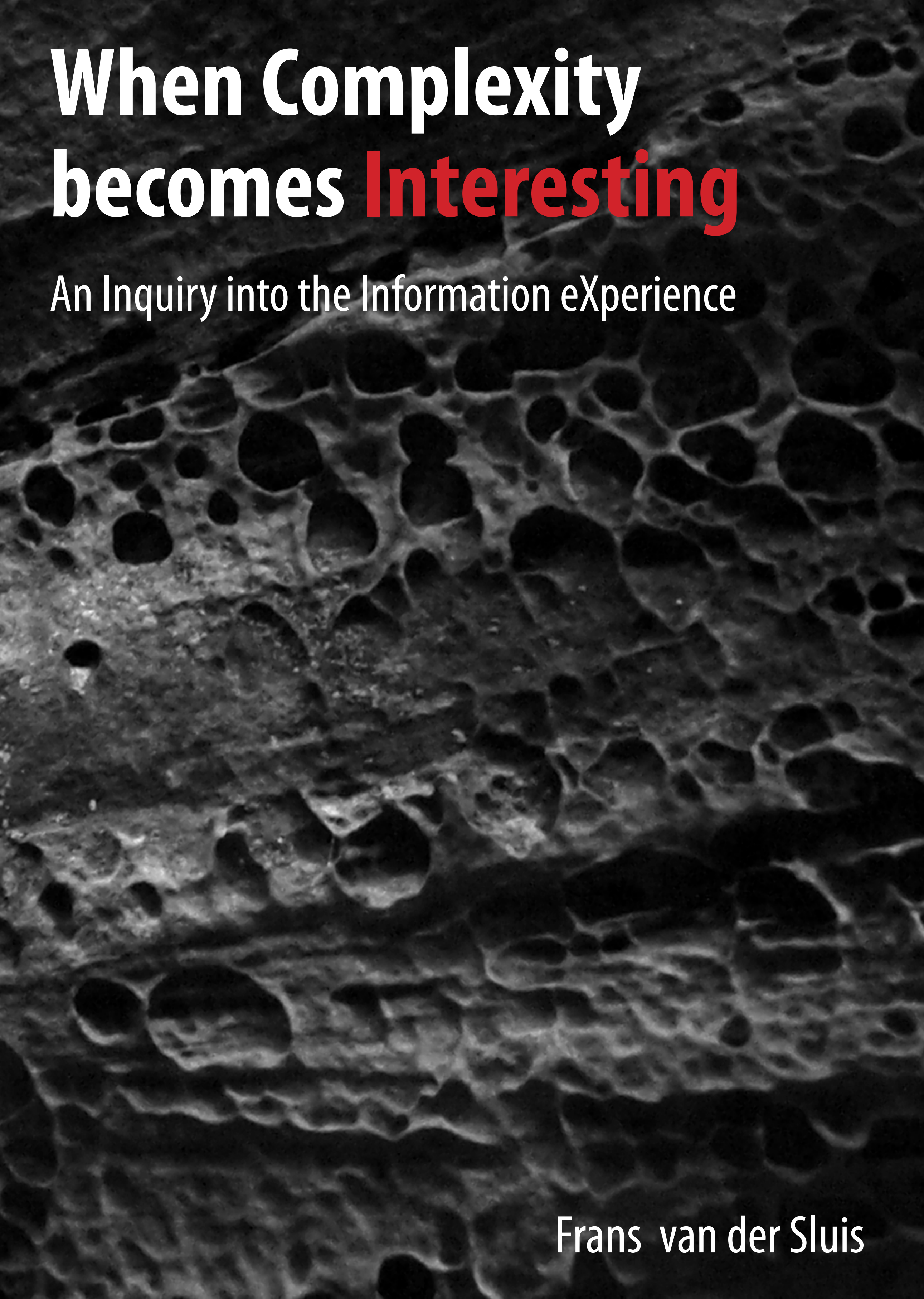
When Complexity becomes Interesting

An Inquiry into the Information eXperience



Frans van der Sluis

Cover story | The cover shows a photograph of a rock formation in Berdorf, Luxembourg: a popular rock climbing destination.

Rock climbing forms an illustration par excellence of the themes in this book.
The depicted structure of the rock yields the complexity of the climbing route. In turn, this complexity determines what the climbing experience will be. The complexity of the climbing route and the ability to cope with this complexity determines whether or not feelings are felt such as interest, enjoyment, flow, frustration, anxiety, or even ecstasy.

When Complexity becomes Interesting

An Inquiry into the Information eXperience

Frans van der Sluis

Ph.D. dissertation committee:

Chairman and Secretary

Prof. dr. ir. A. J. Mouthaan, University of Twente, The Netherlands

Promotores:

Prof. dr. F. M. G. de Jong, University of Twente, The Netherlands / Erasmus University Rotterdam, The Netherlands

Prof. dr. ir. A. Nijholt, University of Twente, The Netherlands

Assistent-promotor:

Dr. dr. E.L. van den Broek, University of Twente, The Netherlands / Radboud University Medical Center Nijmegen

Referent:

Dr. E. M. A. G. van Dijk, University of Twente, The Netherlands

Members:

Prof. dr. T. Dijkstra, Radboud University Nijmegen, The Netherlands

Dr. ir. D. Hiemstra, University of Twente, The Netherlands

Prof. dr. T. Huibers, University of Twente, The Netherlands

Prof. dr. P. Ingwersen, Royal School of Library and Information Science, Denmark

Prof. dr. I. Ruthven, University of Strathclyde, United Kingdom

Paranimfen:

Dr. R. J. Glassey, Robert Gordon University, United Kingdom

L. M. Stam, B.Sc., University of Twente, The Netherlands

When Complexity becomes Interesting

An Inquiry into the Information eXperience

PROEFSCHRIFT

ter verkrijging van
de graad van doctor aan de Universiteit Twente,
op gezag van de rector magnificus,
prof. dr. H. Brinksma,
volgens besluit van het College voor Promoties
in het openbaar te verdedigen
op donderdag 29 augustus 2013 om 16.45 uur

door

Frans van der Sluis

geboren op 16 augustus 1983
te Leeuwarden

This dissertation is approved by:

Promotores:	Prof. dr. F.M.G. de Jong, University of Twente, The Netherlands / Erasmus University Rotterdam, The Netherlands Prof. dr. ir. A. Nijholt, University of Twente, The Netherlands
Assistent-promotor:	Dr. dr. E.L. van den Broek, University of Twente, The Netherlands / Radboud University Medical Center Nijmegen, The Netherlands

ISSN: 1381-3617; CTIT Ph.D.-thesis series No. 13-262
ISBN: 978-90-365-0567-3
DOI: 10.3990/1.9789036505673

"Some men see things as they are and ask why. Others dream things that never were and ask why not."

– George Bernard Shaw (1856-1950)

Colophon

This book was typeset using LaTeX 2$_{\varepsilon}$.

Cover design and graphics: Liesbeth M. Stam, Enschede, The Netherlands.

Cover photograph: Roel Aartsen, Enschede, The Netherlands.

Printing: Ipskamp Drukkers, Enschede, The Netherlands.

CTIT Ph.D.-thesis series No. 13-262 (ISSN: 1381-3617)
Centre for Telematics and Information Technology (CTIT)
P.O. Box 217, 7500 AE Enschede, The Netherlands

SIKS Dissertation series No. 2013-28
The research reported in this thesis has been carried out under the auspices of SIKS, the Dutch Research School for Information and Knowledge Systems.

This work was part of the PuppyIR project, which is supported by a grant of the 7th Framework ICT Programme (FP7-ICT-2007-3) of the European Union.

Summary

To date, most research in information retrieval and related fields has been concerned primarily with efficiency and effectiveness of either the information system or the interaction of the user with the information system. At the same time, understanding the experience of a user during information interaction is recognized as a grand challenge for the development of information systems. There is a widely shared intuition that the value of the retrieved information is dependent on more than system characteristics such as the topical overlap between a query and a document. As it is not obvious how to embrace this intuition, this challenge has mostly been left ignored. This dissertation embarked upon the challenge of describing and developing an operational model of the Information eXperience (IX) – the experience during the interaction with information. This task was decomposed into three sub-challenges:

I Transform the fuzzy concept of the IX into a formalized one.

II Develop a model of textual complexity that enables an information system to influence a user's IX.

III Identify and influence the causes of the experience of interest in text.

The first sub-challenge has been addressed through the introduction of an Information eXperience Framework (IXF), described in Chapter 2. This framework provides structure to the intuitive understanding of an IX. The IXF structures how throughout an information interaction, the IX results from an interplay between the following four angles: on the one hand, the information objects and, on the other hand, the values, affective responses, and (cognitive-affective-motivational) states as seen or experienced by the user. Specifically, the values are approximated through the notion of relevance, which connects the IXF to models of relevance. The proposed IXF allows one to zoom-in on specific relations between each of them. These specific relations show how manipulations of characteristics of information direct a response, and how a user's state can influence judgments of information. In sum, the IXF specifies how information systems can orchestrate the IX.

The second sub-challenge was responded to through the development of a model of textual complexity. Previously, this challenge proved hard due to the innate difficulty of modeling human's comprehension and the data sets' size and variety (e.g., genres). Instead of trying to model the human comprehension ability, in this dissertation those textual features were modelled that are commonly observed in experimental studies to cause psycholinguistic processing difficulty. The resulting model was fine-tuned using a large data set that is distinctive on textual complexity. This gave a unifying model of textual complexity without the necessity of a unifying theory of processing difficulty. Two distinct studies confirmed that this novel model advances the state-of-the-art in textual complexity analysis. First,

encyclopedic texts were classified as either simple or complex, with 93.62% accuracy (see Chapter 3). Second, a user study confirmed that the objective model of textual complexity allowed to predict the mean subjectively appraised complexity of news articles, as indicated by a correlation of $r = .704$ (see Chapter 4). These results show that modeling common observations about psycholinguistic processing difficulty increases validity yet maintains applicability and, accordingly, can yield a next level of models of textual complexity.

The third sub-challenge addresses a key question in psychology: What motivates humans to explore, search, and learn? A user study was conducted that explored the determinants of the emotion of interest (see Chapter 4). This study included both objective and subjective predictors, amongst which the predictions made by the model of complexity (see Chapter 3). The study resulted in a path model that showed the extent to which complexity, familiarity, and an individual's epistemic curiosity trait influenced interest. Users' familiarity with information showed a linear relationship with users' reported interest, partly confirming the effectiveness of so-called filter bubbles in which users often see more of the same. The existence of the "sweet spot" of interest was confirmed; that is, interest peaks where the information is complex yet comprehensible. In line with this, the relation between objective textual complexity and interest provided proof for the existence of the Wundt-curve (or inverted-U shape). These findings highlight the possibility of objective models to reveal the subjective IX and provide a proof-of-concept for the IXF.

A key aspect in addressing the combination of challenges is that it required to interweave multiple disciplines, primarily information science, artifical intelligence, and psychology. This multi-disciplinarity led to the adoption of several unconventional approaches. Each of the approaches can be viewed on a bipolar dimension: the approach to the IXF integrates emotion and cognition, the approach to the model of textual complexity integrates data and theory, and the approach to the study of interest integrates objective and subjective variables. These approaches led to surprising achievements, including the re-establishment of the Wundt-curve. To the author's knowledge, the series of studies presented here is the first since the introduction of the Wundt-curve over 110 years ago that successfully confirms this relation for epistemic textual stimuli such as news articles. The Wundt-curve forms a summary par excellence of the synergy that arises from a comparison between objective and subjective variables.

This monograph took on the grand challenge of understanding and influencing the experience of a user during information interaction. Together, the series of studies presented in this monograph shows that and confirms how information systems can orchestrate the IX: *a)* it offers the IXF that specifies the relation between characteristics of information and the resulting IX, and; *b)* it confirms the value of this framework by unveiling when complexity becomes interesting. Both in terms of efficiency and experience, this line of inquiry can be expected to yield *the* next step in improving our interaction with information.

Samenvatting

Wanneer wordt complexiteit interessant?
Een onderzoek naar de informatieërvaring.

Tot op heden is het meeste onderzoek in de *information retrieval* en aanverwante velden gericht op de effectiviteit en efficiëntie van zowel het informatiesysteem als de interactie tussen de gebruiker en het informatiesysteem. Tegelijkertijd wordt het begrijpen en beïnvloeden van de gebruikerservaring gezien als een belangrijke en grote uitdaging. Er bestaat een breed gedragen intuïtief begrip dat de waarde van informatie afhankelijk is van meer dan alleen de mate van overeenkomst tussen een zoekopdracht en het onderwerp van een document. Vanwege de moeilijkheid van het concretiseren van dit intuïtieve begrip is deze uitdaging veelal genegeerd. Dit proefschrift neemt de taak op zich van het operationaliseren van de Information eXperience (IX) - de ervaring tijdens de interactie met informatie. Deze taak valt uiteen in drie subtaken:

I Formaliseren van het vage begrip IX.

II Ontwikkelen van een model van tekstuele complexiteit waarmee een informatiesysteem de IX van een gebruiker kan beïnvloeden.

III Identificeren en beïnvloeden van de oorzaken van interesse in tekst.

Voor taak I is in hoofdstuk 2 een raamwerk geïntroduceerd, het Information eXperience Framework (IXf). Dit raamwerk geeft structuur aan het intuïtieve begrip IX. Het structureert hoe gedurende de interactie met informatie de informatieërvaring het resultaat is van een samenspel tussen de volgende aspecten: enerzijds de informatieobjecten, en anderzijds de waarden, de affectieve reacties en de mentale staat van de gebruiker. De waarden worden geconcretiseerd door middel van relevantie modellen. Het IXf maakt duidelijk hoe bepaalde eigenschappen van informatie de affectieve reactie beïnvloeden en hoe de mentale staat van de gebruiker de waardering van informatie beïnvloedt. Hiermee laat het IXf zien hoe informatiesystemen de IX kunnen sturen.

Voor taak II is in hoofdstuk 3 een model van tekstuele complexiteit ontwikkeld. Voorheen is dit beperkt haalbaar gebleken door de inherente moeilijkheid van het modelleren van de menselijke begripsfunctie en de verscheidenheid aan teksten en de grootte van datasets. In plaats van de menselijke begripsfunctie zijn de tekstuele kenmerken gemodelleerd waarvan in experimentele studies vaak is vastgesteld dat zij psycholinguïstische verwerkingsmoeilijkheid creëren. Het resulterende model is vervolgens verfijnd aan de hand van een grote dataset bestaande uit teksten die verschillen in complexiteit. Dit resulteerde in een algemeen bruikbaar model zonder dat daarvoor een algemene theorie noodzakelijk was. Met twee studies werd aangetoond dat dit nieuwe model een vooruitgang vormt ten opzichte van de

huidige *state-of-the-art* in tekstuele complexiteitsanalyse. Ten eerste konden encyclopedische teksten als eenvoudig of complex geclassificeerd worden met $93,62\%$ nauwkeurigheid (zie hoofdstuk 3). Ten tweede bevestigde een gebruikersstudie dat voor nieuws artikelen het objectieve model van complexiteit de gemiddelde subjectieve inschatting van complexiteit kon voorspellen met $r = .704$ precisie. Deze resultaten tonen aan dat het modelleren van de oorzaken van psycholinguïstische verwerkingsmoeilijkheid de validiteit van het model verhoogt en tegelijkertijd de toepasbaarheid ervan in stand houdt. Deze aanpak maakt een volgende generatie modellen van tekstuele complexiteit mogelijk.

Taak III richt zich op een belangrijke vraag in de psychologie: Wat motiveert mensen om te exploreren, zoeken, en leren? Er is een gebruikersstudie uitgevoerd naar de determinanten van de interesseëmotie. Hiervoor zijn zowel objectieve als subjectieve variabelen gebruikt, waaronder de voorspellingen die voortvloeien uit het complexiteitsmodel (zie hoofdstuk 3). De studie resulteerde in een padmodel waaruit blijkt in welke mate complexiteit, bekendheid, en individuele epistemische nieuwsgierigheid de interesse in nieuwsartikelen beïnvloeden. De bekendheid van de gebruikers met de artikelen liet een lineair verband zien met interesse. Dit bevestigt de effectiviteit van zogenaamde “filter bubbles” waarin een gebruiker veelal meer van hetzelfde ziet. Verder werd het bestaan van de “sweet spot” van interesse bevestigd: interesse piekt daar waar de informatie complex doch te begrijpen is. In het verlengde hiervan ligt dat de omgekeerde-U relatie (of de Wundt-curve) tussen objectieve complexiteit en interesse is onthuld. De resultaten demonstreren dat objectieve modellen de subjectieve informatieërvaring kunnen beïnvloeden.

Een belangrijk aspect van de gevolgde aanpak is dat meerdere disciplines zijn betrokken, waaronder informatica, kunstmatige intelligentie en psychologie. Dit heeft geleid tot de toepassing van enkele onconventionele benaderingen. Elk van deze benaderingen kan worden bekeken op een bipolaire dimensie: de aanpak van het IXF integreert emotie en cognitie, de aanpak van het tekstuele complexiteitsmodel integreert data en theorie, en de aanpak van de interesse studie integreert objectieve en subjectieve variabelen. Deze onconventionele benaderingen hebben geleid tot verrassende conclusies. Zo is onder meer evidentie geleverd voor dat de Wundt-curve een ere herstel verdient. Voor het eerst sinds de introductie van de Wundt-curve, meer dan 110 jaar geleden, is deze relatie bevestigd voor epistemische tekstuele stimuli zoals nieuwsartikelen.

In het onderzoek dat in deze dissertatie wordt beschreven stond de uitdaging centraal om de ervaring van een gebruiker tijdens de interactie met informatie te begrijpen en te beïnvloeden. De studies demonstreren dat en hoe informatiesystemen de IX kunnen sturen door: *a)* het IXF, waarin de relatie tussen eigenschappen van informatie en de resulterende IX is gespecifieerd, en de waarde waarvan bevestigd is door; *b)* te laten zien wanneer complexiteit interessant wordt. Zowel qua efficiëntie als ervaring kan worden verwacht dat deze lijn van onderzoek *de* volgende stap zal vormen in de verbetering van onze interactie met informatie.

Acknowledgements

It seems we all benefit from a little bit of complexity, although within limits. And so did I during the writing of this dissertation. What started as a complicated venture (luckily :-)) turned into a solid and interesting journey. This turnaround occurred in part by the help of many people, who motivated me to dream, learn, relax, enjoy, and in many other ways, made this journey more interesting. As such, the title of this thesis, "*when complexity becomes interesting*", represents the journey of its creation. I am most grateful to everybody who makes my life more complex at times, yet easier when needed!

There are several people who made it possible for me to pursue my PhD. I would like to thank my promotors, Anton and Franciska, for giving me the freedom to pursue my research. Anton, I am particularly grateful for your best of effort at several key moments. Franciska, I am particularly thankful for your aid in writing and the enjoyable discussions that it led to :-) Betsy, your role in this dissertation should not be underestimated. You facilitated in many aspects, yet also gave me the freedom that I needed. Especially, your social understanding combined with your humbleness is an example to everybody. Finally, I am honoured by the participation of Ton Dijkstra, Djoerd Hiemstra, Theo Huibers, Peter Ingwersen, and Ian Ruthven in my dissertation committee. In particular, I would like to thank Ian Ruthven and Djoerd Hiemstra for their constructive comments.

Several people in particular played a key role in the journey that led to this thesis. Egon, you're undoubtly the most important for this thesis and, perhaps, for my work in general. You've been an inspiration, you gave a kick in the ass when needed, always brought a critical view, and you've supported me in all those things that weren't really my cup of tea. And, most importantly, you've motivated me already a long time ago for engaging with science in the first place. Let's soon drink a beer again on the past, yet also on the future! Ric, I guess it all started with a fussball table and some beers when we discovered our mutual interest was, interest. And I guess that's also how it (will) continue(d) :-) Thank you for the fine discussions, great input, and hosting an awesome and inspiring time in Glasgow. You showed me many of the best things that Scotland has to offer (e.g., beers and burgers :-)). Liesbeth, you've without a doubt significantly increased the quality of this dissertation. Thank you for your sharp eye and coherent thoughts at those moments that I totally lost them. And, above all, thank you for the brilliant cover! Still, your input to the book isn't even close to your input to my life; thanks for being an amazing friend.

My *Liebe Freunde*, luckily there's been a life next to work! Thanks to you guys for the many beers, travels, festivals, climbing, and especially all these enjoyable moments. In particular, thanks to all fun-loving friends who shared great moments at Molly's and many other places, including: Eva, Liesbeth, Julieta, Gijs, Merijn, Ronald, Jorge, Freddy, Johnny, Linn, Desmond, Nadine, Nick, Maria, Noor, . . . , and of course the employees and regulars at Molly's! My awesome colleagues at HMI, thank you for creating such a nice work environment in the past years. Whether it were the nice lunches, sharing a good coffee, the late night drinks, or playing numerous fussball games :-) Thx guys! And btw, please take good care of my precious fussball table ;-)

By the danger of missing out nearly all, I will highlight a few moments from the past four years. JW, Jeroen, Eva, and all the others who joined in the glory of my swimming pool: sweet :-) Sanne and JW, the time that you guys shared my house made some of the most fun and simultaneously remarkable months. And, above all, months with a great sequel! Ric, Sanne, Jw, Koen, Sjoerd, Jeroen, Linda, please picture again the magnificent sunrises at Exit in Serbia last year. Dancing the night away while greeting the new day. Perfect! Maarten B, thank you for being a great friend and having the best couch that served many relaxing moments. Julieta, thanks for often knocking on my door and all the lovely dinners, great conversations, and long evenings that always followed :-) And especially thanks for knocking on my digital door one surprising moment, you've clearly brought Spanish sunshine and enlightment since. Noor, dank je voor het moment van aanbellen en gezelligheid brengen. En het goede (burger-)voorbeeld geven ;P Jaak and Maarten Z, thank you for sharing the time we went to Dublin and arrived in Riga :-) And Jaak! An extra thanks to you, for all the great conversations at sometimes remarkable locations and often weird moments. Matthijs, thanks for the many moments in which you taught me about all the cool nerd-stuff :-) Thanks for being a good teacher, a better friend, and for living in one of the most beautiful areas of NL. Jos, our roads somehow keep crossing, a wonderous thing! Let's keep that up. And thanks for giving me the excuse to buy and enjoy a talking Nijntje doll :D

Climbing was perhaps the biggest inspiration for this work and for many big adventures – a great lifestyle! Thanks to all of you who accompanied me on trips, in the gym, or at the bar :-) Especially, Koen, thanks for making many of these adventures happen. I mention: many travels to Berdorf, nights in Innsbruck, crazy (!) climbing adventures in Arco, chillin' in Osp, partying in Budapest, Novi Sad, Zagreb, Ljubljana, and where not. Het leven was mooi, is mooi, en blijft (!) mooi :-) And, Sander, thanks for bringing me to many new places (e.g., the municipality office ;-)), the fun trips to Berdorf, Finale, and Bleau, and the nice dinners (thanks Michelle ;-)). Even more, endless respect that you got me to go bouldering in between icicles in the snow!

Last but far from least, I would like to thank my family for their support and trust. Mams, paps, dank jullie voor alle hulp en geduld. Jullie inzet heeft mij hier gebracht – jullie inbreng heeft veel van de hierboven beschreven momenten mogelijk gemaakt. Zusje, dank je voor alle gezelligheid, van Enschede tot aan Curacao :-)

To all of you, sorry if I haven't been the best friend, brother, or son in the past year. Sadly, the last miles of the journey of a PhD candidate also partly come with actually being away, if not in location then in thoughts. Now it's time to return again. See you soon!

Contents

List of Figures

List of Tables

List of Abbreviations

AI	Artificial Intelligence.
AUC	Area Under Curve.
CFI	Comparative Fit Index.
DLT	Dependency Locality Theory.
ECT	Epistemic Curiosity Trait.
ESA	Explicit Semantic Analyses.
IF&R	Information Filtering and Recommending.
IR	Information Retrieval.
IX	Information eXperience.
IXF	Information eXperience Framework.
LDA	Latent Dirichlet Allocation.
LRM	Logistic Regression Model.
MANOVA	Multivariate ANalysis Of VAriance.
PCFG	Probablistic Context-Free Grammar.
PMF	probability mass function.
POS	Part-Of-Speech.
RF	Random Forest.
SEM	Structural Equation Model.
SRMR	Standardized Root Mean Square Residual.

SVM	Support Vector Machine.
SWE	Sliding Window Entropy.
TF-IDF	term frequency - inverse document frequency.
UX	User eXperience.
VIRT	Valued Information at the Right Time.

1

General Introduction

Abstract

The Information eXperience (IX) follows an information system, a user, and the interaction between them. Each of these will be given a brief introduction in Section 1.1. Specifically, the relation between current development and evaluation methods and the current experience of information systems will be explored. Based on this short summary, the thesis statement is derived and three key challenges will be introduced in Section 1.2. Given the width of topics discussed and connected throughout this dissertation, this chapter will continue with a concise introduction of the core concepts in Section 1.3. In particular, of: *i)* IX; *ii)* textual complexity, and; *iii)* interest. Section 1.4 will subsequently describe the contributions made in this dissertation for each of the challenges. The introduction will finish with an outline of the remainder of the dissertation.

1.1 Introduction

Understanding the experience of a user during information interaction has been recognized as a grand challenge for information systems. Yet, in parallel, it is this challenge that is ignored the most (Belkin, 2008). This is in contrast with the intuitive understanding that the value of the retrieved information is dependent on more than system characteristics (e.g., topicality). Namely, that the experience with information is inextricably linked with the user's mood, his predisposition, expectations, and so forth (Hassenzahl and Tractinsky, 2006). Whilst it is intuitively clear that the usefulness of information systems is dependent on more than the pragmatic aspects of information, it is unclear how to embrace this intuitive understanding and, subsequently, how to use it to improve the experience during information interaction. This thesis will focus upon the task of describing and developing an operational model of the Information eXperience (IX) – the experience during the interaction with information. Crucial to this model is the subtle relationship between aspects of information and the affective response they cause. This thesis will explore this subtle relationship between textual complexity and the emotion of interest and, accordingly, show that the subjective experience can be evaluated and improved.

The current approach to evaluating information systems is pragmatic and generally neglects the subjective aspects of system performance. This is a rather narrow focus that does not adequately comprise subjective factors that constitute the user experience of the information supplied by an information system. Consequently, the challenge of improving the IX is far from resolved (Belkin, 2008). This is illustrated by several studies that show a range of negative emotions during search tasks (Kuhlthau, 2004; Arapakis et al., 2008; Bowler, 2010). Moreover, this challenge exists across many activities, whether users are making critical professional decisions, or looking for casual social interaction, and across many groups of users, varying from children to information (search) experts. For example, the PuppyIR project set out to improve the experience of children with information systems through supporting the key problems they face during information interaction (Lingnau et al., 2010). The project resulted in numerous tools to support these problems (Van der Sluis and Van Dijk, 2010), such as a query-by-image system that relieves the difficulty of coming up with proper search terms (Van der Sluis et al., 2011) and an interface extension that provides explanations for complicated words (Eickhoff et al., 2012). The many group-specific solutions that have been devised suggest the salience of looking at the experience of information systems.

The current focus on pragmatic aspects for information systems turned out highly succesful. Information systems can effectively retrieve, aggregate, rank, filter, and recommend information. The concept that lays at the basis of these features is relevance: that is, retrieving information (objects) that are likely to be relevant for the user. However, since the experience of a user is not evaluated as part of a system's usefulness, this does not always lead to an optimal experience. We will illustrate this next for Information Retrieval (IR) and Information Filtering and Recommending (IF&R) systems.

IR systems are evaluated on their ability to deliver relevant documents given a query. The ground truth that is used to evaluate relevance in IR is typically generated by information experts who assess whether a document is relevant to a query or not (i.e., the Cranfield paradigm; Voorhees, 2002). Looking solely at the query-document relation has its limitations. This is demonstrated by the continuous search for other relevance predictors (Borlund and Ingwersen, 1998; Demartini and Mizzaro, 2006). Additionally, this is demonstrated by the finding that people use topicality mainly to filter out irrelevant documents and not to make their final selection. Given these limitations it is not unexpected that Kuhlthau (2004), Arapakis et al. (2008), and Bowler (2010) have observed the occurrence of a range of negative emotions (e.g., irritation, anxiety, and despair) during retrieval tasks.

IF&R systems are a similar case. IF&R systems are based on the assumption that selecting information based on its topical similarity, selections made by other users, or the characteristics of the user will lead to a positive experience during information interaction (cf. Konstan and Riedl, 2012). The current approach to evaluating the performance of IF&R systems is by assessing their ability to predict (withheld) (e.g., movie-)ratings. The limitations of this approach are demonstrated by the existence of a "magic barrier" in the prediction of ratings; that is, the finding that there seems to be a ceiling to the achievable performance. People provide inconsistent ratings for the same item when asked at different times (Hill et al., 1995), suggesting that unexplained natural variability creates this "magic barrier" (Herlocker et al., 2004). The focus on similarity can be thought of as mainly selecting "more of the same", which generates filter bubbles that are characterized by a limited degree of novelty (Ricci et al., 2009). Together with the existence of a "magic barrier", it could be argued that "more of the same" may cause "diminishing returns" and be detrimental to a user's experience with the information system.

Those information systems that deal with the retrieval or selection of information play a key role in creating a solution to the non-optimality of the current experience with information systems. They can select information that likely leads to a target experience. Yet, the preceding review of the current evaluation methods applied for IR and IF&R systems has highlighted several limitations, both in terms of efficiency and experience. Seemingly, to evaluate the usefulness of an information system and the user experience, either positive or negative, one cannot rely on system behavior alone.

1.2 Thesis Statement

This thesis aims to include the subjective factors that constitute an IX in the evaluation of information systems. Instead of looking solely at the ability of an information system to retrieve and select relevant documents, an argument will be made for the importance of the IX with respect to a system's usefulness. However, although the concept of IX has an intuitive attractive quality to it, simultaneously it can be critiqued for being vague and even elusive. It seems that experience is a concept which is very difficult to grasp. To fill this gap, the IX will be defined and structured through the introduction of an Information eXperience Framework (IXF). Building upon the basis provided by the IXF, this thesis will show that the IX can be evaluated on a range of subjective factors and, subsequently, the experience of interest be explained. In particular it will be shown that, by applying a model of textual complexity, information systems select information that is likely to lead to the emotional response of interest.

This monograph will undertake the following three core challenges:

I To transform the fuzzy concept of the IX into an amendable target for information systems. This requires the identification and specification of the different aspects that constitute an IX and relate them to the notion of relevance that generally underlies information systems.

II To develop a model of textual complexity that is applicable to a variety of data sets and predictive of subjective appraisals of textual complexity. Moreover, a key challenge for this model is to be able to influence a user's IX.

III To identify the causes of the emotion of interest and to explore whether these causes can be influenced by an information system.

A key aspect of the combination of challenges is that it requires the interweaving of multiple disciplines, primarily information science, Artificial Intelligence (AI), and psychology. Together, these challenges form a road map to prove whether and to show how a user's IX can be changed. As such, the combination aims to show that the IX is an amendable target for information systems and that it forms a valuable aspect of the evaluation of a system's usefulness. The coming section will describe the core constructs that constitute each of these challenges.

1.3 Experience, complexity, and interest

A concise introduction will be given to the remainder of this thesis. Three core constructs will be highlighted: experience, complexity, and interest.

1.3.1 Information eXperience

The current conceptualizations on the experience during information interaction do not allow for it to be used for evaluative purposes. It is unclear what exactly constitutes a fruitful IX: neither is it clear which emotional experiences are desirable or "useful" during interaction, nor what their causes or effects are (Kuhlthau, 2004; Arapakis et al., 2008; Belkin, 2008; Bowler, 2010). There is a clear need to delineate what constitutes a better IX and correspondingly a better User eXperience (UX): that is, the complex fabric of thoughts, feelings, and actions experienced during user interaction (Hassenzahl, 2013).

The scope of inquiry of the IX can be considered a subset of the scope of the UX. The latter describes the experience during interaction with all facets of a product, including its design and interface. The former focuses only on one aspect: the information. The UX and IX are probably not independent. A good UX, as caused by other aspects (e.g., a search interface), can change the perception of the IX. Nonetheless, the focus in this thesis will be solely on the IX.

