semantic and Associative Expansion' hypothesis constitutes an example of such an inference. The underlying aim of formulating the hypothesis consists of contributing the understanding of the meaning of relevance. Given the complexity of the phenomenon of relevance (see Section 4.2), such contributions should be interpreted as small steps forming part of an iterative constructive validation of relevance. This concludes the interpretation of the section's empirical investigations with regard to the the aim of establishing construct validity for relevance.

The next subsection substantiates the demonstration of the application of the nomological network methodology through the formulation of a second hypothesis with regard to the relation of the constructs.

Associative Nature of Recall

The prior section introduced the 'Effectiveness of Semantic and Associative Expansion' hypothesis. Based on an alignment of measurements relating to the constructs of word relation type and relevance, it presented supportive results for the hypothesis. An interpretation of the observations with regard to the aim of establishing construct validity for relevance was provided in the last paragraph of the section.

In this section, a second hypothesis labelled the 'Associative Nature of Recall' hypothesis is introduced.

> **Hypothesis 9.2:** Associative expansion leads to a proportionally larger increase in recall than semantic expansion.

An underlying assumption of query expansion is, that it leads to the retrieval of a larger

number of relevant documents. (Carpineto and Romano, 2012, p. 4) expresses this in the following form.

> *For instance, if the query Al-Qaeda is expanded to Al-Qaeda al-Qaida al-Qaida 'Osama bin Laden' 'terrorist Sunni organization' 'September 11 2001,' this new query does not only retrieve the documents that contain the original term (Al-Qaeda) but also the documents that use different spellings or dont directly name it.*

The above formulated hypothesis expresses, that the expansion with associative terms such as 'terrorist', and 'Sunni' leads to proportionally larger increases in recall than the expansion with a semantic term such as 'organization'. The hypothesis is investigated based on an analysis of the observed performance over the complete parameter space.

Parameter Space Performance Figure 9.6 provides an overview of recall performance for the different retrieval tasks. The listed recall values constitute the arithmetic means with respect to the K value bins. That is, the recall values shown in the top-left plot constitute mean recall performance for the expansion with 1 to 5 terms. An increase in recall can be observed in the first three rows of Figure 9.6. The applied query expansion leads to the retrieval of a higher number of documents for these tasks. No increase in recall is registered for the difficult .GOV TD task.

With regard to the semantic-associative focus of the underlying HAL models the following can be said. Figure 9.4 outlined, that HAL models with window sizes of 1 and 2 are semantically focused. HAL models with a window size larger than 6 are primarily associatively focused. Figures 8.4 and 8.7 show that this observation is also reflected by the neighbourhood assessments. The alignment with measurements associated with the type of word relatedness highlights, that increases in recall can be predominantly observed for associatively focused HAL models. With the exception of the right-most plot in the first row, and the left-most plot in the third row, no increases in recall for expansion with primarily semantic terms are registered. In both cases the increase in recall is proportionally much smaller than that achieved by associatively focused expansion. Figure 9.7 illustrates that this also applies for larger bins of K. Expansion with semantic terms does not lead to an increase in recall relative to the baseline. However, as can be seen in the figure, expansion with semantic terms leads to lower reductions of retrieval performance.

To substantiate these observations, a next step consists of an analysis of the performance with regard to different weighting schemes of HAL models. Figure 9.8 illustrates the performance over the different weighting schemes. The plots in the figure show the aggregated mean performance for expansion with K values in the range of 1 to 15. The

plots in the figure generally confirm the previous observations that associative expansion leads to larger increases in recall. Of interest with respect to the different weighting schemes is the lower performance of HAL models with a geometric weighting scheme. As illustrated in Figure 8.6, the application of geometric weighting leads to a distinctly lower associative focus of HAL models for large window sizes. The observed lower recall performance could be attributed to the less associative nature of geometric HAL models. An alternative interpretation of this observation is given by the explanation that the lower performance could be implied by a lower 'quality' of the HAL models. That is, a failure of the respective HAL models to identify related terms. However, an analysis of the retrieval performance in Tables 9.1 to 9.5 shows a strong performance of geometric weighting based HAL models with large window sizes. It therefore seems unlikely that such models perform worse in the identification of related terms. This indication is strengthened by contrasting GMAP and recall performance for different weighting schemes. Figure 9.5 shows the performance of the three weighting schemes for the TREC 8 Ad Hoc task.

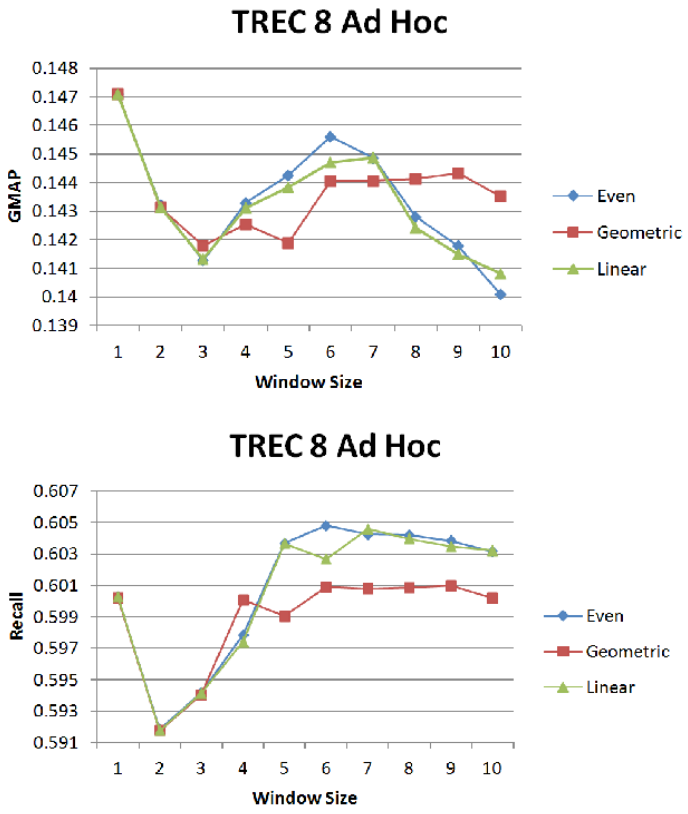


FIGURE 9.5: GMAP and Recall Performance of Even, Linear, and Geometric Weighting Schemes on TREC8 Ad Hoc task.

This indicates, that the observed differences might indeed be attributed to the less associative nature of geometric models. This can be considered as a supporting observation with respect to the 'Associative Nature of Recall' hypothesis.

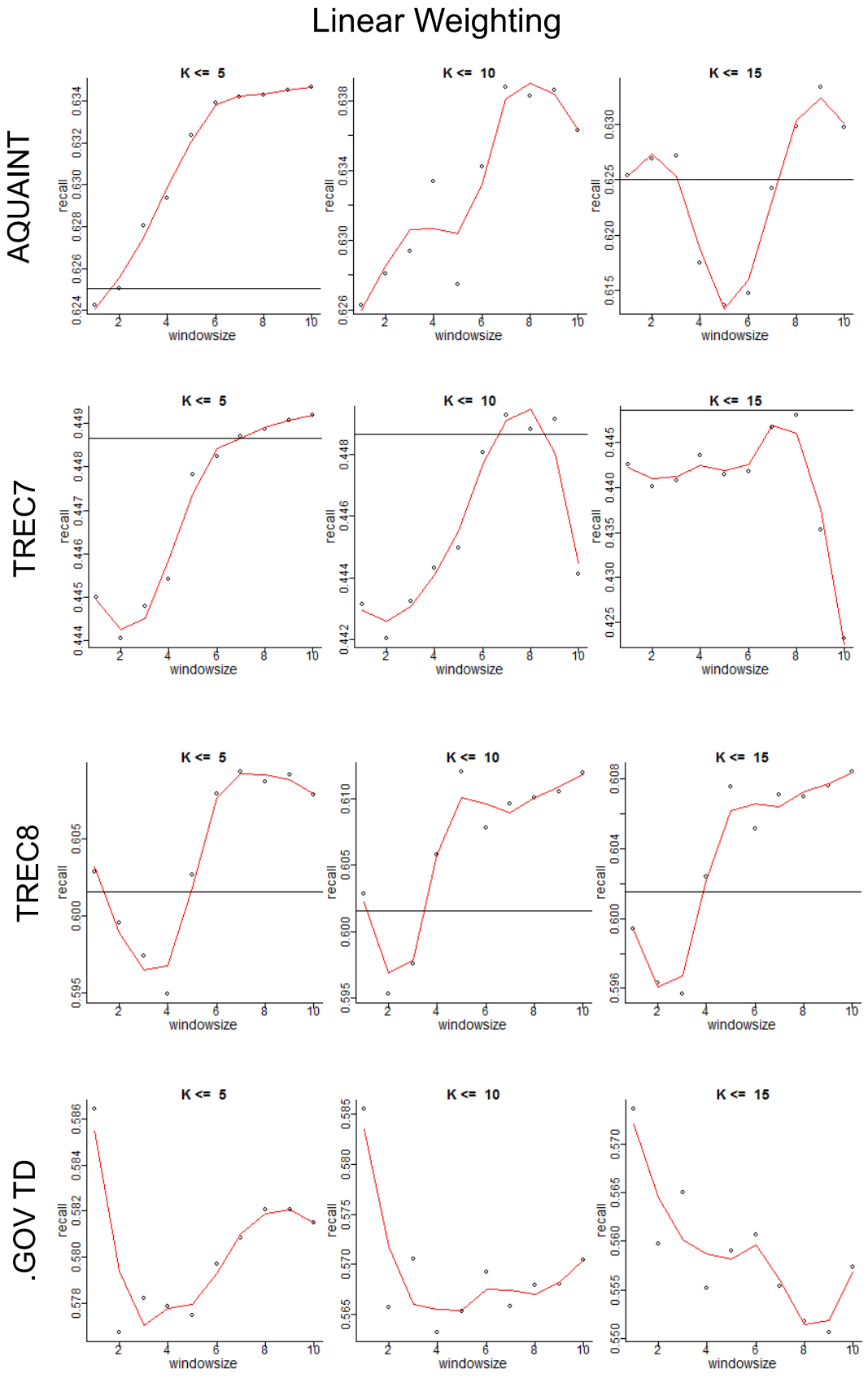


FIGURE 9.6: Arithmetic Mean of Recall Performance for the Expansion with Ranges $K <= 5$, $K <= 10$, and $K <= 15$

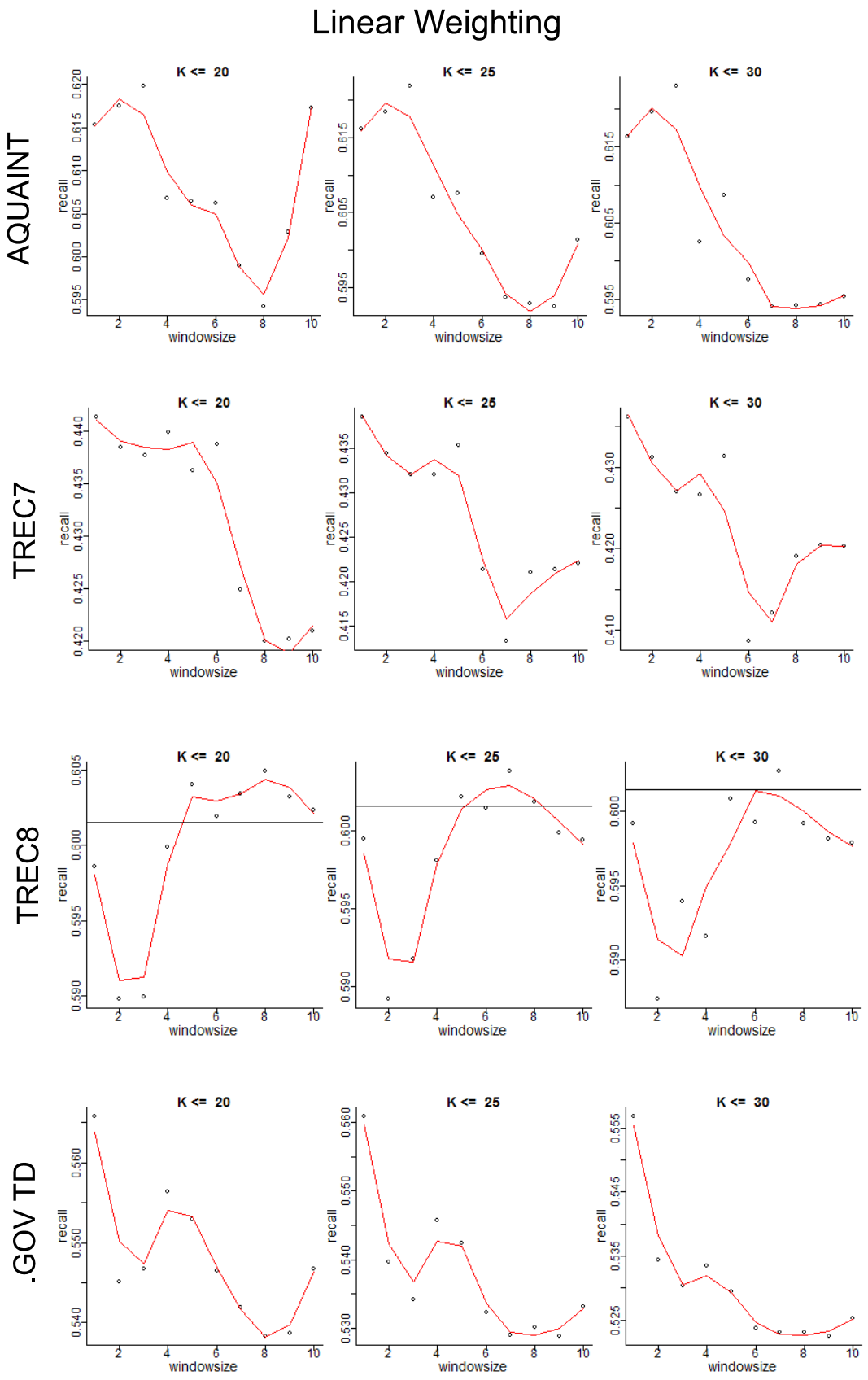


FIGURE 9.7: Arithmetic Mean of Recall Performance for the Expansion with Ranges $K <= 20$, $K <= 25$, and $K <= 30$