Whilst UX is a difficult term to operationalize, it provides guidelines for conceptualizing and utilizing the IX. Several aspects of the UX have been identified: usability, beauty, enjoyment, meaningfulness, emotions, temporality, situatedness, enjoyment, motivation, and challenge (Hassenzahl and Tractinsky, 2006; Csikszentmihalyi and LeFevre, 1989). Together, these aspects describe part of the UX Hassenzahl and Tractinsky (2006) and, possibly, a user's IX. This thesis will define the IX as the values, affective responses, and experiential states that arise during interaction with information. Applying this definition allows us to explain how characteristics of information influence aspects of the IX and how aspects of the IX influence the goal of solving an information need. Hence, it can be used to structure which aspects of relevance are of prime importance for a fruitful IX.

Aside from the influence of aspects of relevance on a user's experience with information, the experience influences the relevance criteria a user applies as well. There is an intricate relation between emotion and knowledge activities, showing that emotions exist ubiquitously when dealing with information. Emotions affect problem solving (Jonassen, 2000), learning (Kort et al., 2001), attention (Lang, 1995), decision making (Pfister and Böhm, 2008), and making inferences (De Sousa, 2008). This intricate relation between emotion and knowledge activities shows that, apart from improving a user's IX as salient primary goal, the IX can contribute to a system's usefulness. The notions of relevance and IX are intertwined. As we will show next, besides topicality the complexity of information is a salient influence on a user's IX.

1.3.2 Textual Complexity

Textual complexity is expected to be an important factor in influencing a user's experience with information. It has been identified as part of the relevance criteria users apply when selecting information (Barry and Schamber, 1998; Xu and Chen, 2006). Its role comes aside from the topical familiarity of the user that is often implemented in IF&R systems. And, its role comes on top of topicality that is generally regarded as a pre-condition to the importance of other types of relevance (Spink and Greisdorf, 2001). When included in the set of information metrics, a metric of textual complexity allows an information system to directly affect a user's IX. However, for a metric of textual complexity to actually influence the IX, the metric needs to have (predictive) validity: that is, it needs to be reflective of subjective, experienced, complexity. To achieve this predictive validity, the model needs to reflect knowledge about what aspects of a text create reading difficulty.

Some studies attend the validity of a metric of textual complexity. However, the actual predictive validity generally remains untested (Benjamin, 2012). Notable exceptions are provided by Collins-Thompson et al. (2011), who use a readability metric to explain the time that a user of an IR system spent on a web page, and by Vor der Brück et al. (2008), who propose a solution in the form of "deep" features (e.g., cohesion) reflective of cognitive constructs (e.g., coherence) and test this solution using subjective judgments of readability for a separate data set. Although some steps have been made towards improving the validity of a model of textual complexity, the ability of such a metric to actually influence a user's IX is unclear. Nonetheless, there is theoretical potential for an indicator of textual complexity to influence the IX. This potential will be described next in relation to the experience of interest that a user can have given a piece of information.

1.3.3 Interest

Interest is regarded as an emotion associated with curiosity, exploration, information seeking, and learning. Interest is believed to be key to a positive experience:

the *“quality of experience seems to be an epiphenomenon of interest”* (Schiefele, 1996, p. 13) and to be part of an engaging experience (O’Brien and Toms, 2008). People who experience an interest response are attracted to the evoking stimulus (Silvia, 2008b). For example, when textual stimuli raise an interest response, people experience a higher level of arousal and process the text more deeply (Schiefele and Krapp, 1996).

According to the contemporary interest-appraisal theory (Silvia, 2008b), interest occurs after two consecutive, subjective, appraisals. The primary appraisals evaluate stimuli by their “novelty-complexity”: assessing whether the stimulus is sufficiently novel and complex, or too predictable and not challenging enough to stimulate interest. The secondary appraisal evaluates the “comprehensibility” of the stimulus, determining the coping potential related to prior knowledge, available resources, and so forth. For example, if a stimulus is too complex, the coping abilities will probably not suffice, leading to a different emotion. A stimulus, then, fosters an interest response if at the first stage appraised as novel and complex, yet at the second stage appraised as comprehensible (Silvia, 2006). Hence, we can define the “sweet spot” between novelty-complexity and comprehensibility in which interest peaks.

Textual complexity is key to both appraisal evaluations, allowing us to approach the “sweet spot” of interest. Whilst the complexity of a text can enhance the primary appraisal, making a text more challenging, it can also impair the secondary coping appraisal, if appraised as too complex. This combination of effects creates a so-called Wundt-curve or inverted-U between textual complexity and interest. However, whereas Wundt (1896) and Berlyne (1970) related complexity as an objective property of the stimulus to the subjective experience of interest, the interest-appraisal theory focuses solely on subjective appraisals (Silvia, 2006). And, whereas several studies confirm a negative effect of complexity on interest via the reduction of comprehensibility (Schraw et al., 1995; Connelly, 2011; Sadoski, 2001), little evidence exists for a positive effect of complexity on interest in text. The model of textual complexity will be used as a proof-of-concept for the ability of information systems to affect the experience of interest. By taking interest as a target experience for information systems, this monograph operationalizes the holistic concept of the IX as the specific concern of predicting if and when a stimulus leads to an interest response.

1.4 Contributions

To solve each of the three challenges, this thesis presents three contributions. The key contributions will be described shortly per challenge.

Challenge I (see Section 1.2) is solved by introducing the IXF. The IXF transforms the fuzzy, ambiguous concept of IX into an amendable target for information systems. If the IXF is implemented, information systems can benefit from a synergy between relevance and the IX. The IXF shows how the prediction of relevance can improve using measurements of both the responses and states of the user, as well as of how a user's IX can be improved by selecting information with certain characteristics (e.g., a particular level of complexity). The IXF deviates from contemporary approaches to information interaction with a focus on "feelings" and experiences instead of "reason" and the fulfillment of an (information) need. Parts of this chapter have been published in Van der Sluis et al. (2010), Van der Sluis et al. (2010), and Van der Sluis and Van Dijk (2010).

Challenge II (see Section 1.2) is fulfilled through the development of a model of textual complexity. The concept of processing difficulty is introduced as subjective counterpart of complexity. A deviation is made from contemporary approaches that focus on, amongst other things, comprehensibility. Subsequently, the model of textual complexity is constituted by a set of features that integrate psycholinguistic findings on the processing difficulty which can arise during the reading of text. Instead of adhering to an overarching theory that explains how these features combine to form processing difficulty, the model is tuned using a large data set distinctive on complexity. As a combination of contemporary data-driven and traditional theory-driven approaches is applied, essentially the best of both worlds is achieved: applicability and validity. Later on, the resulting model is applied to influence the IX and, in particular, the emotion of interest. Parts of this chapter have also been described in Van der Sluis and Van den Broek (2010) and Van der Sluis et al. (ip).

Challenge III (see Section 1.2) is solved by an experimental study that explores the antecedents of the emotion of interest. Besides the influence of an individual's topical familiarity and trait curiosity, the study evaluates if and when textual complexity influences the emotion of interest by applying the objective model of textual complexity. The effects of familiarity, textual complexity, and curiosity trait is evaluated for both consecutive appraisals, appraised complexity and appraised comprehensibility, and in relation to interest. The effects are summarized into an explanatory path model. The resulting path model allows us to reflect on, amongst other things, when complexity becomes interesting. Within current approaches to the study of emotion the comparison between an objective model and a subjective experience is rare and even somewhat abandoned due to a lack of success. However, as this thesis will show, with a proper conceptualization such a combination between objective and subjective can be insightful and successful. Moreover, this is an important step in better understanding the interplay of information interaction and the experience of interest, and, accordingly, a novel step in operationalizing the IX. Parts of this chapter have been published in Van der Sluis et al. (2012) and Van der Sluis et al. (ip).

1.5 Outline

The current section will give an outline of the remainder of this thesis. There are five parts and a set of appendices.

The first part is the **prologue**, which includes the chapter you are currently reading. It offers a short introduction, raison d'etre, and overview of the remainder of the monograph.

Chapter 2 consists of a lengthy exploration of the relation between relevance, a key construct from information science, and the IX. It defines and introduces all the core, subjective, factors that form the IX. Moreover, it highlights the possible synergistic relations between each of the factors. As such, the IXF solves Challenge I (see Section 1.2) of transforming the IX into an amendable target for information systems.

Chapter 3 describes the bases for and implementation of the model of textual complexity that is applied later-on again in the monograph. An extensive description of psycholinguistic findings on the causes of processing difficulty is given. These causes are implemented into a set of features which, together, form the model of textual complexity. The performance of the model to differentiate between texts of different complexity is tested on a large data set of encyclopedic articles. Based on both the method and the results, it is shown that the model adheres to Challenge II (see Section 1.2).

Chapter 4 presents an experimental study that explores the antecedents of the emotion of interest. The model of Chapter 3 is directly applied in Chapter 4 to filter both complex and easy stimuli from a news corpus, making Chapter 4 an additional test-case for the generic model from Chapter 3. The chapter shows that, with this model of textual complexity, the emotion of interest can be influenced. Hence, this chapter will solve challenge III (see Section 1.2) and at the same time give a proof-of-concept that a user's IX can be changed.

Chapter 5 summarizes and discusses the results reported in this monograph and, accordingly, forms the epilogue of this thesis. Aside from attention to the results, the value of the applied approaches will be assessed within the general context of the respective research fields. Moreover, extra attention will be given to the synergy that arises from the combination of topics presented in this monograph.

Finally, four appendices are included. These appendices offer extra analyses and details for each of the content chapters.

2

Modeling Information eXperience

Abstract

The Information eXperience (IX) is a fuzzy concept which needs to be made concrete in order to allow for its application. To transform the IX into a workable concern for information systems, throughout this chapter the notion of relevance, a central notion for many information systems, is compared to the IX. The chapter begins with an introduction in which the potential of the IX is highlighted, for both its practical value and its potential to help understand and predict relevance. It then continues with an elaborate review of the notion of relevance in Section 2.2. Subsequently, a clear definition is provided to the IX and the Information eXperience Framework (IXF) is introduced. The IXF formulates four perspectives that describe an IX: information objects, values, affective responses, and experiential states. By reviewing the connections between each of the perspectives, it is made clear to what extent we can design an experience. In parallel, the connections indicate how aspects of the IX can function as relevance feedback. The IXF forms the blueprint on which the remainder of this dissertation continues. Specifically, it shows the importance of textual complexity in relation to affective responses in general and the emotion of interest in particular.

Parts of this chapter also appear in:

Van der Sluis, F., Van Dijk, E.M.A.G., and Van den Broek, E.L. (2010). Aiming for user experience in information retrieval: Towards User-Centered Relevance (UCR). In Chen, H.-H., Efthimiadis, E. N., Savoy, J., Crestani, F., and Marchand-Maillet, S. (Eds.), *SIGIR 2010: ACM Proceedings of the 33rd International Conference on Research and Development in Information Retrieval*, p. 924 – 924, July 19–23, Geneva, Switzerland.

Van der Sluis, F. and Van Dijk, E.M.A.G. (2010). A closer look at children's information retrieval usage: Towards child-centered relevance. In Serdyukov, P., Hiemstra, D., and Ruthven, I (Eds.), *Proceedings of the Workshop on Accessible Search Systems held at the 33st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, p. 3 – 10, July 23, Geneva, Switzerland.

Van der Sluis, F., Van den Broek, E.L., and Van Dijk, E.M.A.G. (2010). Information Retrieval eXperience (IRX): Towards a human-centered personalized model of relevance. In *WI-IAT 2010: IEEE Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, Vol. 3, p. 322 – 325, August 31 – September 3, Toronto, Canada.

2.1 Introduction

The information available to users is often compared to a landscape; that is, an "information landscape". Characteristic of a landscape is that it is omnipresent. Similarly, in the "information landscape", information is omnipresent and exists in vast and increasing quantities. Early on, the invention of the printing press already inspired the idea of information overload: an ever expanding amount of information that is associated with feelings of unease (Rosenberg, 2003). Several types of information systems have been developed to assist users in their need to make sense of the information landscape. These information systems support their users in goals such as learning (intelligent tutoring systems), finding (Information Retrieval (IR) systems), and encountering (Information Filtering and Recommending (IF&R) systems). Aside from the negative connotation of overload, the abundance associated with information overload can also lead to positive feelings and even to expressions of ecstasy (Rosenberg, 2003). In other words, information overload creates an "information opportunity": the opportunity to provide users with a positive Information eXperience (IX).

Information systems that deal with the retrieval or selection of information play a key role in creating a solution to the negative effects of information overload as well as in utilizing the "information opportunity". The feature that allows information systems to play this role is their potential to distinguish information objects in terms of relevance. Relevance pertains to the relation between a user and a piece information. Sometimes a (weak) form of relevance is explicitly modeled, such as is the case for information retrieval systems. Yet, its invisible hand is almost always present in information systems that are used for in the selection of information (Saracevic, 2007). Denning (2006) suggests that increasing relevance forms the ultimate remedy to the negative effects of information overload. An optimized prediction of relevance allows information systems to select Valued Information at the Right Time (VIRT). To optimize relevance with respect to VIRT it is important to identify which aspects of relevance lead to value. In other words, to determine relevance increasing features such as topicality, novelty, and scope, the "information opportunity" can be utilized to improve the IX.

The IX is the experience during interaction with information. The IX of current information systems is not always optimal. It is often characterized by negative emotions, filter bubbles, and fast thinking (Arapakis et al., 2008; Ricci et al., 2009; Sparrow et al., 2011). Each of these three characteristics will be described concisely. First, Kuhlthau (2004), Arapakis et al. (2008), and Bowler (2010) have observed among users the occurrence of a range of negative emotions (e.g., irritation, anxiety, and despair) during information retrieval tasks (Arapakis et al., 2008; Bowler, 2010) and more complicated information seeking tasks (Kuhlthau, 2004). D'Mello et al. (2007) showed the occurrence of boredom, confusion, and frustration during learning in combination with intelligent tutoring systems. Second, IF&R systems generally base their selections of information on "more of the same", which may cause "diminishing returns". These narrow selections generate filter bubbles with a limited degree of novelty (Ricci et al., 2009) and are thought to be detrimental to the IX (Silvia, 2008b). Finally, the increased access to information due to the advent and success of information retrieval systems causes a change in information behavior. Users remember less of the actual content but instead remember how to access the information (Sparrow et al., 2011). Instead of fostering slow thinking and reasoning about the information, the current IX seems to be characterized by fast and effortless thinking that is not optimal for learning (Kahneman, 2003).

Certain aspects of relevance can counter negative emotions, filter bubbles, and fast thinking. In addition to the finding that a lack of topicality leads to negative emotions (Arapakis et al., 2008), topicality is also one of the factors that contributes to user satisfaction (Gluck, 1996; Al-Maskari and Sanderson, 2010). And, novelty has been posited as an important evaluative criterion of IF&R systems and as a solution to the negative consequences of filter bubbles (Ricci et al., 2009). Finally, a certain degree of irrelevance can cause confusion, counter fast thinking, and be beneficial for learning in combination with intelligent tutoring systems (D'Mello and Graesser, 2012). The effects of topicality, novelty, and even irrelevance shows that certain aspects of relevance are key to the IX. Through implementing and manipulating distinct aspects of relevance it is possible to direct the IX.

Although certain aspects of relevance have been identified by several researchers as key to the IX, the influence of relevance on the IX is generally not conceptualized in full accordance with the width of findings on either relevance or experience. Often, a comparison is made between an indication of topicality (an aspect of relevance) and an indication of positivity (a descriptor of experience). Such is the case when comparing topicality to user satisfaction (Gluck, 1996; Huffman and Hochster, 2007) or to the general affective state of the user (Arapakis, 2010). The limitations of these approaches have been acknowledged. For example, Gluck (1996) concludes that "*neither relevance nor user-satisfaction subsumes the other concept*" (p. 89). And, Arapakis et al. (2008) found that measures of facial expression and psychophysiological reactions could explain no more than 60.4% of singular, binary assessments of document relevance. Notwithstanding, more elaborate studies exist that investigated the constitution and determinants of the IX. Several empirical studies explored a plethora of determinants on particular aspects of the IX (e.g., on user satisfaction; Al-Maskari and Sanderson, 2010). And, more detailed descriptions of the IX exist as well (e.g., O'Brien and Toms, 2008; Hassenzahl and Tractinsky, 2006). To fully benefit from the potential to let relevance direct the IX, a conceptualization is needed that covers the detailed findings on relevance, experience, and their connection.

In addition to utilizing the effect of (aspects of) relevance on the IX, predicting the extent to which a user finds a piece of information relevant remains a challenge as well (Schamber et al., 1990; Saracevic, 2007). Relevance is difficult to predict because relevance is a multifaceted judgment that changes over time and between situations. In other words, relevance is a subjective, multi-dimensional, dynamic and situational phenomenon (Schamber et al., 1990; Ruthven, 2005). Current information systems have difficulties with the multi-dimensional and dynamic nature of relevance (Saracevic, 2007). Usually, the prediction of relevance is implemented via objective or weak relevance which denotes a static indication of the similarity between a query or model of user interests and an information object. The importance of other aspects of relevance, aside from topicality or "more of the same", have been indicated as well. In particular, aspects such as novelty, reliability, understandability, and scope are salient for relevance decisions (Xu and Chen, 2006; Xu, 2007). These aspects show that, although the notion of relevance is intuitively clear, explaining and predicting relevance judgments can be hard.

There is a need for conceptualizations that account for and allow us to handle the multi-dimensional and dynamic nature of relevance. A framework of the IX can help explain and predict what a user finds relevant. Several existing models and findings contribute to this proposition. Namely, Wilson (2006) proposed a model of an individual's information behavior and included motivational, affective, and cognitive layers that create an information need. The motivational layer influences the affective layer and the affective layer influences the cognitive layer in determining the information behavior and, accordingly, the relevance judgments (Nahl, 2005; Wilson, 2006). And, Kuhlthau (2004) showed how these three layers and the resulting information behavior changed throughout an information seeking session. When users proceed with a complex information seeking task their feelings became more certain, thoughts more focused, and their information behavior more directed and exhaustive. Next to the influence of experience on the information need, an emotional need may exist secondary to an information need (Ruthven, 2012; Cosijn and Ingwersen, 2000). This emotional need may even be the primary need (Moshfeghi, 2012). These findings and related theories indicate that the experience of the user influences the information need and, in turn, the applied relevance judgments (Schamber, 1994; Cosijn and Ingwersen, 2000; Wilson, 2006; Nahl, 2005): that is, the momentary experience directs the relevance judgments that a user makes.

Although many authors have stated that the momentary experience is a salient determinant of information behavior, the specificities of how the IX directs relevance judgments are unclear. For example, Saracevic (2007) identified affective relevance, Cosijn and Ingwersen (2000) subsumed the influence of the IX under intentionality, and Wilson (2006) described the three layers underlying information behavior. Although each of these articles noted the existence of an influence of the IX on information behavior they did not explicate this influence. Hence, similar to the necessity for a model of the influence of relevance on the IX, there is the necessity for a model of the influence of the IX on relevance. That is, to show how relevance judgments are influenced by the IX and how this influence can be used to improve the prediction of relevance by information systems.

This chapter will explore the intricate relationship between relevance and IX and examine their synergistic potential: how algorithmic relevance influences the IX and whether and how the IX can inform algorithmic relevance. Put differently, the selected information affects the emotional responses and more general cognitive-affective-physiological state of the user. In turn, this state influences the relevance judgments a user makes. In particular short-lived emotional responses can function as relevance feedback. A model of the IX will be proposed to explicate the mutual dependence between the IX and relevance; the Information eXperience Framework (IXF). An overview of the IXF is shown in Figure 2.1. Figure 2.1 also shows the possible synergistic relations between the IX and relevance.

The remainder of this chapter is organized as follows (see Figure 2.1). First, the two main concepts of this chapter will be introduced, relevance (Section 2.2) and IX (Section 2.3). Second, the IX will be elaborated on using the following three perspectives from the IXF:

(a) Values, (Section 2.4), operationalized through instrumental (Section 2.4.1) and non-instrumental (Section 2.4.2) relevance;

(b) Responses (Section 2.5), in particular feelings that originate during information processing and while resolving an information need; and,

(c) States (Section 2.6), which emerge from the components of the IX and are situated in a user, and exist only in the moment; in particular three states are identified which are particularly salient during information interaction.

Thirdly, Section 2.7 explores how relevance directs the IX and Section 2.8 outlines possibilities on how the IX can direct relevance. Finally, Section 2.9 concludes on the merits and feasibility of harnessing the potential for mutual reinforcement between relevance and the IX.

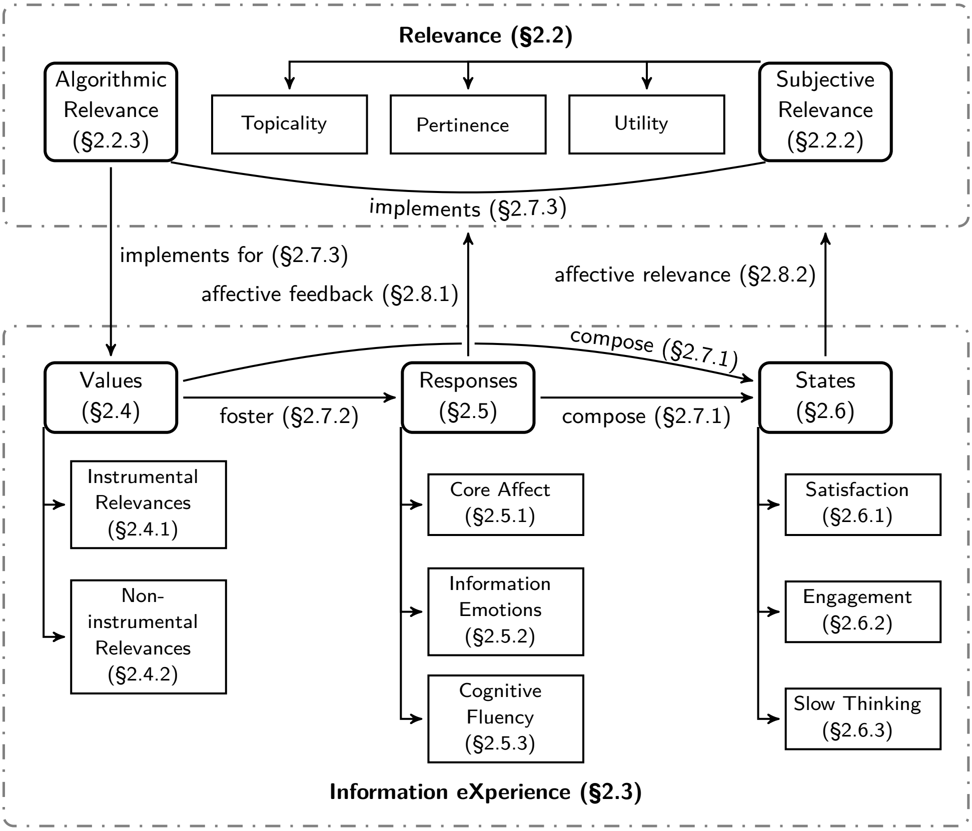

Relevance (§2.2)
Algorithmic Relevance (§2.2.3)
Topicality
Pertinence
Utility
Subjective Relevance (§2.2.2)
implements (§2.7.3)
implements for (§2.7.3)
affective feedback (§2.8.1)
affective relevance (§2.8.2)
compose (§2.7.1)
Values (§2.4)
foster (§2.7.2)
Responses (§2.5)
compose (§2.7.1)
States (§2.6)
Instrumental Relevances (§2.4.1)
Non-instrumental Relevances (§2.4.2)
Core Affect (§2.5.1)
Information Emotions (§2.5.2)
Cognitive Fluency (§2.5.3)
Satisfaction (§2.6.1)
Engagement (§2.6.2)
Slow Thinking (§2.6.3)
Information eXperience (§2.3)

Figure 2.1: Overview of the chapter structure and of the model of Information eXperience (IX) in relation to relevance.

2.2 Relevance

The thinking about relevance is converging (Jansen and Rieh, 2010); a consensus is arising on its definition and its models. An oft-cited definition of relevance is given by Saracevic (1975), stating that "*relevance is the A of a B existing between a C and a D as determined by an E*," where A may be "*measure, degree, estimate* ...;" B may be "*correspondence, utility, fit,* ...;" C may be "*document, information provided, fact* ...;" D may be "*query, request, information requirement* ...;" and E may be "*user, judge, information specialist*" (p. 150). Following this definition is the idea that there are many relevances, as aptly noted by Mizzaro (1998) with the question "*how many relevances in information retrieval?*". This multi-faceted, complicated relevance relation can be made tangible through theories (Section 2.2.1) and models (Section 2.2.2) that aim to delineate its multi-dimensional and dynamic nature.

2.2.1 Relevance Theory

Saracevic (2007) concludes that "*we are still in search of a theory of relevance*" (p. 1923). Nonetheless, there is one (debated) theory on relevance (Sperber and Wilson, 1996) with the potential for application in information science (Harter, 1992; White, 2007a,b). At the center of the theory are two basic principles. First, the cognitive principle of relevance, which claims that cognition tends to maximize relevance. Second, the communicative principle of relevance, which states that every observable stimulus conveys a presumption of being relevant (Sperber and Wilson, 1996). Relevance, then, is derived from a ratio between:

1. Cognitive Effect, denoting a worthwhile change to an individual's representation of the world.

2. Processing Effort, denoting the effort required for individuals to process a stimulus.

Every stimulus is evaluated on these two components relative to other stimuli. Of all stimuli available at a moment in time, the one with the highest ratio between cognitive effect and processing effort is processed.

There are some indications of the value of relevance theory to information science. Harter (1992) used relevance theory to derive psychological relevance, which lead to the identification of pertinence or cognitive relevance as a (sub-)type of relevance (see Section 2.2.2). White (2007a,b) successfully applied the concepts of cognitive effect and processing effort to bibliometric retrieval. Although this shows the value of relevance theory for IR, the study merely had its constructs "*on loan*" (Saracevic, 2007, p. 1923). Namely, it is unclear whether the implementations by White (2007a,b) actually reflect the cognitive effect and processing effort as perceived by the user. Notwithstanding, relevance theory does suggest when a user finds a stimulus relevant. In particular, relevance theory indicates an important aspect of relevance often not implemented in information systems: processing effort.

2.2.2 Subjective Relevance Models

Models on relevance allow us to structure observations about relevance such as its multi-dimensionality and dynamicity. The multi-dimensional nature of relevance has been aptly illustrated by Schamber (1994): "*[the] statement made earlier – that relevance is a multi-dimensional phenomenon – is, of course, a gross understatement. In fact, so many factors have been suggested as affecting relevance judgments that it is not possible to list them all here. The 80 factors listed . . . , however, represent a reasonable sample*" (p. 19). The factors she refers to are criteria that users apply when they judge the relevance of an information object.

The many relevances (i.e., manifestations of relevance) that describe the relevance relation can be classified according to the following terminology (Borlund, 2003; Cosijn and Ingwersen, 2000; Mizzaro, 1998; Saracevic, 1975, 2007):

Classes An overall differentiation between objective, systems-based and subjective, user-based approaches to relevance. The latter is the main focus of this section.

Levels Within the classes, relevance manifests itself at different levels (i.e., types). The typical levels include: algorithmic, topicality, pertinence, and utility (Saracevic, 2007).

Criteria When users make relevance judgments, they base their final judgment on several criteria. Although theoretically expected to belong to the different levels of relevance, empirical evidence supporting this is thin (Borlund, 2003).

Degree The rating or scale used to indicate the strength of relevance, its types, or its criteria. The degree is usually described as either binary or non-binary.

Dimensions The axes along which the levels of relevance change, such as time and affect.

The multi-dimensionality of relevance refers to the set of classes, levels, and criteria (Borlund, 2003), yet not to the dimensions itself (Mizzaro, 1998). On the contrary, the dynamics of relevance depict the dimensions along which the classes, levels, and criteria change[1].

At least four types of relevance can be identified spanning the two classes. The class of objective or systems-based relevance subsumes the type of algorithmic relevance. Algorithmic relevance takes a static and objective approach to the concept of relevance. It is the most used and probably the best understood form of relevance, and will be further described in Section 2.2.3. The remaining types of relevance belong to the user-based class of relevance, which sees relevance as

[1] Mizzaro (1998) identified four interacting dimensions: information type, request type, time, and components. Here, the components dimension consists of the classes, levels, and criteria of relevance.

a subjective individualized mental experience that involves cognitive restructuring (Borlund, 2003). The following three types of subjective relevance can be identified:

1. Topicality or aboutness, which compares a manifestation of information to a request;

2. Pertinence or cognitive relevance, which compares a manifestation of information to the cognitive state of the user, and;

3. Utility or situational relevance, which compares a manifestation of information to the task at hand that underlies the information need.

The last, utility, is often considered the most realistic type of relevance (Borlund, 2003). More types of relevance have been proposed as well. For example, Saracevic (2007) and Cosijn and Ingwersen (2000) suggest the influence of social and cultural context on the relevance judgments and Saracevic (2007) proposes the affective relevance type to capture the user's motivation. The classes, types, and criteria of relevance make clear that relevance is indeed a multi-dimensional concept.

Besides being multi-dimensional, relevance is dynamic as well (Schamber et al., 1990). For example, in the context of recommender systems, users were shown to give inconsistent ratings for the same item when asked at different times (Hill et al., 1995). This is partly because of the dynamic interplay between the types of relevance (Saracevic, 2007), and partly because of several dimensions along which relevance operates. In particular, the dimensions of time (Mizzaro, 1998) and affect (Cosijn and Ingwersen, 2000). The time dimension indicates that relevance changes throughout an information interaction session (Mizzaro, 1998; Cosijn and Ingwersen, 2000; Borlund, 2003). An explanation for the influence of time is given by Harter (1992), who proposes that each relevance judgment is a function of the user's mental state, and that each information object serves as a stimulus that results in cognitive changes to this state.

Compared to the time dimension, the influence of affect is more often disputed. It has been included in models of relevance as a type of relevance (Saracevic, 1996), as another dimension affecting relevance (Cosijn and Ingwersen, 2000), and as a need creating relevance (Wilson, 2006). These different perspectives on the role of affect do not exclude each other. In the case of an affective or emotional need, this need is reflected in a type of relevance (Moshfeghi, 2012). An example of this interrelation is in queries about grief, where emotions influence the expressed information need as well as applied relevance judgments (Ruthven, 2012). Similarly, the current cognitive-affective state of the user influences all types of subjective relevance (Borlund, 2003; Cosijn and Ingwersen, 2000; Saracevic, 2007). Saracevic

Table 2.1: The role of relevance for main Information Retrieval (IR) and Information Filtering and Recommending (IF&R) techniques.

Techniques	Dynamicity	Dimensionality	Degree	Class
Algorithmic relevance	static	singular	binary	system
Relevance feedback	session	singular	binary	user
Cognitive filtering	static	multiple	non-binary	user
Collaborative filtering	static	multiple	non-binary	user

(2007) stated about this influence that "*it can be argued that affective relevance underlies other relevance manifestations, particularly situational relevance*" (p. 1931). This view and related theories indicate that the IX of the user influences the information need and, accordingly, the applied relevance criteria (Schamber, 1994; Cosijn and Ingwersen, 2000; Wilson, 2006; Nahl, 2005). The possible broad influence of affect on the relevance judgments is further explored in Section 2.8, showing that the IX can delineate the affective dimension and, accordingly, explain part of the dynamics in relevance assessments.

2.2.3 Objective Relevance Techniques

Relevance is a concept underlying a wide range of information systems. Yet, often the concept of relevance is not specified as such, but it is there through an "*invisible hand*" (Saracevic, 2007, p. 1916). Exceptions are IR and IF&R systems, which are both explicitly aimed at the delivery of relevant information to their users. The role of relevance in IR and IF&R systems will be discussed with respect to the following techniques: algorithmic relevance, relevance feedback, cognitive filtering, and collaborative filtering. Table 2.1 summarizes the role of relevance for each of the techniques.

Algorithmic relevance, as implemented by IR systems, compares the topics expressed in a query to the topics expressed in documents. The ground truth to evaluate algorithmic relevance on is generated by information experts who assess whether a document is relevant to a query or not (i.e., the Cranfield paradigm; Cleverdon et al., 1966; Voorhees, 2002). This form of relevance is a system-based, binary, singular, and static relation between a document and a query. That binary, singular, and static criteria are not perfect indicators of relevance is illustrated by a constant effort to find better metrics (Demartini and Mizzaro, 2006). For example, Borlund and Ingwersen (1998) developed the relative-relevance measure that indicates the congruence between different measures of (subjective and objective) relevance, supporting a varied set of relevance manifestations. And, Järvelin and

Kekäläinen (2002) developed the normalized discounted cumulative gain measure that takes into account a graded scale of relevance and a decreasing probability users will review lower-ranked search results on a search engine results page. These examples show that numerous attempts have been made to incorporate a non-binary and multi-dimensional model of relevance in (evaluating) algorithmic relevance.