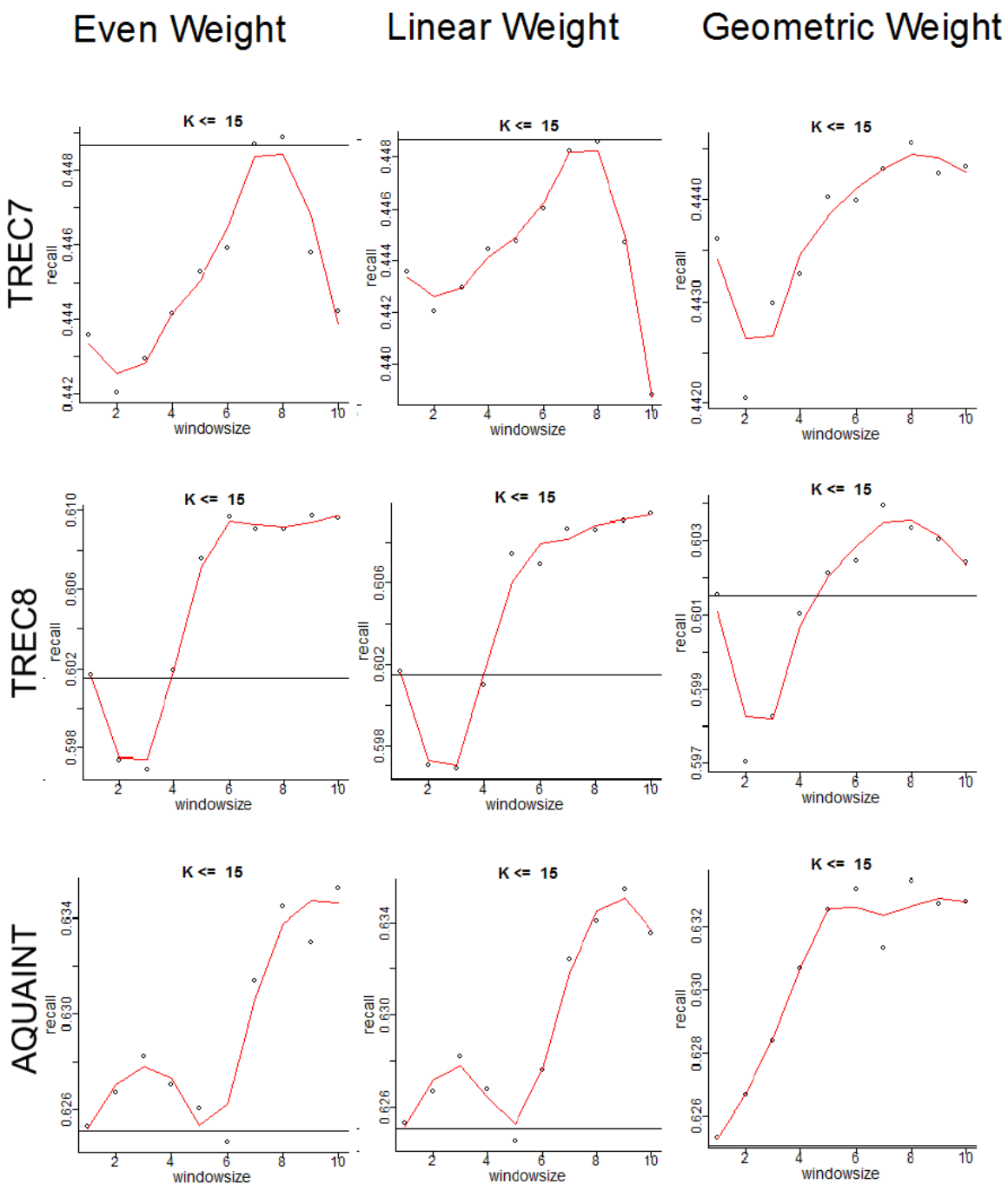


FIGURE 9.8: Mean Recall Performance over Window Size Parameter. TREC 8 Ad Hoc, TREC 7 Ad Hoc, AQUAINT Robust 05

Discussion The exploration of the performance over the parameter space constituted a second example of the application of the network. The alignment of measurements from the constructs and relevance and type of word relatedness indicated support for the hypothesis that associative expansion leads to a proportionally larger increase in recall than semantic expansion. However, these conclusions are of a tentative character. Figure 9.5 outlined that the observed performance differences are small. This might be implied by the fact, that neither HAL nor LSA constitute purely semantic or associative models of word relatedness. This is outlined by the neighbourhood assessment-based evaluations in Section 8.2.2. Adding supportive evidence to the observations can be based on the consideration of prior work regarding query expansion. The number of published studies investigating query expansion with regard to semantic and associative aspects is limited. To the best of our knowledge only one study was concerned with the investigation of such aspects. Greenberg (2001) conducted an investigation of query expansion using a controlled vocabulary mapping synonyms and partial-synonyms, narrower terms, broader terms, and related terms. The first three categories can be attributed to the definition of semantic relatedness. Greenberg's (2001) notion of related terms is interpreted to be closely related to the definition of associative relatedness. Greenberg (2001) reported, that the highest gains in recall were achieved based on expansion with related terms. The highest precision was recorded for expansion with synonymous terms. This is congruent with the observations in Figure 9.5. In the next subsection a third hypothesis is explored.

Robustness of Semantic Expansion

To outline how inferences about the characteristics of the relations of constructs in a nomological network can be derived, a third hypothesis is formulated. The 'Robustness of Semantic Expansion' hypothesis assumes the following.

> **Hypothesis 9.3:** Expansion with semantically related terms is more robust than expansion with associatively related terms

This expresses the notion, that the probability of 'query drift'. Mitra et al. (1998, p.204) describe query drift as 'the alteration of the focus of a search topic caused by improper expansion'. The fact that improper expansion leads to a reduction of retrieval performance is a phenomenon described in many query expansion focused studies (Xu and Croft, 1996; Collins-Thompson, 2009; Carpineto and Romano, 2012). The hypothesis assumes, that the expansion with semantically related terms (e.g. whale–dolphin) bears a lower risk of query drift, than the expansion with associatively related terms (e.g. spider–web).

Parameter Space Performance A first analysis with regard to the hypothesis can be based on the previously discussed Figures 9.1 and 9.2. The development of retrieval performance over the bins of K shows, that MAP for all retrieval tasks degrades more rapidly for HAL models with larger window sizes (i.e. associatively focused HAL models). Specifically, in Figure 9.2, it becomes evident that performance for expansion with semantic terms exhibits higher stability. This is a first indication that the occurrence of query drift is more likely when expanding with associated terms. To substantiate this indication, the next step of the analysis consists of the consideration of the standard deviation of the retrieval results. The hypothesis assumes, that expansion with associated terms is more 'risky'. That is, it is more likely to result in query drift caused by improper expansion. It is presumed, that a higher risk of improper expansion results in a higher variation of query performance. This should be reflected in a higher standard variation. Figure 9.9 shows standard variation of MAP plotted over the window size parameter for the full parameter space (i.e. all weighting schemes and $K \in 1, 2, ..., 30$). As can be seen in the figure, standard deviation steadily increases with larger window sizes. This supports the assumptions of the hypothesis. Another supportive indication is given by the lower observed standard deviation of geometric weighting based HAL models. As previously noted, such models exhibit a lower associative focus. Figure

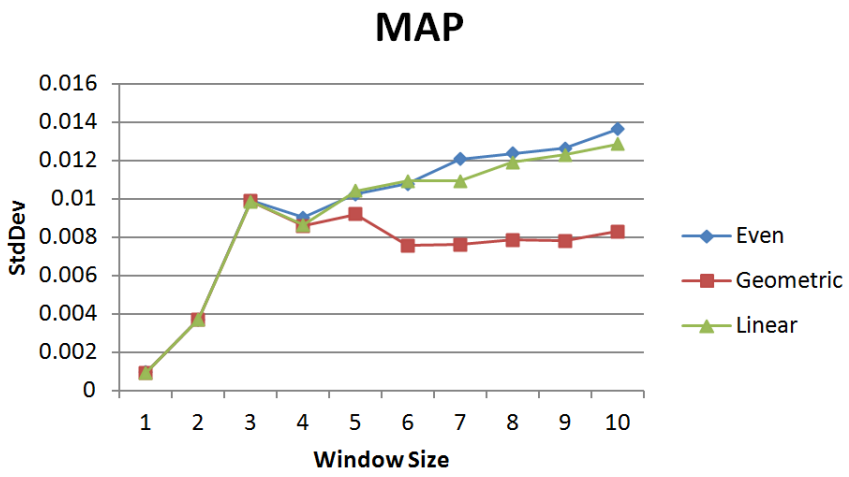


FIGURE 9.9: StdDev of MAP over full parameter space. TREC 8 Ad Hoc Task

9.10 shows that the same observations can be made based on the standard deviation of the GMAP measure. Standard deviation is highest for larger window sizes, and lowest for a window size of 1 (i.e. semantically focused HAL model). The interpretation of the impact of the associative-semantic factor on robustness can be confirmed by examining the standard deviation of recall. Figure 9.11 shows a plot of the standard deviation of recall over window sizes. As can be seen in the figure, the same relation between window size and standard deviation is exhibited. With respect to the previously made observations, this is indicated to be supportive of the hypothesis that the expansion with

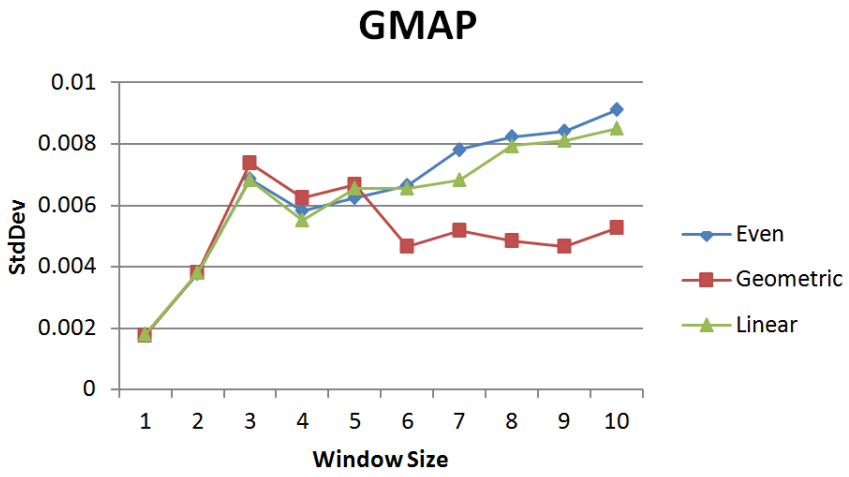


FIGURE 9.10: StdDev of GMAP over full parameter space. TREC 8 Ad Hoc Task

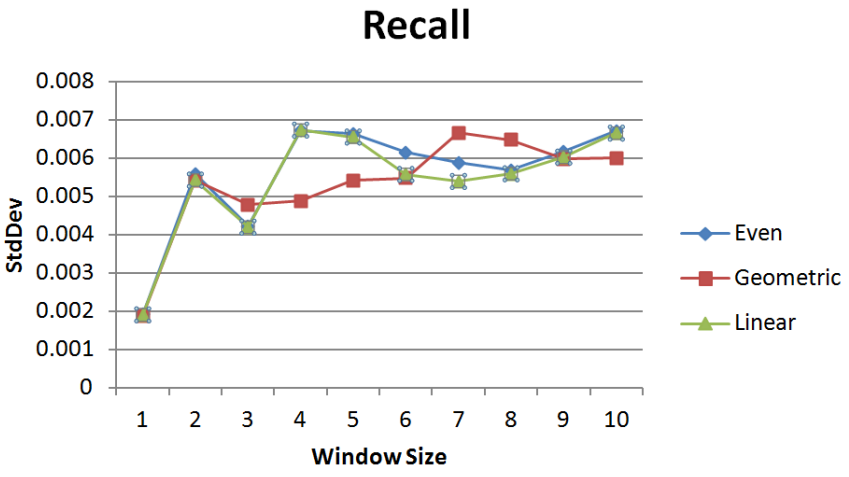


FIGURE 9.11: StdDev of Recall over full parameter space. TREC 8 Ad Hoc Task

associative terms is less robust than semantic expansion.

Discussion The previous paragraph reported on an analysis with regard to the 'Robustness of Semantic Expansion' hypothesis. The observations with regard to retrieval performance and its variance were interpreted to be supportive of the hypothesis that associative expansion bears a greater risk of causing query drift. A general conclusion that can be drawn is that retrieval performance degrades with increasing numbers of expansion terms. This conclusion can also be in interpreted in terms of Fang and Zhai's (2006) term similarity function s in $STMC1$. Fang and Zhai (2006, p. 117) defined s as follows.

> "*Without loss of generality, we assume that term t is semantically more similar to term u than to term v if and only if $s(t,u) > s(t,v)$, i.e., a large value of s indicates a high similarity.*"

This definition implies that term relatedness is a monotonic function. Its integration in $STMC1$ (see Section 9.1.3) suggests, that Fang and Zhai (2006) assume that term relatedness and relevance are positively correlated. That is, the more closely related two terms are, the higher their contribution to the retrieval score should be. And vice versa, the less related a term is, the less it should contribute. The observations of this section support the assumption. In the empirical data this is expressed in terms of a degrading retrieval performance for larger K. The query expansion mode underlying the experiments starts the expansion with the closest related term. The observed retrieval performance indicates that less related terms are indeed less likely to make a positive contribution. The observed retrieval performance and standard deviations suggest, that the steepness of the s might differ with respect to semantic and associative relations. This might be implied by the nature of associative and semantic relations. Semantic relations require the existence of shared features between terms. No such constraints exist for associatively related terms. It can be assumed, that associatively related terms are more likely to form part of distinctly different topical context. However, such considerations are of speculative character. They are meant to outline potential future hypotheses that could be based on the observations of this section. This concludes the empirical investigations within this chapter. The next section summarizes the findings and previous discussions.

9.3.6 Discussion

Three hypotheses with respect to the relation of relevance and word related constructs were formulated in this section. The underlying aim for their formulation consists of contributing insights to the construct of relevance, and to demonstrate the application of the nomological network. In general, the observations resulting from the conducted analysis support each of the three hypotheses. However, at this point, this is interpreted as constituting only tentative evidence for the assumptions underlying the hypotheses. This is based on the following reasoning.

A cautious approach to the interpretation of the presented results is taken with regard to the complexity underlying the phenomenon of relevance. Exemplary, this can be based on observations of previous studies on query expansion. Carpineto and Romano (2012) identified great variation with regard to the results obtained from query expansion based experimentation. He attributed the variability of the results to differences in test collections with respect aspects such as size, noise, heterogeneity of contents, and difficulty of topics. With regard to this empirical investigation, varying results can be observed for the .GOV Topic Distillation (TD) task Aquaint Robust 05 task. The TD task differed from the results obtained on the other tasks by showing no improvements based on query expansion. In the case of the Robust 05 task, the observed retrieval performance differs from the observations made with regard to the TREC 8 and TREC 7 Ad Hoc tasks. This can be seen in Figure 9.1 where, relative to the other tasks, retrieval performance degrades much faster for associatively focused HAL models. Both observations might be related to Carpineto and Romano's (2012) reference of query difficulty. However, it seems reasonable to assume, that the exploration of such hypotheses requires an extension of the presented experimentation with respect to the following aspects.

Manipulation of the Semantic-Associative Variable Constraints on the manipulation of the semantic-associative variable are identified as a first potential limitation induced by the experimental setup underlying the empirical investigations of this chapter. The manipulation of the variable is interpreted to be constrained with regard to two aspects. The first is given by the observation, that neither HAL nor LSA exhibit purely semantically or associatively focused models. As outlined in Sections 8.2.2 and 8.3.2, the term neighbourhoods of both models are represented by a ratio of semantic and associative terms. A second constraint results from the restrictions induced by the window size parameter. A more fine grained, and potentially continuous way of manipulating the semantic-associative axis might benefit the empirical investigation of the relation of the constructs. With respect two both constraints, a potential solution is given by the development of novel computational models of word relatedness that exhibit distinctly semantic or associative focus. This consideration is related to a second identified constraint of the current experimental setup.

Measurement on the Semantic-Associative Axis The conducted empirical investigations indicated distinct differences with regard to the impact of semantic and associative relations on retrieval performance. With regard to the evaluation of these observations, the availability of more sensitive measurement instruments for the type of word relatedness would be beneficial. As noted in Section 8.1 only one study dedicated to the assessment of semantic associative word relations has been published. It was noted, that certain limitations apply to the study with regard to the definition of the measured variable. It is assumed, that the development of better measurement instruments would strengthen the ability to make inferences based on the nomological network. Further, it is assumed that this would be beneficial to the construction of the above described purely semantically or associatively focused models. The existence of such instruments would also allow for a more fine-grained analysis of variance and render the application of techniques such as factor analysis more feasible.

Expansion of Experimental Context Thirdly, the formulated hypotheses could be strengthened by an expansion of the experimental context. This concerns two main aspects. The first is given by expanding the empirical investigations to additional retrieval tasks, topics, and domains. Specifically, an increase in the number of queries is considered beneficial. Mitigating the effects of query drift was approached by binning over values of K. A more optimal form of compensating for potential outliers is given by an expansion of the experiments on a larger number of queries. Further, as noted in Section 9.2.7, a more ideal testbed for the investigation of the relation of the constructs would consist of a test collection featuring complete relevance judgements. A final aspect with regard to the expansion of the experimental setup is given by the consideration of additional constructs. Additional, and perhaps more concrete inferences, could result from the introduction of constructs such as term importance, and document authority.

Nevertheless, meaningful observations can be derived from the reported empirical results. While none of the three specific hypotheses are considered to be conclusively confirmed, the observed empirical results provide strong support for the hypothesis that the impact of associative and semantic term relations on the estimation of relevance differs distinctly. This outlines two important aspects with regard to the application of the nomological network. Firstly, it demonstrates that meaningful inferences can be derived by relating measurements associated with word relatedness constructs and the relevance construct. Secondly, these inferences offer insights with regard to both constructs. This exemplary demonstrates the principle of convergent validation. The observed distinct differences with respect to the impact of semantic and associative relations convergently validates earlier conclusions about the validity of the measurement instruments and the semantic-associative focus of HAL models of varying window sizes. This illustrates, how validity can be based on the establishment of a chain of interlocking inferences. Generally, the investigation of the three hypotheses illustrates how the relation of con-

structs on a lower level of abstraction can lead to the contribution of insights for the construct of relevance. This concludes the consideration of the empirical part of the chapter. The next section presents concluding remarks and the answer to RQ 6.

9.4 Chapter Conclusions and Answer to RQ 6

The investigations in this chapter were guided by RQ 6.

> **RQ 6** What are characteristics of the relation of the postulated constructs of relevance and grade and type of word relationships?

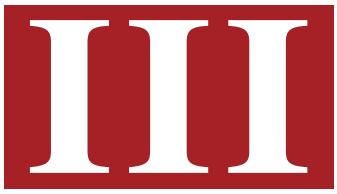
Approaching RQ 6 was pursued based on the following two sub-questions.

> **RQ 6a** Does a relation between the postulated constructs of relevance and grade and type of word relationships exist?
>
> **RQ 6b** What are characteristics of the relation between the postulated constructs of relevance and grade and type of word relationships?

Section 9.3.3 investigated RQ 6a. Based on the analysis of retrieval results obtained on four different tasks, and the alignment of these observations with measurements of type and grade of word relatedness it was concluded that the constructs are related to each other.

The investigation of RQ 6b was guided by the definition of three hypotheses. They are interpreted as potential characteristics of the relation between the three constructs. Converging observations from the empirical investigation of these hypotheses strongly suggest that semantic and associative relations distinctly differ with regard to their impact on relevance estimation.

As outlined in the introduction of the chapter, one of the underlying motivations for the formulation of RQ 6 consists of demonstrating the application of a nomological network. With regard to this aspect, the drawn conclusions illustrate that meaningful inferences can be derived by relating measurements associated with different constructs. Further, the conducted analysis exemplary outlined the application of the concept of convergent validation. Finally, the demonstration of the application of the network, and the reported discussions emphasized the underlying complexity of the task of validating relevance.



CONCLUSION

After having investigated theoretical considerations regarding a validation of relevance in Part I, and having explored empirical concerns of the application of an IR focused nomological network in Part II, the thesis comes to a conclusion in this final part. The final part is structured as follows. First, the research questions of the dissertation are recalled and a summary of the answers to each of them is provided. Then a listing of what are believed to be the main contributions of this work is provided. Finally, the dissertation concludes with recommendations for future research.

CHAPTER 10
CONCLUSION

The investigations of this thesis focused on considerations with regard to the validation of relevance. The focus of the dissertation lead to the formulation of the following problem statement in Chapter 1.

PS *How can Information Retrieval centric constructs be validated?*

This chapter concludes the thesis. Section 10.1 provides a summary of the answers to the six main research questions of the dissertation. In Section 10.2 the potential usefulness and applicability of the research approach of the thesis is discussed. The discussion focuses on identifying the differences of the chosen approach to other research approaches on relevance, and outlines potential scenarios for its application. The main contributions of this work are summarized in Section 10.3. The chapter is concluded in Section 10.4, where future research directions are outlined.

10.1 Answers to Research Questions

The defined problem statement led to the formulation of six main research questions. Two additional subquestions were formulated during the course of the investigations. In this section, a summary of the answers to the eight research questions is provided. The first three research questions focused on theoretical considerations of the validation of IR constructs.

RQ 1 What constitutes a principled approach to construct validation in IR?

The first research question addresses the need of identifying a principled approach to the validation task set by the problem statement. The question is of analytical character and focused on an exploration of the learnt lessons in cognitive science with respect

to the validation of cognitive constructs. It identified that construct validity constitutes the dominant approach to the validation of constructs in cognitive science. Construct validity constitutes a proposal for the conduction of validation in scenarios, where no direct observation of a phenomenon is possible. With regard to the limitations on direct observation of a construct, it bases validation on the evaluation of its relations to other constructs. Based on the interpretation of relevance as a product of cognitive processing it was concluded, that the limitations with regard to direct observation apply to its investigation. In light of this, construct validity was identified as a principled approach to the validation of the construct of relevance. The evaluation of its applicability to an IR context, focused on the exploration of the nomological network methodology. A nomological network constitutes an analytically constructed set of constructs and their relations. The construction of such a network forms the basis for establishing construct validity through investigation of the relations between constructs. The suggestion of the method of construct validity was based on an exploration of the intrinsic challenges of validating cognitive phenomena.

An analysis focused on contemporary insights to the nomological network methodology identified two important aspects with regard to its application in IR. The first aspect is given by a choice of context and the identification of a pool of candidate constructs for the inclusion in the network. These considerations led to the formulation of the second research question of the thesis. The second aspect addresses the requirements for 'realism' identified by Kane (2006) and Borsboom et al. (2004). RQ 3 was formulated to guide the investigation of these concerns.

RQ 2 What are potential constructs for the formulation of an IR focused nomological network?

Research question two addresses the initial step underlying the construction of a nomological network: The identification of the pertinent constructs for the network. The context for the exploration of this question was set by the definition of the Correlation to Cognition analogy. The analogy interprets relevance as a phenomenon emanating from the interaction between two systems: An IR system, and the cognitive processing system of the user. In light of the defined context, the investigation of RQ 2 was based on a review of the principles of cognitive exploration, and an analysis of the state of the art in text based discourse processing and reasoning. On that basis, a listing of known sub-processes contributing to the pertinent cognitive processing was presented. The listing demarcates the pool of potential constructs from the cognitive processing side of the Correlation to Cognition analogy. The pool of potential candidates for the network was complemented, by relating the cognitive constructs to the IR side of the analogy. Based on the cognitive bands defined by Newell (1994), the resulting pool of candidate constructs was categorized with respect to the level of abstraction of the constructs. It was emphasized, that the initially defined pool is of preliminary character, and with respect to the nature of construct validation, is meant to be iteratively refined.

Based on the identification of a large number of potential candidates, the next step consisted of the inference of criteria for the selection of an initial set of constructs for the network.

RQ 3 What are criteria for the selection of constructs?

The third research question represents a central and fundamental element of the research approach underlying this work. The investigation of these criteria focused on the consideration of pragmatic and meta-theoretical aspects. Based on a survey of experimental means in cognitive science and IR, five pragmatic criteria for the selection of constructs were presented. These criteria focus on pragmatic aspects concerning the measurement of observables associated with the constructs in a nomological network. Consideration of meta-theoretically motivated criteria required to investigate the question, of what constitutes the specific challenges with regard to the validation of highly abstract constructs. This question was explored based on the underlying considerations of the Information Processing paradigm and Newell's (1994) cognitive bands. This led to the identification of a set of three meta-theoretical criteria for the selection of constructs. The three criteria provide a basis for the selection of constructs with regard to the call for 'realism', that is motivated by observed limitations of nomological networks consisting only of high level constructs.

Based on the criteria and the demarcated candidate pool, an IR focused nomological network was defined. The network consists of the constructs of relevance and type and grade of word relatedness. The choice of the word related constructs was motivated by the criteria identified by RQ 3. The presentation of the network concluded the theoretically focused part of the thesis.

The empirically focused part of the dissertation was guided by three main research questions. The first two research questions are concerned with the validity of measurement instruments associated with the word related constructs of the network. The empirically and analytically approached research questions are motivated by the observation, that inferences based on the analysis of construct relations, are dependent on the validity of the measurements associated with these constructs.

RQ 4 What are valid instruments for the measurement of the grade of relatedness between words?

RQ 5 What are valid instruments for the measurement of the type of relatedness between words?

The foundation for the investigation of these research questions was set by outlining the theoretic background of the constructs, and by the conduction of a survey of available measurement instruments. With respect to the inconclusive results of a preliminary analysis with regard to the validity of these instruments, a strategy for their validation was formulated. Methodologically the strategy was based on the concept of convergent

validation. Convergent validation describes the idea, that a conclusion can be validated based on the convergence of several different kinds of data. Based on the formulated strategy, an experimental setup encompassing different kinds of measurements and a large experimental space was defined. Two validation studies were conducted based on this setup.

The first validation study concluded, that the FS353 assessment method constitutes a valid instrument for the measurement of graded word similarity. It was observed, that the question of validity in the cases of the R&G and M&C assessment methods cannot be conclusively confirmed. The second validation study concluded that supportive evidence exists for the validity of the neighborhood assessment and priming simulation methods as measurement instruments for the type of word relatedness. The conclusion was based on the observation of concordance of the respective measurements and the plausibility of these measurements with regard to the algorithmic characteristics of the models and the theoretic basis of the construct.