Relevance feedback allows the extension of algorithmic relevance and make it user-based and dynamic. Proxies of relevance are used for implicit relevance feedback, serving as direct input that (de)emphasizes certain terms in a future query (Salton and Buckley, 1990). This type of feedback allows a deviation from a static relation between a query and a search result to a dynamic, personalized adjustment of this relation. For example, by using previous behavioral data such as click-through data, browsing features, and query-text features, Agichtein et al. (2006) could improve relevance performance by up to 31%. This shows what can be achieved by directly incorporating behavioral data in giving real-time, implicit support to the prediction of relevance, a method suggested in Section 2.8.1 as well. This also shows the importance of user-based adjustments of relevance predictions. However, although relevance feedback is user-based, it is often restricted to a singular and binary notion of relevance. An exception is explicit relevance feedback where users are asked to indicate the relevance of a search result. This type of relevance feedback generally includes a non-binary and multi-dimensional measure of relevance (Maglaughlin and Sonnenwald, 2002), but is considered a burden on the user.

Lacking the notion of a query or an active information need, IF&R systems predict the relevance of an information object to a passive information (or emotional) need of a user: "*Recommender systems recommend items based on the likelihood that they will meet a specific user's taste or interest*" (Herlocker et al., 2004, p. 23). The classic approach to evaluating the performance of IF&R systems is by their accuracy to predict (withheld) ratings of a user for an information object. Hence, user-based and non-binary relevance assessments are common in IF&R evaluation. Usually the implemented type of relevance is static - the user models are created incrementally (i.e., the cold-start problem) but are not (or barely) refined during an information interaction session. An exception is when a decay function is used, decrementing irrelevant parts of the user model and increasing the impact of recent interactions on the user model. Which types of relevance are modeled by IF&R systems is dependent on the filtering technique: cognitive filtering or collaborative filtering.

Cognitive filtering is one method to predict relevance without an active information need. Either information is filtered in that matches an area of interest (i.e., content-based) or that reflects a set of user properties (i.e., properties-based) (Hanani et al., 2001). User features typically modeled are individual traits, context of work, tasks and goals, background, knowledge, and interests. Interests are the most commonly modeled attribute by IF&R systems (Brusilovsky and Millán, 2007). It is clear that these techniques relate to types of relevance beyond topicality. For example, content-based filtering captures the (subjective) aboutness, whereas property-based methods can reflect parts of pertinence such as cognitive correspondence and novelty (Hanani et al., 2001). Collaborative filtering infers the preferences of a user based on his/her similarity to or connections with other users (Furner, 2002). In essence, algorithms underlying collaborative filtering systems try to predict if a user will view an item based on the viewing behavior of similar users for similar content. This is the most-used method of filtering partly because it connects directly to the tastes and interests of a user, without the need for extensive (cognitive) modeling (Herlocker et al., 2004).

Overall, IR and IF&R systems struggle with a dynamic, multi-dimensional, non-binary, and user-centered implementation of relevance. The current techniques circumvent the problem of incorporating a complicated model of relevance and instead implement a weak form of relevance, possibly complemented by user models or relevance feedback. This approach led to a great success for IR and IF&R systems. However, information systems can further address the dynamic and multi-dimensional nature of relevance (Schamber et al., 1990; Borlund, 2003; Saracevic, 2007). As we will illustrate next, the experience of the user has value to both aspects.

2.3 Information eXperience

The IX follows from the application of the concept of User eXperience (UX) to information interaction. It will be defined as follows:

Definition 1: The IX constitutes the values, affective responses, and experiential states during interaction with information.

The IX focuses on the information, which includes the information object (e.g., document) that contains the information. The following sections describe the concept of UX (Section 2.3.1) and operationalizes the concept of IX by presenting an IXF (Section 2.3.2).

2.3.1 User eXperience

The notion of an experience has an intuitive understanding. Yet, transforming this understanding into a definition and model of the UX proves to be a challenge. Hassenzahl (2013) describes the UX as a complex fabric of feelings, thoughts, and actions. The use of the three attributes of thoughts, feelings, and actions to describe a UX during information interaction is not new. Kuhlthau (2004) already outlined the process of a complex information seeking task using these three categories. She describes that as users proceed with an information seeking task, their thoughts become more focused, feelings more certain and confident, and their activity shifts from explorative to exhaustive information behavior. The use of thoughts, feelings, and actions in the description of an experience is reminiscent to Plato's tripartite structure of the soul: cognition, affect, and motivation (i.e., action tendencies). Although this structure gives a poor representation of the mind in general (Scherer, 1995), it shows that besides feelings other factors also are involved in an experience.

The number of factors coming together in an IX shows the need for models of the IX that structure its elements. One of the leading attempts to operationalize the concept of UX is by Hassenzahl and Tractinsky (2006), who view the UX as the non-instrumental aspects of technology use. The UX, then, is viewed from three (partly overlapping) perspectives (See Figure 2.2):

(a) Beyond the instrumental, describing values derived from a general human need for being stimulated, to perfect one's skills and knowledge, and to grow (Csikszentmihalyi, 1991);

(b) Emotions and affect, addressing the antecedents and consequences of, ideally, positive affective states;

(c) Experiential perspective, describing the holistic experience emerging from all aspects of the interaction, including the values and emotions, situated in a user and a moment in time; that is, the present experience.

Other models of the UX exist as well. For example, McCarthy and Wright (2004) differentiate between four threads of experience. Namely, the sensual, the emotional, the spatio-temporal, and the compositional thread. The sensual thread is concerned with the sensual and bodily engagement with a situation. The emotional thread constitutes the emotions arising during an experience. The spatio-temporal thread contains the perception of space and time. And, finally, the compositional thread views the experience as a narrative relating the parts of an experience to a holistic experience that has a beginning and an end (Mc-

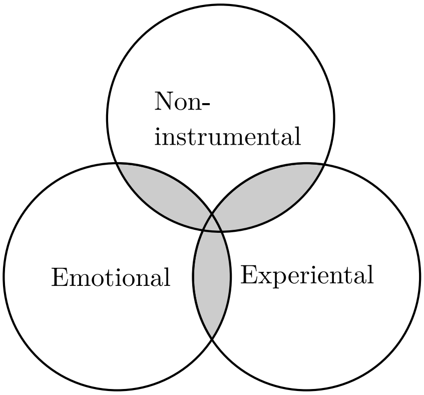


Figure 2.2: The three perspectives on the User eXperience (UX) beyond instrumental value, adapted from Hassenzahl and Tractinsky (2006, p. 95).

Carthy and Wright, 2004). This thread can be seen as an overarching dimension, embedding all the other threads in the structure of a narrative (O'Brien and Toms, 2008). The threads identified by McCarthy and Wright (2004) overlap considerably with the perspectives indicated by Hassenzahl and Tractinsky (2006). The main difference is between the non-instrumental perspective and the sensual thread, where the former delineates that the sensual experiences connect to a general (non-instrumental) human need. Furthermore, the experiential perspective subsumes more than the spatio-temporal thread. Although the foregoing models by McCarthy and Wright (2004) and Hassenzahl and Tractinsky (2006) indicate several aspects that possibly constitute an UX, it does not yet transform the UX into a concrete target for (information) systems. Namely, a consensus seems to be lacking on which aspects constitute the UX and the definition of those aspects.

The models of UX and, in particular, the perspective of a narrative, create a distinction between the momentary experience and the retrospective evaluation of that experience. The momentary experience is generally viewed as a holistic combination of the cognitive, affective, and motivational aspects that emerge during interaction; it is situated in a user, and exists only in the moment (Hassenzahl and Tractinsky, 2006; McCarthy and Wright, 2004). A plethora of descriptors for this momentary experience have been suggested. For example, Csikszentmihalyi and LeFevre (1989) listed affect, potency, concentration, creativity, motivation, satisfaction, and relaxation to describe momentary experiences. The number of available descriptors for an experience illustrates a difficulty in actually describing the momentary, holistic experience.

The retrospective evaluation follows from a set of values, responses, and states. For example, a mood is a result of short-term emotions (see Section 2.5). Although this suggests that an experience follows from an accumulation of emotions, this experience is not simply a sum of its parts. This holds more generally for all aspects of an experience. A retrospective evaluation of an experience is governed by at least three effects: peaks, slopes, and endings. The peak-effect indicates that particularly significant moments contribute very heavily to the retrospective evaluation (Fredrickson and Kahneman, 1993). Similarly, the slope-effect shows the importance of a gradual increase in intensity of positive outcomes (Loewenstein and Prelec, 1993). Finally, the ending-effect of experiences shows that a summary assessment of an experience is heavily influenced by experience near the end of an episode (Kahneman et al., 1993). These effects clearly show that the relation between the elements of an experience and the experience as seen retrospectively is not trivial.

Current studies on the UX already indicate which aspects combine to form a UX and how this UX is viewed either momentarily or retrospectively. However, partly due to the lack of a consensus, the current conceptualizations are still somewhat difficult to operationalize. Nonetheless, the current models of the UX already give some guidance to the challenge of defining and modeling the IX. In particular the UX model by Hassenzahl and Tractinsky (2006) will be used as basis for the IXF.

2.3.2 Information eXperience Framework

To define the IX, an abstraction from the different UX models is made. This abstraction is an extension of the conceptualization by Hassenzahl and Tractinsky (2006) and defines the following perspectives that describe an IX: instrumental value, non-instrumental value, affective responses, and experiential states. Essentially, this abstraction is in line with the aspects covered by the cognitive-appraisal theories of emotion (Ellsworth and Scherer, 2003). According to these theories, a stimulus is appraised, these appraisals give rise to emotional responses, and these responses in turn influence an individual's state[2] (see Section 2.5.1). Furthermore, McCarthy and Wright (2004) are followed by portraying an IX as a narrative; that is, as a story consisting of momentary values, responses, and states with a beginning and an end, which can be summarized retrospectively. This abstraction

[2] For comparison, one can intepret these perspectives as the derivatives of an experience: states (first derivative), responses (second derivative), and appraisals (third derivative). Similarly, the integral can be taken from higher derivatives all the way back to an experience.

forms the basis for the IXF.

Figure 2.3 gives an illustration of the four perspectives of the IX during an information interaction session. The figure depicts how several information objects (o) have a certain instrumental (i) and non-instrumental (n) value to the user that leads to (affective) responses (r) and changes in the experiential state (x). In Section 2.4.1 we will show that, regarding information interaction, the notion of (situational) relevance is a proxy for the value ("utility") of information, allowing the value to be denoted as relevance in Figure 2.3. The outer layer in Figure 2.3 consists of information objects ($o_{1...4}$) that at different moments during an information interaction session serve as stimuli for relevances[3], responses, and states. Considering that the information objects are supplied by an information system, the outer layer can be seen as algorithmic relevance (see Section 2.2.3).

Figure 2.3 shows a series of possible relations between information and the different perspectives (circles) on the IX, where significant changes in any of the perspectives are denoted by a dot. Furthermore, Figure 2.3 shows how an IX is temporal by having a beginning and an end and is situated by continuing on previous states. Connected to Figure 2.3 is Table 2.2, which shows the relations that exist between the different perspectives. Each of the relations denoted in Table 2.2 will be described in the next sections. Essentially, Figure 2.3 forms a summary of the IXF by illustrating the perspectives, their relations, and the narrative structure. Accordingly, the figure shows the potential of the IXF to transform the IX into an actionable concern for information systems.

Figure 2.3 illustrates that the analysis of the IX can be performed at multiple levels of granularity. Namely, per information interaction session and per manifestation of information. The latter consists of multiple levels of granularity ranging from an information object to smaller entities such as paragraphs or sentences. As an information interaction session consists of multiple manifestations of information, multiple significant changes in the values, responses, and states may arise during an information interaction session. This shows an important difference between the approach taken by Hassenzahl and Tractinsky (2006) and the framework proposed here. The three perspectives as identified by Hassenzahl and Tractinsky (2006) are generally evaluated for a singular product having a set of values and evoking particular emotions and experiences dependent on the user and context. In contrast, information systems that deal with the selection and retrieval of information generally serve a multitude of "products". Each manifestation of information can be seen as a "product" having its own set of values and evoking particular emotions and experiences. Accordingly, in Figure 2.3 the starting point

[3] The plural of relevance, relevances, will be used for a variety of descriptors of the relevance relation (see Section 2.2.2).

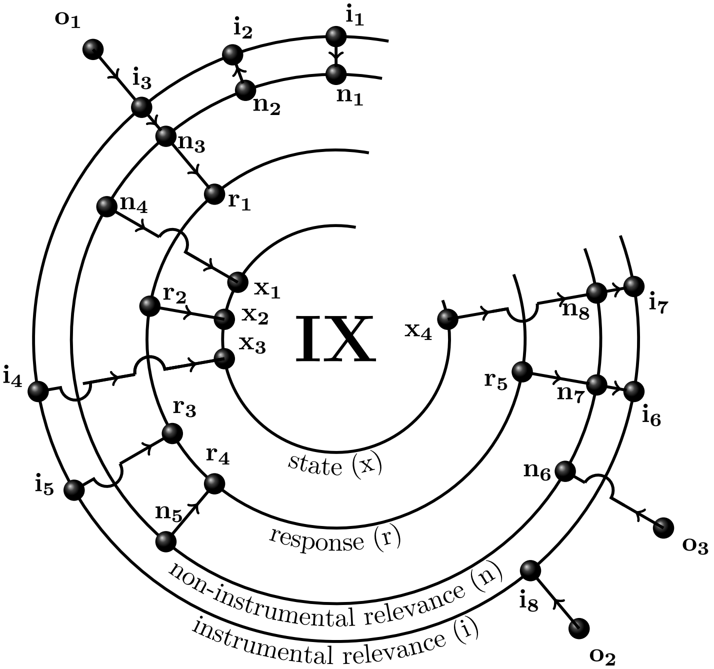


Figure 2.3: The Experience Wheel: Four perspectives and their interconnections which describe the narrative of an Information eXperience (IX).

of analysis is a manifestation of information.

Inherent to the multitude of information manifestations is that an experience evolves over time, as suggested by viewing an experience as a narrative (McCarthy and Wright, 2004). This implies that all facets of the IX are continuous. Figure 2.3 only depicts the significant changes in any of the facets of the IX as a description of this continuum. For example, relevance judgments fuel the decision to start, continue, or stop processing an information object. This idea is in line with relevance theory (Sperber and Wilson, 1996) which states that of all stimuli available to a user at a moment in time, the one with the highest ratio between cognitive effect and processing effort is processed. Hence, relevances are judged continuously to decide on which stimulus to attend to. Similarly, Wang and Soergel (1998, p. 117) stated that "*for real users, the meaningful task is to make a decision, not merely a relevance judgment*", acknowledging the importance of relevance judgments comes from having a significant impact on the decision start, continue, or stop processing an information manifestation: that is, causing a significant change. The whole pattern of significant changes describes the IX of an information interaction session. The retrospective evaluation derives from this pattern as well, with a primacy for the slope, peaks, and endings (see Section 2.3.1).

The notion of a narrative and the accompanying notions of a beginning and an end signify the importance of the context in which an IX is embedded. Partly, this context is covered by the starting point of each thread. Each starting point influences the subsequent IX: values are dependent upon existing needs, affective responses on the initial affective state, and the experiential state evolves from the initial state (see Table 2.2). In particular, an IX starts with a specific motivation to search; whether it is an extrinsic motivation such as an information seeking task or an intrinsic motivation such as curiosity. Often, this motivation is dependent on the broader context in which the interaction takes place, such as a geographical location or the surroundings of a user (e.g., in an office or a waiting room). Although the broader context is clearly of influence on the IX, it is not part of the IXF. Nonetheless, as Figure 2.3 suggests, the starting state of the user is regarded as part the IXF and, indirectly, supplies a window on this broader context as well.

The preceding overview of the IXF indicates how the different models and aspects of the UX were transformed into a coherent model of the IX. The resulting IXF supports the narrative structure of the IX and allows us to explain retrospective evaluations of an IX. Moreover, relevance was included as a perspective in the IXF to assure applicability for information systems. The next sections will elaborate on each of the perspectives and their role in forming an IX. Furthermore, the constructs that are subsumed by each perspective and their relations with other perspectives are shown in Table 2.2.

2.4 Relevances and Values

Values form an important aspect of the IX. Within the context of information interaction, this value can be approximated by relevance. More precisely, two partly overlapping types of relevance will be discerned: instrumental (Section 2.4.1) and non-instrumental (Section 2.4.2) relevance.

2.4.1 Instrumental Relevance

The instrumental value of a system describes its ability to fulfill its primary task: that is, to be useful and usable. It is the starting point for any system: without instrumental value there is no need to look at the UX (Hassenzahl and Tractinsky,

Table 2.2: The constructs of and relations between each of the perspectives on the Information eXperience (IX).

		Constructs		
Connection	Section	Antecedents	Consequents	References
$i_1 - n_1$	2.4.3	Topicality	Non-instrumental relevances	Spink and Greisdorf (2001)
		Functional and conditional value	Epistemic value	Wang and Soergel (1998)
$n_2 - i_2$	2.4.3	Non-instrumental aspects	Perceived usefulness, ease of use, intention to use	Moon and Kim (2001); Van der Heijden (2003); Yi and Hwang (2003); Liaw and Huang (2006)
		Aesthetic value	Usability	Tractinsky et al. (2000)
$o_1 - i_3 - n_3 - r_1$	2.5.4	Intrinsic pleasantness and novelty, coping potential, and motivational appraisals	Emotional response	Ellsworth and Scherer (2003)
		Control-value	Information emotions	Pekrun (2006); Pekrun et al. (2010); Silvia (2008b)
		Linguistic, conceptual, and visual complexity	Processing fluency	Alter and Oppenheimer (2009); Song and Schwarz (2008)
$n_4 - x_1$	2.7	interactivity, vividness, and attractiveness	Flow	Finneran and Zhang (2005); Hoffman and Novak (2009)
		Stimulation / Challenge	Flow	Csikszentmihalyi (1991); Higgins (2006)
		Aesthetic value	Engagement	O’Brien and Toms (2008)
		Hedonic value, complexity	Engagement	Higgins (2006)

$r_2 - x_2$	2.7	Positive affect	Satisfaction	Oliver (1993)
		Interest	Engagement (start of)	O'Brien and Toms (2008); Reeve (1989)
		Positive affect	Engagement (continuation of)	Reeve (1989); Higgins (2006)
		Negative affect, cognitive fluency	Slow thinking	Schwarz and Clore (2007); Bless et al. (1996)
		Emotions of uncertainty	Slow thinking	Tiedens and Linton (2001)
$i_4 - x_3$	2.7	Objective Relevance	User satisfaction	Gluck (1996); Huffman and Hochster (2007)
		Goal attainment	Engagement / Flow	Higgins (2006); O'Brien and Toms (2008); Chen (2007)
		Coherence	Slow thinking	McNamara et al. (1996)
		Goal obstruction, goal relevance*	Slow thinking	Schwarz and Clore (2007)
$i_5 - r_3$	2.7.2	Objective Relevance*	Negative emotions	Arapakis et al. (2008)
		Goal relevance, goal conduciveness	Positive emotions	Kreibig et al. (2010, 2012)
$n_5 - r_4$	2.7.2	Hedonic values	Positive affect	Jordan (1998); Hassenzahl et al. (2010)
		Hedonic and aesthetic values	Pleasure and arousal	Fiore et al. (2005); Mummalaneni (2005)
		Complexity	Interest	Silvia (2005)
$o_2 - i_8$	2.7.3	Manifestation of information	Utility	Saracevic (2007); Borlund (2003)
		Document	Functional and conditional value	Wang and Soergel (1998)

$o_3 - n_6$	2.7.3	Document	Epistemic, emotional, and social value	Wang and Soergel (1998)
		Simplicity	Aesthetic value	Karvonen (2000); Michailidou et al. (2008)
		Interactivity	Enjoyability	Blythe et al. (2004)
		Seduciveness and vividness (of text)	Attractiveness	Schraw and Lehman (2001)
$r_5 - n_7 - i_6$	2.8.1	Affect-as-information (valence)	(Relevance) judgments	Schwarz and Clore (2007)
		Affect-as-spotlight (valence)	Value and control*	Alhakami and Slovic (1994); Peters (2006)
		Cognitive fluency	Truth, liking, certainty, familiarity	Alter and Oppenheimer (2009)
$x_4 - n_8 - i_7$	2.8.2	Engagement	Value	Higgins (2006); Csikszentmihalyi (1991)
		Positive mood	Value and control*	Isen et al. (1978)
		“Deep diving” information search pattern	Depth/scope, quality, source reputation (value*)	Heinstrom (2002)

Note. * Negative relation.

2006). The instrumental value of an information system is in providing information that has utility to a situation, task, or problem. This can be derived from the definition of situational relevance by Borlund (2003, see Section 2.2), which states that this highest type (level) of relevance describes the utility of a manifestation of information to the situation, task, or problem at hand. The connection between situational relevance and instrumental value allows us to directly connect models of relevance into the IXF. It allows the exploration of the connection between relevance and the IX. Given that situational relevance has been noted as the most realistic type of user relevance (Harter, 1992; Saracevic, 1996; Schamber et al., 1990; Borlund, 2003) this a salient connection.

Wang and Soergel (1998) provided evidence for the relation between relevance and instrumental value by showing the existence of two instrumental values of documents: functional value and conditional value. The first pertains to the utility of a manifestation of information to the situation, task, or problem at hand. The second closely relates to the first as it pertains to potential functional value which can become actual functional value when the information need changes (i.e., dynamic relevance; see Section 2.2). Users employ a specific set of relevance criteria to infer instrumental and functional values, as compared to the relevance criteria used to infer non-instrumental values (Wang and Soergel, 1998). Generally, the relevance criteria that users apply when making a relevance decision represent the values that a user wants to attain.

Instrumental value is a prerequisite for and has a positive influence on non-instrumental values (Figure 2.3, $i_1 - n_1$; Xu, 2007; Spink and Greisdorf, 2001). At least two findings confirm this. First, topicality is a pre-condition before other relevance criteria are evaluated (Spink and Greisdorf, 2001), suggesting that general needs are secondary to an information need. This is illustrated by qualitative descriptions such as "*very informative, wrong topic*" (Spink and Greisdorf, 2001, p. 169). Second, documents that have instrumental relevance often, if not always, have epistemic value as well (Wang and Soergel, 1998). This epistemic value contributes to personal growth through an increase of knowledge and skills, clearly a non-instrumental value (Hassenzahl, 2003).

2.4.2 Non-Instrumental Relevance

In addition to the instrumental value of information, non-instrumental values can be identified as well. Instead of connecting to an information need, these non-instrumental values connect to general human needs: that is, the well-being of the user. A plethora of human needs can be identified (e.g., Maslow's hierarchy of needs; Maslow, 1943). Here, we identify two values that are well-supported in the context of information systems: aesthetics and hedonics. Aesthetic factors concern the beauty or visual appeal of interactive systems and are detected almost instantly (Lindgaard et al., 2006). Hedonic factors are functions and attributes that have a strong potential for pleasure. These hedonic factors can be subdivided into aspects about stimulation (e.g., personal growth, an increase of knowledge and skills), identification (e.g., self-expression, social interactions), and evocation (e.g., self-maintenance, memories) (Hassenzahl, 2003).

Wang and Soergel (1998) confirm the existence of three non-instrumental values next to two instrumental values of documents: epistemic, emotional, and social. These values roughly correspond to: stimulation, evocation, and identification, respectively[4]. In comparison to the relevance criteria applied to select instrumental values, non-instrumental values are reflected by a specific set of relevance criteria that has overlapping as well as distinct elements (Wang and Soergel, 1998). Considering the overlap only occurs for epistemic values, which per definition overlaps with instrumental values within the context of information-intensive tasks, the non-instrumental values can be seen as relevances not related to an information need but to general human needs. These values contribute, just as instrumental values do, to the decision to use a document.

Besides their importance to relevance decisions, non-instrumental values influence the instrumental value as well. The influence of non-instrumental values on instrumental aspects (Figure 2.3, $n_2 - i_2$) is primarily approached using the technology acceptance model[5]. The studies using this model define non-instrumental aspects of technology use that relate to intrinsic motivations. And, these are contrasted to instrumental aspects of technology use that relate to extrinsic motivations, in particular perceived usefulness and ease of use. These studies confirm that non-instrumental values influence and contribute to instrumental values and heighten the explanatory capacity of the technology acceptance model for the intention to use (Moon and Kim, 2001; Yi and Hwang, 2003; Liaw and Huang, 2006) and actually use (Van der Heijden, 2003) of web-based information systems. In particular aesthetic values have been shown to influence perceived usability both before and after use, in a way that what is beautiful is also perceived as usable (Tractinsky et al., 2000; Lindgaard and Dudek, 2003). For example,

[4] Epistemic values relate to a desire for knowledge and can be seen as aspects of stimulation. Emotional values describe the capacity of an information object to arouse feelings or affective states. Although not the same as evocation, the results described in Wang and Soergel (1998) do refer to documents that evoke memories and associations that, in turn, evoke an affective response. And, social values relate an information object to specific social groups and can be interpreted as aspects of identification.

[5] The Technology Acceptance Model (TAM) is an information systems theory that models how users come to accept and use a technology.

beauty has been identified as an important predictor of overall preference for web sites, supporting the influence of non-instrumental values on a relevance judgment (Schenkman and Jönsson, 2000). These findings confirm that the notions of instrumental and non-instrumental value are interweaved and, accordingly, that non-instrumental relevance forms an important aspect of (instrumental) relevance.

Non-instrumental relevances can become the instrumental value as well, blurring the distinction between the values and confirming the close connection between instrumental and non-instrumental values. This is the case during a hedonic or epistemic search in which users search information for entertainment or for its epistemic value instead of for solving an information problem (Xu, 2007). The intricate relation between instrumental and non-instrumental relevance indicates the importance of utilizing the former to fully benefit from the "information opportunity". That is, the ever increasing availability of information allows the inclusion of non-instrumental criteria which, in turn, contribute to instrumental criteria as well.

Whereas non-instrumental values are generally not included in relevance models, they are an intrinsic part of the IX and, more generally, the UX. This shows a discrepancy between relevance models and the IX. Notwithstanding, numerous relevance criteria have been identified that confirm the existence of relevances not related to an information need. For example, Barry and Schamber (1998) identified affectiveness (i.e., stimulation) and relationship with author (i.e., evocation). And Xu (2007) confirmed the role of novelty and understandability in creating affective relevance (i.e., stimulation). Hence, an argument can be made that relevance with respect to non-instrumental values and needs should be noted in relevance models. This notion of non-instrumental relevance will be included in the IXF.

2.4.3 Value Perspective

As the preceding outline of instrumental and non-instrumental relevance has shown, the value of information can be approximated by its relevance either to an information need or to general needs. This connects the IXF directly to contemporary models of relevance, showing the importance of relevance for the IX. Moreover, the foregoing showed that non-instrumental value is interweaved with instrumental value, suggesting their reciprocal dependency.

Essential to the notion of value is that a stimulus is valued: that is, a stimulus is appraised according to numerous criteria. How these appraisals combine into denoting a value is subject to debate. Numerous models exist for the relevance or value of information, including: Wang and Soergel (1998), who connects document characteristics via relevance criteria to values; Saracevic (2007) who relates relevance criteria via numerous stratified types of relevance to, eventually, utility, and; Sperber and Wilson (1996) who connect processing effort and cognitive effect to relevance decisions in their relevance theory. No explicit requirement is put on the model. Instead, the underlying criteria and their appraisals are of main concern, both in terms of applicability and for their importance to the other perspectives of the IXF. As we will argue in Section 2.6, the relative importance of these appraisals and, accordingly, the way in which these appraisals combine is dependent on the goal of the experience.

By using relevance to denote value, this thesis applies a rather liberal definition of relevance. This rather liberal definition is in line with its use in information science, where it encompasses the judgments a user makes about a piece of information on numerous criteria (i.e., relevance criteria). In comparison, a more strict definition is that relevance only describes the significance of a stimulus to an individual (Scherer, 2004) (cf. the topicality of information). Following this strict definition, the actual value or utility of the stimulus is denoted separately under the notion of implications. Similarly, control is also treated separately from strict relevance (Scherer, 2004). Notwithstanding, this thesis applies the broad definition of relevance as proxy for value, which includes both relevance criteria for implications (e.g., utility; Borlund, 2003; Saracevic, 2007) and control (e.g., understandability; Xu and Chen, 2006).

In sum, the value of information with respect to an information need and to general needs is reflected in the following definition:

Definition 2: Next to (instrumental) relevance in relation to an information need, the value of information can be approximated by non-instrumental relevances that exist in relation to general human needs.

2.5 Feelings and Responses

A plethora of feelings exists, not all easily definable. During information interaction numerous emotions have been identified. Amongst others: happiness, surprise, anger, interest, confusion, familiarity, and frustration (Kort et al., 2001; Kuhlthau, 2004; Arapakis et al., 2008). Some of these feelings are rather vague and, overall, the role of these emotions during information interaction is unclear. This section aims to structure the “forest” of (information related) feelings.

Figure 2.4 shows the “feeling tree”, which gives a broad overview of the plethora of feelings that exists. Following Schwarz and Clore (2007), they are divided into three categories: (meta)cognitive, affective, and bodily. Feelings like hunger and pain are bodily feelings, happiness and fear are affective feelings, and feelings such

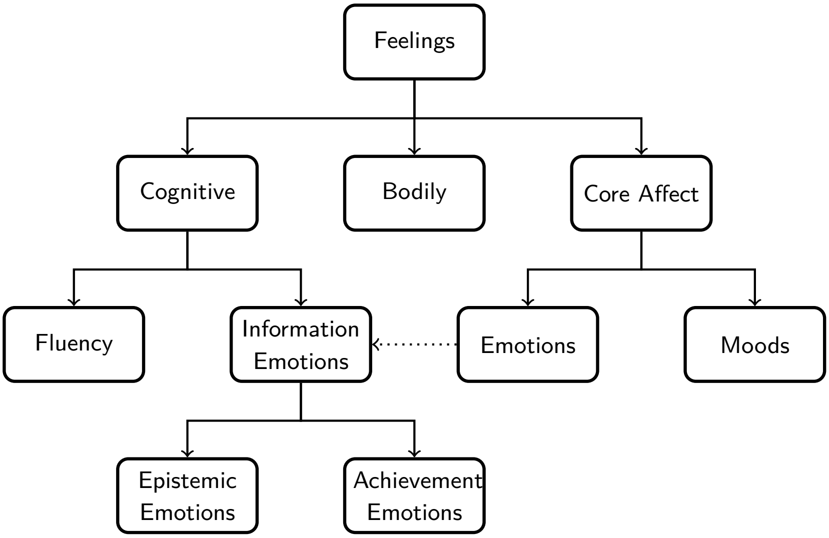


Figure 2.4: Feeling Tree of affective responses salient to information interaction.

as familiarity and confusion are related to knowledge and, therefore, cognitive feelings. For cognitive feelings an extension will be made upon the categorization by Schwarz and Clore (2007) through a further differentiation between cognitive fluency (Schwarz, 2010) and information emotions (Silvia, 2008b, 2009; Pekrun and Linnenbrink-Garcia, 2012). Given that all these feelings have an affective component (i.e., create a subjective experience), they are referred to as affective responses in the IXF. Because of their relevance to information interaction the categories of affective and (meta)cognitive feelings will be described.

2.5.1 Emotions and Moods

Both emotions and moods can be described along two dimensions: that is, by the dimensions of valence (i.e., pleasantness) and arousal (i.e., intensity) (Lang, 1995; Russell, 2003). Emotions are assumed to have an identifiable referent (Schwarz, 2000) and a short time-span (Watson, 2000). Mood, a longer term concept, will result from shorter term emotional reactions (Watson, 2000). Thus, an emotion is about something whereas a mood usually lacks a referent. Both do reflect an appraisal of a situation in its broadest sense. Given that an emotion describes the feelings towards a stimulus, emotions have a higher information value than described by the dimensions of valence and arousal alone. The intricate relation between the (cognitive) appraisal of a stimulus and the resulting emotion favors a description of discrete emotions more reflective of the underlying appraisals.