The clarification of the question of the validity of the measurement instruments enabled the application of the nomological network. The examination of the relation between the constructs of relevance and word relatedness is addressed by the last research question.

RQ 6 What are characteristics of the relation of the postulated constructs of relevance and grade and type of word relationships?

Approaching RQ 6 was pursued based on the following two sub-questions.

RQ 6a Does a relation between the postulated constructs of relevance and grade and type of word relationships exist?

RQ 6b What are characteristics of the relation between the postulated constructs of relevance and grade and type of word relationships?

The investigation of these two research questions is based on the strategy to induce variation with respect to the word related constructs, and to observe the impact of this variation with respect to the construct of relevance. Based on this strategy, RQ 6a was investigated by analyzing retrieval results obtained on four different tasks. With regard to the results obtained from statistical tests, it was concluded that the constructs are related to each other.

The investigation of RQ 6b was guided by the definition of three hypotheses relating to different aspects of the relations between the word related constructs and relevance. The empirical investigation of RQ 6b focused on the evaluation of these hypotheses. The evaluation was based on the alignment of observed retrieval results with word related measurements. It was concluded, that the conclusive confirmation of the hypotheses requires an extension of the experimental means underlying the study. Based on con-

verging observations from the empirical investigation of the three hypotheses, and with regard to the results of the previous validation studies it was concluded, that semantic and associative relations distinctly differ with regard to their impact on relevance estimation.

The drawn conclusions illustrated, that meaningful inferences can be derived by relating measurements associated with different constructs.

10.2 Applicability of the Paradigm

Based on the previously presented work in the dissertation, this section explores considerations with regard to the applicability of the proposed research approach. The section focuses on outlining suitable research scenarios for its application in IR and a discussion of the potential usefulness of such applications. A basis for the discussion of these aspects is set by investigating how the proposed research approach differs from existent research approaches in IR.

Differentiation to Existent Research Approaches in IR

The research approach followed in the dissertation differs from existent approaches with regard to two main aspects: The interpretation of the cognitive system of the user and the implications of this view on the investigation of relevance.

With regard to the first point, this work differs from previous user focused research in IR (Ingwersen, 1994; Belkin et al., 1993; Arapakis et al., 2008) through its adherence to the Information Processing (IP) paradigm (Palmer and Kimchi, 1984) of cognitive science. Based on the IP paradigm's decomposition tenet, the user is represented as a system of hierarchically structured cognitive processes. The processes in the system are characterized by the completeness of informational description tenet of the IP paradigm. The tenet expresses, that more elementary processes within the hierarchical process chain exhibit a higher completeness of informational description. That is, the knowledge about the variables influencing the system is more complete. The IP paradigm is motivated (Chomsky, 1959) by the assumption, that it is more feasible to measure and model such processes. This outlines, that these principles of cognitive exploration not only result in the description of cognition in terms of hierarchical models, but also that these models categorize cognitive observations with respect to meta-theoretical considerations. The consideration of the IP paradigm and Newell's (1994) cognitive bands results in a representation of the user in form of a hierarchical system of processes that are categorized with respect to the level of abstraction and measurability.

This perspective influences the research approach underlying the investigations in this dissertation. Level of abstraction and measurability are inversely correlated in the above described interpretation of the user 'system'. This expresses, that highly abstract concepts, exhibit low measurability. Due to the complexity of the processes, and the larger number of contributing subprocesses, the knowledge about the involved variables is less complete. Within this work, relevance is interpreted as a highly abstract concept, and consequently assumed to exhibit low measurability. Section 5.3 outlined, how this can be interpreted to limit the ability to make inferences based on relevance measurements. These observations constituted the basis for the choice of the research approach followed in the dissertation. Within this work, the investigation of relevance is based on the principles of construct validity. Construct validity expresses, that the validity of a construct (i.e. a postulated theoretic concept) can be established based on the evaluation of its relation to other constructs. This is a reflection both of constraints on direct observation of cognitive phenomena, as well as of the low measurability of highly abstract concepts. Methodologically, the investigation of the relations of constructs is realized based on the concept of the nomological network. Section 5.3.1 outlined, how research based on relating relevance to concepts such as pertinence, utility, and usefulness can be interpreted as the act of constructing a nomological network. This work differs from this line of research with regard to its focus on validity. The evaluation of relations between constructs is based on relating measurements associated with these constructs. With regard to the focus on validity, a prerequisite is given by ensuring the validity of those measurements. Inferences about the relation between two constructs are considered meaningful only, if the associated measurements are validated. This has been reflected by this work with regard to the choice of word related constructs for the proposed nomological network, and the conduction of validation studies for these word related constructs. With respect to the validation studies, this work differs from related research through the application of the principle of convergent validation. Convergent validation expresses, that conclusions can be validated based on converging observations made on measurements of different kinds of data. The concept emphasizes that validation is an iterative process. It expresses, that validity cannot be determined in a single act, but rather is established through continued addition of converging observations. Within this work, this was reflected by the structuring of the work in Chapters 7, 8, and 9, and the considerations expressed with regard to the reported results.

This illustrates how this work differs in its approach to the investigation of relevance. In light of this, it can also be illustrated how this work differs from the system paradigm in IR. Within the system paradigm, the validity of relevance is presupposed. It is assumed, that relevance can be reliably and objectively measured. Whereas in the research approach of the thesis, the question of the validity of relevance constitutes the main focus. A second distinction can be made with regard to the mode of evaluation. In the system paradigm, evaluation is based on evaluating modelled variables (e.g. term importance, document importance) based on their integration in an IR system and the measurement

of application performance. Contrary to this, with respect to the IP paradigm, this work considers the direct evaluation of elementary processes. The differences in the chosen research approach can be summarized as follows. With regard to research focusing on relevance and cognition, this work differs in the interpretation of the user system and the considered levels of abstraction. With regard to the system paradigm, the work differs with regard to the interpretation of the validity of relevance and the empirical evaluation approach.

Potential Application and Usefulness

This subsection aims at assessing the usefulness and application scenarios of the research approach followed by the dissertation. The discussion focuses on two points: The potential usefulness of insights about the relation of relevance to other constructs, and the utility for the identification processes contributing to the estimation of relevance.

The evaluation of the first point can be based on the results obtained from the alignment of the measurements of relevance and word related constructs. The alignment resulted in the formulation of three tentative hypotheses, and the conclusion that semantic and associative terms differ in their impact on relevance estimation. With regard to the viewpoint of Saracevic (1997, p. 17) that 'unlike art IR is not there for its own sake, that is, IR systems are researched and built to be used', an evaluation of the usefulness of this insight requires to evaluate if the insight can be beneficially applied to a retrieval scenario. That is, if knowledge of the differing impact of semantic and associative terms can be beneficial with regard to the application of query expansion, pseudo relevance feedback, concept modelling, and query drift mitigation. While this seems plausible, the question cannot be conclusively answered within the scope of this work. Illustrating the potential for an application of the approach on other scenarios is based on consideration of the use of hyperlink information in IR. A widely applied technique for the utilization of hyperlink information is given by the PageRank algorithm (Page et al., 1999). In the original publication, the algorithm was suggested as a measure of the importance of a web page. The assumption was evaluated based on integrating the algorithm into a retrieval system, and an assessment of the performance of the system. The research approach followed by this dissertation could be applied to this scenario based on the decomposition of 'importance' into constructs with an assumed lower level of abstraction such as authority, popularity, and quality. Insights with regard to the relation of these constructs and relevance could be based on the alignment of measurements of these construct and relevance. This would require the creation of computational models and measurement instruments for the constructs. Further, a validation study of the measurement instruments would be necessary. Insights gained with respect to such experimentation could then be applied with regard to the aims of

creating more accurate models of the constructs and to achieve better retrieval performance. This example illustrates a potential scenario for the application of the approach. It also serves to illustrate potential disadvantages and limitations of the method. A disadvantage exists with regard to the necessary effort for the creation of models for the constructs and the validation of their measurement instruments. Within this work, the necessary effort is documented by the validation studies conducted in Chapters 7 and 8. A limitation of the approach exists with respect to the feasibility of measuring constructs. A prerequisite for basing inferences on relating constructs is given by the availability of validated measurement instruments for these constructs. It is unclear if validated instruments can be constructed for higher level constructs such as popularity, quality, and authority. The creation of PageRank was motivated by the intuition that the importance of a web page constitutes an important factor, and can be modelled based on the use of hyperlinks. Evaluating these assumptions within an application represents a more a more straightforward and less effort requiring approach to evaluation. The utility of the research approach of this work depends on whether the potentially gained insights justify the required necessary effort.

The following can be said with regard to the utility of the research approach for the identification of processes contributing to the estimation of relevance. Evaluating this aspect can be based on the work of Cuadra and Katter (1967). The study of Cuadra and Katter (1967) empirically outlined, that measurements of relevance vary dependent on the instructions given to the assessor. The study was motivated by the aim to improve the measurement of relevance by identifying which factors influence its judgement. The obtained results led to the assumption that 40 to 50 variables might influence the judgement of relevance. It was suggested, that these variables should be researched through a winnowing and iterative process. With regard to this research proposal, the interpretation of the user 'system' portrayed in this work might form a potential source for the identification of processes and variables impacting the estimation of relevance. As outlined in Section 4, models of discourse comprehension and reasoning document the large influence of processes such as source-credibility bias, heuristic information processing bias, and self-confirmation bias on the interpretation of textual of information. As such, it is conceivable that these factors also influence the judgement of relevance. The consideration of these and other factors of the cognitive processing system of the user might be beneficial to the understanding and measurement of relevance.

Summary

This section outlined in what regard this work can be interpreted to differ from previous research approaches in IR. It was concluded that this work differs with respect to the considered level of abstraction, the applied empirical research approaches, and the interpretation of the validity of relevance. With respect to Thagard's (2005) state-

ment that the 'best way to grasp the complexity of human thinking is to use multiple methods' (p. 10), this work is seen as a small complementary addition to existing approaches targeting the investigation of relevance. With regard to the potential usefulness and application of the research approach two potential scenarios were illustrated. It was concluded, that a potential disadvantage of the approach is given by the larger required effort, and that its general utility is dependent on whether the potentially gained insights justify this necessary effort.

10.3 Summary of Contributions

In this thesis we have performed a principled investigation with regard to the validation of Information Retrieval constructs. We have made the following contributions:

1. We have devised a principled approach for the validation of relevance. The formulation of the approach was based on an analysis of the learnt lessons of cognitive science regarding the validation of cognitive phenomena. An investigation of its applicability to IR was based on a survey of the state of the art of text based discourse comprehension and reasoning.

2. Under consideration of the principles of cognitive exploration, a methodology for the construction of IR focused nomological networks was developed. It is based on the consideration of pragmatic and meta-theoretical aspects with regard to the construction of networks.

3. A strategy for evaluating the validity of word related measurement instruments was formulated. Its formulation was based on the concept of convergent validation, and the consideration of the principles of the validation of cognitive phenomena.

4. A validation study of measurement instruments of the grade of word relatedness was conducted. The conduction of the study was based on the use of four test collections, and several word similarity focused measurement instruments from cognitive science and linguistics. The study identified that the validity of assessment based measurement instruments is dependent on the associative-semantic focus of the assessed word relations. Based on this finding it suggested that data sets covering larger numbers of related word pairs exhibit higher validity.

5. A validation study of measurement instruments of the type of word relatedness was conducted. A novel assessment method for the measurement of the semantic-associative focus of computational word models was introduced. Based on the results obtained through this method it was concluded, that the semantic-associative

focus of computational models can be measured through the simulation of priming effects.

6. An evaluation based on the alignment of validated measurements of type and grade of word relatedness and relevance was conducted. The evaluation was conducted based on the utilization of four standard test collections and measurement instruments from cognitive psychology and linguistics. Based on converging observations resulting from the empirical investigation of three hypotheses, it was concluded that the impact of associative and semantic relations on the estimation of relevance differs.

10.4 Future Research Directions

Subsequently potential future research directions of the presented work are outlined.

10.4.1 Measurement Instruments

With regard to measurement instruments focused on word relationship effects several sensible extensions of the presented work can be proposed. On grounds of the identified need to interpret measurements with respect to the associative and semantic type of a relation, a first extension consists of the conduction of assessment procedures under consideration of this aspect. A first step in this direction consisted of the FS353-SimSt dataset utilized in this study. More sophisticated approaches are described in the following subsections.

Measurement instruments of grade of relations

As outlined in Chapter 8 inferences with regard to the validity of computational models are limited when not considering the underlying semantic-associative nature of the word pairs used in an assessment procedure. A proposed future research direction therefore consists of the development of datasets created under consideration of the following aspects:

- Semantic, associative, or semantic-associative type of relation of word pairs.
- Frequency of word pair elements in representative corpora.
- Part-of-Speech type of word pair elements.

In this form the creation of such datasets enables the validation of computational models in terms of their ability to correctly estimate the grade of semantic, associative, or semantic-associative relations between words.

Measurement instruments of semantic-associative degree

With regard to the measurement of the semantic-associative degree of a relationship several future extensions can be explored. A first extension consists of extending the neighborhood assessment method by consideration of the frequency of neighbors. Such an extension would allow for better control of this independent variable with regard to analysis conducted on grounds of such assessments. A second extension consists of the development of better methods for the normalization of computational model output over the respective model parameter spaces. A third extension consists of the consideration of other kinds of measurement instruments. Potential candidate sources for the development of such instruments are presented by data stemming from free association norm studies (Nelson et al. (2004)) and semantic feature production norms (McRae et al. (2005),Vinson and Vigliocco (2008)).

10.4.2 Computational Models

The availability of better measurement instruments as proposed in the prior sections enables several interesting future directions for the development of computational models. As outlined in course of the validation study neither LSA nor HAL constitute purely associative or semantic relation focused models.

On basis of more sophisticated instruments to measure semantic/associative focus, a future direction consists of developing algorithms and statistical methods that are focused on the identification of a specific type of relationship. A re-evaluation of existing computational models (Rohde et al. (2004),Durda et al. (2009), Newman and Welling (2009)) on basis of such measurement instruments represents a viable first step to such developments. Of high interest, specifically with regard to the identification of semantic relations, are structured computational models (Padó and Lapata (2007), Baroni et al. (2010)).

10.4.3 Relevance and Word Relation Dependencies

A primary limitation with regard to researching the relevance–word-relation dependencies is given by the lack of purely semantic or associative focused computational models. A future direction of the presented research therefore consists of the conduction of experimentation following the availability of such models.

Supposing the availability of measurement instruments created under consideration of POS type and term frequencies another extension consists of extending the analysis with regard to the impact of these factors. In principle such an evaluation could be conducted by an analysis of variance.

10.4.4 Consideration of Additional Cognitive Processes

Lastly a future direction with regard to the presented work consists of the application of the paradigm and the associated methodology on additional cognitive processes. A potential candidate is provided by targeting the 'source credibility bias' cognitive process. While the consideration of authority is well established in contemporary IR, the assumption can be drawn that valuable insights could be gained through dedicated modelling of the cognitive process. A first exploration with regard to such a dedicated exploration of the construct of source credibility is given by the study of Liu (2004).

With regard to the focused cognitive processing within this study an interesting extension is also provided by the consideration of formal cognitive models of the word and sentence meaning identification phases (Kintsch and Mangalath (2011)).

List of Figures

List of Tables

BIBLIOGRAPHY

James S. Adelman, Gordon D.A. Brown, and José F. Quesada. Contextual diversity, not word frequency, determines word-naming and lexical decision times. *Psychological Science*, 17(9):814–823, 2006. URL http://pss.sagepub.com/content/17/9/814.short.

Eneko Agirre, Enrique Alfonseca, Keith Hall, Jana Kravalova, Marius Paca, and Aitor Soroa. A study on similarity and relatedness using distributional and WordNet-based approaches. *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics on - NAACL '09*, pages 19–27, 2009. doi: 10.3115/1620754.1620758. URL http://portal.acm.org/citation.cfm?doid=1620754.1620758.

Anne Anastasi. *Psychological Testing*. MacMillan, New York, NY, USA, first edition, 1954. ISBN 0023030852.

John R. Anderson. A spreading activation theory of memory. *Journal of Verbal Learning and Verbal Behavior*, 22(3):261–295, June 1983. ISSN 00225371. doi: 10.1016/S0022-5371(83)90201-3. URL http://linkinghub.elsevier.com/retrieve/pii/S0022537183902013.

John R. Anderson. Spanning seven orders of magnitude: A challenge for cognitive modeling. *Cognitive Science*, 26(1):85–112, February 2002. ISSN 03640213. doi: 10.1016/S0364-0213(01)00062-3. URL http://linkinghub.elsevier.com/retrieve/pii/S0364021301000623.

John R. Anderson. *Cognitive psychology and its implications*. Worth publishers, sixth edition, 2005. ISBN 0716701103, 9780716701101.

John R. Anderson, Daniel Bothell, Michael D. Byrne, Scott Douglass, Christian Lebiere, and Yulin Qin. An integrated theory of the mind. *Psychological review*, 111(4):1036–1060, October 2004. ISSN 0033-295X. doi: 10.1037/0033-295X.111.4.1036. URL http://www.ncbi.nlm.nih.gov/pubmed/15482072.

Ioannis Arapakis, JM Jose, and PD Gray. Affective feedback: an investigation into the role of emotions in the information seeking process. In *31st Annual International*

ACM SIGIR Conference on Research and Development in Information Retrieval, number January, pages 20–24, 2008. URL http://dl.acm.org/citation.cfm?id=1390403.

Mark Baillie, Leif Azzopardi, and Ian Ruthven. Evaluating epistemic uncertainty under incomplete assessments. *Information Processing & Management*, 43(2):811 – 837, 2008. URL http://www.sciencedirect.com/science/article/pii/S0306457307001057.

Peter Barker, Xiang Chen, and Hanne Andersen. Kuhn on Concepts and Categorization. In Thomas Nickles, editor, *Thomas Kuhn*, pages 212–245. Cambridge University Press, 2003.

Marco Baroni and Alessandro Lenci. Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*, 36(4):673–721, 2010. URL http://www.mitpressjournals.org/doi/pdf/10.1162/coli_a_00016.

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43(3):209–226, February 2009. ISSN 1574-020X. doi: 10.1007/s10579-009-9081-4. URL http://www.springerlink.com/index/10.1007/s10579-009-9081-4.

Marco Baroni, Brian Murphy, Eduard Barbu, and Massimo Poesio. Strudel: a corpus-based semantic model based on properties and types. *Cognitive science*, 34(2):222–254, March 2010. ISSN 1551-6709. doi: 10.1111/j.1551-6709.2009.01068.x. URL http://www.ncbi.nlm.nih.gov/pubmed/21564211.

Lawrence W Barsalou. Grounded cognition. *Annual review of psychology*, 59:617–45, January 2008. ISSN 0066-4308. doi: 10.1146/annurev.psych.59.103006.093639. URL http://www.ncbi.nlm.nih.gov/pubmed/17705682.

Jon Barwise and Jerry Seligman. *Information flow: the logic of distributed systems*. Cambridge University Press, 1997.

Antoine Bechara and Antonio R. Damasio. The somatic marker hypothesis: A neural theory of economic decision. *Games and Economic Behavior*, 52(2):336–372, August 2005. ISSN 08998256. doi: 10.1016/j.geb.2004.06.010. URL http://linkinghub.elsevier.com/retrieve/pii/S0899825604001034.

Antoine Bechara, Hanna Damasio, and Antonio R. Damasio. Emotion, decision making and the orbitofrontal cortex. *Cerebral cortex*, 10(3):295, March 2000. ISSN 1047-3211. URL http://www.ncbi.nlm.nih.gov/pubmed/10731224http://cercor.oxfordjournals.org/content/10/3/295.short.

William Bechtel. *Philosophy of mind: An overview for cognitive science*. Lawrence Erlbaum Associates, 1988.

Nicholas J. Belkin. Some(what) grand challenges for information retrieval. *ACM SIGIR Forum*, 42(1):47–54, June 2008. ISSN 01635840. doi: 10.1145/1394251.1394261. URL http://portal.acm.org/citation.cfm?doid=1394251.1394261.

Nicolas J. Belkin, Robert N. Oddy, and Helen M. Brooks. ASK for information retrieval: Part II. Results of a design study. *Journal of Documentation*, 38 (3):145–164, 1993. URL http://www.emeraldinsight.com/journals.htm?articleid=1649968&show=abstract.

Robert P. Benedict. *Fundamentals of temperature, pressure, and flow measurements*. Wiley-Interscience, 1984.

Shlomo Bentin, Gregory McCarthy, and Charles C. Wood. Event-related potentials, lexical decision and semantic priming. *Electroencephalography and clinical Neurophysiology*, 60(4):343–355, 1985. URL http://www.sciencedirect.com/science/article/pii/0013469485900082.

David C. Blair. *Language and representation in information retrieval*. Elsevier North-Holland, Inc., 1990. ISBN 0-444-88437-8. URL http://portal.acm.org/citation.cfm?id=78085.

Paolo Boldi and Sebastiano Vigna. MG4J at TREC 2005. *The Fourteenth Text REtrieval Conference (TREC 2005)*, 500:266, 2005. URL http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.61.1052&rep=rep1&type=pdf.

Denny Borsboom. *Conceptual issues in psychological measurement*. PhD thesis, University of Amsterdam, 2003. URL http://dare.uva.nl/document/195741.

Denny Borsboom. The attack of the psychometricians. *Psychometrika*, 71 (3):425–440, 2006. URL http://link.springer.com/article/10.1007/s11336-006-1447-6.

Denny Borsboom, Gideon J. Mellenbergh, and Jaap van Heerden. The concept of validity. *Psychological review*, 111(4):1061–1071, October 2004. ISSN 0033-295X. doi: 10.1037/0033-295X.111.4.1061. URL http://www.ncbi.nlm.nih.gov/pubmed/15482073.

Andrei Broder. A taxonomy of web search. *SIGIR Forum*, 36:310, 2002. doi: http://doi.acm.org/10.1145/792550.792552.

Chris Buckley, Darrin Dimmick, Ian Soboroff, and Ellen Voorhees. Bias and the limits of pooling for large collections. *Information Retrieval*, 10(6):491–508, August 2007. ISSN 1386-4564. doi: 10.1007/s10791-007-9032-x. URL http://link.springer.com/10.1007/s10791-007-9032-x.

Alexander Budanitsky and Graeme Hirst. Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures. In *Workshop on WordNet and Other Lexical Resources*, volume 2. Citeseer, 2001. URL http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.29.2985&rep=rep1&type=pdf.

Alexander Budanitsky and Graeme Hirst. Evaluating WordNet-based Measures of Lexical Semantic Relatedness. *Computational Linguistics*, 32(1):13–47, March 2006. ISSN 0891-2017. doi: 10.1162/coli.2006.32.1.13. URL http://www.mitpressjournals.org/doi/abs/10.1162/coli.2006.32.1.13.

Raluca Budiu, Christiaan Royer, and Peter Pirolli. Modeling information scent: A comparison of LSA, PMI and GLSA similarity measures on common tests and corpora. In *RIAO'2007 Large Scale Semantic Access to Content (Text, Image, Video, and Sound)*, pages 314–332, Pittsburgh, Pennsylvania, 2007. Le Centre de hautes Etudes internationales d'Informatique Documentaire. URL http://dl.acm.org/citation.cfm?id=1931422.

Vannevar Bush. As we may think. *Atlantic Monthly*, 176:101–108, July 1945.

Claudio Carpineto and Giovanni Romano. A Survey of Automatic Query Expansion in Information Retrieval. *ACM Computing Surveys*, 44(1):1–50, January 2012. ISSN 03600300. doi: 10.1145/2071389.2071390. URL http://dl.acm.org/citation.cfm?doid=2071389.2071390.

Shelly Chaiken, Akiva Liberman, and Alice H. Eagly. Heuristic and systematic information processing within and beyond the persuasion context. In James S. Uleman and John Bargh A., editors, *Unintended thought*, volume 16, pages 212–252. Guilford Press, New York, NY, USA, 1989.

Christine Chiarello, Curt Burgess, Lorie Richards, and Alma Pollock. Semantic and associative priming in the cerebral hemispheres: some words do, some words don't ... sometimes, some places. *Brain and language*, 38(1):75–104, January 1990. ISSN 0093-934X. URL http://www.ncbi.nlm.nih.gov/pubmed/2302547.

Noam Chomsky. A review of BF Skinner's Verbal Behavior. *Language*, 35(1):26–58, 1959. URL http://cogprints.org/1148/.

Cyril Cleverdon. The Cranfield tests on index language devices. *Aslib proceedings*, 19(6):173–194, 1967. URL http://www.emeraldinsight.com/journals.htm?articleid=1692671&show=abstract.

I. Bernard Cohen and George. E. Smith. *The Cambridge Companion to Newton*. Cambridge University Press, 1st edition, 2002.

Jacob Cohen. *Statistical power analysis for the behavioral sciences*. Lawrence Erlbaum Associates, 1988.

Kevin Collins-Thompson. Reducing the risk of query expansion via robust constrained optimization. In *Proceedings of the 18th ACM conference on Information and knowledge management*, pages 837–846, 2009. ISBN 9781605585123. URL http://dl.acm.org/citation.cfm?id=1646059.

Jerry A. Colliver, Melinda J. Conlee, and Steven J. Verhulst. From test validity to construct validity ...and back? *Medical education*, 46(4):366–371, April 2012. ISSN

1365-2923. doi: 10.1111/j.1365-2923.2011.04194.x. URL http://www.ncbi.nlm.nih.gov/pubmed/22429172.

William S. Cooper. On selecting a measure of retrieval effectiveness. *Journal of the American Society for Information Science*, 24(2):87–100, 1973. URL http://onlinelibrary.wiley.com/doi/10.1002/asi.4630240204/full.

Irene Cramer. How well do semantic relatedness measures perform? A meta-study. In *Proceedings of the 2008 Conference on Semantics in Text Processing*, pages 59–70. Association for Computational Linguistics, 2008. URL http://acl.eldoc.ub.rug.nl/mirror/W/W08/W08-2206.pdf.

Nick Craswell and David Hawking. Overview of the TREC 2004 Web Track. In *Proceedings of the Thirteenth Text Retrieval Conference (TREC 2004)*, Gaithersburg, Maryland USA, 2004. URL http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Overview+of+the+TREC-2004+web+track#8.

Lee J. Cronbach. Construct validation after thirty years. In Robert L. Linn, editor, *Intelligence: Measurement, theory, and public policy*, pages 147–171. University of Illinois Press, 1st edition, 1989.

Lee J. Cronbach and Paul E. Meehl. Construct validity in psychological tests. *Psychological Bulletin*, pages 281–302, 1955. doi: 10.1037/h0040957. URL http://psycnet.apa.org/doi/10.1037/h0040957.

Alan D. Cruse. *Lexical semantics*. Cambridge University Press, 1997.

Carlos A. Cuadra and Robert V. Katter. Opening the black box of 'Relevance'. *Journal of Documentation*, 23(4):291–303, 1967. ISSN 0022-0418. doi: 10.1108/eb026436. URL http://www.emeraldinsight.com/10.1108/eb026436.