Contemporary accounts of emotion view emotion as a process, having a set of input and output systems. A commonly used subdivision of the output is into three response systems: influencing behavior, affective experience, and changing the physiological state. The input of the emotional process includes percepts and physiology (Cannon, 1927; Niedenthal et al., 2007). The cognitive-appraisal theories on emotion further define the pattern of the emotional process. Cognitive-appraisal theories link emotion to cognitive processes of causal attribution, evaluation of meaning, and assessment of coping capabilities. A stimulus (i.e., input) is appraised according to two consecutive appraisals (Ellsworth and Scherer, 2003; Scherer, 2004):

1. A primary appraisal that evaluates a situation's significance regarding *1)* its intrinsic pleasantness and novelty, and, *2)* the goals and needs of a person.
2. A secondary appraisal that assesses our ability to deal with the situation: that is, the coping potential.

Often, another group of appraisals is included as well that denotes the significance of an event with respect to one's self-concept and social norms and values (Ellsworth and Scherer, 2003; Scherer, 2004). Due to its probably limited weight this group will not be discussed further.

The appraisals of the significance of a situation to an individual's well-being take center-stage in several appraisal theories of emotion and are, therefore, sometimes referred to as "primary appraisals" (Ellsworth and Scherer, 2003). These primary appraisals describe the significance and implications of an event to an individual. Whereas at a first level this is a highly automated evaluation of stimulus characteristics, at a second level this involves the motivational bases of the individual (Ellsworth and Scherer, 2003). At the first level, the intrinsic pleasantness and novelty are appraised. This includes an evaluation of stimulus intensity, of the predictability or familiarity of a stimulus, and of the valence or intrinsic pleasantness of the stimulus. At a second level the implications of the stimuli with respect to the motivational bases of a user are not as highly automated as the intrinsic pleasantness and novelty, although the specificities of their computation is unknown (Ellsworth and Scherer, 2003). Given the significance of a situation, the secondary appraisals assess the ability of a person to change the situation and its consequences determines the response (e.g., fight or flight) (Ellsworth and Scherer, 2003). This appraisal is proactive, going beyond the current stimulus and evaluating possible outcomes.

The appraisals of the implications of a stimulus with respect to the motivational bases of a user are likely to be particularly relevant to information interaction, given that most models of relevance assume an underlying information need (see Section 2.2.2). At least two (interacting; Kreibig et al., 2012) appraisals have been identified that describe the motivational bases: goal relevance and goal conduciveness. Goal relevance denotes the significance of the goal or need and the relevance of a stimulus to that goal. Goal conduciveness describes the value of the stimulus to the goal, either furthering or obstructing it. Valence or intrinsic pleasantness already reflects part of the motivational bases. Although the intrinsic pleasantness of a stimulus is often congruent with goals and needs, this is not a necessity. An intrinsically pleasant stimulus might distract from and even obstruct achieving a significant goal (Ellsworth and Scherer, 2003).

The output or result of the emotional process depicted by the cognitive-appraisal theories cannot be defined by the input alone, the cognitive appraisal of the input specifies the emotion experienced. Cognitive processes take a central role in that they specify the (significance of) the experienced emotion. Hence, the experienced emotions (i.e., the outputs) in relation to a stimulus (i.e., the inputs) form a window on the underlying cognitive appraisals.

2.5.2 Information Emotions

The emotions that arise during information interaction are sometimes referred to as cognitive-affective states (D'Mello and Graesser, 2012), signifying the importance of cognition to this subset of emotions. Because of this cognitive reliance, these emotions do not reduce easily to basic structures. This increases the difficulty of structuring the "forest of feelings" for information emotions in particular. Nonetheless, an extension is proposed upon the categorization by Schwarz and Clore (2007) through the inclusion of information emotions as a subset of cognitive feelings. Similar to regular emotions, the information emotions can be described and related to the appraisal structure of emotions. To illustrate this, Figure 2.4 shows a dotted line that subsumes information emotions amongst emotions. Two categories of information emotions (also referred to as academic emotions) are deemed relevant to information interaction: achievement emotions (e.g., anxiety, frustration) and epistemic emotions (e.g., surprise, confusion) (Pekrun and Linnenbrink-Garcia, 2012; Schwarz, 2010). More categories have been identified as well, in particular topic emotions (e.g., empathy for a protagonist in a novel) and social emotions (e.g., pride, shame). Although they can profoundly influence an information interaction episode, they are not of primary interest at this moment (cf. Xu, 2007).

Achievement emotions relate to activities and their outcomes and can be explained by the control-value theory (Pekrun, 2006). The control-value theory brings the cognitive appraisal theories within the context of information needs and information problem solving. In the control-value theory, the value pertains to both intrinsic values, where one appreciates the activity itself (i.e., non-instrumental value), and extrinsic values, representing the instrumental value of the activity with respect to its outcomes (Pekrun, 2006). This distinction between intrinsic and extrinsic value reflects the division between the primary appraisals of intrinsic pleasantness and novelty and the appraisal of motivational bases. The control pertains to: the action, having confidence in initiating and performing the activity (i.e., self-efficacy expectations); the outcome, the belief that the action leads to a positive outcome, and; the situation, the extent to which the situation supports the outcome (Pekrun, 2006). This definition of control reflects the ability to achieve the outcome value and is similar to the secondary appraisal that assesses our ability to cope with a situation. The quadrant of high versus low control and value predicts the experience of a range of achievement emotions such as frustration, boredom, and enjoyment (Pekrun, 2006; Pekrun et al., 2010).

Epistemic emotions are emotions associated with thinking and comprehending (Silvia, 2009). They arise because of qualities of information and during the processing of such information (Pekrun and Linnenbrink-Garcia, 2012). The emotions motivate learning, thinking, exploring, and can lead to a general growth of knowledge (Silvia, 2009). Examples are surprise, confusion, and interest. Surprise is an emotion that is mainly due to a primary appraisal of novelty. Interest is an emotion caused by a primary novelty-complexity appraisal and a secondary dimension of comprehensibility or coping potential (Silvia, 2008b). And, confusion, although generally not viewed as an emotion, can be interpreted using a similar appraisal structure of novel and complex, yet incomprehensible. The appraisals of these epistemic emotions are metacognitive: they all involve an appraisal of what individuals know, expect, and think to understand.

2.5.3 Metacognitive Feelings and Cognitive Fluency

Metacognitive fluency describes the feelings that arise from monitoring our own cognitive processes. They all pertain to a subjective experience of cognitive ease or strain, although related to different faculties (Alter and Oppenheimer, 2009). Two types will be described: processing fluency, the ease with which information is processed, and; accessibility, the ease with which information is brought to mind (Schwarz, 2010).

Processing fluency can be influenced by many variables, including: perceptual fluency, conceptual fluency, and linguistic fluency. Perceptual fluency arises from aspects such as easy-to-read fonts (Song and Schwarz, 2008), duration of presentation, and amount of previous exposure (Schwarz, 2010). Conceptual fluency describes the consistency between a stimulus and its context. For example, semantically primed words and concepts are more easily processed (Alter and Oppenheimer, 2009). Furthermore, linguistic fluency describes the regularity of words, where more regular words are processed more easily. This type of fluency indicates the processing difficulty at phonological, lexical, orthographic, and syntactic levels (Alter and Oppenheimer, 2009, Chapter 3). Similar to processing fluency, accessibility describes the ease or difficulty of recall and thought generation (Tversky and Kahneman, 1973). Several causes of this are one's expertise, possible situational distractions, and the recency and importance of the memory itself (Schwarz, 2010). Hence, these are feelings of cognitive ease and knowing.

High fluency is generally experienced as pleasant and causes a stimulus to be experienced as familiar, (aesthetically) pleasing, true, effortless, and gives the impression of goal conduciveness (Winkielman et al., 2003; Reber et al., 2004). Contrary to emotions, which have an identifiable referent, the exact cause of processing fluency and accessibility is not consciously known to the individual. Misattribution of the cause leads to biases in the interpretation of these metacognitive feelings that can (wrongly) influence (relevance) judgments and decision making (Schwarz and Clore, 2007).

2.5.4 Affective Perspective

The preceding outline of feelings and affective responses transformed the "forest" of information related feelings into a structured "feeling tree", which allows for its inclusion in the IXF. The corresponding perspective in the IXF looks at the antecedents and consequents of short-lived affective responses, in particular emotions. This focus allows us to identify stimulus-response contingencies[6] (Niedenthal et al., 2007), where the stimulus pertains to a manifestation of information. Moreover, although in the case of cognitive feelings (see Figure 2.4) there is not necessarily an identifiable referent, information emotions do arise while processing a piece of information (Figure 2.3, $o_1 - r_1$). This also implies there can be multiple, even opposing, affective responses. For example, reading might foster positive emotions such as the emotion of interest while the content might not further or may even obstruct the information seeking task and eventually lead to the emotion

[6] Stimulus-response contingencies denote the likely occurence of a response given a stimulus. This pairing is reminiscent to Skinner's three-term contingency, in which a behavior (i.e., a response type) is either punished or reinforced; that is, antecedent-behavior-consequents.

of frustration.

There is a seeming reciprocity between relevances, cognitive fluency, and emotions. A particular manifestation of information gives rise to various subjective relevances. These evaluations are influenced by the cognitive fluency that the manifestation of information induces. And, emotions can partly be explained by these evaluations; given the overlap between relevance (in its broad definition) and the primary and secondary appraisals of emotion. Hence, cognitive fluency can influence the relevance evaluations whereas, in turn, relevances influence the experienced emotions again. The specificities of this reciprocity between (relevance) judgments and affective responses are unclear (Scherer, 2004; Scherer et al., 2006; Schwarz and Clore, 2007). Nonetheless, these specificities do not need to be solved in order to apply these notions in the IXF.

Understanding the appraisal structure of emotions allows information systems or affective systems to influence emotions via their appraisals. And, the experienced emotions in relation to a stimulus form a window on the underlying cognitive appraisals (see Section 2.5). As subjective relevances are an appraisal of the relationship between a manifestation of information (i.e., the stimulus) and a user (e.g., his needs and goals), affective responses follow from relevances. On the one hand, information systems can apply a combination of objective metrics that promote a specific affective response. On the other hand, an emotion describes what the resulting experience of a (lack of a) relevant information for a user is. The intricate relation between cognitive information processing and (information) emotions makes this perspective a particularly useful framework for research in which the antecedents and consequents can be identified and related to relevance. This promotes the following definition:

Definition 3: The affective responses that arise during information interaction are intrinsically related to the relevance(s) of a manifestation of information.

2.6 Prototypical Experiences and States

During information interaction, numerous mental states can arise. Since an overall description of the mental state is infeasible (see Section 2.3.1), it is common to highlight numerous aspects of an experience. Within the context of information interaction these aspects are often described via several principles. Kuhlthau (1993) derived the uncertainty principle that unified her observations on the thoughts, feelings, and actions during an information search, stating that uncertainty causes feelings of anxiety and frustration and vague and unclear thoughts. Continuing on this is the affective load principle, "*uncertainty multiplied by felt time pressure*"

(Nahl, 2004, p. 193), which indicates the reciprocity of positive emotions and available time in the experience of the user (Nahl, 2004). However, even given these principles, the notion of mental state remains elusive.

Instead of the many principles and attributes that exist to describe aspects of a mental state, a simpler, discrete approach will be proposed that classifies these dimensions. Specifically, three prototypical experiences are identified in the following sections: a positive, engaging, and learning experience. These experiences exist in relation to three goals (finding, encountering, and learning) that are supported by three types of information systems (IR, IF&R, and intelligent tutoring systems). Through describing their narratives, three typical momentary states are derived. These typical, discrete states counter the elusiveness of the notion of a mental state. The narratives, states, and importance of each of the prototypical experiences will be illustrated in the next sections (2.6.1 – 2.6.3). Furthermore, the role of their typical momentary states within the IX is discussed in Section 2.6.4.

2.6.1 Positive Experience

Queries submitted to an IR engine can be grouped according to three categories: navigational, transactional, and informational (Broder, 2002). In general, most of these queries solve a more or less precise information need or request for which a precise answer is optimal (Taylor, 1962; Byström and Hansen, 2005; Saracevic, 2007). Retrospectively, to find a specific web page, a product, or general information (Manning et al., 2009). However, the ideal situation of a precise information need or a precise answer is unrealistic. For example, informational queries are sometimes about a broad topic and are, therefore, often not solvable by a single web page (Manning et al., 2009). This suggests that an engaging experience or even learning experience is more suitable for less precise information needs. Generally, a positive experience can be seen as a target experience for common information interaction such as generally with an IR system. A positive experience is considered key to information use intentions and the overall success and adoption of information systems (DeLone and McLean, 2002; Hassenzahl, 2013).

The primary state of a positive experience is satisfaction or "*the sum of one's feelings*" (Gluck, 1996, p. 90). This closely resembles a mood, which reflects a situation and results from an accumulation of responses (Watson, 2000). Although satisfaction is a complex phenomena (Lindgaard and Dudek, 2003), several key determinants have been identified. Namely, it is characterized by positive affect, limited effort, and high effectiveness (Fulton, 2009; Hassenzahl and Tractinsky, 2006; Oliver, 1993; Lindgaard and Dudek, 2003; Al-Maskari and Sanderson, 2010).

Following the causes of satisfaction, a positive experience is a narrative consisting of cognitive ease and both instrumental and non-instrumental relevance. Some indications exist for the importance of a slope effect (Loewenstein and Prelec, 1993, see Section 2.3.1) on the retrospective evaluation of a positive experience. Namely, Gwizdka and Lopatovska (2009) investigated subjective variables during the information search process and found that a positive mood before the search correlated with mood after the search, but also with lower performance and less satisfaction. This suggests an ideal narrative of increasingly high relevance and positive affective responses.

2.6.2 Engaging Experience

As already indicated in the previous section for the positive experience, in some cases an engaging experience is more apt. For example, for hedonic and epistemic information interaction (Xu, 2007) or information encountering such as via IF&R systems. A flow/engaging experience contributes to communication, exploratory behavior, learning, and use (Finneran and Zhang, 2005; Hoffman and Novak, 2009). A state of engagement/flow increases task performance (Schaik and Ling, 2012) and leads to positive emotions (Chen, 2006). Engagement or flow have even been stated as prime targets for information systems. As phrased by O'Brien and Toms (2008): "*Successful technologies are not just usable; they engage users*" (p. 938).

Higgins (2006) defined engagement as: "*to be involved, occupied, and interested in something. Strong engagement is to concentrate on something, to be absorbed or engrossed with it*" (p. 442). Hence, (strong) engagement or flow indicate a highly concentrated mental state with an intrinsic motivation for an activity (Csikszentmihalyi, 1991; O'Brien and Toms, 2008). In particular flow is characterized by concentration, time distortion, a loss of self-consciousness, and a merging of action and awareness (Finneran and Zhang, 2005; Chen, 2007). This is in line with the spatio-temporal thread of experience that describes the present experience of space and time (McCarthy and Wright, 2004) which typically changes during an engagement/flow experience (Csikszentmihalyi, 1991; O'Brien and Toms, 2008).

An engaging experience follows a specific narrative (O'Brien and Toms, 2008). It starts with an aesthetic or otherwise attention grabbing quality and an emotional interest response to the information. An engaging experience continues with positive affect such as enjoyment and an appropriate level of challenge and goal directedness. The end of an engaging experience can occur when an information need is fulfilled or, for example, in the case of boredom (Csikszentmihalyi, 1991; O'Brien and Toms, 2008). Hence, the engaging narrative is clearly composed of and explained by specific relevances (e.g., aesthetic and instrumental) and affective responses (e.g., interest and enjoyment).

2.6.3 Learning Experience

A learning experience might actually be an opposite of a positive experience. Whereas a positive experience is characterized by cognitive fluency and "fast thinking", learning requires "slow thinking" and a state of cognitive disequilibrium (Kahneman, 2003; Schwarz and Clore, 2007). The former, "fast thinking", refers to a processing style consisting of fast, automatic, or unconscious processes. The latter, "slow thinking", constitutes slow, effortful, and conscious processes (Evans, 2008). Both the occurrence of cognitive disequilibrium and the application of "slow thinking" have been noted as pivotal for deep learning to occur (Berlyne, 1978; Evans, 2008; Kahneman, 2003; Schwarz and Clore, 2007). The state of a cognitive disequilibrium has been noted as a target of intelligent tutoring systems and more generally of information systems aimed at learning (D'Mello and Graesser, 2012).

Key to a learning experience is the occurrence of a state of cognitive disequilibrium. The uncertainty associated with a cognitive disequilibrium leads to compensatory slow thinking (Tiedens and Linton, 2001), essentially tuning cognition to meet the situational requirements (i.e., the tuning hypothesis[7]) (Schwarz and Clore, 2007). The differentiation between "fast thinking" and "slow thinking" is based on dual-processing accounts of reasoning and judgments and relates to two distinct types of mental processes. The former, Type 1 processes ("fast thinking"), are used to make judgments and process information. The latter, Type 2 processes ("slow thinking"), can monitor and control Type 1 processes (Evans, 2008; Kahneman, 2003; Schwarz and Clore, 2007). When using Type 1 processes, individuals rely on existing knowledge (i.e., top-down processing) and on emotions. In contrast, with Type 2 processes individuals focus more on the details in the stimulus (i.e., bottom-up processing) (Schwarz and Clore, 2007). Hence, Type 2 processes, or slow thinking, reflect a processing style that is particularly important for a learning experience (Evans, 2008; Kahneman, 2003; Schwarz and Clore, 2007).

A learning experience consists of a specific narrative of states, responses, and relevances. The pattern of states is a repetitive cycle between engagement and a cognitive disequilibrium. A state of engagement is interrupted by a cognitive disequilibrium when an individual is confronted with impasses, contradictions, cog-

[7] The tuning hypothesis states that human cognition is situated and adaptively tuned to meet situational requirements (Schwarz and Clore, 2007). Following this hypothesis is that Type 2 processes associated with slow thinking become more salient when in problematic or goal obstructing situations, whereas Type 1 processes associated with fast thinking are salient in normal situations.

nitive dissonance, incongruities, novelty, and obstacles to goal attainment. This points to the importance of specific relevances and the processing difficulty they create. The uncertainty and concentration associated with a cognitive disequilibrium fosters slow thinking, which in turn allows us to solve the disequilibrium. However, the uncertainty can also lead to negative states such as frustration, confusion, and eventually, boredom (D'Mello and Graesser, 2012), showing the challenge of finding the ideal narrative for a learning experience.

2.6.4 Experiential Perspective

The experiential perspective gives a holistic view on information interaction. It describes the momentary thoughts, feelings, and actions or, in other words, the current cognitive-affective-motivational state[8]. This state emerges from the components of an interaction, is situated in a user, and exists only in the moment; that is, the momentary experience. The components subsume at least the relevances and affective responses of a user. All these components interact with each other and are themselves already hard to describe, explain, and influence. This shows the vast width of influences that come together in forming a mental state and the challenge of describing, explaining, and influencing it. Moreover, the optimal experience is dependent on the context and activity.

To tackle the elusiveness of describing a cognitive-affective-motivational state, the three typical states of satisfaction, engagement, and slow thinking were derived from the preceding outline of the positive, engaging, and learning experience. First, satisfaction shows the role of an overall positive state that reflects an easy, conducive, situation. Second, engagement or flow indicate a highly concentrated mental state with an intrinsic motivation for an activity (Csikszentmihalyi, 1991; O'Brien and Toms, 2008). Third, slow thinking is an analytic processing style which is particularly important for solving a cognitive disequilibrium and, accordingly, for explicit learning (Evans, 2008; Kahneman, 2003; Schwarz and Clore, 2007). Each of these aspects are well-researched and connect to different types of information activities and systems. Nonetheless, the complexity of the cognitive-affective-motivational state of the user suggests the importance of understanding the values and responses first.

The following definition summarizes the experiential perspective on the IX:

Definition 4: The experiential perspective highlights the cognitive-affective-

[8] The affective aspects of the experience are already partly covered by the emotional perspective, which focuses on affective responses. The distinction between short-term responses and experiential affective states is somewhat ambiguous. Hassenzahl and Tractinsky (2006) differentiate between emotions and mood, respectively.

motivational state that emerges from (non-)instrumental relevances and emotions, is situated in a user and exists only in the moment.

2.7 Designing an Information eXperience

By applying indicators that reflect a distinct set of relevance criteria, information systems can influence the IX. The extent to which this holds will be explored in this section: that is, the extent to which algorithms can influence the relevances, responses, and even states of the user. Considering the difficulty of directly influencing the states, the focus will be on designing for values that foster certain responses ("foster" in Figure 2.1, p. 19) which, in turn, can affect the user's mental state ("compose" in Figure 2.1, p. 19).

2.7.1 Composing States

The framework of the IX suggests that the momentary state of the user originates, amongst other aspects, from the components of interaction. Figure 2.3 (p. 30) highlights the relation of the momentary state with instrumental ($i_4 - x_3$) and non-instrumental ($n_4 - x_1$) relevances and with affective responses ($r_2 - x_2$). The extent to which the states, which belong to each of the prototypical experiences, can be governed by the elements of the IX will be explored next.

Satisfaction

Satisfaction is a complex phenomena to which many contributing variables have been identified. These variables include instrumental value, non-instrumental values such as aesthetic value, and positive affect. In particular aesthetic value has been found to influence satisfaction strongly and instantly (Lindgaard et al., 2006). In addition to the clear contribution of values and affective responses to satisfaction, more antecedents to satisfaction exist that are not covered by the IXF. Specifically, certain user traits (Al-Maskari and Sanderson, 2010) and the expectations of the user (Oliver, 1993) are important in explaining the occurence of satisfaction.

Within the context of information retrieval the determinants of satisfaction are, partly, known. Satisfaction is positively dependent on system effectiveness (i.e., objective relevance), user effectiveness (i.e., perceived success), negatively on user effort, and on several user traits (Al-Maskari and Sanderson, 2010), showing instrumental relevance promotes satisfaction. The positive contribution of instrumental relevance is further supported by findings showing that a higher objective relevance strongly leads to more satisfied users (Gluck, 1996; Huffman and Hochster, 2007). However, the supporting evidence is mainly indirect. In both Gluck (1996) and Huffman and Hochster (2007) the relevance of each search result relative to a query was indicated retrospectively by a judge, whereas satisfaction was measured directly (Gluck, 1996) or was also inferred by a judge (Huffman and Hochster, 2007).

Engagement

Regarding the engagement/flow experiences their common antecedents can be identified. Instrumental value is found key to flow through the importance of clear goals and immediate feedback in attaining those goals (Chen, 2007) and to engagement via the importance of current need fulfillment and goal achievement (Higgins, 2006). Regarding non-instrumental value, challenge is perhaps the distinctive determinant of flow and engagement. For flow, the right level of challenge in comparison to the skills of the user is central to its theory (Chen, 2007; Csikszentmihalyi, 1991). For engagement, Higgins (2006) identified how internal (e.g., overcoming personal resistance) and external (e.g., complexity) opposing forces increases value and engagement. From an affective perspective, interest has been shown as pivotal for the onset of engagement (O'Brien and Toms, 2008) intrinsic motivation (Reeve, 1989), a characteristic aspect of flow (Csikszentmihalyi, 1991; Higgins, 2006). Furthermore, the experience of flow/engagement is associated with positive affect in general (Higgins, 2006; Reeve, 1989; Csikszentmihalyi, 1991; Chen, 2007). The foregoing overview of the core antecedents of flow and engagement (for a full review, see Csikszentmihalyi, 1991; Higgins, 2006) clearly shows that the IXF covers the key antecedents to this experience.

The influence of instrumental value, the non-instrumental value of challenge, the emotion of interest, and of overall positive affective responses for engagement and flow has been confirmed within the context of information interaction as well. In particular studies have confirmed the importance of the non-instrumental relevances of aesthetic quality, such as vividness and attractiveness, and of hedonic quality, such as stimulation, novelty and playfulness (O'Brien and Toms, 2008; Finneran and Zhang, 2005; Hoffman and Novak, 2009; Moon and Kim, 2001). Moreover, the epistemic emotion of interest has been shown key to the onset of engagement during various activities on the internet (O'Brien and Toms, 2008). The importance of the non-instrumental relevances is often not reflected in relevance models. As previously noted, non-instrumental relevance criteria such as complexity, understandability, and effort are often of little focus in contemporary relevance models; even though they are pivotal to identify whether the amount of challenge is in balance with the abilities of the user.

Slow Thinking

Type 2 processes become more salient when in problematic or goal obstructing situations (i.e., the tuning hypothesis[9]; Schwarz and Clore, 2007). This indicates that a lack of instrumental value or, more specifically, (subtle) goal obstruction, can enhance slow thinking (Schwarz and Clore, 2007). Evidence for a direct effect of non-instrumental values on the processing style is lacking. An indirect effect, however, is possible, such as by affecting processing fluency. For example, the positive affect generated by hedonic value can inhibit slow thinking by signaling a (goal) conducive situation (Song and Schwarz, 2008; Schwarz and Clore, 2007). Emotions also affect the cognitive processing style that individuals adapt. Generally, a negative affective response signals a problematic environment and leads to the adoption of a systematic, bottom-up, processing style with considerable attention to detail (i.e., slow thinking) (Schwarz and Clore, 2007). The extent of this influence is governed by the information value of the affective response. With an identifiable referent, emotions have a clear information value. In particular emotions that entail a high uncertainty appraisal induce slow thinking (Tiedens and Linton, 2001). Moods and cognitive fluency lack a clear referent and therefore their influence on the processing style is smaller and more susceptible to the (mis)attribution of its causes (Bless et al., 1996).

Within the context of information interaction, a lack of instrumental relevance can enhance slow thinking (Schwarz and Clore, 2007). This only applies to certain relevance criteria. For example, when reading a text the background knowledge of the reader interacts with the coherence of the text, where deep comprehension improves for highly knowledgeable readers when reading less coherent texts. This suggests that poorly written texts force knowledgeable readers to engage in compensatory, slow, thinking (Kahneman, 2003) to infer unstated relations from a text (McNamara et al., 1996). Also, non-instrumental relevances can influence the adoption of slow thinking. For example, the cognitive fluency caused by easy-to-read fonts inhibits slow thinking (Song and Schwarz, 2008). Aside from the signals related to relevances, numerous other signals can foster slow thinking. Perhaps in some situations more easily controllable, task framing and explicit instructions can cause individuals to adopt slow thinking without signals indicating a problematic environment, also when individuals are in a happy mood state (Schwarz and Clore, 2007).

[9] See definition on p. 47

Although the framework of IX covers key antecedents for all prototypical experiences and their accompanying momentary states, an implementation of the experiential perspective is difficult. This is testified to by the numerous values and affective responses that come together, the situatedness and temporality, and by being the result of an interactive process spanning multiple smaller information interactions. As Hassenzahl and Tractinsky (2006) state, "*it is not clear whether we can 'design' an experience*" (p. 95). To promote an experience, control over its elements and their interactions is requisite. Hence, the focus should be on the components of an experience before tuning to the (holistic) experience.

2.7.2 Fostering Responses

For the feasibility of designing an IX the central question is whether a certain affective response can be made probable. To address this question, the antecedents of information emotions and cognitive fluency will be identified and their influenceability will be explored.

Information Emotions

The core antecedents of information emotions are summarized by the control-value theory (Pekrun, 2006). This theory clearly connects the antecedents of information emotions to the IXF perspective on relevances and values. Furthermore, also from the general cognitive appraisal theories of emotion this connection becomes clear. The effect of instrumental value on emotions illustrates this. When a stimulus is high on extrinsic value (control-value theory) it describes a positive effect with respect to the current goals and needs of a user; that is, the motivational bases (i.e., cognitive appraisal theories). Similarly, also the effect of non-instrumental value is evident. When a stimulus is high on intrinsic value (i.e., control-value theory) it is characterized by intrinsic pleasantness and novelty (i.e., cognitive appraisal theories), denoting a positive effect.

Whereas the influence of value on emotions is clear, both via the control-value theory and the underlying appraisals, the antecedents of information emotions also overlap with findings on and models of relevance. The extrinsic value clearly overlaps with situational relevance: that is, the utility of information to a situation, task, or problem (e.g., topicality). The intrinsic value is only partly covered by models of relevance, in particular cognitive relevance or pertinence and affective relevance. Cognitive relevance describes criteria such as novelty and cognitive correspondence (i.e., familiarity), which are primary appraisals for epistemic emotions (Silvia, 2008b, 2009). Criteria of control or coping potential have, to some extent, been suggested as part of the relevance criteria individuals apply (e.g., understandability, clarity; Barry and Schamber, 1998; Xu and Chen, 2006; Xu, 2007). However, although both relevance theory and theories about emotions suggest the importance of control or coping potential, it is often not included in models of relevance. Numerous suggestions have been made to include it as part of (situational) relevance (Barry and Schamber, 1998; Xu and Chen, 2006; Xu, 2007). Accordingly, we do consider it part of the relevance perspectives (see Section 2.4.3).

Evidence confirms the expected effect of instrumental relevance on information emotions. The effects of the (positive) appraisal of motivational bases on emotions were explored for the goal conduciveness and goal relevance appraisals (Kreibig et al., 2010) and their interaction (Kreibig et al., 2012). A higher appraisal of motivational bases was found to lead to a range of positive emotions among which interest, joy, pride, surprise (Kreibig et al., 2010, 2012). Moreover, Arapakis et al. (2008) showed a direct relation between objective relevance and emotions (Figure 2.3, $i_5 - r_3$). By reducing the possible number of relevant search results per task, the task difficulty was increased, leading to a range of negative emotions. These findings are in line with the control-value theory of information emotions, which predicts that a lack of extrinsic (goal or need related) value leads to anger (in the case of a high control appraisal) or frustration (in the case of a low control appraisal) (Pekrun, 2006).

Hedonic and aesthetic relevances have been confirmed to influence affective responses ($n_5 - r_4$). This is suggested by the definition of hedonic values as pleasurable attributes (Hassenzahl, 2003) and the identification of intrinsic pleasantness as part of the primary appraisals in the emotional process (Ellsworth and Scherer, 2003). This relation is supported by an exploratory study, finding that pleasurable products are associated with a range of positive feelings (Jordan, 1998). More precisely, a product its hedonic qualities were shown to correlate with positive affect and need fulfillment (Hassenzahl et al., 2010). And, for web sites, hedonic and aesthetic relevances correlate with both the valence and arousal appraisals of emotions (Fiore et al., 2005; Mummalaneni, 2005). The specific effects of stimulation have been found to be pivotal for the emotion of interest, where a higher complexity is related to an increased interest response for stimuli such as poems (Silvia, 2005).

Cognitive Fluency

The appraisals for information emotions are metacognitive and, therefore, probably affected by (meta)cognitive fluency. The sources of this fluency give clear indications for the development and use of algorithms that can denote its antecedents.

For example, fluency can be enhanced via a high base frequency of words, easy-to-read fonts, regular word forms, and semantic priming (e.g., cohesion) (Alter and Oppenheimer, 2009). In particular the linguistic aspects that influence processing fluency can be approximated using algorithms (see Chapter 3). Processing fluency is indicative of complexity and influences judgments of familiarity and confidence, which can, in turn, affect the primary novelty-complexity and secondary comprehensibility appraisals that underlie key epistemic emotions such as interest (Silvia, 2008b). Moreover, processing fluency influences the adopted processing style as well (Schwarz and Clore, 2007, see Section 2.5.4). These antecedents directly link epistemic emotions and cognitive fluency to the appraisals of a specific set of relevance criteria. These appraisals can be approximated by algorithms, the feasibility of which will be explored in the next section (Section 2.7.3).

Implementation

Several studies explored implementing emotion-aware information related technology. Mooney et al. (2006) examined the role of searchers' emotional states to extend data indexing for IR. They used physiological responses to a range of emotional stimuli. Moshfeghi (2012) explored the possibility of inferring the emotional value from text. This indication is used to satisfy users' emotional need, as opposed or complementary to their information need. Arapakis (2010) explored the possibility of facial expressions and physiological responses to serve as relevance feedback. The current emotion-aware information related technologies tend to follow the ideas underlying objective relevance, assuming that similar content will lead to similar emotional responses, possibly non-personalized, while both the research on subjective relevance and on emotions show that both relevances and affective responses are personal and dynamic. Notwithstanding, information systems or affective systems can influence affective responses via the cognitive appraisals of information (see Definition 3).

Affective responses can serve as a framework for giving weights to diverse algorithms that reflect appropriate relevance criteria: that is, to a relevance profile. This allows a system to target a specific response by selecting information congruent with the accompanying set of relevance criteria. An example is for the emotion of interest, which can be fostered by selecting information of an appropriate level of complexity. Currently, models of relevance only partly cover the non-instrumental value and often under-estimate the importance of control (e.g., understandability) of information. To give users an optimal IX requires the inclusion of non-instrumental value as well as control. Notwithstanding, even though the models of relevance are, obviously, dissimilar to the appraisals underlying the information emotions, this does not exclude information systems from affecting the appraisals and, accordingly, the responses: that is, relevance directs experience (denoted as “foster” in Figure 2.1).