Carlos A. Cuadra, Robert V. Katter, Emory H. Holmes, and Everett M. Wallace. *Experimental Studies of Relevance Judgments: Final Report. Volume 1: Project Summary*. System Development Corp., Santa Monica, CA., 1st edition, 1967.

Roy G. D'Andrade. *The Development of Cognitive Anthropology*. Cambridge University Press, 1995. URL http://www.jstor.org/stable/30131990.

Adrie H. G. De Vries-Griever and Theodore F. Meijman. The impact of abnormal hours of work on various modes of information processing: a process model on human costs of performance. *Ergonomics*, 30(9):1287–1299, September 1987. ISSN 0014-0139. doi: 10.1080/00140138708966023. URL http://www.tandfonline.com/doi/abs/10.1080/00140138708966023.

Scott Deerwester, Susan T. Dumais, George W. Furnas, and Thomas K. Landauer. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407, 1990. URL http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.62.1152&rep=rep1&type=pdf.

Daniel C. Dennett. *Elbow room: The varieties of free will worth wanting*. MIT Press, Cambridge, MA, 1984.

Keith J. Devlin. *Logic and information*. Cambridge University Press, Cambridge, MA, 1995.

Fred Dretske. *Knowledge and the Flow of Information*. MIT Press, Cambridge, MA, 1983. URL http://mitpress.mit.edu/catalog/item/default.asp?tid=7275&ttype=2.

Susan T. Dumais. Improving the retrieval of information from external sources. *Behavior Research Methods*, 23(2):229–236, 1991. URL http://www.springerlink.com/index/M628723M7151N676.pdf.

Kevin Durda and Lori Buchanan. WINDSORS: Windsor improved norms of distance and similarity of representations of semantics. *Behavior Research Methods*, 40(3): 705–712, August 2008. ISSN 1554-351X. doi: 10.3758/BRM.40.3.705. URL http://brm.psychonomic-journals.org/cgi/doi/10.3758/BRM.40.3.705.

Kevin Durda, Lori Buchanan, and Richard Caron. Grounding co-occurrence: Identifying features in a lexical co-occurrence model of semantic memory. *Behavior research methods*, 41(4):1210–23, November 2009. ISSN 1554-3528. doi: 10.3758/BRM.41.4.1210. URL http://www.ncbi.nlm.nih.gov/pubmed/19897830.

Robert A. Fairthorne. Implications of test procedures. In *Information retrieval in action: proceedings of the conference held on April 16-18, 1962 Center for Documentation and Communication Research*, page 109, Cleveland, Ohio, 1963. Press of Western Reserve University.

Hui Fang. A re-examination of query expansion using lexical resources. In *Proceedings of ACL-08: HLT*, pages 139–147. Citeseer, 2008. URL http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.164.8998&rep=rep1&type=pdf.

Hui Fang and ChengXiang Zhai. An exploration of axiomatic approaches to information retrieval. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '05*, pages 480–487, New York, New York, USA, 2005. ACM Press. ISBN 1595930345. doi: 10.1145/1076034.1076116. URL http://portal.acm.org/citation.cfm?doid=1076034.1076116.

Hui Fang and ChengXiang Zhai. Semantic term matching in axiomatic approaches to information retrieval. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '06*, pages 115–122, New York, New York, USA, 2006. ACM Press. ISBN 1595933697. doi: 10.1145/1148170.1148193. URL http://portal.acm.org/citation.cfm?doid=1148170.1148193.

Hui Fang, Tao Tao, and ChengXiang Zhai. A formal study of information retrieval heuristics. *Proceedings of the 27th annual international conference on Research and development in information retrieval - SIGIR '04*, page 49, 2004. doi: 10.1145/1008992.1009004. URL http://portal.acm.org/citation.cfm?doid=1008992.1009004.

Michael R. Fehling. Unified Theories of Cognition: modeling cognitive competence. *Artificial Intelligence*, 59(1-2):295–328, February 1993. ISSN 00043702. doi: 10.1016/0004-3702(93)90197-J. URL http://linkinghub.elsevier.com/retrieve/pii/000437029390197J.

Christine Fellbaum. *WordNet: An electronic lexical database*. The MIT press, Cambridge, MA, 1998.

Ludovic Ferrand and Boris New. Semantic and associative priming in the mental lexicon. In Patrick Bonin, editor, *Mental lexicon: Some words to talk about words*, chapter 2, pages 25–45. Nova Publishers, 2004. URL http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.3.705&rep=rep1&type=pdf.

Lev Finkelstein, Evgeniy Gabrilovich, and Yossi Matias. Placing search in context: The Concept Revisited. *ACM Transactions on Information Systems*, 20(1):116–131, January 2002. ISSN 10468188. doi: 10.1145/503104.503110. URL http://portal.acm.org/citation.cfm?doid=503104.503110http://portal.acm.org/citation.cfm?id=372094.

Luciano Floridi. Open problems in the philosophy of information. *Metaphilosophy*, 35 (4):554–582, 2004. URL http://onlinelibrary.wiley.com/doi/10.1111/j.1467-9973.2004.00336.x/abstract.

George W. Furnas, Thomas K. Landauer, Louis M. Gomez, and Susan T. Dumais. The vocabulary problem in human-system communication. *Communications of the ACM*, 30(11):971, 1987. URL http://portal.acm.org/citation.cfm?id=32206.32212.

Evgeniy Gabrilovich and Shaul Markovitch. Feature Generation for Text Categorization Using World Knowledge. In *International joint conference on artificial intelligence. Vol. 19. , 2005*, volume 19, page 1048. Lawrence Erlbaum Associates, 2005.

Evgeniy Gabrilovich and Shaul Markovitch. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *Proceedings of the 20th international joint conference on Artifical intelligence*, pages 1606–1611, 2007. URL http://www.aaai.org/Papers/IJCAI/2007/IJCAI07-259.pdf.

Peter Gardenfors. Cognitive science: From computers to anthills as models of human thought. *Human IT*, 3(2):2–99, 1999. URL http://www.ida.liu.se/~729G01/mtrl/computers_to_anthill.pdf.

Peter Gardenfors. *Conceptual spaces: The geometry of thought*. The MIT Press, Cambridge, MA, 2004.

Wendell R Garner, Harold W Hake, and Charles W Eriksen. Operationism and the Concept Of Perception. *Psychological Review*, 63(3):149–159, 1956.

Merrill Garrett. Thinking across the boundaries: Psycholinguistic perspectives. In Gareth M. Gaskell, editor, *The Oxford Handbook of Psycholinguistics*, chapter 49, pages 815–820. Oxford University Press, 1 edition, 2007.

Gareth M. Gaskell and Gerry Altmann, editors. *The Oxford Handbook of Psycholinguistics*. Oxford University Press, 2007. ISBN 0198568975. URL http://books.google.com/books?id=QZxc9WRODyoC&pgis=1.

Wilhelm R. Glaser and Franz-Josef Düngelhoff. The time course of picture-word interference. *Journal of Experimental Psychology: Human Perception and Performance*, 10(5):640, 1984. URL http://psycnet.apa.org/journals/xhp/10/5/640/.

Gene H. Golub and Charles F. Van Loan. *Matrix computations*. Johns Hopkins University Press, 1996.

Arthur C. Graesser, Keith K. Millis, and Rolf A. Zwaan. Discourse comprehension. *Annual review of ...*, 48(1):163–189, 1997. URL http://www.annualreviews.org/doi/abs/10.1146/annurev.psych.48.1.163.

Arthur C. Graesser, Jose A. León, and Jose Otero. Introduction to the psychology of science text comprehension. In *The psychology of science text comprehension*, chapter 1, pages 1–15. Lawrence Erlbaum Associates, first edition, 2002.

Jane Greenberg. Automatic query expansion via lexicalsemantic relationships. *Journal of the American Society for Information and Technology*, 52 (5):402–415, 2001. URL http://onlinelibrary.wiley.com/doi/10.1002/1532-2890(2001)9999:9999%3C::AID-ASI1089%3E3.0.CO;2-K/full.

David Z. Hambrick and Randall W. Engle. Effects of domain knowledge, working memory capacity, and age on cognitive performance: an investigation of the knowledge-is-power hypothesis. *Cognitive psychology*, 44(4):339–87, June 2002. ISSN 0010-0285. doi: 10.1006/cogp.2001.0769. URL http://www.ncbi.nlm.nih.gov/pubmed/12018938.

Trevor A. Harley, Lesley J. Jessiman, and Siobhan B. G. Macandrew. Decline and fall: A biological, developmental, and psycholinguistic account of deliberative language processes and ageing. *Aphasiology*, 25(2):123–153, January 2011. ISSN 0268-7038. doi: 10.1080/02687031003798262. URL http://www.tandfonline.com/doi/abs/10.1080/02687031003798262.

Steven P. Harter. A probabilistic approach to automatic keyword indexing. Part II. An algorithm for probabilistic indexing. *Journal of the American Society for Information Science*, 26(5):280–289, 1975. URL http://onlinelibrary.wiley.com/doi/10.1002/asi.4630260504/abstract.

William Hersh, Ravi T. Bhuptiraju, Laura Ross, Phoebe Johnson, Aaron M. Cohen, and Dale F. Kraemer. TREC 2004 genomics track overview. In *The Thirteenth Text Retrieval Conference (TREC 2004)*, volume 2004, Gaithersburg, MD, 2004. National Institute for Standards and Technology.

Charles R. Hildreth. Accounting for users' inflated assessments of on-line catalogue search performance and usefulness: an experimental study. *Information Research*, 6 (2), 2002. URL http://informationr.net/ir/6-2/paper101.html.

John J. Hopfield. Neurons with graded response have collective computational properties like those of two-state neurons. *Proceedings of the national academy of sciences*, 81(10):3088–3092, 1984. URL http://www.pnas.org/content/81/10/3088.short.

David Hull. Using statistical testing in the evaluation of retrieval experiments. *Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 329–338, 1993. URL http://dl.acm.org/citation.cfm?id=160758.

Hutchins, Edwin and Gavan Lintern. *Cognition in the Wild*, volume 262082314. MIT press Cambridge, MA, 1995. URL http://www.ida.liu.se/~nilda/CST-papers/Hutchins.pdf.

Adele Hutchinson, Douglas R. Whitman, Chris Abeare, and Jennifer Raiter. The unification of mind: Integration of hemispheric semantic processing. *Brain and language*, 87(3):361–368, 2003. URL http://www.sciencedirect.com/science/article/pii/s0093934x03001330.

Keith A. Hutchison. Is semantic priming due to association strength or feature overlap? A microanalytic review. *Psychonomic bulletin & review*, 10(4):785–813, December 2003. ISSN 1069-9384. URL http://www.ncbi.nlm.nih.gov/pubmed/15000531.

Peter Ingwersen. Polyrepresentation of information needs and semantic entities: elements of a cognitive theory for information retrieval interaction. In *SIGIR '94 Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, pages Pages 101 – 110, 1994.

Peter Ingwersen and Kalervo Järvelin. The Turn: Integration of Information Seeking and Retrieval in Context (The Information Retrieval Series). In W. Bruce Croft, editor, *The Turn: Integration of Information Seeking and Retrieval in Context (The Information Retrieval Series)*, chapter 3, pages 87–102. Springer-Verlag New York, Inc., Secaucus, NJ, USA, first edition, 2005. ISBN 140203850X.

Bernard J. Jansen. Search log analysis: What it is, what's been done, how to do it. *Library & Information Science Research*, 28(3):407–432, September 2006. ISSN 07408188. doi: 10.1016/j.lisr.2006.06.005. URL http://linkinghub.elsevier.com/retrieve/pii/S0740818806000673.

Nick Jardine and Cornelis J. van Rijsbergen. The use of hierarchic clustering in information retrieval. *Information storage and retrieval*, 7(5):217–240, December 1971. ISSN 00200271. doi: 10.1016/0020-0271(71)90051-9. URL http://linkinghub.elsevier.com/retrieve/pii/0020027171900519.

Mario Jarmasz and Stan Szpakowicz. Roget's thesaurus and semantic similarity. *In: Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP-2003)*, pages 212–219, 2003. URL http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.146.5568.

Todd D. Jick. Mixing qualitative and quantitative methods: Triangulation in action. *Administrative science quarterly*, 24(4):602–611, 1979. URL http://www.jstor.org/stable/2392366.

Michael N. Jones, Walter Kintsch, and Douglas J. K. Mewhort. High-dimensional semantic space accounts of priming. *Journal of Memory and Language*, 55(4): 534–552, 2006. ISSN 0749596X. doi: 10.1016/j.jml.2006.07.003. URL http://linkinghub.elsevier.com/retrieve/pii/S0749596X06000829.

Marcel A. Just, Vladimir L. Cherkassky, Sandesh Aryal, and Tom M. Mitchell. A neurosemantic theory of concrete noun representation based on the underlying brain codes. *PloS one*, 5(1):e8622, January 2010. ISSN 1932-6203. doi: 10.1371/journal.pone.0008622. URL http://www.ncbi.nlm.nih.gov/pubmed/20084104.

Michael Kane. In Praise of Pluralism. A Comment on Borsboom. *Psychometrika*, 71 (3):441–445, September 2006. ISSN 0033-3123. doi: 10.1007/s11336-006-1491-2. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2779426&tool=pmcentrez&rendertype=abstract.

Truman Lee Kelley. *Interpretation of Educational Measurements*. Measurement and adjustment series. MacMillan, New York, NY, 1927.

Maurice G. Kendall. A new measure of rank correlation. *Biometrika*, 30(1):81–93, 1938. URL http://www.jstor.org/stable/10.2307/2332226.

Walter Kintsch. The role of knowledge in discourse comprehension: a construction-integration model. *Psychological review*, 95(2):163–182, April 1988. ISSN 0033-295X. URL http://www.ncbi.nlm.nih.gov/pubmed/3375398.