2.7.3 Implementing for Values

Table 2.2 summarizes the instrumental and non-instrumental values that are important to foster affective responses and compose an experiential state. Whether it is possible to design (algorithms for) each of these values will be explored next.

Values

During information interaction, values originate from a manifestation of information. For instrumental value (Figure 2.3, o_2 -– i_7), this follows directly from the stratified model of relevance (Saracevic, 2007) as well as from empirical evidence (Wang and Soergel, 1998). The stratified model of relevance indicates which aspects of information give instrumental value: topicality, pertinence, and utility. These aspects contribute to the functional, conditional, and epistemic value of information and are operationalized through criteria such as topicality, coherence, novelty, and complexity. These criteria relate to a positive effect of the information, which contributes to positive affect and a state of satisfaction. Furthermore, these aspects relate to the effort required to attain the value of the information. In particular the complexity of information, reflected by expected reading time and expected effort, can reduce the instrumental value (Wang and Soergel, 1998; Song and Schwarz, 2008). Spink and Greisdorf (2001) show (and related studies suggest: e.g., Barry and Schamber, 1998; Beresi et al., 2010) that all levels of relevance contribute to explaining relevance decisions and, thus, the value of information is reflected by all levels of relevance.

Aside instrumental relevance, manifestations of information have aesthetic and hedonic value as well (Figure 2.3, $o_3 - n_6$). Non-instrumental value is important for the IX. Namely, aesthetic values contribute to cognitive fluency and are important for the start of an engaging experience. Hedonic values, although not always in line with goal attainment, can foster the emotion of interest, increase engagement, and even improve learning. Those aspects of web sites, information, or media that give it aesthetic and hedonic value are known to some extent. Regarding aesthetic values, visual simplicity is generally associated with aesthetic appeal (Karvonen, 2000). This is confirmed for web sites in relation to the prevalence of functional elements (Michailidou et al., 2008). Regarding hedonic values, aspects of interactivity contribute to the enjoyability of interactive systems, indicating their hedonic value (Blythe et al., 2004). Similarly, for textual information, seductiveness and vividness have been identified as aspects contributing to the attractiveness of a text (Schraw and Lehman, 2001). And, textual complexity has been associated with the primary appraisal of intrinsic pleasantness for the emotion of interest, indicating its hedonic value as well (Silvia, 2008b). However, the value of textual complexity and other aspects that increase the challenge and effort of information only exists as long as the challenge is in balance with the abilities of the user: that is, the information remains understandable.

An overlap exists between the non-instrumental and instrumental values. Some relevance criteria contribute to both values, in particular for those criteria that belong to the cognitive relevance type. For example, novelty and understandability are both cognitive relevance criteria that can increase the utility of the information (Saracevic, 2007) and also have non-instrumental value (Silvia, 2008b). The instrumental value of information towards an information need might even contradict its instrumental value with respect to a task. If a document is highly relevant towards an overall information need, it is not necessarily valuable towards less granular task representations. For example, while textual complexity has non-instrumental value through stimulation and is producing challenge (Silvia, 2008b; Csikszentmihalyi, 1991), a decrease in complexity likely to increase the instrumental value through aspects such as (expected) reading time (Wang and Soergel, 1998). Which value perspective is preferred depends on the activity and task. In an epistemic and hedonic information search non-instrumental values are prevalent (Xu, 2007), whereas in regular search activities the instrumental values will be most important.

Implementation

Given that relevance is dynamic, multi-dimensional, and non-binary, the implementation (detection) of instrumental value is inherently difficult. To circumvent this difficulty, both IR and IF&R systems implement a weak form of relevance which models the topicality of a document in relation to either a query or a user model of long-term interests. This normally leads to a static relation between a document and a query or a model. Although these methods indeed reflect the topicality of a document, there is already a considerable disagreement in the test collections used to train the algorithms. In the creation of these collections, there is less than 50% overlap between two judges and 30.1% overlap between three judges on whether a document is relevant given an information problem (Voorhees, 2002). Moreover, other levels of relevance are not reflected by algorithmic relevance (Borlund and Ingwersen, 1998). Aspects such as novelty and complexity or not explicitly taken into account in the creation of test collections, although they can be understood as partly responsible for the low inter-rater agreement. The problems that are associated with the static relation between a document and a query or a model can be alleviated by using relevance feedback or related techniques (see Section 2.2.3), which includes the possibility of using the IX as input to algorithmic relevance (see Section 2.8).

Non-instrumental values are usually implemented indirectly. For example, tagging systems are often used to indicate hedonic or aesthetic qualities such as "*funny*" or "*inspirational*" (Golder and Huberman, 2006, p. 203). For IR systems an argument can be put forward that hedonic and aesthetic values, although not included in models of relevance, are already implemented indirectly. Namely, the famous PageRank algorithm uses the graph structure of the web to measure the centrality of web pages. This centrality is primarily interpreted as an indicator of relative importance and quality (Page et al., 1999). Hedonic and aesthetic values can be expected to be part of this quality and importance of a web site. For IF&R systems based on collaborative filtering a similar argument holds. Since collaborative filtering uses the ratings of other (related or similar) users for media items, such ratings reflect a range of relevance criteria likely reflecting foremost hedonic and aesthetic values. For example, musical taste, an aesthetic and hedonic preference, can be inferred from the ratings of music by other, similar, users (Uitdenbogerd and van Schnydel, 2002). Although indirectly, most information systems do include some notion of non-instrumental values via a collaborative approach. Considering that an IX is personalized partly explains the success of collaborative methods to indicate the hedonic and aesthetic aspects of the IX.

The related relevance criteria of readability, textual complexity, and coherence have been modeled as well. Traditional accounts of predicting the complexity of information mostly consisted of readability formulas. These formulas use basic word and sentence features (e.g., length) to generate ordinal scores allowing the ranking of texts on their complexity (Benjamin, 2012). Textual coherence, which is subsumed by the general notion textual complexity, has been modeled as well. Namely, the Coh-Metrix system applies cognitively inspired indexes to indicate the cohesion of a text (Graesser et al., 2004). Combined with a few lexical measures, such techniques explained 76.3% of variance in textual cohesion (McNamara et al., 2010).

Either directly or indirectly and in weak or strong form, most instrumental, aesthetic, and hedonic values can be modeled. This opens up possibilities to use these models to select and retrieve information which is likely to be appraised in accordance with a response: that is, information systems can foster responses using a diverse set of algorithmic indicators which reflect salient values. An implementation of this can be in the form of a relevance profile. This profile denotes an objective model containing features that describe the (relation between the) information, query, and the user. The notion of a relevance profile is not new. Current implementations of a relevance profile often already exist in the form of a feature vector (Liu, 2009). Notwithstanding, connecting the objective relevance profile to a subjective appraisal profile is a new step that allows us to influence the affective responses of a user.

2.8 Directing Relevance

The relation between relevance and IX can be reversed as well, when the IX informs (algorithmic) relevance. Cosijn and Ingwersen (2000) were a clear proponent of this, arguing that affective relevance (dealing with the affective responses of a user to a document) influences all other types of relevance. Two pathways along which the IX informs relevance can be derived from the framework presented in Section 2.3: either as relevance feedback on specific target responses (Arapakis, 2010) (described as “affective feedback” in Figure 2.1), or as the current state of the user that affects relevance criteria (Cosijn and Ingwersen, 2000) (described as “affective relevance” in Figure 2.1).

2.8.1 Affective Feedback

This section explores the value of affective responses for relevance feedback. An affective response can, hypothetically (Silvia, 2008b), be measured using various proxies derived from interaction data (Agichtein et al., 2006), query analysis (Ruthven, 2012), eye-tracking, or psychophysiological measures (Arapakis, 2010). Because the experienced emotions (i.e., the outputs) in relation to a stimulus (i.e., the inputs) form a window on the underlying cognitive appraisals (see Section 2.5), these affective responses can serve as affective feedback. Affective feedback can serve at least two roles: either as feedback about a manifestation of information or as input to a (relevance) decision.

Affect as Feedback

As first role, the affective response can serve as feedback about the relevance of a manifestation of information. One of the first to use affective responses to inform algorithmic relevance was Arapakis (2010). In a controlled study it was shown that sensory channels regarded indicative of a user’s affective responses can predict the topical relevance (feedback) of a document. Compared to a baseline accuracy of 50.00%, binary relevance could be predicted with 57.90% based on facial expressions and with 60.40% using physiological signals (Arapakis et al., 2009). The possibility of this is grounded in the effect that satisfying an information need has on a user's IX. However, the effect that reaching closer to (or deviating from) satisfying an information need has on the affective responses of the user is not straightforward. The extrinsic value or motivational bases operate next to the intrinsic value and control, where the appraisals of novelty/pleasantness (i.e., intrinsic values) and coping potential (i.e., control) have a more solid empirical support (see Section 2.5). Moreover, relevance is neither binary nor singular (see Section 2.2; Borlund, 2003; Saracevic, 2007). The value of using affective responses as relevance feedback is likely not for utility in particular, but can still function as more general affective feedback.

An affective response tells that a user appraises a manifestation of information according to a specific pattern of relevance criteria. Hence, a response gives an indication of the underlying set of appraisals that caused it. The emotion of interest illustrates this. When experienced this indicates that the stimulus was seen as novel and complex yet comprehensible (Silvia, 2008b). Furthermore, when an IF&R system tries to select information according to this pattern of criteria in order to foster an interest response, the affective feedback can be used to verify whether the information was indeed appraised accordingly. Affective feedback can thus indicate whether a stimulus was appraised according to a specific pattern of relevance criteria, which is a form of non-binary and multi-dimensional relevance feedback. A further understanding of the relation between relevance criteria and information emotions allows us to fully utilize the possibility of affective feedback. This relation is partly denoted by the control-value theory, which shows how the appraisals of value and control affect the experienced information emotion. However, more research is needed to confirm its predictions (Pekrun, 2006).

Affect as Information and Spotlight

As second role, an affective response informs relevance decisions as well, influencing both instrumental and non-instrumental relevance criteria (Figure 2.3, $r_5 - n_7, i_6$). People use their emotional response to a target stimulus to decide whether to engage with the target, essentially asking themselves "How do I feel about this?" (Schwarz and Clore, 2007). This so-called affect-as-information paradigm also operates for more general affective responses that are not emotions. For example, cognitive fluency has been shown to influence, amongst others, judgments of truth, liking, certainty, and familiarity (Alter and Oppenheimer, 2009). The affective response has been posited to function as a common currency based on which distinct alternatives are compared to each other (Peters, 2006). This idea is in line with relevance theory, which states that relevance is only judged relative to alternative stimuli, and shows the importance of the initial affective response for relevance (judgments). Hence, in a situation where a relevance decision is made, the affect-as-information mechanism shows how the affective response to each of the alternatives gives immediate feedback on the relevance judgments of each of the alternatives and on the subsequent relevance decision.

Besides being a direct source of information, immediate affective responses guide attention. Positive affect functions as a spotlight directing attention to certain positive aspects and avert attention from certain negative aspects (Peters, 2006; Pfister and Böhm, 2008). This affect-as-spotlight function particularly relates to directing attention to the value of and away from the control over a target stimulus (Alhakami and Slovic, 1994), reducing the importance of relevance criteria such as understandability. Some findings suggest that cognitive ease changes the ratio between value and control as well. Namely, when induced using easy-to-read fonts, cognitive fluency decreases the expected effort and increases the motivation for engaging with a text (Song and Schwarz, 2008).

The initial affective response that guides attention and judgments is detectable (Pfister and Böhm, 2008). This suggests that the affective response can serve as relevance feedback. This role for affective feedback is likely most salient when a relevance decision changes (e.g., when a user starts or stops viewing an information object) or when alternative choices are compared (e.g., which happens on a search engine results page). However, the extent to which the affect-as-information and affect-as-spotlight mechanisms are valuable for relevance feedback is yet unclear.

Two approaches were identified in which affective responses can inform (algorithmic) relevance: as affective feedback and via the affect-as-information and affect-as-spotlight mechanisms. Both approaches can give feedback on the selected and retrieved information, either about whether the information indeed fostered a desired affective response or as detailed and immediate feedback that underlies a relevance decision. Moreover, in the case of an undesired response this response is reflective of a (lack of) value or control (Pekrun, 2006) and can subsequently be used to change the constitution of the relevance profile. Both approaches help delineate the multi-dimensionality of relevance by giving feedback on aspects of a relevance profile. Hence, both approaches can serve as relevance feedback and can make (algorithmic) relevance dynamic in a similar fashion as traditional relevance feedback can (Salton and Buckley, 1990); in addition, affective feedback also support its multi-dimensional nature.

2.8.2 Affective Relevance

Aside the potential of responses as feedback, the user's state can also predict the importance of certain relevance criteria ($x_4 - n_8 - i_7$). The affective state has been posited as a different dimension of relevance that affects all other dimensions of relevance: that is, affective relevance. Although the influence of the affective state has been noted by numerous authors (Cosijn and Ingwersen, 2000; Borlund, 2003; Saracevic, 2007; Wilson, 2006), the specificities of this influence are not clear. Nonetheless, the two aspects of control and value that are important for information emotions can also partly summarize the influence that a particular state has on the applied, subjective relevance criteria. As described next, positive affect, engagement, and slow thinking change the ratio between the importance of the appraisals of control and value.

Satisfaction, when defined as an overall positive (affective) state, can influence (relevance) judgments via the affect-as-information mechanism (see Section 2.8.1). Due to misattribution of the affective state judgments are sometimes flawed (Isen et al., 1978). Individuals are likely to evaluate any target more positively when in a happy rather than a sad mood. Furthermore, individuals in a happy mood become more risk averse than individuals in a negative mood (Isen et al., 1978). This suggests that, besides decreasing the importance of value, a happy state increases the importance of appraisals of control and, accordingly, relevance criteria such as understandability. Using similar indicators as for affective feedback, the mood state can likely be measured as well. However, measuring the state of a user is more difficult than measuring a change in a state. This is related to the baseline available to compare a measured signal with. In the case of a response, this change is in comparison to a previous measured signal. In the case of a state, the baseline to compare the measured state to is a previous, different state, which requires longer periods of time and, likely, personalization (Van den Broek, 2011).

Strength of engagement leads to increased value as signified by an increased attraction to or repulsion from a target. This increase in value occurs aside the already existing value of a target (Higgins, 2006; Csikszentmihalyi, 1991). This idea is in line with the description of a Flow experience (Csikszentmihalyi, 1991), which turns an activity into an autotelic activity, essentially giving value to the activity itself (i.e., an intrinsic motivation). Hence, a state of engagement likely has a similar effect on the ratio of importance of value and control as a positive affective state has, although hypothetically the importance of control might lessen due to the increased motivation and perseverance (cf. Higgins, 2006). A first step to the measurement of engagement has been explored using various proxies derived from interaction data (Lehmann et al., 2012), suggesting its feasibility.

Slow thinking influences the applied relevance criteria as well. The influence of affect and heuristics decreases when individuals are utilizing Type 2 processes. In this strategy, although slow, individuals will put more effort in correctly appraising the relevance of an information object. For example, a systematic processing style can undo the effects of processing fluency such as heightened truth, preference, and beauty, and reduced effort appraisals (Schwarz and Clore, 2007). The information search pattern of ”deep diving“ gives a more precise indication of the effect on the applied relevance criteria. This search pattern is characterized by a high effort and focused attention, signaling slow, conscious, bottom-up processing, and is associated with a preference for information that is thorough (high in depth and scope), of high quality, and written by well respected authors (Heinstrom, 2002). This indicates that the processing style of slow thinking increases the importance of relevance criteria such as depth, scope, quality, and source reputation towards an overall goal of understanding. Hence, it is likely that the criteria related to value become more important whereas the criteria related to control become less important. Whether or not the extent to which an individual applies Type 2 processes can be measured is unclear. A likely candidate is through an individual's pupil size, which reflects the cognitive processing load of the individual (Beatty, 1982). The value of pupil size for indicating the state of a user during information interaction has been noted as well (Granka et al., 2008). However, processing load is likely more related to cognitive strain than to the application of Type 2 processes.

The previous outline of affective relevance indicates that, although the idea of affective relevance has been suggested by numerous authors, it is not pursued much yet. Possible connections between a user's state and the applied relevance criteria were identified, in particular in light of the ratio of importance of control and value. The highlighted user's states were derived from prototypical experiences that occur regularly during current information interaction. The identified connections indicate how changes in the state change the set of relevance criteria applied and, accordingly, change the optimal constitution of the applied relevance profile. As such, knowledge about the current state of the user can help delineate the multi-dimensionality and dynamicity of relevance as well. However, whether this contribution is possible is yet unclear. Affective relevance (Saracevic, 2007; Cosijn and Ingwersen, 2000; Wilson, 2006) has a great potential which has yet to materialize.

2.9 Conclusion

The information overload argument already indicated a pivotal role for relevance in changing the experience with the information landscape. In particular, in changing the feelings associated with information. The IX resulting from information interaction can benefit a range of goals, including: finding, encountering, and learning. For an information system to orchestrate an IX, target states need to be identified first. These targets will be dependent on the activity, such as: satisfaction, engagement, and slow thinking. By attending the IX, the utility of information systems can be expanded beyond usability and become more supportive to an activity.

Saracevic (2007) noted: "*Relevance is a feature of human intelligence. Human intelligence is as elusive to "algorithmize" for IR as it was for AI*" (p. 1925). Notwithstanding, a further understanding of the (causes of the) IX can partially delineate relevance. This does not require a claim of intelligence. Instead, an implementation of context-awareness can suffice. Actors indicative of textual complexity, familiarity, and novelty can already influence aspects of the IX, while psychophysiological sensors and interaction data have been studied as sensors that can serve as affective feedback. Furthermore, acknowledgment of the IX of the user can indicate how relevance judgments change dependent on the mental state of the user and, consequently, explain part of the dynamics in relevance judgments. Hence, the IXF shows the importance of relevance criteria not included in objective relevance and supports an implementation of a multi-dimensional and dynamic model of relevance; that is, a relevance profile.

The interplay between the different perspectives of information objects, values, affective responses, and (cognitive-affective-motivational) states shows the possibility of information to foster a fruitful IX. The IXF allows us to zoom-in on specific relations between each of the perspectives. It shows how manipulations of aspects of information affect appraisals, direct a response, and influence a state. Simultaneously, the framework shows how a state can influence judgments of information. The IXF allows us to solve the information overload and simultaneously fully utilize the "information opportunity". Perhaps, it can even foster a sense of ecstasy during information interaction.

3

Predicting Textual Complexity

Abstract

The Information eXperience Framework (IXF) identified a central role for textual complexity, both for its importance to values and its significance in determining an affective response. To implement this role, this chapter introduces a computational model of textual complexity. To attain its validity, a set of features is introduced based on common observations about the causes of our processing difficulty. This allowed to create a unifying model of textual complexity without the necessity of a unifying theory of processing difficulty. To attain its applicability, the resulting model is trained on a large data set distinctive on textual complexity. Furthermore, by minimizing the influence of text length and semantics on the features, the model becomes applicable to a variety of data sets. The model achieved a maximum performance of 93.62% on classifying encyclopedic articles of two classes of complexity, confirming the success of the model. Combined with the focus on both validity and applicability, this success makes it probable that the model can influence the Information eXperience (IX).

Parts of this chapter also appear in:

Van der Sluis, F., Van den Broek, E.L., Glassey, R.J., Van Dijk, E.M.A.G., and De Jong, F.M.G. When complexity becomes interesting. *Journal of the American Society for Information Science and Technology. [in press]*

Van der Sluis, F. and Van den Broek, E.L. (2010). Using complexity measures in information retrieval. In Belkin, N.J., Kelly, D. (Eds.) *IIiX 2010: ACM Proceedings of the 3th symposium on Information Interaction in Context*, p. 383–388, Augustus 18–22, New Brunswick, USA.

3.1 Introduction

Nearly a century ago the first readability indicator was introduced by Lively and Pressey (1923). Since then, many indicators have been developed, mainly from an educational perspective. The importance of text readability for education is clear: different levels of expertise (e.g., literacy) require different levels of educational material. For example, a novice is better of with an expository text, whereas a competent learner is likely to learn most from a more challenging text (Kintsch, 1994). Besides the relevance of readability to learning, at least two more applications can be identified. First, readability and related notions such as understandability, cognitive correspondence, and processing difficulty have been theorized and shown to be part of the relevance judgments that users of Information Retrieval (IR) systems apply (Xu and Chen, 2006). Hence, people appraise the readability of a document when considering its relevance, and decide partly on this appraisal whether or not to explore a document. This was confirmed by Collins-Thompson et al. (2011), who showed that the difference between document and snippet readability strongly (negatively) predicted dwell time (till 120 seconds), explaining 69% of its variation. Second, textual complexity[1] fosters the emotional experience of interest that a user has with an information object: more complex information becomes more challenging and, therefore, heightens interest (Silvia, 2008b). Yet, information can also be too complex and be appraised as incomprehensible and lead to a profoundly different response than interest (see Chapter 4).

In solving Artificial Intelligence (AI) problems, generally two approaches can be distinguished (Manhart, 1996): bottom-up, data-driven (Breiman, 2001b), and top-down, theory-driven (Chomsky, 1956). Although the former has become the standard this is not without critique, for it provides little insight into the workings of the underlying system[2]. Notwithstanding, the data-driven approach has become dominant and led to most successes of AI, such as in IR, Part-Of-Speech (POS) tagging, sentence structure parsing, and coreference resolution (e.g., Klein and Manning, 2003; Lee et al., 2011). The focus in this chapter, the challenge of identifying the textual complexity of a text, has historically been approached data-driven: that is, based on observations of differences in text styles[3] (Dubay, 2004).

[1] Textual complexity and readability will be used interchangeably. Nonetheless, textual complexity is the preferred term for it underlines that the influence goes beyond reading ease alone.

[2] http://www.theatlantic.com/technology/archive/2012/11/noam-chomsky-on-where-artificial-intelligence-went-wrong/261637/

[3] The old readability formulas were based on historical work in which texts were analyzed statistically (e.g., Thorndike, 1921). In its purest sense this is data-driven, as rules were derived from these statistical analyses. Yet, it can also be interpreted as slightly theory-driven, as the statisticians performing the content analysis had an intuition of where to look.

For example, the first readability indicator (Lively and Pressey, 1923) counted the number of distinct words per 1 000 words, the number of non-common words that are not on a list of 10 000 common words in English writing (Thorndike, 1921), and the median of the index of a common word on that list. The challenge of classifying articles on their complexity is part of a more general challenge of text categorization. Although historical exceptions exist (e.g., rule-based approaches), the data-driven approach is the dominant, if not exclusive, approach taken to this problem (Sebastiani, 2002).

Although there is a clear value for indicators of textual complexity, its use is currently limited to a descriptive (reflective) approach. When used as norm, texts do not necessarily become more readable (as attribute of the text) when their readability score (as analyzed by an algorithm) increases. The shortcomings to the readability indicators when used as a norm were demonstrated by Davison and Kantor (1982). They showed with four texts that a clear increase of (subjective) readability, achieved through a higher discourse cohesion, actually led to a decrease in (objective) readability scores. This is due to the exclusive use of shallow, data-driven features such as sentence length. However, shortening sentences does not necessarily lead to lower complexity, yet statistically speaking more readable passages use fewer words (Benjamin, 2012).

To operationalize the notion of textual complexity, a distinction is proposed between comprehension and processing difficulty. Comprehension is usually the aim of readability indicators, which originates from a long history of educational application. Within this setting, readability indicators are compared to the results of a Cloze test in which the x^{th} word is left out for test subjects to fill in. This test gives an indication of whether participants comprehend a text. LaBerge and Samuels (1974) perfectly illustrate the innate difficulty of predicting comprehension, stating that *"the complexity of the comprehension operation appears to be as enormous as that of thinking in general."* (p. 320). It is of little surprise that shallow, data-driven proxies of this operation provide limited (predictive) validity.

To overcome the challenge of predicting comprehension, this chapter introduces the notion of processing difficulty as a proxy in between textual complexity and comprehension. Processing difficulty refers to the measurable effort required to process a new token of information (Jaeger and Tily, 2011). The complexity of a text, then, is defined by the processing difficulty it creates. Three levels of processing can be discerned: word, sentence, and discourse. At a word-level, word decoding and lexical access influence processing difficulty; at a sentence-level syntactical analysis and semantic interpretation; and, at a discourse-level sentence integration and inference processing (Perfetti, 1988).

The relation between processing difficulty and comprehension is somewhat equivocal: less-than-optimal word processing may already be sufficient for good discourse comprehension (Long et al., 2006b). This indicates a difference between processing difficulty and comprehension, where word-level effects may contribute to processing difficulty but not per se hurt comprehensibility. Reflecting on the wide scope of applications of a metric of textual complexity, comprehensibility is just one of the possible consequences: learning, interest, and relevance judgments all take processing difficulty as a determinant, separately from comprehensibility.

Congruent with the paraphrased statement of LaBerge and Samuels (1974), processing difficulty is better understood than comprehension (Long et al., 2006a) and, therefore, provides a viable alternative account of textual complexity. The use of processing difficulty as an intermediate goal allows for the implementation of common findings about the causes of processing difficulty: that is, to bridge the gap between between readability and psycholinguistics.

This chapter presents a model of textual complexity that deviates from the purely data-driven approach that is usually taken for the identification of textual complexity in particular and the categorization of texts in general. Yet, the model also deviates from a traditional theoretical approach. Figure 3.1 illustrates this position. Instead of creating a model of the differences between easy and difficult texts, and instead of deriving a model directly from an extensive theory on processing difficulty, this intermediate approach models the known causes of (subjective) ease or difficulty that a reader can experience during the processing of a text; that is, it unifies contemporary findings or "small data". A similar approach has been posited before; for example, Kaplan (1972) proposed a model of sentence comprehension "*based on some common observations about our sentence processing abilities*" (p. 80). Furthermore, a similar approach has been shown to increase the precision of psychiatric document retrieval, where subjective constructs such as negative life events, depressive symptoms, and semantic relations between symptoms were analyzed and added to an otherwise data-driven retrieval model (Yu et al., 2009).

To contrast the contemporary findings to the data driven approach ("big data"), they are referred to as "small data" in Figure 3.1. An example of this small data is the finding that words of high lexical familiarity, which indicates how familiar a reader is with a word, decrease a reader's fixation duration and are more likely to be skipped (Inhoff and Rayner, 1986; Reichle et al., 1998). These findings in themselves are inspiration for and explained by numerous theories and (deep) models, including Morton (1969)'s Logogen, Coltheart et al. (2001)'s DRC (dual route cascade), and Seidenberg and McClelland (1989)'s PDP (parallel distributed processing). This shows that an intermediate model and related theories are essentially two sides of the same coin: both take the same observations on

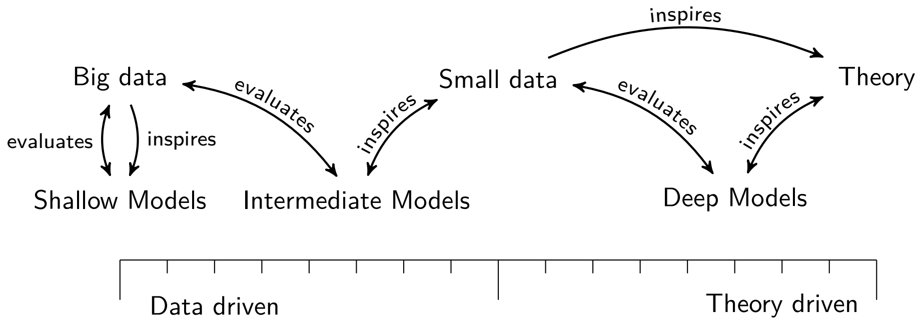


Figure 3.1: The position of the intermediate approach in between data-driven and theory-driven approaches, and its relations to big data, small data, and theory.

stimulus-response contingencies[4] as input. However, the latter tries to explain the response, whereas the former models those aspects of a stimulus (i.e., the data) that are associated with a response. In turn, the focus on the stimulus allows its application to "big data" (see Figure 3.1).

Contrary to small data is big data, which involves data sets of ever increasing size (Jacobs, 2009). Although currently sizes up to several exabytes are common, this is not achievable for the extensive analysis required to detect textual complexity. As Jacobs (2009) notes: "*The pathologies of big data are primarily those of analysis*" (p. 39). The reviewed applications of textual complexity indicate the data sizes that are common for these applications. The sizes range from several books to (a subset of) the internet. Although the intermediate approach is applicable to big data, the bottleneck for any model that is closer to the theory-driven approach will be the speed of analysis. This will inevitably form a limit to the current approach as well, creating a trade-off between speed and accuracy. To account for this limit, Figure 3.1 places the proposed model in between big and small data, implying that the model is applicable to big data but its application adheres to limits in processing power.

Academically, the intermediate approach can, eventually, lead to stronger (weak) AI. This idea follows from the theory-driven approach, which states that it can more aptly represent (human) intelligence and, therefore, allows us to build more intelligent systems. Essentially, this approach is a nuanced approach to AI, adding some findings about intelligence to the statistical approach. Hence, no theory about intelligence, but a model of some observations about intelligence. Or, in

[4] See p. 43 for a definition.

other words, not actual intelligence, but a resemblance of intelligence. In some ways this acknowledges the difficulty of actually describing and modeling intelligence, and circumvents this problem not by purely a bottom-up approach but circumvents it by modeling basic findings about intelligence instead of intelligence itself. Furthermore, the intermediate approach suffers less from the problems of the theory-driven "*toy systems*" (Wilks, 1999, p. 166) that operate well only on small data and the data-driven statistical systems that lack theoretical underpinnings (see Figure 3.1). The approach is aimed to be a next step towards "*a well-designed well-motivated state of the art system*" (Wilks, 1999, p. 166), that operates on big data, yet incorporates (some findings about human) intelligence.

The performance of the intermediate approach can be evaluated against the shallow approach, for both the indication of textual complexity in specific and the categorization of texts in general. From the previous outline it follows that the intermediate approach should have a performance benefit with comparison to shallow models for the identification of textual complexity. However, as King (1996) states: "*each evaluation is very specific to a particular system, and, perhaps even more importantly, to a specific environment in which the system should work*" (p. 79). This implies that the state-of-the-art performance on text categorization only marginally applies as benchmark for the performance of models of textual complexity. Nevertheless, in order to have practical value the intermediate approach should have a comparable performance with respect to text categorization as well.

This chapter will introduce a model of textual complexity based on psycholinguistic findings about (the causes of) processing difficulty. The outline of the remainder of the chapter is as follows: Section 3.2 describes related work, in particular the state-of-the-art achievements in readability detection in order to set a benchmark performance for the model. Section 3.3 reviews observations about human processing difficulty and derives a set of features from these key determinants. The resulting set of features is then described in detail in Section 3.4. These features will be trained and tested on a large, split data set, the methodology of which is described in Section 3.5 and the results in Section 3.6. Besides the overall performance (Section 3.6.1), the results section will delineate the effect of text length on the model performance (Section 3.6.2). And, to examine whether the features do indeed measure different aspects of the data set, a dimensionality reduction will be performed and interpreted in view of the theoretical underpinnings (Section 3.6.3). Finally, Section 3.7 discusses and interprets the results.

3.2 Related Work

The state-of-the-art of automatic readability detection is difficult to concretize due to the incomparability of different, often proprietary, data sets. However, the data-dependent performance can be reviewed, as well as the theories and techniques used to achieve this performance. Following the categorization by Benjamin (2012), three types of techniques will be reviewed: (1) traditional methods, (2) methods based on the use of statistical language modeling tools, and (3) methods inspired by cognitive science. Moreover, the performance of a combination of these techniques will also be reviewed.

Traditional accounts of predicting the processing difficulty of information mostly consisted of readability formulas. These formulas use basic word and sentence features (e.g., length) to generate ordinal scores allowing the ranking of texts. Mostly used within an educational context, these scores often approximate a required reading grade level. The Flesch-Kincaid Grade Level formula is a popular example that uses the average number of words per sentence and the average number of syllables per word as a measure for readability. The formula is defined as follows: $-15.59 + .39\frac{\text{number of words}}{\text{number of sentences}} + 11.8\frac{\text{number of syllables}}{\text{number of words}}$ (Kincaid et al., 1975). Congruent with other studies (e.g., Schwarm and Ostendorf, 2005), this formula will later on be used as baseline (see Section 3.6.3). Although not actively pursued anymore, the most recent versions of traditional methods explain 71% to 85% of variance (R^2) in Cloze test results (Dubay, 2004).