Walter Kintsch and Praful Mangalath. The Construction of Meaning. *Topics in Cognitive Science*, 3(2):346–370, April 2011. ISSN 17568757. doi: 10.1111/j.1756-8765.2010.01107.x. URL http://doi.wiley.com/10.1111/j.1756-8765.2010.01107.x.

Walter Kintsch and Cathleen Wharton. An overview of the construction-integration model: a theory of comprehension as a foundation for a new cognitive architecture. *ACM SIGART Bulletin*, 2(4):169–173, 1991. URL http://portal.acm.org/citation.cfm?id=122379.

Philip Kitcher. Explanatory unification. *Philosophy of Science*, 329(5990): 399–400, July 1981. ISSN 1095-9203. doi: 10.1126/science.1189416. URL http://www.ncbi.nlm.nih.gov/pubmed/21670322http://www.jstor.org/stable/186834.

Jon M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)*, 46(5):604–632, September 1999. ISSN 00045411. doi: 10.1145/324133.324140. URL http://portal.acm.org/citation.cfm?doid=324133.324140http://portal.acm.org/citation.cfm?id=324140.

Lubomir Krcmár, Miloslav Konopik, and Karel Jezek. Exploration of Semantic Spaces Obtained from Czech Corpora. *Proceedings of the Dateso*, pages 97–107, 2011. URL http://textmining.zcu.cz/publications/Dateso_final.pdf.

Thomas S. Kuhn. *The Structure of Scientific Revolutions*. University of Chicago Press, Chicago, first edition, 1962.

Thomas S. Kuhn. *The Structure of Scientic Revolutions*. SO Source: University of Chicago Press: Chicago. (1962). Princeton University Press, Princeton, second edition, 1970. ISBN 0226458032.

Thomas S. Kuhn. The Road since Structure. In A Fine, M Forbes, and L Wessels, editors, *PSA Proceedings of the Biennial Meeting of the Philosophy of Science Association*, volume 1990, pages 3–13. The University of Chicago Press on behalf of the Philosophy of Science Association, 1990. URL http://www.jstor.org/stable/193054.

Roy Lachman, Janet L. Lachman, and Earl Butterfield. *Cognitive psychology and information processing: An introduction*. Lawrence Erlbaum Associates, Hillsdale, New Jersey, 1979.

Thomas K. Landauer and Susan T. Dumais. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, 104(2):211, 1997. URL http://psycnet.apa.org/journals/rev/104/2/211/.

Shuang Liu and Clement Yu. UIC at TREC2005 : Robust Track. In Ellen M. Voorhees and Lori P. Buckland, editors, *Proceedings of the Fourteenth Text REtrieval Conference (TREC 2005)*, Gaithersburg, Maryland, USA, 2005. National Institute of Standards and Technology (NIST). URL http://dblp.uni-trier.de/db/conf/trec/trec2005.html#LiuY05.

Ziming Liu. Perceptions of credibility of scholarly information on the web. *Information Processing & Management*, 40(6):1027–1038, November 2004. ISSN 03064573. doi: 10.1016/S0306-4573(03)00064-5. URL http://linkinghub.elsevier.com/retrieve/pii/S0306457303000645.

Zhong-Lin Lu and Barbara A. Dosher. Cognitive psychology. *Scholarpedia*, 2(8):2769, 2007.

Hans P. Luhn. A Statistical Approach to Mechanized Encoding and Searching of Literary Information. *IBM Journal of Research and Development*, 1(4):309–317, October 1957. ISSN 0018-8646. doi: 10.1147/rd.14.0309. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5392697.

Hans P. Luhn. The Automatic Creation of Literature Abstracts. *IBM Journal of Research and Development*, 2(2):159–165, April 1958. ISSN 0018-8646. doi: 10.1147/rd.22.0159. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5392672.

Kevin Lund and Curt Burgess. Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods*, 28(2):203–208, 1996. URL http://www.springerlink.com/index/W06U365573X83884.pdf.

Yuanhua Lv and ChengXiang Zhai. Adaptive term frequency normalization for bm25. *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 1985–1988, 2011. doi: 10.1145/2063576.2063871. URL http://dl.acm.org/citation.cfm?doid=2063576.2063871http://dl.acm.org/citation.cfm?id=2063871.

Norman Malcolm. *Problems of mind: Descartes to Wittgenstein*. Allen & Unwin, 1971. URL http://www.getcited.org/pub/101450638.

Dominic W. Massaro and Nelson Cowan. Information processing models: microscopes of the mind. *Annual review of psychology*, 44:383–425, January 1993. ISSN 0066-4308. doi: 10.1146/annurev.ps.44.020193.002123. URL http://www.ncbi.nlm.nih.gov/pubmed/8434893.

Margaret Masterman. The Nature of a Paradigm. In Imre Lakatos and Alan Musgrave, editors, *Criticism and the growth of knowledge*, Criticism and the growth of knowledge, pages 59–89. Cambridge University Press, 1970. ISBN 0521078261. URL http://books.google.com/books?id=Vutfm5n6LKYC.

Danielle S. McNamara and Walter Kintsch. Learning from texts: Effects of prior knowledge and text coherence. *Discourse processes*, 22:247–288, 1996.

Timothy P. McNamara. *Semantic priming: Perspectives from memory and word recognition*. Psychology Press, 2005a.

Timothy P. McNamara. Cognitive Neuroscience of Semantic Priming. In *Semantic priming: perspectives from memory and word recognition*, chapter 18, pages 116–125. Psychology Press, 2005b.

Ken McRae and Stephen Boisvert. Automatic semantic similarity priming. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 24(3):558, 1998. URL http://psycnet.apa.org/journals/xlm/24/3/558/.

Ken McRae, George S. Cree, Mark S. Seidenberg, and Chris McNorgan. Semantic feature production norms for a large set of living and nonliving things. *Behavior*

research methods, 37(4):547–59, November 2005. ISSN 1554-351X. URL http://www.ncbi.nlm.nih.gov/pubmed/16629288.

David E. Meyer and Roger W. Schvaneveldt. Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. *Journal of experimental psychology*, 90(2):227, 1971. URL http://psycnet.apa.org/journals/xge/90/2/227/.

David E. Meyer, Roger W. Schvaneveldt, and Margaret G. Ruddy. Activation of lexical memory. In *Meeting of the Psychonomic Society*, St Louis, Missouri, 1972. URL http://www.umich.edu/~bcalab/documents/MeyerSchvaneveldtRuddy1972.pdf.

George A. Miller. WordNet: a lexical database for English. *Communications of the ACM*, 38(11):39–41, November 1995. ISSN 00010782. doi: 10.1145/219717.219748. URL http://portal.acm.org/citation.cfm?doid=219717.219748.

George A. Miller. The cognitive revolution: a historical perspective. *Trends in Cognitive Sciences*, 7(3):141–144, March 2003. ISSN 13646613. doi: 10.1016/S1364-6613(03)00029-9. URL http://linkinghub.elsevier.com/retrieve/pii/S1364661303000299.

George A. Miller and Walter G. Charles. Contextual correlates of semantic similarity. *Language and cognitive processes*, 6(1):1–28, 1991. URL http://www.tandfonline.com/doi/abs/10.1080/01690969108406936.

Mandar Mitra, Amit Singhal, and Chris Buckley. Improving automatic query expansion. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 206–214, 1998. ISBN 1581130155. URL http://dl.acm.org/citation.cfm?id=290995.

Stefano Mizzaro. Relevance: The whole history. *Journal of the American Society for Information Science*, 48(9):810–832, September 1997. ISSN 0002-8231. doi: 10.1002/(SICI)1097-4571(199709)48:9⟨810::AID-ASI6⟩3.0.CO;2-U. URL http://doi.wiley.com/10.1002/(SICI)1097-4571(199709)48:9<810::AID-ASI6>3.0.CO;2-U.

Stefano Mizzaro. How many relevances in information retrieval? *Interacting with computers*, 10:305–322, 1998. URL http://www.sciencedirect.com/science/article/pii/S0953543898000125.

Edward F. Moore. Gedanken-experiments on sequential machines. *Automata studies*, 34:129–153, 1956.

Lynn Nadel. *Encyclopedia of cognitive science*. John Wiley, fourth edition, 2005. URL http://www.stanford.edu/class/linguist156/Boroditsky_2003.pdf.

James H. Neely. Semantic priming effects in visual word recognition: A selective review of current findings and theories. *Basic processes in reading: Visual word recognition*, pages 264–336, 1991.

Ulric Neisser. *Cognitive Psychology*. Appleton-Century-Crofts, New York, NY, USA, 1967.

Douglas L. Nelson, Cathy L. McEvoy, and Thomas A. Schreiber. The University of South Florida free association, rhyme, and word fragment norms. *Behavior Research Methods*, 36(3):402–407, 2004. URL http://www.springerlink.com/index/U81L55373371RN04.pdf.

Nancy J. Nersessian. Kuhn, conceptual change, and cognitive science. In Thomas Nickles, editor, *Thomas Kuhn*, pages 178–211. Cambridge University Press, 2003. URL http://books.google.co.uk/books/about/Thomas_Kuhn.html?id=QJ5z5MfrCn0C.

Allan Newell. *Unified theories of cognition*. Harvard University Press, Cambridge, MA, 1994.

David Newman and Max Welling. Distributed Algorithms for Topic Models. *Journal of Machine Learning Research*, 10:1801–1828, 2009.

Daniel Nolan. Quantitative parsimony. *The British journal for the philosophy of science*, 48(3):329, 1997. URL http://bjps.oxfordjournals.org/content/48/3/329.short.

Anthony J. Onwuegbuzie, Ann E. Witcher, Kathleen M. T. Collins, Janet D. Filer, Cheryl D. Wiedmaier, and Chris W. Moore. Students' Perceptions of Characteristics of Effective College Teachers: A Validity Study of a Teaching Evaluation Form Using a Mixed-Methods Analysis. *American Educational Research Journal*, 44(1): 113–160, March 2007. ISSN 0002-8312. doi: 10.3102/0002831206298169. URL http://aer.sagepub.com/cgi/doi/10.3102/0002831206298169.

Sebastian Padó and Mirella Lapata. Dependency-Based Construction of Semantic Space Models. *Computational Linguistics*, 33(2):161–199, June 2007. ISSN 0891-2017. doi: 10.1162/coli.2007.33.2.161. URL http://www.mitpressjournals.org/doi/abs/10.1162/coli.2007.33.2.161.

Larry Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The PageRank citation ranking: Bringing order to the web. In *World Wide Web Internet And Web Information Systems*, pages 1–17. Stanford InfoLab, 1999. URL http://ilpubs.stanford.edu:8090/422.

Stephen E. Palmer and Ruth Kimchi. The information processing approach to cognition. In T. J. Knapp and L. C. Robertson, editors, *Approaches to cognition: Contrasts and controversies*, pages 37–77. Lawrence Erlbaum Associates, Hillsdale, New Jersey, 1984. URL ftp://sunsite.berkeley.edu/pub/techreps/COGSCI-84-28.html.

Charles A. Perfetti, Nicole Landi, and Jane Oakhill. The acquisition of reading comprehension skill. In Margaret J. Snowling and Charles Hulme, editors, *The Science of Reading: A Handbook*, chapter 13, pages 227–247. Blackwell Publishing Ltd,

2005. ISBN 9780470757642. doi: 10.1002/9780470757642.ch13. URL http://onlinelibrary.wiley.com/doi/10.1002/9780470757642.ch13/summary.

Chanthika Pornpitakpan. The persuasiveness of source credibility: A critical review of five decades evidence. *Journal of Applied Social Psychology*, 34(2):243–281, 2004. URL http://onlinelibrary.wiley.com/doi/10.1111/j.1559-1816.2004.tb02547.x/abstract.

Martin F. Porter. An algorithm for suffix stripping. *Program: electronic library and information systems*, 14(3):130–137, 1980. URL http://www.emeraldinsight.com/journals.htm?articleid=1670983&show=abstract.

Terry Pratchett. *Guards! Guards!* Corgi, Ealing, 1989. ISBN 0-552-13462-7.

R. A language and environment for statistical computing (Version 2.12.0) [Computer Software] Retrieved June 15, 2012, 2012. URL http://www.r-project.org/.

Gabriel Recchia and Michael N. Jones. More data trumps smarter algorithms: comparing pointwise mutual information with latent semantic analysis. *Behavior research methods*, 41(3):647–56, August 2009. ISSN 1554-351X. doi: 10.3758/BRM.41.3.647. URL http://www.ncbi.nlm.nih.gov/pubmed/19587174.

Philip Resnik. Using information content to evaluate semantic similarity in a taxonomy. *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, 1:448–453, 1995. URL http://arxiv.org/abs/cmp-lg/9511007.

Cornelis Joost Van Rijsbergen. *Information Retrieval*. Butterworth-Heinemann, Newton, MA, USA, 1979. ISBN 0408709294.

Anna Ritchie, Simone Teufel, and Stephen Robertson. Creating a test collection for citation-based IR experiments. In *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pages 391–398, Morristown, NJ, USA, 2006. Association for Computational Linguistics. doi: http://dx.doi.org/10.3115/1220835.1220885.

Stephen E. Robertson. The probability ranking principle in IR. *Journal of documentation*, 33(4):294–304, 1977. URL http://www.emeraldinsight.com/Insight/ViewContentServlet?contentType=Article&Filename=/published/emeraldfulltextarticle/pdf/2780330404.pdf.

Douglas L.T. Rohde, Laura M. Gonnerman, and David C. Plaut. An Improved Method for Deriving Word Meaning from Lexical. *Cognitive Psychology*, 7:573–605, 2004.