Language models generally perform best for text categorization. It gives the probability that a particular word or sequence of words is written using a certain language (model). Distinguishing 5 grade levels performance peaked at 79% classification accuracy when applied to the same data set. This performance was only achieved for the biggest group of data (i.e., the fifth grade level) and precision and recall were skewed (Schwarm and Ostendorf, 2005). Performance ranged from 63% to 67% when trained on as many as 12 grade levels and tested on 6 grade levels using different, although similar, data sets (Collins-Thompson and Callan, 2005). The problem of this dependency of language models on the genre of text (e.g., educational material, scientific, news, etc.) has been approached by training genre-specific language models. Within genres, the maximum correlation with a 5 points readability scale was $r = .817$ (Kate et al., 2010). Since language models are trained on a specific data set they are dependent on the availability of a representative data set. A language model only gives the chance that a text is created using the modeled language. Hence, it is not a model of readability, though it can be applied as such when, for a certain group of users and a certain genre of texts, a representative data set is available.

Cognitively inspired methods have been employed in just a few systems to detect textual complexity, often profiting from contemporary systems for syntactic-semantic analysis. Syntactic-semantic analysis can be used for analyzing more specific criteria such as co-reference resolution, POS analysis, and topic detection. Coh-Metrix is a well-known system, applying cognitively inspired indexes to indicate the cohesion of a text (Graesser et al., 2004). Combined with a few lexical measures, such techniques explained 76.3% of variance in textual cohesion using discriminant analysis on a two-class problem with a small data set of 38 items (McNamara et al., 2010). Applied to readability research, Crossley et al. (2008) shows how a model based on three criteria of the Coh-Metrix system correlates highly ($r = .925$) with Cloze test results for 31 texts. Another more recent system is DeLite (Vor der Brück et al., 2008). It uses syntactic-semantic analysis to heighten its content validity. Compared to the traditional Flesch–Kincaid formula, DeLite's readability predictions correlated more highly with participants' difficulty ratings ($r = .43$ vs. $r = .53$, respectively). However, the predictions only accounted for 28% of the variance among ratings. The difference in performance between Coh-Metrix and DeLite can be explained by the data sets used (31 vs. 500 texts), as well as by the applied statistical methods which can lead to overfitting.

The most successful approaches combine traditional methods, language models, and syntactic-semantic analysis to predict readability. Feng et al. (2010) combined a very wide range of criteria to achieve a classification accuracy of 74%, based on a skewed data set with a baseline accuracy of 37.8%. However, an F-measure or other weighted precision-recall measure, both important for classification performance on skewed data sets, was not reported. Similarly, Kate et al. (2010) achieves a correlation of $r = .817$ using features based on both syntactic-semantic analysis and language models.

Although the state-of-the-art on predicting readability is already achieving high correlations and classification performance, it is difficult to conclude on the actual state-of-the-art performance. Three evaluative problems explain this difficulty:

- The size of the data set. A small data set can heighten the relative importance of some characteristics, causing the effectiveness of just a few shallow features.
- The length of the texts. Besides being a factor that correlates with textual complexity itself (easier texts are often shorter as well), the length of the text is highly influential on the achieved performance.
- A lack of a separation between training and test sets. This leads to an overfitting of the metric to the data at hand, implying poor predictive performance on other data sets (i.e., predictive validity) (Babyak, 2004).

To enhance the comparability of the efforts in classifying textual complexity, three guidelines are identified. Namely, to have a large data set, control for text length, properly split between training and test sets. To truly test the effectiveness of the proposed features these guidelines will be adhered to in the evaluation performed in Section 3.6. Furthermore, to contribute to the comparability a large publicly available data set will be introduced in Section 3.5.1.

Although there is a clear potential of using knowledge about the causes of processing difficulty to increase the predictive validity, the previous review shows that this potential has yet to materialize (Benjamin, 2012). The best performing methods have low content validity, which questions what is modeled: text or complexity. Often, the features used are data-driven and can therefore be expected to be more representative of language than of complexity. Considering the complexity assessment is seen as part of text categorization, increased content validity has to be able to compete with shallow models on the performance on text categorization tasks.

3.3 Textual Determinants of Processing Difficulty

In order to be able to predict processing difficulty its determinants need to be outlined first. Theories and (deep) models of reading and understanding have been developed at various levels of granularity. There are models focusing on words, connections between words, sentences, and discourses (Rayner and Reichle, 2010). These models will only be touched upon briefly. Instead, the common observations underlying these models will be described and features will be presented which reflect these observations. Describing the technical details of these features will be postponed to Section 3.4. As the goal is to give an indication of processing difficulty over the whole discourse, when possible, word-level and sentence-level effects will be extrapolated to the discourse level.

3.3.1 Word Effects

The word is a logical starting point in explaining variation in the reading process, being a particularly well-defined and tractable token of text; higher levels such as sentences and discourses will inherently lead to more variation (see Section 3.3.3 and Section 3.3.4). Throughout the following descriptions, references will be made to two commonly used tasks to assess word-level abilities: the naming task, the rapid pronunciation of words and pseudowords; and the lexical decision task, the decision which of two letter strings is correctly spelled.

Word Length

Word length is a classic approach to inferring readability, having a central role in nearly all readability formulae. The importance of word length is confirmed with longer words taking more time for perceptual identification (McGinnies et al., 1952), giving higher fixation durations during reading (Just and Carpenter, 1980), decreasing naming performance (Balota et al., 2004), and decreasing lexical decision performance (New et al., 2006), and shorter words being more likely to be skipped while reading (Brysbaert et al., 2005). These results are, however, moderated by lexical familiarity (explained furtheron) (Balota et al., 2004).

Word length is generally defined as the number of characters per word. For completeness with the traditional readability formulae the number of syllables per word will be defined as well:

$\text{len1} = |c \in w|$, word length in characters c per word w;

$\text{len2} = |s \in w|$, word length in syllables s per word w.

Lexical Familiarity

Lexical familiarity indicates how familiar a reader is with a word. It influences a reader's fixation duration, where more frequent words take less initial processing time (Inhoff and Rayner, 1986). This effect is even observed when controlling for word length, number of syllables, and bigram and trigram frequency. And, high-frequency words are more likely to be skipped than less frequent words (Reichle et al., 1998).

The most salient measure of lexical familiarity is printed word frequency. Other measures are age of acquisition and subjective familiarity. Both were shown to capture unique but overlapping aspects of word reading time (Juhasz, 2005). Moreover, the effect of word frequency has been found to correlate and interact with variables such as conceivability and concreteness, in a way that for less frequent words other aspects become more apparent (Balota et al., 2006), confirming the robust role of printed word frequency. Hence, printed word frequency will be defined as a proxy for how familiar a reader is likely to be with a word; that is, as a measure of lexical familiarity:

$\text{fam} = \log_{10}\text{cnt}(w)$, the logarithm of the term count cnt per word w.

For the term count function cnt, a representative collection of writing is needed. In this study the Google Books N-Gram corpus will be used, see Section 3.5.2. The use of a logarithm is congruent with Zipf's law of natural language, stating that the frequency of any word is inversely proportional to its rank in a frequency table

(Zipf, 1935). Although this measure resembles the inverse document frequency (idf) metric that is common in IR, word frequency metrics are common in psycholinguistics and studied extensively in relation to processing difficulty. Hence, to ensure construct validity the printed word frequency feature is preferred.

Measures related to word frequency have been implemented in other systems as well. Notable are the Coh-Metrix system, which also takes a logarithm of the word frequency derived from four different sources of frequency counts (Graesser et al., 2004), and the Dale list of 3000 common words for English as applied in the Dale-Chall readability formula (Dubay, 2004).

Connectedness

Word identification is not only grounded in lexical variables. The semantic interpretation of a word has been shown to influence word identification on unique aspects, different from measures of lexical familiarity (Balota et al., 2004). These effects are generally explained from a connectionist theory, such that better connected items are more easily retrieved from memory. One functional explanation of these effects comes from applying graph theory to (partly) model semantic memory. Steyvers and Tenenbaum (2005) show a correlation between the time at which a word (node) first joins a graph network and the number of connections it ultimately acquires. Since any new word added to the network is connected to at least one existing word, older words will have more connections. This effect is mediated by the frequency (utility) of a word, such that frequent words are more likely to connect with new words. Using a lexical decision and naming task, the connectedness of words explain a large and unique proportion of variance.

The degree of a node is indicative of its connectedness. It represents the number of connections a node has to other nodes in the network. Similarly, the indegree is used for directed graphs and counts the number of incoming connections a node has from other nodes in the network. Based on two different semantic models (WordNet as semantic lexicon (Miller, 1995) and Explicit Semantic Analyses (ESA) as topic space (Gabrilovich and Markovitch, 2009)) two different features of connectivity will be defined:

con1 $= |A_n(w)|$, the node degree within n steps of a word w in a semantic lexicon (Equation 3.10, p. 86).

con2 $= C \circ T(w)$, the in-degree of the topic $T(w)$ of a word w in topic space (Equation 3.15, p. 88 and Equation 3.19, p. 90).

3.3.2 Inter-word Effects

Besides characteristics of the words themselves influencing their processing difficulty, relations between nearby words and the target word influence its processing difficulty as well. The connections between words have been found on multiple representational levels. Here, two representational levels will be described: orthographic and semantical.

The inter-word effects have a strong relation to the discourse effects of cohesion. In particular on a pragmatic level, measures of cohesion are generally implemented as the similarity of a sentence to nearby sentences (Lapata and Barzilay, 2005; Graesser et al., 2004), whereas the inter-word measures look at the similarity between (a sequence of) words and nearby words, or even between characters and nearby characters. Hence, overlap between the two types of measures can be expected. Section 3.3.4 further explains and compares cohesion measures.

Character and Word Density

Numerous studies of priming have shown that a target string is better identified when it shares letters with the prime. This holds for both identity priming (repeating the prime) as well as form priming (using a partly different string). These effects remain when the prime is masked (i.e., unconscious) and are most effective for actual words (as compared to nonwords), indicating a lexical interpretation of the effects (Humphreys et al., 1982). Although more vulnerable when extrapolated to a sentential or discourse context, the lexical repetition effects remain. This is confirmed by eye-tracking studies as well, where, within a meaningful context, word repetition decreases early eye fixation measures indicative of lexical access (Ledoux et al., 2006).

From an information theoretic point of view, repetition creates a form of redundancy which can be measured in terms of entropy. Entropy is a measure of the uncertainty within a random variable. It defines the number of bits needed to encode a message, where a higher uncertainty requires more bits (Shannon, 1948). Since the aim is not to measure text size but instead, to measure uncertainty, a sliding window will be applied within which the local uncertainty will be calculated. This Sliding Window Entropy (SWE) gives a size invariant information rate measure, or in other words, an information density measure. Text with a higher repetition of symbols will have a lower entropy rate. Using a sliding window f_w (Equation 3.4, p. 84) of entropy H_n (Equation 3.2, p. 83) and probability mass function (PMF) $p(x)$ (Equation 3.6, p. 85) two features are defined using either characters or words as symbols:

cha$_n$ = $f_w(X)$ with $f(X) = H_n(X)$, a sliding window of n-gram entropy using PMF $p(x)$ where X is an ordered collection of characters x.

wor$_n$ = $f_w(X)$ with $f(X) = H_n(X)$, a sliding window of n-gram entropy using PMF $p(x)$ where X is an ordered collection of words x.

The SWE has several benefits over other size-corrected measures. Firstly, when correcting for the size by calculating the entropy ratio (i.e., the ratio between measured entropy and the entropy of a uniform distribution), the influence of text size on the distribution is still profound: longer samples will have an inherently different distribution compared to shorter ones. Secondly, psycholinguistic effects of priming are vulnerable to distance: further away primes are less effective (Ledoux et al., 2006), making SWE a measure more in accordance with psycholinguistic priming effects.

Semantic Density

In a seminal study, Meyer and Schvaneveldt (1971) showed that subjects were faster in making lexical decisions when word pairs were related (e.g., cat-dog) than when they were unrelated (e.g., cat-pen). Since then, studies have shown consistent findings that words are better recognized when embedded in a semantically related context (Hutchison, 2003). During reading, these findings are not replicated unless the distance between prime and target is small (e.g., the same part of the sentence). However, extrapolated to a discourse level the effects are strong: a word congruent with the discourse is recognized and processed faster (Ledoux et al., 2006).

For indicating the congruence of words with the discourse, a measure of entropy is proposed within the ESA topic space. In a topic space each dimension represents a topic. Within this topic space, the discourse is described as the centroid of the concept vectors of each of the individual words Abdi (2009). Based on the resulting concept vector, the entropy can be calculated using the topics (i.e., dimensions) as symbols. The entropy of the centroid concept vector indicates the semantic information size of the discourse; that is, the number of bits needed to encode the discourse in topic space. For example, with less overlap between individual concept vectors, the uncertainty about the topic(s) is higher, resulting in a higher entropy.

Since an increase in text size will lead to a higher uncertainty and, thus, a higher entropy, a metric of the global discourse is mainly a measure of the (semantic) text size. Similar to Section 3.3.2, the aim is not to measure text size but, instead, only to measure information rate. Hence, a sliding window will be applied. The resulting SWE describes topical uncertainty within the local discourse. Using the local discourse assures size-invariance and, accordingly, gives the (average) relatedness of the words to their local discourse.

Using a sliding window f_w (Equation 3.4, p. 84) of entropy H (Equation 3.1, p. 83), a converter to topic space $T(X)$ (Equation 3.17, p. 89), and PMF $p(t)$ (Equation 3.20, p. 90), a measure of entropy in topic space can be defined:

> sem $= f_w(X)$ with $f(X) = H \circ T(X)$, a sliding window of topical entropy using PMF $p(t)$ over topics t conveyed in $T(X)$, where X is an ordered collection of words x.

3.3.3 Sentence Effects

Theories of sentence-level effects on processing difficulty fall into two broad categories: memory-based accounts, explaining difficulty due to some limited resource, and constraint-satisfaction accounts, explaining difficulty by the probability of a processed structure: infrequent or unexpected words or structures are harder to process. A prominent effect will be discussed for each category: the effect of dependency-locality and of surprisal. Although progress has been made in more fine-grained theories of sentence complexity (e.g., derivational entropy; (Hale, 2003; Roark et al., 2009)), those are not as easily scalable as the effects discussed next.

Dependency-Locality

The Dependency Locality Theory (DLT) states that a reader, while reading, performs a moment-by-moment integration of new information sources. This implies there is an evolving structure kept in mind, keeping track of what has just been just read (i.e., storage costs). Next to keeping the evolving structure in mind, it is the integration of new information into the current structure which requires resources (i.e., integration costs). Hence, the bigger the structure, or the larger (longer) the connections within the structure, the more is used of a limited resource (Gibson, 2000). The theory has been shown to account for differences in reading time across a range of linguistic effects (Lewis et al., 2006). The DLT is a particularly interesting theory since it explains the processing cost of commonly used sentences, contrary to syntactically ambiguous ones which are often used in scientific settings. Moreover, its computation is fast and accurate, using state-of-the-art POS taggers and dependency resolvers.

For commonly used sentences, integration costs are the main cause of difficulty: "*reasonable first approximations of comprehension times can be obtained from the integrations costs alone, as long as the linguistic memory storage used is* *not excessive at these integration points*" (Gibson, 1998, p. 19). In other words, when the load of remembering previous discourse referents is not exceeding storage capacity, memory costs will not be significant. Normally, such excessive storage requirements will be rare. Hence, the focus will be on integration costs alone. Integration costs were found to be dependent on two factors. First, the type of the element to be integrated, where new discourse elements require more resources than established ones. Second, the distance between the to be integrated head and its referent, where distance is measured by the number of intervening discourse elements. Section 3.4.4 operationalizes these observations about integration costs:

> loc = $I(D)$, sentential integration costs where D is the collection of dependencies within a sentence (Equation 3.13, p. 87).

There is a relation between locality and sentence length. A sentence consisting of more words is likely to have more dependencies to connect these words. Hence, sentence length is expected to correlate with integration cost. To allow a comparison between the traditional feature of sentence length and the proposed feature of sentential integration costs, the following feature is defined:

> wps = $|w \in X|$, the number of words w in a sentence X.

Surprisal

Constraint-satisfaction accounts use the informativeness of a new piece of information to predict its required processing effort. The resources needed to process a piece of information are related to its informativeness: more information takes more resources to process. Surprisal assumes multiple options are activated simultaneously, where each new piece of information constrains the array of possibilities: the probability mass decreases. A higher decrease in probability mass then leads to higher processing complexity and longer reading times (Jaeger and Tily, 2011). Translated to words, the information a word adds to a text is given by its surprisal: if a word is very unlikely to occur, its surprisal and thus its information value is high. On the other hand, the more predictable a new token is, the less information it adds, and the less demand it puts on processing the new information.

Models that focus on words can be used to base a measure of surprisal on. Such models capture both lexical and syntactical effects and give an important simplification from more a sophisticated representation. Because of computational problems arising with calculating the probability of very long sequences of words (e.g., sentences), typically the probability of a word only depends on the n previous words (e.g., three), also referred to as the "trigram assumption". In practice, this simplification has been shown to work well. Already bigram probabilities are capable of partly predicting reading times (McDonald and Shillcock, 2003). Perplexity is a common metric of inverted sentence probability: the more predictable the sentence is, the less the surprisal will be. A normalized version will be reported, giving the surprisal (alternatives) per word:

> $\text{sur}_n = PP_n(X)$, n-gram perplexity, where X is a sentence consisting of N words x, $X = \{x_i : i = 1, \ldots, N\}$ (see Equation 3.3, p. 84).

Key to perplexity is the training corpus used to base the PMF $p(w)$ on (see Equation 3.6, p. 85). As a representative collection of writing, this study uses the Google Books N-Gram corpus (see Section 3.5.2).

3.3.4 Discourse Effects

On a discourse level, a reader interacts with a text to form a mental representation of the described situation. The creation of this mental representation, or situation model, is an interplay between the information provided by the text and the background knowledge of a user (Kintsch and van Dijk, 1978). The integration of incoming information with the current situation model can be facilitated either by lexical cues showing how it should be integrated (connectives, Section 3.3.4) or by facilitating the degree to which a reader can connect incoming information with prior information either in the stimulus or in memory (coherence, Section 3.3.4).

Connectives

Connectives, such as “although”, “as”, “because”, and “before” are linguistic cues helping the reader integrate incoming information. They give cues as to what association should be built between linguistic units. A text supplying more cues on how the reader should integrate the information is more readable, influencing reading times and comprehension. In particular causal associations have been shown to benefit from explicit connectives (Sanders and Noordman, 2000).

Most connectives belong non-exclusively to three syntactic categories: conjunction, adverbs, and prepositional phrases. Of these, subordinate conjunctions give the best approximation of connectives. A subordinate conjunction explicitly connects an independent and a dependent clause. Although this excludes most additive connectives, where there is (often) not a clear independent clause making these words categorized as (more general) adverbs, this includes the most beneficial type of causal connectives (Fraser, 1999).

con $= p(\{\text{subordinate conjunction}\}, X)$, the ratio of subordinate conjunctions to words in a text X (see Equation 3.11, p. 87).

Cohesion

Coherence is a direct function of the degree to which comprehenders can connect information they are processing, with prior information either in the linguistic stimulus or in memory. Cohesion is a discourse characteristic linked to coherence (Benjamin, 2012). Morris and Hirst (1991) argued that cohesion is formed by lexical chains; that is, sequences of related words spanning a discourse topic. A cohesive text can then be formalized as having dense lexical chains. Such chains can be identified on a semantic level and on a linguistic level (i.e., cohesion). On a semantic level this can be operationalized as semantic relatedness: the more related a new piece of information is to the foregoing information, the easier it is to integrate the new information. On a linguistic level anaphora indicate which instances refer to the same (linguistic) entity. They provide cues on how to relate incoming information to the active mental representation of a text.

Coherence has been related to the comprehension of a text as well as its reading time. High cohesion requires less reading time, presumably because fewer inferences have to be made, and lead to heightened comprehension. However, this effect is not always straightforward: the background knowledge of the reader interacts with this effect, such that deep comprehension improves for highly knowledgeable readers when reading less coherent texts. This holds at a semantic level (McNamara and Kintsch, 1996; Kintsch, 1994) as well as at a linguistic level (Degand et al., 1999; Ozuru et al., 2009).

Equation 3.5 (p. 84) provides a generic way to calculate the local cohesion of a text ($C_n(X)$). Defining two types of similarity $sim(s_i, s_j)$, Equation 3.5 (p. 84) can be used to identify:

coh1$_n$ $= C_n(X)$, local cohesion over n foregoing sentences s in a discourse X using anaphora-based connections $sim(s_1, s_2)$ (Equation 3.14, p. 88).

coh2$_n$ $= C_n(X)$, local cohesion over n foregoing sentences s in a discourse X using semantic-based relatedness $sim(s_1, s_2)$ (Equation 3.18, p. 90).

As already explained in Section 3.3.2, an overlap exists between the measures derived from inter-word effects and the measures of cohesion. On a theoretical level there is a clearer distinction between the two. On the one hand, cohesion models the connections between pivotal discourse referents important for building a coherent mental representation of the text. On the other hand, the inter-word effects are caused by the similarity or overlap between a target word and nearby words, which eases the processing of the target word. The anaphora-based cohesion measure is a direct measure of the connections between discourse referents and can, therefore, be expected to better reflect discourse cohesion instead of an inter-word effect.

3.4 Models and Equations

The features suggested in the previous section (Section 3.3) used numerous equations and models, the details of which will be described next. Four representational models will be described: n-grams, semantic lexicons, phrase structure grammars, and topic models. Each model represents a different aspect of information, creating a complementary set of representations. Furthermore, a few common equations will be described first, which can be defined irrespective of the underlying representational model.

3.4.1 Common Methods

Three types of common methods will be described: entropy, sliding window, and cohesion.

Entropy

Entropy is a measure of the uncertainty with a random variable. It defines the amount of bits needed to encode a message, where a higher uncertainty requires more bits. Entropy can be directly calculated from any probability distribution. Consider the random variable X with PMF $p(x)$ for every value x. Then, the entropy is defined as (Shannon, 1948):

$$H(X) = -\sum_{x \in X} p(x) \log_2 p(x) \tag{3.1}$$

For longer sequences entropy can be defined as well. If we define a range of variables $X_1, \ldots, X_n$ with PMF $p(x_1, \ldots, x_n)$ giving the probability for a sequence of values $x_1, \ldots, x_n$ occurring together, then the joint entropy is given by (Cover and Thomas, 2006):

$$H_n(X_1, ..., X_n) = -\sum_{x_1 \in X_1} ... \sum_{x_n \in X_n} p(x_1, ..., x_n) \log_2 p(x_1, ..., x_n) \tag{3.2}$$

The range of variables $X_1, \ldots, X_n$ can be equal to the variable X (i.e., $H_n(X_1, \ldots, X_n) =$

$H_n(X)$), such that the PMF $p(x_1, \ldots, x_n)$ indicates the probability of the sequence $x_1, \ldots, x_n$ in X (see Section 3.4.2).

Perplexity is a different notation for entropy which is most commonly used for language modeling purposes (Goodman, 2001). Perplexity is an indication of the uncertainty, such that the perplexity is inversely related to the number of possible outcomes given a random variable. It is defined as following:

$$\mathrm{PP_n}(X) = 2^{H_n(X)} \tag{3.3}$$

Generally, perplexity is normalized by the number of (sequences of) symbols.

Sliding Window

For entropy calculations, the size N of the variable X will inevitably lead to a higher entropy: more values implies more bits are needed to encode the message. A sliding window, calculating the average entropy per window over the variable X, creates a size-invariant measure; that is, the (average) information rate. Given the variable $X = \{x_i : i = 1, \ldots, N\}$, any function f over X can be rewritten to a windowed version f_w:

$$f_w(X) = \sum_{i=w}^{N} \frac{1}{N-w} f \circ \{x_j : j = i - w + 1, \ldots, i\} \tag{3.4}$$

Depending on the type of entropy, different functions f can be used. Here, three implementations will be given: standard entropy $f(X) = H(X)$, n-gram entropy $f(X) = H_n(X)$, and entropy in topic space $f(X) = H \circ T(X)$. Both standard entropy and n-gram entropy use the PMF $p(x)$ defined in Section 3.4.2, whereas topical entropy is defined with PMF $p(t)$ (see Equation 3.20, p. 90).

Local Cohesion

Independent of the level of analysis, cohesion measures share a common format based on a similarity function $sim(x_i, x_j)$ between two textual units x_i and x_j. Although the type of units does not require a definition, only sentences will be used as units. Let X be an ordered collection of units, then the local cohesion over n nearby units is:

$$C_n(X) = \sum_{i=1}^{|X|} \sum_{j=\max(1,i-n)}^{i-1} \frac{1}{i-j} \mathrm{sim}(x_i, x_j) \tag{3.5}$$

This includes a weighting factor ($\frac{1}{i-j}$), set to be decreasing with increasing distance between the units: connections between closeby units are valued higher. This is in line with Coh-Metrix (Graesser et al., 2004) and the DLT (Gibson, 2000), who posit that references spanning a longer distance are less beneficial to the reading experience.

Two similarly measures $sim(x_i, x_j)$ will be used for Equation 3.5, both measuring a similarity between sentences: coreference similarity (see Equation 3.14, p. 88) and semantic similarity (see Equation 3.18, p. 90).

3.4.2 N-Grams and Language Models

The goal of a language model is to determine the probability of a sequence of symbols $x_1 \ldots x_m$, $p(x_1 \ldots x_m)$. The symbols are usually words, where a sequence of words usually models a sentence. The symbols can be more than words: for example, phonemes, syllables, and the like. n-Grams are employed as a simplification of this model, the so-called trigram assumption. n-Grams are subsequences of length n consecutive items from a given sequence. Higher values of n in general lead to a better representation of the underlying sequence. Broken down into components, the probability can be calculated and approximated using n-grams of size n as following:

$$p(x_1 \ldots x_m) = \prod_{i=1}^{m} p(x_i | x_1, \ldots, x_{i-1}) \approx \prod_{i=1}^{m} p(x_i | x_{i-(n-1)}, \ldots, x_{i-1}) \tag{3.6}$$

The probability can be calculated from n-gram frequency counts as follows, based on the number of occurrences of a symbol x_i or a sequence of symbols $x_1 \ldots x_m$ of n-gram size n:

$$p(x_i) = \frac{c(x_i)}{\sum_x c(x)} \quad \text{if n=1} \tag{3.7}$$

$$p(x_i | x_{i-(n-1)}, \ldots, x_{i-1}) = \frac{c(x_{i-(n-1)}, \ldots, x_{i-1}, x_i)}{c(x_{i-(n-1)}, \ldots, x_{i-1})} \quad \text{if } n > 1 \tag{3.8}$$

The frequency counts can be based on a separate set of training data (e.g., Google Books N-Gram corpus, See Section 3.3.3) or on an identical set of training and test data (e.g., a random variable X). The former can lead to zero probabilities for a (sequence) of values, when a value from the test set does not occur in the training set (i.e., the model). To this end, smoothing techniques are often employed. For more information on language models and smoothing techniques we refer to Goodman (2001).

3.4.3 Semantic Lexicon

A semantic lexicon is a dictionary with a semantic network. In other words, not only the words but also the (type of) relationships between words are indexed. In a semantic lexicon a set of synonyms can be defined as φ; then the synonym sets related to a word w is:

$$A_0(w) = \{\varphi \in W | w \in \varphi\} \tag{3.9}$$

where W stands for the semantic lexicon and the 0 indicates no related synsets are included.

Node Degree

Continuing on Equation 3.9 (p. 86), the node degree of a word can be defined. All the synonym sets related in n steps to a word w represent the connectiveness, and are given by:

$$A_n(w) = A_{n-1}(w) \cup \{\varphi \in W | r(\varphi, \varphi') \wedge \varphi' \in A_{n-1}(w)\}. \tag{3.10}$$

where $r(\varphi, \varphi')$ is a Boolean function indicating whether there is any relationship between synonym set φ and synonym set φ'.

The number of synonym sets a word is related to within 1 step is the node degree $A_1(w)$ of that word (Steyvers and Tenenbaum, 2005). The definition supplied in Equation 3.10 (p. 86) is different from the node degree as defined by Steyvers and Tenenbaum (2005), for it combines polysemic word meanings and, therefore, is the node degree of the set of synonym sets $A_0(w)$ related to a word w (see Equation 3.9, p. 86) instead of the node degree of one synonym set (cf. Steyvers and Tenenbaum, 2005). Moreover, the node degree as defined in Equation 3.10 (p. 86) is generalized to n steps.

3.4.4 Phrase Structure Grammar

A phrase structure grammar describes the grammar of (part of) a sentence. It describes a set of production rules transforming a constituent i of type ξ_i to constituent j of type ξ_j: $\xi_i \rightarrow \xi_j$, where i is a non-terminal symbol and j a non-terminal or terminal symbol. A terminal symbol is a word, non-terminal symbols are syntactic variables describing the grammatical function of the underlying (non-)terminal symbols. A phrase structure grammar begins with a start symbol, to which the production rules are applied until the terminal symbols are reached; hence, it forms a tree.

For automatic parsing a Probablistic Context-Free Grammar (PCFG) will be used (see Section 3.5.2), which defines probabilities to each of the transitions between constituents. Based on these probabilities, the parser selects the most likely phrase structure grammar; that is, the parse tree. The parse tree will further on be denoted as P.

Syntactic Categories

Each terminal node is connected to the parse tree via a non-terminal node denoting its POS. The POS indicates the syntactic category of a word, for example, being a verb or a noun (Marcus et al., 1993). The constitution of a text in terms of POS tags can simply be indicated as following. Let u be the unit of linguistic data under analysis (e.g., a text T), let $\mathrm{cnt}(u, y)$ be the number of occurrences of POS tag y in u, and let $\mathrm{cnt}(u)$ be the total number of POS tags in u. Then, the ratio of POS tags Y compared to all POS tags in u is:

$$p(Y, u) = \sum_{y \in Y} \mathrm{cnt}(u, y)/\mathrm{cnt}(u) \tag{3.11}$$

Locality

Dependencies exists between different nodes in the parse tree P, indicating a relation between parts of the sentence. The collection of dependencies in a parse tree P will be denoted as D and a dependency between node a and node b as d. Using the definition of the DLT (Gibson, 2000), the length of (or integration cost of) a dependency d is given by the number of discourse referents in between node a and node b, inclusive, where a discourse referent can be (pragmatically) defined as a noun, proper noun, or verb (phrase). Defining u as the POS tags of the terminal nodes between and including node a and b of a dependency d, the dependency length is given by:

$$L_{\mathrm{DLT}}(d) = \mathrm{cnt}(u, \{\mathrm{noun}, \mathrm{proper\ noun}, \mathrm{verb}\}) \tag{3.12}$$

where $\mathrm{cnt}(u, Y) = \sum_{y \in Y} \mathrm{cnt}(u, y)$ is the number of occurrences of POS tag y in u.

The integration costs of a whole sentence containing dependencies D is then defined as:

$$I(D) = \sum_{d \in D} L_{\mathrm{DLT}}(d) \tag{3.13}$$

Co-references

Essentially, co-reference resolution identifies and relates mentions. A mention is identified using the parse tree P, usually a pronominal, nominal, or proper noun phrase. Connections are identified using a variety of lexical, syntactic, semantic, and discourse features, such as their proximity or semantic relatedness. Using the number of referents shared between sentences, a similarity measure can be defined. Given the set of referents R and a Boolean function $m(r, s)$ denoting true if and only if a referent r is mentioned in a sentence s, a sentence similarity metric based on co-references is then:

$$\mathrm{sim}(s_i, s_j) = |\{r \in R | m(r, s_i) \wedge m(r, s_j)\}| \tag{3.14}$$

Using Equation 3.5 (p. 84), this similarity metric can indicate textual cohesion.

3.4.5 Topic Model

A topic model is a model of the concepts occurring in a (set of) document(s). This can be derived without any previous knowledge of possible topics; for example, as is done with Latent Dirichlet Allocation (LDA), which discovers the underlying, latent, topics of a document. This approach has a few drawbacks: it derives a preset number of topics, and giving a human-understandable representation of the topics is complicated (Blei and Lafferty, 2009). On the contrary, a “fixed” topic model, consisting of a set of pre-defined topics, can also be used. Such a model does not suffer the mentioned drawbacks but is less flexible in the range of topics it can represent.

ESA will be used for a fixed topic model. It supports a mixture of topics and it uses an explicit preset of possible topics (dimensions). This preset is based on Wikipedia, where every Wikipedia article represents a topic dimension. A single term x (word) is represented in topic space based on its term frequency - inverse document frequency (TF-IDF) value for each of the corresponding Wikipedia articles d_n:

$$T(x) = [\mathrm{ti}(x, d_1)\, \mathrm{ti}(x, d_2)\, \ldots\, \mathrm{ti}(x, d_n)], \tag{3.15}$$

where n is the number of topics (i.e., articles) and $\mathrm{ti}(x, d_j)$ is the tf-idf value for term x. It is given by:

$$\begin{aligned}
\mathrm{ti}(x, d_j) &= \mathrm{tf}(x, d_j) * \mathrm{idf}(x) \\
\mathrm{tf}(x, d_j) &= \begin{cases} 1 + \frac{\log c(x,d_j)}{|d_j|} & \text{if } c(x, d_j) > 0 \\ 0 & \text{else} \end{cases} \\
\mathrm{idf}(x) &= \log \frac{n}{|\{d_j : x \in d_j\}|},
\end{aligned} \tag{3.16}$$

where $c(x, d_j)$ gives the number of occurrences of term x in document d_j and $|d_j|$ gives the number of terms in document d_j. Hence, it is a regular inverted index of a Wikipedia collection which underlies the topic model: the topic vector $T(x)$ is a TF-IDF vector of a word (query term) x.

Using Wikipedia as the basis for the topics gives a very broad and up-to-date set of possible topics, which has been shown to outperform state-of-the-art methods for text categorization and semantic relatedness (Gabrilovich and Markovitch, 2009).

Topic Centroid

The topics covered in a text fragment is defined as the centroid of the vectors representing each of the individual terms. Given a text X containing n words $\{x_i : i = 1, \ldots, n\}$, the concept vector is defined as following (Abdi, 2009; Gabrilovich and Markovitch, 2009):

$$T(X) = \sum_{i=1}^{n} \frac{1}{n} T(x_i) \tag{3.17}$$

The centroid of topic vectors gives a better representation of the latent topic space than each of the individual topic vectors. Combining vectors leads to a synergy, disambiguating word senses. Consider two topic vectors $T_1 = [a, b]$ and $T_2 = [b, c]$ which each have two competing meanings yet share one of their meanings (i.e., b). This shared meaning b will then be favored in the centroid of the two vectors $[\frac{1}{2}a, b, \frac{1}{2}c]$, essentially disambiguating the competing senses (Gabrilovich and Markovitch, 2009).

Semantic Relatedness

The ESA cosine similarity measure has been shown to reflect human judgments of semantic relatedness, with a $r = .75$ correspondence as compared to a $r = .35$ for WordNet-based semantic relatedness (Gabrilovich and Markovitch, 2009). Hence, for $\mathrm{sim}(x_1, x_2)$ the ESA semantic relatedness measure is defined:

$$\mathrm{sim}(x_1, x_2) = \frac{T(x_1) \cdot T(x_2)}{\|T(x_1)\| \, \|T(x_2)\|}, \tag{3.18}$$

where $\|T(x)\|$ is the norm of topic vector $T(x)$ for a linguistic unit x. The exact unit is undefined here, for it can be words or combinations of words (e.g., sentences). A sentence-level semantic relatedness measure can be used as input for Equation 3.5 (p. 84), defining a semantic cohesion measure.

In Degree

A measure of connectedness based on the ESA topic model is implemented using the number of links pointing to each topic (i.e., Wikipedia article). Thus, if a topic is central to a wide range of topics (i.e., having a lot of incoming links), it is considered a common, well-connected, topic (Gabrilovich and Markovitch, 2009). Let I be a vector in topic space containing for each topic t the number of links i_t pointing to that topic. The connectedness of a topic vector T is then defined as:

$$C(T) = \log_{10}(T \cdot I) \tag{3.19}$$

A logarithmic variant is used based on graph theory, stating that with any self-organized graph few nodes are highly central and many nodes are in the periphery, showing a log-log relation (Barabási and Albert, 1999).

Probability Distribution

The probability distribution over topics is easily derived from a topic vector (e.g., a centroid vector). Considering each of the elements of a topic vector are TF-IDF weights indicating the relevance of the element, the relative importance of an element can be derived by comparing the TF-IDF weight of the element to the TF-IDF weights of all elements in the topic vector. That is, the probability of an element t (i.e., a topic) in a topic vector $T = [t_1, \ldots, t_n]$ is defined as its relative weight:

$$p(t) = \begin{cases} \frac{t}{\sum_{i=1}^{n} t_i} & \text{if } t \in T \\ 0 & \text{otherwise} \end{cases} \tag{3.20}$$

Using the probability distribution $p(t)$ for every topic t in a topic vector T, the entropy $H(T)$ can be calculated (see Equation 3.2, p. 83).

3.5 Evaluation Methodology

To indicate the power of the proposed features, we compared two data sets which are overall similar yet highly distinctive in their expected processing difficulty. Namely, Simple English Wikipedia and (normal) English Wikipedia. Moreover, the size of these data sets is a first indication of the scalability of the proposed features.

3.5.1 Data Set

In order to evaluate the proposed features, a data set with a clear diversity in expected processing difficulty was needed. One relatively large data set perfectly suited for this aim is the Wikipedia encyclopedia. The Wikipedia encyclopedia is among the best encyclopedias available, containing user-generated content with a quality close to that of, for example, the Encyclopedia Britannica (Giles, 2005). Wikipedia is available in many languages, among which normal English and simple English. For the latter, the authors are instructed to write using easy words and shorter sentences, but not to be less informative. This data set is expected to represent how the authors viewed processing difficulty. However, one cautionary comment should be made. The articles tend to be smaller than their English Wikipedia counterparts, leading to less depth by which a topic is discussed.

The Wikipedia dump of August 3, 2011 was used. Only articles were selected which were found in both languages, allowing for a pair-wise comparison. The resulting data set consisted of 69 395 pairs of both English and simple English articles. That is, a total of 138 790 articles, 398 718 sections, and 1 459 370 paragraphs. Of this data set, only articles which were neither a stub (i.e., an incomplete article) nor a special, redirect, or disambiguation page were selected. Moreover, only the oldest 20 000 articles were used for classification purposes, the underlying assumption being that more matured articles better reflect the actual intended writing style of both Wikipedia versions.

The following preprocessing steps were performed on the data set. First, the original obtained data consisted of two dumps, containing all articles encoded as wiki-text for each language. Both sets were imported into a MySQL database, using JWPL. All dumps were retrieved on August 29, 2011. Second, all articles were parsed to plain text using JWPL. All templates and links to files and images were removed. Third and finally, all feature were extracted using a Hadoop cluster of 16 64-bit dual-core 3.00GHz machines each with 8GB memory.

3.5.2 Feature Extraction

Each of the features were extracted from one or more of the representational models: n-grams, semantic lexicon, phrase structure grammar, and topic model (see Section 3.4). The common implementation details and the implementation details per representational model will be described.

For all features, the Stanford CoreNLP word and sentence tokenizers were used (cf. Toutanova et al., 2003). Words were split up into syllables using the Fathom toolkit (Ryan, 2012). Where possible (i.e., for n-grams, semantic lexicons, and topic models), the Snowball stemmer was applied (Porter, 2001), reducing each word to its root form. As model representative for common English the Google Books N-Gram corpus was used (Michel et al., 2011). For each (sequence of) words, the n-gram frequencies were summed over the years starting from the year 2000.

Concerning n-grams, entropy was calculated on n-grams of length $n = 1 \dots 5$ and windows of size $w = 25$ for words and $N = 100$ for characters. The window size was based on a limit to the minimal required text length (in this case 100 characters or 25 words) and a trade-off between, on the one hand, psycholinguistic relevance (i.e., a stronger effect for nearer primes) and, on the other hand, a more reliable representation. As input for the SWE algorithm, next to characters, stemmed words were used. The stemmer was applied in order to reduce simple syntactical variance and, hence, give more significance to the semantic meaning of a word.

As a semantic lexicon (W in Equation 3.9, p. 86) we used WordNet (version 3.0). WordNet is a collection of 117 659 related synonym sets (synsets), each consisting of words of the same meaning. The synsets are categorized according to their part-of-speech, either being a noun, verb, adverb, or adjective. Each synset is connected to other synsets through (one of) the following relations: an antonymy (opposing-name), hyponymy (sub-name), meronymy (part-name), holonymy (whole-name), troponymy (manner-name; similar to hyponymy, but for verbs), or entailment (between verbs) (Miller, 1995). Before matching a word to a synset, each word was stemmed. Moreover, stopwords were removed from the bag of words. Due to memory limitations a maximum n of 4 was used for the connectedness features based on the semantic lexicon (see Section 3.3.1). For computing the semantic lexicon based cohesion measure (see Section 3.3.4), the St-Onge distance measure was calculated using as parameters $C = 6.5$ and $k = .5$.

For parsing sentences to a PCFG the Stanford Parser was used (Klein and Manning, 2003). Before PCFG parsing, the words were annotated with their POS tags. POS detection was performed using the Stanford POS tagger. The resulting PCFG trees are used for co-reference resolution using the Stanford's Multi-Pass Sieve Coreference Resolution System (Lee et al., 2011). Due to computational complexity, co-reference resolution is limited to smaller chunks of text and, therefore, was only calculated for paragraphs.

The ESA topic model was created using a Lucene (Hatcher et al., 2010) index of the whole English Wikipedia data set. This data was preprocessed in a similar fashion as described in Section 3.5.1, with the exception that stopwords were removed and all terms were lemmatized using the Snowball stemmer (Porter, 2001). This led to a total of 3,734,199 articles or topic dimensions. As described by Gabrilovich and Markovitch (2009), each term in the index was normalized by its L2-norm. Moreover, the index was pruned as follows: considering the sorted TF-IDF values related to a term, if over a sliding window of 100 values the difference is less than five percent of the highest TF-IDF value for that term, the values are truncated below the first TF-IDF value of the window (Gabrilovich and Markovitch, 2009, p. 453).

The features were computed on article, section, and paragraph level. Wikipedia's built-in structure was used to select each. For those features extracted at a smaller granularity, such as at a sentence or word level, the results were aggregated by deriving their statistical mean. This assured that parameters indicative of the length of the text (e.g., the number of observations, the sum, or the sum of squares) were not included, reducing the influence of article length.

3.5.3 Classification

The classification pipeline consisted of four steps: preprocessing, feature selection, classification, validation. As preprocessing, features containing more than 25% of missing values were removed, after which any observations containing missing values were removed.

A lenient feature selection was performed mainly to speed-up the classification process. It consisted of removing features having zero or near-zero variance and those features too highly correlated with other features. Firstly, features were removed if the frequency ratio was above 95/5, that is the frequency of the most prevalent value over the second most frequent value. Secondly, features were removed if the percent of unique values was below 10%, that is the number of unique values divided by the total number of samples (times 100%). Thirdly, from all pairs of features correlating more than $r = .9$ with each other, the feature having the lowest mean absolute correlation with all other features was removed.

Three types of classifiers will be used: Random Forest (RF), Support Vector Machine (SVM), and Logistic Regression Model (LRM). RFs and SVMs are among the best classifiers, showing state-of-the-art performance in benchmark studies (e.g., Meyer et al., 2003).

A RF is a bagging technique, building many decision trees based on random selections of features (Breiman, 2001a). As such it is a feature selection technique as well, making separate feature selection unnecessary. However, the employed (lenient) feature selection speeds up the tuning of hyperparameters. The classifier was tuned on two hyperparameters: the number of features randomly sampled as candidates at each split, and the minimum size of terminal nodes. Respectively, the ranges 1 to $log_2(k+1)$, where k is the number of features, and 1 to 10, were used. The number of trees was set to 100, found to be an optimal amount (cf. Meyer et al., 2003).

SVM places observations as vectors in a feature space, after which it uses a hyperplane to separate two classes of input. The optimal hyperplane, creating the largest space between the two classes of observation, is solved using Lagrange multipliers minimizing the deviation from the linearly separable case. Considering not all features are relevant to this optimization, the so-called support vectors remain. For non-linear cases a kernel function is applied. We applied a radial basis kernel function, defined as:

$$k_G(x_i, x^l) = \exp\left(-\gamma |x_i - x^l|^2\right) \tag{3.21}$$

where x_i is a feature vector that has to be classified, x^l is a feature vector assigned to a class (i.e., the training sample). Please note that the radial basis function is a variant on the Gaussian kernel function. The classifier is tuned on two parameters: γ, valued $10^{-6 \cdots -1}$, and cost, range $10^{1 \cdots 2}$.

As benchmark classifier a LRM was created. A LRM explains how p independent variables (predictors, x) predict the dependent, binary, variable y via the probability p on y being either true or false. In order to do so, m optimal weighting factors Beta (B) over each of the $i = 1, \ldots, n$ observations are determined:

$$\text{logit}(p_i) = \log \frac{p}{1-p} = B_0 + B_1 x_{1,i} + \cdots + B_m x_{m,i} \tag{3.22}$$

The left hand side of the equation, the logit function, is used to map the odds of y being true or false to a variable that ranges between $(-\infty, +\infty)$, matching the potential range of the right hand side of the equation. The method used to determine the average optimal weighting factors over all n observations is the iteratively reweighted least squares method. To select the best possible subset of predictors, both forward and backward stepwise search through feature space was applied. Through an iterative process, this algorithm removes and adds predictors optimizing the Bayesian information criterion, a measure of information loss (either through bias or variance) of a model (Schwarz, 1978). As tuning parameter $k = log_{10}(n)$ was set, where n is the number of observations, specifying a penalty for the number of variables included in a model.

There is a multitude of possibilities to evaluate the classification performance, of which four will be reported: accuracy, the area under the receiver operator curve, the F1-score, and the Phi or Matthew's correlation (cf. Powers, 2011). This collection of performance metrics is in line with often used (e.g., accuracy) and state-of-the-art (e.g., Matthew's correlation) metrics. Accuracy is the most common measure of performance, simply giving the percentage of correctly classified instances compared to the total of test instances. However, this measure does not look at precision, recall, or skewness. The F1-score is a weighted harmonic mean of precision, the number of true positives divided by the number of all positives, and recall, the number of true positives divided by the number of results that should have been returned (true positives and false negatives). The F1-score still leaves out any indication of how well the classifier handles negative cases. The Area Under Curve (AUC) is an all-round measure, giving the probability that the classifier will score a randomly drawn positive sample higher than a randomly drawn negative sample. However, its practical value has been called into question. The final, Matthew's correlation, includes both true and false positives and negatives and is robust against skewed class-distributions (Powers, 2011). Hence, Matthew's correlation is preferred. However, if all measures indicate similar results, common accuracy can be used in line with common practice.

All steps were implemented using R, a statistical package (Ihaka and Gentleman, 1996), with the packages 'randomForest' for classification, the 'e1071' package for machine learning tools, the 'caret' package for data preprocessing and feature selection, and the 'rocr' package for classifier evaluation. The SVM was based on the LibSVM implementation (Chang and Lin, 2011).

3.6 Results

The developed model of textual complexity was tested on its ability to classify texts correctly as originating either from Simple English Wikipedia or from (normal) English Wikipedia. The following results will be presented: the classification performance (Section 3.6.1), an analysis of the influence of text length on the classification performance (Section 3.6.2), and an analysis of the power of each of the features in relation to each other and in relation to the baseline Flesch-Kincaid formula (Kincaid et al., 1975) (Section 3.6.3).

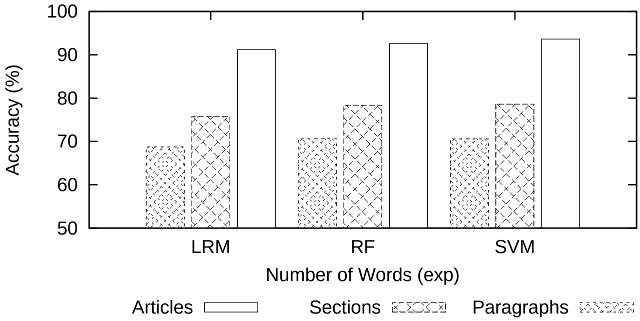


Figure 3.2: Accuracy per data set and classifier.
Note. Used abbreviations: Logistic Regression Model (LRM), Random Forest (RM), Support Vector Machine (SVM).

3.6.1 Classification Performance

To validate the performance a classifier was trained on 80% and tested on 20% of the data. As data, 20 000 observations were randomly selected of both articles, sections, or paragraphs from both data sets. The data was balanced to contain an equal portion of both simple and normal observations. The 80%-20% split was chosen to reduce computational load, and will likely not have affected the performance significantly due to the still relatively large size of the data set.

The data was split and, after pre-processing and balancing the data, three classifiers were trained and tested on the articles ($N = 38\,773$), sections ($N = 39\,642$), and paragraphs ($N = 39\,806$). The classification performances are shown in Figure 3.2. The best performance was achieved on classifying articles on their complexity with an accuracy of 93.62% using a SVM. Classification accuracy for sections and paragraphs was substantially less, respectively: 78.60%, also using a SVM classifier, and 70.60% using a RF classifier. Overall, the best performance was achieved using a SVM classifier and there seems to be a profound influence of text length on the performance.

Several tests confirm the results depicted in Figure 3.2. The F1-score, the AUC, and Phi or Matthew's correlation are shown in Table 3.1. Please note the scale of Phi or Matthew's correlation is between -1 and 1. All indicators of classification performance show a similar result in comparison to the accuracy measure, confirming the success of the model to differentiate between two levels of complexity on a large scale.

Table 3.1: Classification performance descriptors per data set and classifier.

		Descriptors*				
Data set	Classifier**	N	Acc	AUC	F1	Phi
articles	LRM	38773	91.20%	.965	.912	.824
articles	RF	38773	92.61%	.975	.927	.852
articles	SVM	38773	93.62%	.979	.936	.872
sections	LRM	39642	75.80%	.832	.756	.516
sections	RF	39642	78.36%	.867	.780	.568
sections	SVM	39642	78.60%	.866	.793	.573
paragraphs	LRM	39806	68.72%	.744	.684	.375
paragraphs	RF	39806	70.60%	.781	.689	.415
paragraphs	SVM	39806	70.55%	.777	.721	.414

Note. * Descriptors: Data set size (N), Accuracy (Acc), Area Under Curve (AUC), F1-score (F1), and Phi or Matthew's correlation (Phi)

Note. ** Classifier abbreviations: Logistic Regression Model (LRM), Random Forest (RM), Support Vector Machine (SVM).

3.6.2 Influence of Length

The difference in performance between articles, sections, and paragraphs suggests a profound effect of text length. This is to be expected, as more data leads to a more reliable estimation of the constructs measured by the different features. Moreover, text length is a differentiating factor itself between the two data sets. Overall, the articles of the simple data set are smaller than those of the normal data set. Although this difference is theoretically supported, such that an increase in the amount of information can be expected to approach some sort of a limited resource (Jaeger and Tily, 2011), the robustness of the features to analyze smaller texts is of interest (see Section 3.1).

The relation between text length, as defined in number of words, and classification performance, defined as accuracy, is shown in Figure 3.3. The graph shows the accuracy per bin, where each bin consists of texts of a certain range of sizes: that is, articles, sections, or paragraphs having a number of words coherent with a specified range. The lower and upper limit of words per bin is determined by ordering all texts according to their size and selecting each subsequent range containing at least 500 "simple" texts and 500 "normal" texts. The size of the bins is depicted in Figure 3.3 on the horizontal axis using error bars to denote the standard deviation around the mean text length of the texts in the bin. Since the

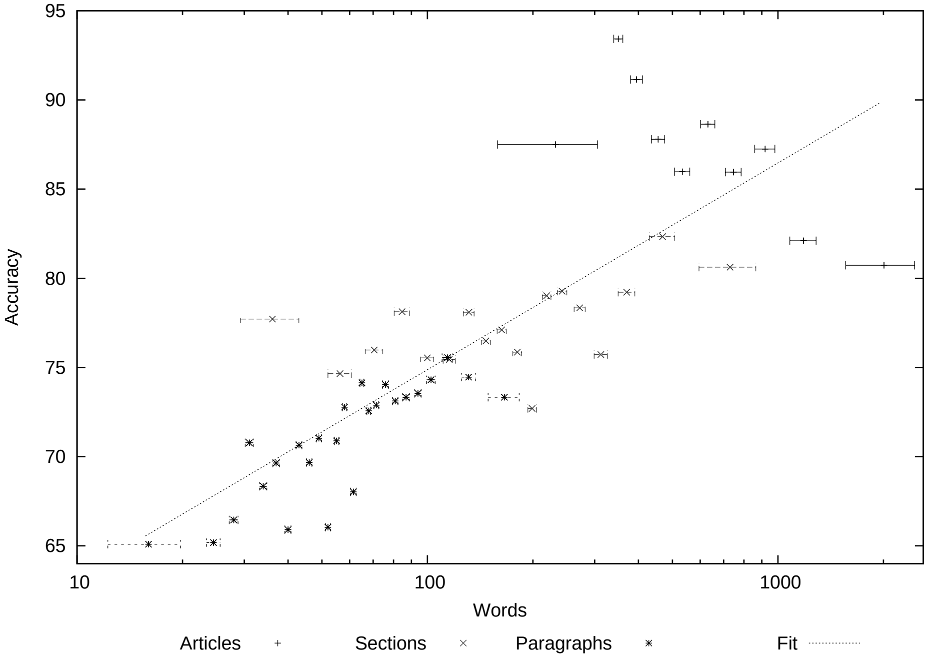


Figure 3.3: Relation between classification accuracy and text length.

distributions of length between the "simple" and "normal" texts differ, the head and tail bins for each of the text types (i.e., articles, sections, and paragraphs) cover a relatively wide range of size in order to have sufficient overlap between the distributions (i.e., 500 texts for each). Hence, the head and tail bins show a relatively high standard deviation on the horizontal axis.

Figure 3.3 shows a log-linear relation between text length and classification accuracy. A fit-line was included to illustrate this relation. The fit-line is approximated using a least-squares regression analysis converging in four iterations to the following formula:

$$\mathrm{acc}(x) = 51.71 + 5.03 * \log_{10}(len(x)). \tag{3.23}$$

Here, acc refers to classification accuracy for a bin x, len is a function giving the average number of words for all texts in the bin x. The fit-line clearly shows, over all types of texts, an increase in accuracy with an increase in text length.

Although the graph indicates an increase with length, this increase seems to reach a ceiling at a peak performance of 93.42%. This is a remarkable peak performance, as it is for the bin with the biggest overlap between the distributions of the sizes of the “simple” and “normal” texts. This is signified by one of the smallest standard deviations in size for all article bins. Hence, with the least benefit from differences in text length, it still reaches the highest performance. Seemingly, articles with a length of around 350.50 words have a most profound difference in complexity.

3.6.3 Evaluation of Features

The classification performance indicates that the data differs profoundly on the proposed features. A more precise indication of the power of subsets of the features can be given via the explained variance or R-square metric. In particular, this affords to compare the feature set underlying the model of textual complexity to the feature set used in the Flesch-Kincaid formula (Kincaid et al., 1975). Since the outcome variable of the classification is a binary logistic variable (denoting either English Wikipedia or Simple English Wikipedia), only a pseudo-R-square can be reported. The contemporary Nagelkerke-R-squared (R^2_N; Nagelkerke, 1991) will be used as pseudo-R-square metric. For articles, the R^2_N is .333 for the Flesch-Kincaid formula and .800 for all features, indicating a substantial increase in explained variance over this popular baseline. For sections R^2_N is .305 and .420 and for paragraphs R^2_N is .168 and .231, both for the Flesch-Kincaid formula and all features respectively. The explained variance confirms the heightened challenge in classifying the subsets of the data containing the sections and paragraphs. Furthermore, it shows the value of the proposed intermediate features in comparison to the traditional approach.

To analyze how the features relate to each other a principal component analysis with varimax rotation was employed (Abdi and Williams, 2010). To keep a level of clarity in reporting, only the analysis of the articles will be reported. The principal component analysis was performed after the data was preprocessed (see Section 3.5.3). Table 3.2 shows the principal components and their feature loadings. A total of 34 components were needed to fully cover all the variance in the data. Of these 34 components, 10 had an Eigen value of more than 1, together explaining 90.13% of the total variance. Figure 3.4 shows the explained variance per component and accumulated over components.

The first seven components show a clear division of features over the components. Per component: semantic density and coherence; short sequenced orthographic density; anaphoric coherence; WordNet-based connectedness; long sequenced orthographic density; surprisal; and word length. Noteworthy is the com-

Table 3.2: The loadings of the features for each principal component.

	Component									
Feature	1	2	3	4	5	6	7	8	9	10
Word features										
len1	.226	.133	-.085	-.038	.006	.088	**.894**	-.055	.123	-.108
len2	.197	.084	-.098	-.138	.031	.070	**.916**	.004	.090	-.063
fam	-.042	-.116	-.172	.103	.012	-.113	-.274	-.188	-.141	**.779**
con1_0	.046	-.002	.050	**.556**	-.003	-.056	-.253	-.640	-.039	.159
con1_1	-.030	-.006	.044	**.836**	.014	-.030	-.108	-.148	-.026	.104
con1_2	.033	-.057	.057	**.929**	.002	-.025	-.043	-.079	.035	-.004
con1_3	-.018	-.094	.056	**.918**	.041	-.009	.061	.185	.002	-.049
con2	-.036	-.180	-.192	.157	-.030	-.072	-.023	**.794**	-.025	.194
Inter-word features										
cha_1	-.107	.136	-.121	-.298	-.152	.007	-.444	**.694**	.000	-.194
cha_2	.066	**.849**	-.031	-.192	-.019	-.001	-.114	.316	.061	-.147
cha_3	.125	**.957**	.056	-.030	.138	-.010	.094	-.017	.074	.019
cha_4	.103	**.932**	.048	.013	.216	-.015	.026	-.038	.059	.119
cha_5	.090	**.903**	.026	.030	.277	-.012	-.014	-.063	.036	.145
wor_1	.109	**.849**	.042	-.082	.134	.035	.161	-.125	.030	-.167
wor_2	.051	**.760**	-.028	-.003	.538	.009	.124	-.076	-.013	.051
wor_3	.029	.536	-.074	.035	**.797**	-.004	.080	-.076	-.029	.040
wor_4	.020	.353	-.125	.042	**.912**	-.008	.028	-.054	-.013	.009
wor_5	.006	.227	-.205	.006	**.917**	-.017	-.041	.013	.003	-.004
sem	**.726**	.336	.073	-.013	-.001	.023	.219	-.032	.316	-.046
Sentence features										

loc	.197	.066	.035	.003	-.009	.025	.082	.020	**.904**	-.098
wps	.236	.089	.012	.002	-.019	-.068	.108	.002	**.886**	.106
sur$_1$	.016	.010	-.010	-.023	.012	**.922**	.045	-.015	-.022	-.016
sur$_2$	.014	-.004	-.010	-.022	-.016	**.982**	.044	-.013	-.005	-.038
sur$_3$	.011	-.009	-.004	-.030	-.018	**.955**	.053	-.014	-.009	-.066
Discourse features										
con	.014	.222	.029	-.001	.043	-.048	.096	.441	.201	**.681**
coh1$_1$	.119	.085	**.891**	.062	.064	-.018	-.092	-.098	.002	-.019
coh1$_2$	.063	.023	**.964**	.043	-.013	-.002	-.050	-.050	.014	-.049
coh1$_3$	.040	.001	**.962**	.043	-.151	-.004	-.027	-.050	.026	-.041
coh1$_4$	.031	-.017	**.881**	.043	-.323	-.003	.019	-.058	.019	-.014
coh2$_1$	**.989**	.059	.048	.001	.015	.008	.071	-.020	.073	-.001
coh2$_2$	**.990**	.060	.048	.002	.015	.008	.072	-.021	.073	-.001
coh2$_3$	**.990**	.061	.049	.002	.014	.008	.072	-.021	.073	-.001
coh2$_4$	**.990**	.062	.049	.002	.014	.008	.072	-.021	.073	-.002
coh2$_5$	**.989**	.063	.049	.002	.014	.008	.073	-.021	.073	-.002

Note. For each feature, the highest load factor is highlighted in bold.

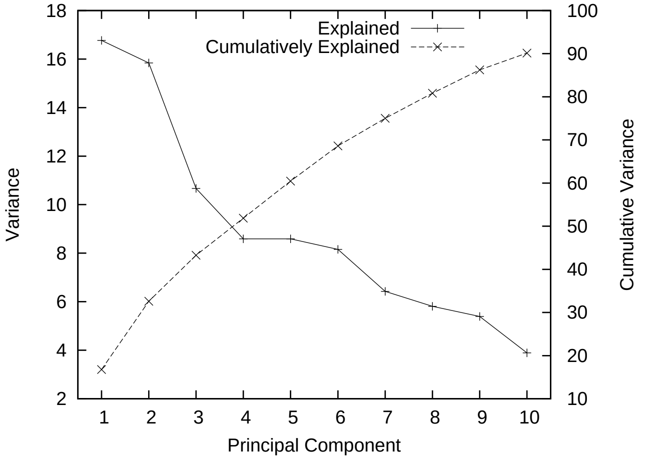


Figure 3.4: Explained variance by each of the principal components.

bination of semantic coherence and semantic density in one component. Already explained in the description of the inter-word features (see Section 3.3.2), there is a considerable overlap between inter-word features and measures of cohesion. The first seven components explain the heap of variance, a total of 75.03%, confirming the value of the intermediate features[5].

The final three components show a less clear division of features. The components consist of: 1-gram character density and ESA-based connectedness; locality; and word familiarity and connectives. For the eighth component, the combination of character density with ESA-based connectedness is surprising. Seemingly, the connectedness of words correlates with the distribution of characters over these and neighboring words. Some relation between the two features can be expected, since both are of a lexical nature. The ninth component shows a combination of features in line with expectations (see Section 3.4.4). Namely, locality combined with words per sentence. It is remarkable that this suggests that the current implementation of locality has little benefit over the shallow feature of words per sentence. For the tenth component, the relation between word familiarity and connectives is somewhat remarkable. This relation can be explained, since it is not unlikely words used as connectives have a higher printed word frequency.

The division of features over components confirms the content validity of the features: with a few exceptions, each of the components measures a unique as-

[5] Note that this does compare directly to the Nagelkerke-R-square previously reported.

pect of textual complexity in line with expectations. Figure 3.4 shows a healthy distribution of the explained variance over the primary components, where all ten components have an important contribution to explaining the total of 90.13% of the variance. Together, Figure 3.4 and Table 3.2 confirm the validity as well as the value of the intermediate features in explaining differences in the data.

3.7 Discussion

In the preceding sections, an approach was taken to model textual complexity intermediate of data-driven and theory-driven approaches. This intermediate model was developed based on common observations (i.e., “small data”) about (the causes of) processing difficulty and trained and tested on a large data set distinctive in complexity (i.e., “big data”). The effectiveness of this approach to differentiate between different levels of complexity was attested to by a classification performance of 93.62%, confirming the value of using the proposed features. This performance was achieved on a set of 38 773 articles. Furthermore, comparing the explained variance of 80.00% by all proposed features to an explained variance of 33.30% by the oft-used Flesch-Kincaid formula (Kincaid et al., 1975), the need to evaluate with large data sets was shown. Seemingly, more subtle differences than those modeled by the shallow features of the Flesch-Kincaid formula become apparent for large, real-world, data sets. Hence, the results suggest a clear advantage of the intermediate approach in tackling the innate difficulty of predicting the complexity of a text.

3.7.1 Interpretation

Although better results have been obtained by other studies (see Section 3.2), these results do not compare in terms of scale or content validity. A combination of three factors show the value of the reported results in perspective to other, similar, studies. First, the large scale of the evaluation assured a representative importance of relevant textual aspects. In comparison, a small test set can heighten the importance of some aspects, leading to a large proportion of explained variance of up to 85% by using shallow features alone (Dubay, 2004). This is a multitude of the achieved performance of 33.30% of the Flesch-Kincaid formula in this study. Second, this study adhered to the standard practice of splitting the data set into a train and a test set. Although standard for general machine learning and textual classification tasks, often merely explanatory figures are reported for readability metrics. Third, the content validity of the used features allows for application to other genres of data as well as a normative indicator. Language models can give a comparable or better performance in textual classification tasks, the focus of this study was not on text classification but on indicating textual complexity. Consequently, the scale, method, and content validity make this a state-of-the-art performance.