Daniel E. Rose and Danny Levinson. Understanding user goals in web search. In *WWW'04: Proceedings of the 13th international conference on World Wide Web*, pages 13–19, New York,NY,USA, 2004. ACM Press. doi: http://doi.acm.org/10.1145/988672.988675.

Arturo Rosenblueth and Norbert Wiener. The role of models in science. *Philosophy of Science*, 12(4):316–321, 1945. URL http://www.jstor.org/stable/184253.

RStudio. Integrated development environment for R (Version 0.96.122) [Computer Software] Retrieved June 15, 2012, 2012. URL http://www.rstudio.org/.

Herbert Rubenstein and John B. Goodenough. Contextual correlates of synonymy. *Communications of the ACM*, 8(10):627–633, October 1965. ISSN 00010782. doi: 10.1145/365628.365657. URL http://portal.acm.org/citation.cfm?doid=365628.365657.

Wheeler Ruml, Alfonso Caramazza, Jennifer R. Shelton, and Doriana Chialant. Testing Assumptions in Computational Theories of Aphasia. *Journal of Memory and Language*, 43(2):217–248, 2000. URL http://www.sciencedirect.com/science/article/pii/S0749596X0092730X.

Edward J. Russo, Victoria H. Medvec, and Margaret G. Meloy. The distortion of information during decisions. *Organizational Behavior and Human Decision Processes*, 66(1):102–110, 1996. URL http://forum.johnson.cornell.edu/faculty/russo/TheDistortionofInformation.pdf.

David M. Sanbonmatsu and Frank R. Kardes. The effects of physiological arousal on information processing and persuasion. *Journal of Consumer Research*, 28(7):379–385, July 1988. ISSN 1050-9135. URL http://www.ncbi.nlm.nih.gov/pubmed/21918149http://www.jstor.org/stable/2489472.

Mark Sanderson, Monica L. Paramita, Paul Clough, and Evangelos Kanoulas. Do user preferences and evaluation measures line up? In *Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 555–562. ACM, 2010. URL http://portal.acm.org/citation.cfm?id=1835542.

Tefko Saracevic. The concept of relevance in information science: A historical review. *Introduction to information science*, pages 111–151, 1970.

Tefko Saracevic. Relevance: A review of and a framework for the thinking on the notion in information science. *Journal of the American Society for Information Science*, 26 (6):321–343, 1975. ISSN 00028231. doi: 10.1002/asi.4630260604. URL http://doi.wiley.com/10.1002/asi.4630260604.

Tefko Saracevic. Users lost: Reflections on the past, future, and limits of information science. *ACM Sigir Forum*, 31(2):16–27, 1997. doi: http://dx.doi.org/10.1145/270886.270889. URL http://portal.acm.org/citation.cfm?id=270889.

Tefko Saracevic. Relevance: A review of the literature and a framework for thinking on the notion in information science. Part II: nature and manifestations of relevance. *Journal of the American Society for Information Science and Technology*, 58(13): 1915–1933, 2007. doi: 10.1002/asi. URL http://onlinelibrary.wiley.com/doi/10.1002/asi.20682/full.

Linda Schamber. A re-examination of relevance: toward a dynamic, situational definition. *Information Processing & Management*, 26(6):755–776, 1990. ISSN 03064573. doi: 10.1016/0306-4573(90)90050-C. URL http://linkinghub.elsevier.com/retrieve/pii/030645739090050C.

Herbert Schriefers, Antje S. Meyer, and Willem J. M. Levelt. Exploring the time course of lexical access in language production: Picture-word interference studies. *Journal of Memory and Language*, 29(1):86–102, 1990. URL http://www.sciencedirect.com/science/article/pii/0749596X9090011N.

Robert C. Sinclair. Mood, categorization breadth, and performance appraisal: The effects of order of information acquisition and affective state on halo, accuracy, information retrieval, and evaluations. *Organizational Behavior and Human Decision Processes*, 42(1):22–46, 1988. URL http://linkinghub.elsevier.com/retrieve/pii/0749597888900180.

Burrhus Frederic Skinner. *Verbal Behavior*. TheCentury psychology series. Appleton-Century-Crofts, 1957. ISBN 0139415912. doi: 10.1037/11256-000.

Mark D. Smucker, James Allan, and Ben Carterette. A Comparison of Statistical Significance Tests for Information Retrieval Evaluation. In *Proceedings of the sixteenth ACM conference on information and knowledge management - CIKM '07*, pages 623–632, New York, New York, USA, 2007. ACM Press. ISBN 9781595938039. doi: 10.1145/1321440.1321528. URL http://portal.acm.org/citation.cfm?doid=1321440.1321528.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng. Cheap and Fast - But is it good?: Evaluating Non-Expert Annotations for Natural Language Tasks. In *Proceedings of the conference on empirical methods in natural language processing*, pages 254–263. Association for Computational Linguistics., 2008. URL http://dl.acm.org/citation.cfm?id=1613751.

Ian Soboroff. Do TREC web collections look like the web? *SIGIR Forum*, 36(2):23–31, 2002. ISSN 01635840. doi: 10.1145/792550.792554.

Ian Soboroff. A comparison of pooled and sampled relevance judgments in the TREC 2006 Terabyte track. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 785–786, 2007. URL http://research.nii.ac.jp/ntcir/ntcir-ws6/OnlineProceedings/EVIA_Preprint_Papers/11.pdf.

Karen Spärck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 28(1):11–21, 1972. URL http://www.emeraldinsight.com/journals.htm?articleid=1649768&show=abstract.

Karen Spärck Jones. Meta-reflections on TREC. In E.M. Voorhees and D.K. Harman, editors, *TREC: experiment and evaluation in information retrieval*, pages 421–448. MIT Press, Cambridge, MA, 2005.

Mark Steyvers. Modeling semantic and orthographic similarity effects on memory for individual words. *Unpublished doctoral dissertation, Indiana University,*, 2000. URL http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.21.8885&rep=rep1&type=pdf.

Ron Sun. Theoretical status of computational cognitive modeling. *Cognitive Systems Research*, 10(2):124–140, June 2009. ISSN 13890417. doi: 10.1016/j.cogsys.2008.07.002. URL http://linkinghub.elsevier.com/retrieve/pii/S1389041708000429.

Ron Sun, Andrew L. Coward, and Michael J. Zenzen. On levels of cognitive modeling. *Philosophical Psychology*, 18(5):613–637, October 2005. ISSN 0951-5089. doi: 10.1080/09515080500264248. URL http://www.informaworld.com/openurl?genre=article&doi=10.1080/09515080500264248&magic=crossref||D404A21C5BB053405B1A640AFFD44AE3.

William B. Swann and Stephen J. Read. Self-verification processes: How we sustain our self-conceptions. *Journal of Experimental Social Psychology*, 17(4):351–372, 1981. URL http://www.sciencedirect.com/science/article/pii/0022103181900433.

Don R. Swanson. Subjective versus objective relevance in bibliographic retrieval systems. *The Library Quarterly*, 56(4):389–398, 1986. URL http://www.jstor.org/stable/10.2307/4308045.

David Swinney. Lexical access during sentence comprehension: (Re)consideration of context effects. *Journal of Verbal Learning and Verbal Behavior*, 18(6):645–659, December 1979. ISSN 00225371. doi: 10.1016/S0022-5371(79)90355-4. URL http://linkinghub.elsevier.com/retrieve/pii/S0022537179903554.

Paul Thagard. *Mind: Introduction to cognitive science*. The MIT press, 2005.

Paul Thagard. Being Interdisciplinary: Trading Zones in Cognitive Science. In S. J. Derry, C. D. Schunn, and M. A. Gernsbacher, editors, *Interdisciplinary collaboration: An emerging cognitive science*, pages 1–31. Lawrence Erlbaum Associates, 2006.

Paul Thagard. Why Cognitive Science Needs Philosophy and Vice Versa. *Topics in Cognitive Science*, 1(2):237–254, April 2009. ISSN 17568757. doi: 10.1111/j.1756-8765.2009.01016.x. URL http://doi.wiley.com/10.1111/j.1756-8765.2009.01016.x.

Paul Thagard. Cognitive Science. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Stanford University, summer 201 edition, 2010. URL http://plato.stanford.edu/archives/sum2010/entries/cognitive-science.

Paul Thagard. Cognitive Science. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Stanford University, fall 2012 edition, 2012. URL http://plato.stanford.edu/archives/fall2012/entries/cognitive-science.

Zakary L. Tormala, Pablo Brinol, and Richard E. Petty. When credibility attacks: The reverse impact of source credibility on persuasion. *Journal of Experimental Social Psychology*, 42(5):684–691, September 2006. ISSN 00221031. doi: 10.1016/j.jesp.2005.10.005. URL http://linkinghub.elsevier.com/retrieve/pii/S0022103105001265.

Andrew Turpin and Falk Scholer. User performance versus precision measures for simple search tasks. In *Proceedings of the 29th annual international ACM SIGIR conference on research and development in information retrieval*, pages 11–18. ACM, 2006. ISBN 1595933697. URL http://portal.acm.org/citation.cfm?id=1148176.

Andrew H. Turpin and William Hersh. Why batch and user evaluations do not give the same results. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 225–231. ACM, 2001. ISBN 1581133316. URL http://portal.acm.org/citation.cfm?id=383992.

Teun van Dijk and Walter Kintsch. *Strategies of discourse comprehension*. Academic Press Inc, New York, NY, USA, first edition, 1983. URL http://www.citeulike.org/group/1868/article/973620.

Gabriella Vigliocco and David P. Vinson. Semantic Representation. In Gareth M. Gaskell, editor, *The Oxford Handbook of Psycholinguistics*, chapter 12, pages 195–216. Oxford University Press, Oxford, 1 edition, 2007. ISBN 978-0198568971.

Gabriella Vigliocco, David P. Vinson, William Lewis, and Merrill F. Garrett. Representing the meanings of object and action words: the featural and unitary semantic space hypothesis. *Cognitive psychology*, 48(4):422–88, June 2004. ISSN 0010-0285. doi: 10.1016/j.cogpsych.2003.09.001. URL http://www.ncbi.nlm.nih.gov/pubmed/15099798.

David P. Vinson and Gabriella Vigliocco. Semantic feature production norms for a large set of objects and events. *Behavior Research Methods*, 40(1):183–190, February 2008. ISSN 1554-351X. doi: 10.3758/BRM.40.1.183. URL http://www.springerlink.com/index/10.3758/BRM.40.1.183.

David P. Vinson, Gabriella Vigliocco, Stefano Cappa, and Simona Siri. The breakdown of semantic knowledge: Insights from a statistical model of meaning representation. *Brain and Language*, 86(3):347–365, 2003. URL http://www.sciencedirect.com/science/article/pii/S0093934X03001445.

Ellen M. Voorhees. The Philosophy of Information Retrieval Evaluation. In *CLEF '01: Revised Papers from the Second Workshop of the Cross-Language Evaluation Forum on Evaluation of Cross-Language Information Retrieval Systems*, pages 355–370, London, UK, 2002. Springer-Verlag. ISBN 3-540-44042-9. URL http://www.springerlink.com/content/vnq9w8q6lbe5plue/fulltext.pdf.

Ellen M. Voorhees. The TREC robust retrieval track. *ACM SIGIR Forum*, 39(1):11, June 2005. ISSN 01635840. doi: 10.1145/1067268.1067272. URL http://portal.acm.org/citation.cfm?doid=1067268.1067272.

Ellen M. Voorhees. The TREC 2005 robust track. *ACM SIGIR Forum*, 40(1):41–48, 2006. URL http://portal.acm.org/citation.cfm?id=1147205.

Ellen M. Voorhees and Donna K. Harman. *TREC: Experiment and Evaluation in Information Retrieval*. Digital Libraries and Electronic Publishing. MIT Press, Cambridge, MA, September 2005. ISBN 0262220733.

William H. Walker and Walter Kintsch. Automatic and strategic aspects of knowledge retrieval. *Cognitive Science*, 9(2):261–283, June 1985. ISSN 03640213. doi: 10.1016/S0364-0213(85)80016-1. URL http://linkinghub.elsevier.com/retrieve/pii/S0364021385800161.

William Webber and LAF Park. Score adjustment for correction of pooling bias. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 444–451, 2009. URL http://dl.acm.org/citation.cfm?id=1572018.

Wikipedia Foundation. Wikipedia — The Free Encyclopedia, 2009. URL http://wikipedia.org.

Florian Wolf and Edward Gibson. Representing discourse coherence: A corpus-based study. *Computational Linguistics*, 31(2):249–287, 2005. URL http://www.mitpressjournals.org/doi/abs/10.1162/0891201054223977.

Jinxi Xu and Bruce W. Croft. Query expansion using local and global document analysis. In *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 4–11, 1996. URL http://dl.acm.org/citation.cfm?id=243202.

Eviatar Zerubavel. *Social mindscapes: An invitation to cognitive sociology*. Harvard University Press, 1999.

Torsten Zesch and Iryna Gurevych. Automatically creating datasets for measures of semantic relatedness. In *Proceedings of the Workshop on Linguistic Distances - LD '06*, pages 16–24, Morristown, NJ, USA, 2006. Association for Computational Linguistics. ISBN 1932432833. doi: 10.3115/1641976.1641980. URL http://portal.acm.org/citation.cfm?doid=1641976.1641980.

Rolf A. Zwaan and Gabriel A. Radvansky. Situation models in language comprehension and memory. *Psychological Bulletin*, 123(2):162–185, 1998. URL http://psycnet.apa.org/index.cfm?fa=fulltext.journal&jcode=bul&vol=123&issue=2&format=html&page=162&expand=1http://psycnet.apa.org/?fa=main.doiLanding&doi=10.1037/0033-2909.123.2.162.