Regarding the influence of length, the model of textual complexity achieved a peak performance of 93.62% for articles with a mean length of 350.50 words, showing the model reached an optimal level of performance relatively quickly. Smaller texts showed a decrease in classification performance, suggesting text length has a singular effect on the ability to infer the complexity of a text. Overall, a paragraph could be classified correctly with 70.60% accuracy, a significantly lower percentage than for articles. Two competing explanations exist for the influence of text length on performance. On the one hand, this influence can be explained by a natural ceiling in achievable performance. Comparing the "simple" and "normal" data sets, the difference in complexity of the paragraphs is likely less profound than of the articles. On the other hand, the influence of length can be explained by a lower number of data points to derive a reliable decision from, suggesting the proposed features are not robust for very small portions of texts. Because of the two competing explanations, the robustness of the intermediate model for smaller texts is an open research question. Nonetheless, the model was shown to be already robust for texts around 350.50 words.

The dimensionality of the data was explored using a principal component analysis. The possibility to find dimensions in the data sets is limited to the variance uncovered by the features. Although there are differences in and between the data sets not modeled by the current feature set, the classification performance of 93.62% does indicate a large proportion of key aspects was uncovered. The principal components can therefore be interpreted as key dimensions of the data set. Overall, the dimensions were congruent with the constructs underlying the features, confirming the content validity of the features. Doubts were raised on only two of the features: ESA-based connectedness and 1-gram character SWE. The former is unexpected, since evidence to the contrary exists as well (Gabrilovich and Markovitch, 2009). The latter was expected. Larger n-grams in general lead to a better representation of the underlying model. In this case, higher n-grams give a better approximation of the overlap between characters in the text. Furthermore, the principal component analysis confirmed that inter-word features and cohesion measures are strongly related. The current oft-used implementation of cohesion is as a measure of similarity between a sentence and its neighboring sentences (see Section 3.4.1). This implementation can, depending on the similarity metric used, be considered more closely related to findings on inter-word effects (see Section 3.3.2) than cohesion. This was found to be the case for the cohesion measures based on the ESA similarity metric. Only when a more elaborate similarity measure such as co-reference resolution is used can the results be interpreted as an indication of cohesion.

3.7.2 Limitations

All features were devised in order to exclude any knowledge of semantics as much as possible, assuring its applicability to other data sets and genres of text. Yet, there is a dependency on a more or less common form of the English language. Namely, the English language practiced in: the Penn Treebank corpus used as training data for the PCFG; the Google Books N-Gram corpus used as language model; the vocabulary of WordNet for the semantic lexicon; and Wikipedia, the training data of the ESA model. For most features uncommon has been equaled to complex. For example, lexical familiarity states more frequent words take less initial processing time (see Section 3.3.1). And, constraint-satisfaction accounts of sentence processing state that more frequent sentence constructions are less informative and, therefore, take less processing time (see Section 3.3.3). This suggests the dependency is an important part of any model predictive of processing difficulty, instead of limiting the applicability. However, for the ESA model, the training data of the ESA model overlaps with the test data used in the experiment, suggesting the reliability of the ESA model as used in this experiment might be higher than with other data sets. Overall, although there are some dependencies, the model is applicable to a wide range of data sets.

A dilemma exists between large-scale content-based and small-scale user-centered evaluations. The former assures a correct representation of the importance of those textual aspects which are relevant to the detection of textual complexity. The latter, although it can heighten the importance of some (irrelevant) textual aspects, indicates the predictive validity of the model. By modeling the content, the model of textual complexity represents a writer's perspective on complexity rather than a user's experience of it. This writer's perspective can be expected to be of high quality, considering the quality of Wikipedia is found to be almost as good as that of professional encyclopedias such as Britannica (Giles, 2005). Since everybody is allowed to discuss and revise the user-generated content on Wikipedia, including end users and experts on writing style, it can even be expected that the authors' perspective on complexity is closely coupled to a user's experience of complexity. Furthermore, the features underlying the model are based on an extensive body of user studies on the core determinants of processing difficulty (see Section 3.3). Hence, the large-scale content-based evaluation of this study does not guarantee the predictive validity of the model. Yet, the writer's perspective, in particularly being user-generated content, combined with the content validity of the features makes it likely the model does indeed reflect a user's experience of complexity.

The use of intermediate features for modeling textual complexity has one drawback: the computational complexity is higher than with traditional features. Dependent on the available computational capacity and the size of the data sets to be analyzed, this creates a trade-off between added value and added computational complexity. The performed principal component analysis gives insight into this trade-off. For example, for the feature of locality, the added value in comparison to the traditional feature of words per sentence was shown to be limited. The number and length of dependencies in a sentence is, in the current data sets, highly related to the length of the sentence. The rest of the features did uncover unique aspects of the data sets. Hence, the trade-off between computational complexity and the added value by using intermediate features is, mostly, positive.

3.7.3 Theoretical Implications

This chapter introduced an intermediate approach to solve the challenge of accurately identifying textual complexity. Instead of a purely data-driven model, the intermediate model was based on common observations about processing difficulty. The inclusion of observations about human intelligence increases the validity of a model while still allowing its application to big data. And, by using a large data set to train the model on, the necessity for a unifying theory that explains and connects these observations was removed. Moreover, the intermediate approach allows for reflection on the importance of these observations as well. The amount of variance that is explained by a feature relative to the other available features gives an indication of its significance and the significance of the observations it is based on. For example, although the feature of locality was expected to be more useful to differentiate between commonly used sentences, the feature of surprisal performed better. This finding provides support to the explanatory value of surprisal, given the current data set and setup. In sum, the intermediate approach increases validity in comparison to a purely data-driven approach, increases applicability in comparison to a theory-driven approach, and gives insight into the relevance of the different common observations underlying intermediate and deep models.

To commence the intermediate approach for the detection of readability, processing difficulty was introduced as subjective equivalent to textual complexity. In doing so, the notions of "easy to read" and "easy to comprehend" that define the concept of readability were concretized. It introduced a vast body of research on the causes and effects of processing difficulty (i.e., the stimulus-response contingencies), allowing us to model the causes of processing difficulty into features of textual complexity. The theoretical framework of processing difficulty can serve as a primary framework for further enhancing the validity of features of textual complexity. Aside from the stimulus-response contingencies currently modeled, characteristics of the user can also be included in this relation. The causes of processing difficulty include individual differences such as working memory and topical knowledge (Long et al., 2006a). These individual differences can be used to personalize a model of textual complexity, improving the "depth" of its features and, accordingly, improving its prediction of processing difficulty. With a proper "depth", models of textual complexity can be used in a normative way. This requires that revisions to the processing difficulty of a text are actually reflected in the model, a requirement which does not hold for models that only use shallow features (Davison and Kantor, 1982). Hence, the use of processing difficulty as a proxy offers a unique theoretical framework for work on readability metrics.

3.7.4 Practical Implications

Textual complexity is key to the Information eXperience (IX). A correct level of complexity fosters the experience of interest, has an important role in determining the relevance of a document, and is an important factor for learning. The use of processing difficulty broadens this perspective. The relations between the different levels of processing (i.e., word, sentence, and discourse) and the IX can be explored. For example, regarding comprehension, less-than-optimal word processing is found to already be sufficient for good discourse comprehension (Long et al., 2006a). Similarly, the effect of difficulty at a lower level representation may be beneficial for the experience of the emotion of interest, being easily detectable and, therefore, primarily influencing the initial appraisal of perceived challenge (i.e., pleasantness) of the information. Yet, difficulty at a higher representational level may be detrimental to the experience of interest by negatively influencing its secondary appraisal of comprehensibility (Silvia, 2008b). In general, the use of features based on small data allows for a fine-grained indication of expected processing difficulty. This creates the ability to target a specific experience, but also to cope with specific individual differences.

The relation between processing difficulty and IX can benefit information systems such as information retrieval systems, information filtering systems, and (adaptive) educational systems. The model of textual complexity allows such systems to provide easier or more challenging information aligned to the needs of their users and, accordingly, leading to a more pleasant IX. In order to apply a model of textual complexity into contemporary information systems, the model has to be scalable. The current model was already shown to scale to a large data set consisting of 138 790 articles. However, for really big data sets such as indexed by information retrieval systems (e.g., terabytes of documents) as well as for nearly real-time analysis such as needed by some information filtering systems, the computational complexity of the intermediate model becomes an issue. Moreover, for the PCFG model, more noisy out-of-domain data significantly decreases parsing performance (Petrov et al., 2010). Hence, although some applications are restricted to a less detailed analysis of textual complexity due to computational constraints or noisy data, the model of textual complexity presented here can give a new level of fine-grained readability analyses beneficial to contemporary information systems.

3.8 Conclusion

The state-of-the-art on readability analysis is already achieving high correlations and classification accuracies. Notwithstanding, the scale of analysis, the evaluative method and results, and the content validity of the underlying features make the presented model of textual complexity unique. To achieve an excellent performance on text categorization, yet at the same time enhance the validity of readability prediction, an intermediate approach was taken to the detection of textual complexity and the categorization of text. This approach lies intermediate in between the contemporary data-driven and theory-driven approaches, sharing (part of) the validity of the theory-driven approach and (part of) the applicability of the data-driven approach.

The basis of the presented model was formed by common observations about the causes of processing difficulty (i.e., “small data”), while the model was trained on a large data set distinctive on textual complexity (i.e., “big data”). This allowed for the creation of a unifying model of textual complexity without the necessity of a unifying theory of processing difficulty. By using common observations about processing difficulty, the presented model can not only overcome some of the limitations of popular readability models, but also gives rise to a new set of possibilities for exploring the complicated relationship between textual complexity, individual differences, and IX. Accordingly, the introduced concepts and features pave the way for information systems to improve the IX of their users. The use of the intermediate approach with processing difficulty as a proxy for textual complexity can yield a next level of models of textual complexity, bridging the gap between readability research and psycholinguistics.

4

Explaining Epistemic Interest

Abstract

Interest is one of the most important information emotions: it is associated with curiosity, exploration, information seeking, and learning. The extent to which information systems can promote interest is explored in a user study. This study investigated how complexity, familiarity, and an individual's Epistemic Curiosity Trait (ECT) influenced interest in news articles. As indicator of complexity, the model of textual complexity, developed in Chapter 3, was used. This study served as the final step in the evaluation of this model and, with a correlation of $r = .704$, confirmed its ability to average subjective complexity. The influence of objective textual complexity on interest followed an inverted-U shape, which is a unique confirmation of the Wundt-curve for epistemic textual stimuli. Furthermore, familiarity showed a linear relationship with interest, confirming that the often assumed notion of "more of the same" indeed increases interest. With a total explained variance of 37.80% in interest responses, the study shows the importance of objective indicators as well as subjective appraisals for explaining part of the Information eXperience (IX). Consequently, this chapter forms a proof-of-concept of both the Information eXperience Framework (IXF) and the feasibility of changing the IX.

Parts of this chapter also appear in:

Van der Sluis, F., Glassey, R.J., and Van den Broek, E.L. (2012). Making the news interesting: Understanding the relationship between familiarity and interest. In J. Kamps, W. Kraaij, and N. Fuhr (Eds.), *IIiX 2012: ACM Proceedings of the 4th symposium on Information Interaction in Context*, p. 314–317. August 21–24, Nijmegen, The Netherlands.

Van der Sluis, F., Van den Broek, E.L., Glassey, R.J., Van Dijk, E.M.A.G., and De Jong, F.M.G. When complexity becomes interesting. *Journal of the American Society for Information Science and Technology. [in press]*

4.1 Introduction

When a user employs a query to an Information Retrieval (IR) system or accesses an information feed from an Information Filtering and Recommending (IF&R) system, the user has a motivation to do so. This motivation is usually described as an information need: "*a need for information not existing in the remembered experience of the investigator*" (Taylor, 1962, p. 392). An information need is expected to result from a task, leading to a relatively clear information need (Saracevic, 2007). However, when Taylor (1962) coined the term information need he already identified the existence of a visceral information need in which the need is subconscious and vague.

Typical examples of information behavior, where the information need is less clear are epistemic information interaction and serendipitous information encountering. When users show epistemic information behavior they search information to satisfy their desire for knowledge, instead of bridging an explicit knowledge gap. Hence, users have an intrinsic goal to search information; that is, to learn (Xu, 2007). Serendipity is related to two types of information behavior: information may be encountered unexpectedly, or information may be of unexpected impact to the user (Foster and Ford, 2003). Similar to epistemic information behavior, serendipitous information behavior shows a clear motivation unrelated to the actual information need. These examples evidently raise the question to the causes of exploratory behavior when there is not a clear information need.

Curiosity, the state of being curious, leads to exploratory behavior (Berlyne, 1960; Litman, 2005; Silvia, 2006). Berlyne (1954) defined epistemic curiosity as the "*drive to know*" (p. 187) and went on to define two distinct types of explorative behavior, specific and diversive (Berlyne, 1966). The former, specific curiosity, originates from a perceived knowledge deficit and leads to a clear information need: that is, a "need to know". It is associated with feelings of uncertainty and deprivation, yet the fulfillment of the need can be enjoyable. The latter, diversive curiosity, originates from a lack of stimulation and leads to a visceral information need: that is, a "like to know". It is associated with feelings of interest and anticipated positive stimulation (Berlyne, 1966; Litman, 2005). Both types of curiosity lead to a specific pattern of information behavior and a particular Information eXperience (IX) (see Chapter 4). Hence, according to Berlyne (1966), exploratory behavior that lacks a clear information need is instead motivated by a need for stimulation and the emotion of interest.

Interest, the momentary feeling of being curious, is the primary target of IF&R systems that "*recommend items based on the likelihood that they will meet a specific user's taste or interest*" (Herlocker et al., 2004, , p. 23). And, interest is pivotal to IR systems because it either reflects a visceral information need or a primary reason to search, such as with an epistemic information search (Xu, 2007). Contemporary filtering and recommending techniques are based on the assumption that selecting information based on its topical similarity, selections made by other users, or the characteristics of the user should improve the IX (cf. Konstan and Riedl, 2012). Although such personalization forms an indication of the long-term interests (Ruthven et al., 2007), which has a confirmed importance in explaining relevance decisions, the focus on topicality leads to the so-called serendipity problem in which IF&R systems *"produce recommendations with a limited degree of novelty"* (Ricci et al., 2009, p. 79). And, the filtering methods applied in both IR and IF&R systems give rise to filter bubbles, where filtering primarily produces more of the same. It can be questioned whether or not "more of the same" and "a limited degree of novelty" provide the stimulation that diversive curiosity searches.

A more proactive system might actively attempt to affect the emotional responses of the user, by directly manipulating the determinants of the emotion of interest. As a starting point, this work investigates how the experience of the emotion of interest can be influenced by modifying the complexity and familiarity of the information presented to users.

Interest is regarded to be an emotion associated with curiosity, exploration, information seeking, and learning. People who experience an interest response are attracted to the evoking stimulus (Silvia, 2008b). For example, when textual stimuli raise an interest response, people experience a higher level of arousal and process the text more deeply (Schiefele and Krapp, 1996). Defining interest as an emotion allows one to identify the causes of interest. In other words, it allows one to learn the relationships between stimuli and responses (Section 4.2; Silvia, 2008b), and in particular the relationship between an interest response and the familiarity and complexity of the information.

Following that diversive epistemic curiosity is motivated by a need for stimulation, the emotion of interest can be explained by the extent to which information is stimulating: that is, by the arousal potential or stimulus intensity of information (Berlyne, 1966; Wundt, 1896). However, whereas at first stimulating properties will attract individuals, too much stimulus intensity will make a stimulus less attractive and even aversive. This idea is illustrated by the Wundt-curve in Figure 4.1. Berlyne (1960) identified several collative variables that contribute to the stimulus intensity. In particular, he hypothesized that novelty (i.e., unfamiliarity) and complexity describe the horizontal axis of the Wundt-curve (Berlyne, 1970). Evidence for the Wundt-curve has been found for aesthetic preference in relation to the complexity of visual and auditory stimuli (Munsinger and Kessen, 1964; Berlyne, 1970; Hargreaves and North, 2010). However, often this evidence

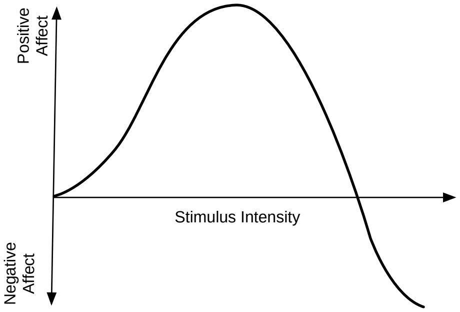


Figure 4.1: Wundt-curve or inverted-U, denoting the relation between objective stimulus intensity (e.g., unfamiliarity and complexity) and subjective valence (e.g., interest).

is equivocal or even in favor of a linear relationship (Day, 1967; Martindale et al., 1990). Moreover, although evidence for the Wundt-curve has been found for the relation between preference and complexity of aesthetic stimuli, the Wundt-curve seems difficult if not impossible to prove for epistemic curiosity and epistemic interest (Silvia, 2006). As Martindale et al. (1990) concludes: "*Although it is widely believed that the Wundt-curve relationship between preference and arousal potential is a well-established "law" this is not the case*" (p. 55).

Part of the difficulty in demonstrating the inverted-U is that a text cannot easily be categorized as interesting or uninteresting by looking solely at objective features. Instead, interest varies between people (not everybody finds the same information interesting) and it varies within people (something previously found to evoke an interest response does not need to do so now) (Schraw and Lehman, 2001). The appraisal theory of interest allows us to accommodate for this variation. According to the appraisal theory of interest (Silvia, 2008b), interest occurs after two consecutive, subjective, appraisals. The primary appraisals evaluate stimuli by their "novelty-complexity": assessing whether the stimulus is sufficiently novel and complex, or too predictable and not challenging enough to stimulate interest. The secondary appraisal evaluates the "comprehensibility" of the stimulus, determining the coping potential related to prior knowledge, available resources, and so forth. For example, if a stimulus is too complex, the coping abilities will probably not suffice, leading to a different emotion.

The two appraisals of interest have been confirmed with a variety of stimuli, such as visual art, polygons, poetry, and films (for an overview, see Silvia, 2006), indicating that the prediction of interest is not straightforward and goes beyond stimuli that are similar to previous interests (i.e., "more of the same"). The combination of the two appraisals explains the shape of the Wundt-curve (see Figure 4.1). The primary appraisals of novelty-complexity will increase preference until the secondary appraisals of coping potential become affected and preference will decrease again. However, stressing the importance of the subjective interpretation of stimuli, Silvia (2006) argues that "*an appraisal approach to interest implies that the inverted-U function should be laid to rest*" (p. 63) Instead, we can define the "sweet spot" between novelty-complexity and comprehensibility in which interest peaks (Silvia, 2006).

Individual differences in the appraisals of stimuli can be delineated by differences in trait curiosity. As a trait, curiosity represents a baseline of the diversive and specific components. It indicates the intensity to which individuals react to a state of uncertainty (i.e., specific motivations) or the extent to which they require stimulation (i.e., diversive motivations) (Litman, 2005). Some evidence exists for an indirect influence of curiosity trait on interest, via the secondary appraisal of comprehensibility (Silvia, 2008a). Given that the current study investigates the emotion of interest without a clear information need and, accordingly, focuses on the intrinsic motivation to know, the diversive component of epistemic curiosity is likely most salient in affecting (the appraisals of) interest. However, it is unclear whether the expected difference caused by the two components of epistemic curiosity holds. As Silvia (2008a) states "*the psychological processes that constitute trait curiosity are not well understood*" (Silvia, 2008a, p. 96).

The study presented in this chapter will combine the foregoing constructs in influencing and explaining interest. In particular the focus will be on the appraisal theory of interest while acknowledging individual differences. We aim to approach the sweet spot of interest through the manipulation of topical familiarity and textual complexity. Additionally, we try to explain the occurrence of interest by measuring trait curiosity. The hypothesis is that both complexity and novelty (i.e., familiarity) contribute to the emotion of interest as predicted by the appraisal theory of interest and as illustrated by the Wundt-curve. Furthermore, diversive curiosity trait and not specific curiosity trait is expected to influence interest, yet only through its appraisals. By taking interest as a primary goal, this chapter operationalizes the holistic concept of the IX as the amendable goal of predicting if and when a stimulus leads to an interest response.

To confound any effects of miscellaneous collative variables (i.e., surprisingness, incongruity, variability, and puzzlingness (Berlyne, 1960)), all stimuli are selected from a single source. Furthermore a news feed will be used as data source to assure mostly novel stimuli. The use of news articles assures the recency of the stimuli. Although recency reflects a publication date while novelty is a subjective evaluation of the content, a recent document has a higher chance of being novel to the reader yet does not necessarily have to be (Barry, 1994; Xu and Chen, 2006).

Furthermore, the model of textual complexity from Chapter 3 will be applied to manipulate interest through its two consecutive subjective appraisals as well as to explain why and when interest occurs. This chapter will complete the evaluation of the generated model of Chapter 3. This study will check the model for overfitting and test its predictive validity by applying the model to a different data set and comparing its predictions to appraised complexity. In doing so, this study combines an algorithmic approach to textual complexity with a user-centered approach to interest in explaining an emotional aspect of the IX.

4.2 Explaining Interest

This section introduces the necessary background knowledge on the hypotheses guiding this study. Namely, on the theories on curiosity (Section 4.2.1), perspectives of interest (Section 4.2.2), and the influence of complexity (Section 4.2.3), familiarity (Section 4.2.4), and individual differences on interest (Section 4.2.5). Moreover, the importance for interest to the IX is explored in Section 4.2.6.

4.2.1 Theories of Curiosity

Berlyne (1954) identified between two types of curiosity: perceptual and epistemic curiosity. The former is defined by Berlyne (1954) as "the curiosity which leads to increased perception of stimuli" (p. 180), the latter as the "drive to know" (p. 187). Furthermore, Berlyne (1966) differentiated between diversive or "like to know" and specific or "need to know" exploratory behavior. In this framework, perceptual curiosity is a typical example of specific curiosity (Berlyne, 1966). Both types of exploratory behavior are driven by specific biological needs (Berlyne, 1966) and reminiscent of two major theoretical accounts of curiosity and motivation: the optimal arousal model and curiosity-drive theory(Litman, 2005).

Specific curiosity is in accordance with the drive-reduction theory which states that individuals seek to reduce any internal conflict or tension (Hull, 1943). It assumes that the optimum level of arousal for organisms is essentially zero. For example, hunger drives an organism to behavior that leads to diminish its hunger (i.e., eating). Following this theory, novelty and complexity can cause internal tension and lead to a state of uncertainty that motivates an individual to resolve this tension and return to a state of equilibrium.

Diversive curiosity is reminiscent of the optimal-arousal theory (Litman, 2005) which states that individuals aim to be at a certain optimum level of arousal that is generally above zero (Hebb, 1955). In the case of over-arousal this leads to avoidance behavior, whereas in the case of under-arousal this leads to approach and stimulation-seeking behavior. What level of arousal is optimal varies between motives and moments. For example, for hunger this optimum level is likely rather low. According to this theory, curiosity has an optimum level of arousal that is above zero. Below-optimal arousal or a lack of stimulation, then, motivates curiosity.

It should be noted that Berlyne (1954) first followed Hull (1943) in assuming that the optimal level of arousal was low. However, in later writings (Berlyne, 1971) he acknowledged that increases in arousal can be pleasant and that the optimal level of arousal was likely higher than low. Furthermore, specific and diversive curiosity were based on antagonistic motivations (i.e., drives). For specific curiosity its reduction is rewarding, whereas for diversive curiosity its induction is rewarding. Berlyne (1971) explained this through a reward system and a (delayed) aversion system, which both respond to arousal potential or stimulus intensity (Silvia, 2006).

Recently, to merge the two types of curiosity, Litman (2005) proposed the Interest/Deprivation model of curiosity in which he states that both the reduction and induction of curiosity can lead to rewarding exploratory behavior. When viewing the example of hunger again, hunger can be stimulated by both nutritional deficits and a pleasing perception of food. In both cases, it is the actual consumption that is pleasing. Hence, instead of the induction or reduction of curiosity, it is the exploratory behavior that follows which is rewarding (Litman, 2005).

4.2.2 Perspectives on Interest

Interest and related epistemological feelings have a long history. At the top of his list of passions, Descartes (1649) listed the feeling of wonder, stating its role in motivating people towards certain actions, such as the desire to learn. A tamer version of wonder, interest, has received considerable attention: its determinants, consequences, and components have received considerable study (Silvia, 2006). Three approaches can be discerned: a behavioristic, an educational, and an affective approach.

Early theories on interest originated from a behavioristic paradigm in psychology. Within this paradigm, the focus was exclusively on the observable: that is, it excluded unobservable events within the mind. Accordingly, the relation between stimulus characteristics and exploratory behavior was studied (Hull, 1943). The

Wundt-curve (see Figure 4.1) is a clear example of this relation, where the horizontal axis depicts the intensity of the stimulus and the vertical axis the probability of approach and avoidance behavior of the organism (i.e., the liking or hedonic tone) (Walker, 1981). Berlyne (1960) concretized the axes of the Wundt-curve when he comprised a list of collative variables, properties of stimuli associated with an interest response: novelty, surprisingness, incongruity, complexity, variation, and puzzlement. Interest, however, depends on more than stimulus characteristics as indicated by the variation in interest responses within and between individuals (Silvia, 2006).

Later theories view (epistemic) interest as the "*liking and willful engagement in a cognitive activity*" (Schraw and Lehman, 2001, , p.23). As a cognitive or educational construct, the role of subjective judgments and related personal states and traits is emphasized. Schraw and Lehman (2001) categorized the causes for interest in text, differentiating between personalized interests (i.e., long-term interests) and situational interest (i.e., short-term interest), where situational interest depends on text characteristics, task context, and the knowledge of the reader. Of these causes, the text characteristics reflect the collative variables identified by Berlyne (1960). Conjointly, when seen as a cognitive construct, interest is an interplay between numerous characteristics of the text, context, and user.

With a focus solely on short-term interest, the appraisal theory of interest provides insight into the complex interaction between user, task, and event. The appraisal theory of interest regards interest as an emotional response. Whether or not interest is an emotion is disputed by some (Ortony and Turner, 1990). Notwithstanding, interest shows all the typical features of emotions (Lazarus, 1991), supporting its position as an emotion (Silvia, 2008b). Interest is characterized by a cognitive component (appraisal), a subjective component (feeling), and by physiological and expressive components (movement of muscles in the forehead and eyes, faster speech rate and greater frequency range) (Hess and Polt, 1960; Banse and Scherer, 1996). As an emotion, interest is dependent on a specific set of cognitive appraisals which have already been described in Section 4.1. Moreover, this allows interest to be explained within a broader framework of the cognitive-appraisal theory of emotion, which highlights the importance of intrinsic pleasantness and novelty (cf. the collative variables), coping potential (e.g., the knowledge of the user), and the motivational bases (e.g., the task or need) (Section 2.5; Ellsworth and Scherer, 2003).

Considering that long-term interests are commonly assumed to be the primary determinant of interest responses (Silvia, 2001; Schraw and Lehman, 2001), each of the perspectives on interest shows that ample evidence exists on the importance of other determinants in evoking an interest response. The importance of textual complexity, familiarity, and individual differences will be reviewed next.

4.2.3 Influence of Complexity

Complexity is key to both the primary and secondary appraisal evaluations, allowing one to approach the “sweet spot” of interest. For example, whilst the complexity of a text can enhance the primary appraisal, making a text more challenging, it can also impair the secondary coping appraisal, when appraised as too complex. This combination of appraisals explains the occurrence of the Wundt-curve (Figure 4.1). This section will review evidence for the influence of complexity on interest, although this evidence is mostly indirect via an influence of comprehensibility. The evidence will be reviewed with regards to epistemic interest for aesthetic stimuli and educational, epistemic texts.

For complex pieces of visual art, poetry, and music the coping potential was an important predictor of interest, as illustrated by three studies. First, Silvia (2005) shows in four studies that interest for both visual art and poetry increased when participants had a higher ability to understand the stimuli, either because of their own pre-existing ability (for visual art) or due to instructions (for poetry), and this effect was only salient for complex stimuli (for visual art). Second, Zyngier et al. (2007) show an interaction effect between repeated reading and complexity on the reported interest in three poems, where repeated reading increased interest only for the complex poems. This suggests an effect of coping potential, although the data was inconclusive for some groups (cultures) of readers. Third, Eerola (1997) shows that (objectively) complex music albums, in comparison to simpler music albums, have a later peak in sales and sustain high levels of sales longer. This suggests the interrelation between familiarity, coping potential, and interest for music as well. The foregoing three studies show that when coping potential is taken into account, the results are supportive of a positive effect of (objective) complexity that interacts with coping potential.

In addition to the importance of comprehensibility as a mediator of the relation between complexity and interest for aesthetic stimuli, its role has also received considerable attention from an educational perspective for textual, epistemic stimuli. The relation between textual complexity and interest focused mainly on one side of the complexity spectrum: for complex stimuli, comprehensibility enhances interest and learning. For example, Schraw et al. (1995) found that comprehensibility alone accounted for 36.63% of variance in interest and 12.30% of variance in text recollection, being the highest predictor for both interest and recollection. Furthermore, the importance of comprehensibility for interest has been confirmed with expository science texts in which people were more interested when they were better able to understand the texts (Connelly, 2011), and in a study on learning from text, where interestingness was found to correlate with comprehensibility, familiarity, and concreteness (Sadoski, 2001). Similar results were obtained by numerous other studies (cf., Hidi, 1990; Schraw and Lehman, 2001). Few studies have investigated the other side of the complexity spectrum where simplistic stimuli can induce boredom. A notable exception for textual complexity is a study by Schiefele (1996), who had high school students read texts below their reading grade level and found a negative relation between verbal abilities and interest, explaining the finding by noting the texts were "*somewhat easy for highly able readers*" (p. 15). Hence, regarding interest in text, little direct evidence exists for a positive effect of textual complexity on interest.

The preceding overview of the influence of complexity suggests that complexity can enhance and reduce epistemic interest. However, the evidence for this conclusion is slim, given that most support for it is indirectly via comprehensibility or coping potential. In contrast to the findings on perceptual curiosity (Munsinger and Kessen, 1964; Berlyne, 1970), there seems to be little to no direct evidence for a Wundt-curve or inverted-U relation between objective complexity and epistemic interest. Notwithstanding, the reviewed findings are generally in support of the appraisal theory of interest.

4.2.4 Influence of Familiarity

The notion of "more of the same" can be described as the familiarity a reader has with a topic. Similar to the effect of complexity on interest, the effect of familiarity can be explained through the appraisal theory of interest. Familiarity can decrease the primary appraisals of novelty-complexity and increase the secondary appraisal of comprehensibility or coping potential. Moreover, Berlyne (1970) identified novelty as one of the collative variables that increases stimulus intensity and has an inverted-U relation with preference (see Figure 4.1). Similar to the research on complexity, the research on familiarity can be described for aesthetic and epistemic stimuli.

With regards to aesthetic stimuli the influence of familiarity has been studied extensively for musical preference. A multitude of studies show that musical preference increases with repeated listening, especially for somewhat complicated pieces of music (for an overview, see Finnäs, 1989). However, some studies show monotonous or opposite effects for the extremes of complexity: very complex or simple pieces of music. The results of repeated listening are in general supportive of an inverted-U relation between familiarity and musical preference when the number of repetitions is large enough (Finnäs, 1989; Verveer et al., 1933; Hargreaves, 1984). Alternative to the manipulation of repetitions, studies compared subjects' own ratings of familiarity to their preference for pieces of music. Such studies mostly report a positive relation between familiarity and preference, which can be explained by a tendency not to listen to the same piece of music until it becomes unpleasant (Russell, 1986; Finnäs, 1989).

The relation of familiarity or prior knowledge with interest for educational, epistemic stimuli is somewhat equivocal, with studies revealing positive as well as neutral relations (for an overview, see Schraw and Lehman, 2001). For example, in a recent study Rotgans and Schmidt (2011) did not find a significant role for topical familiarity on interest, whilst familiarity did lead to increases in learning. And, Schraw et al. (1995) found only a marginal direct relation between prior knowledge and interest. Although in this study most readers had a high and similar degree of familiarity with the topics of the stimuli. On the contrary, in a large study Alexander et al. (1994) found a large positive relation between both domain and topic knowledge and interest. Furthermore, Sadoski (2001) reports on the direct relation between familiarity on interest in two studies, where in the first study there was a substantial effect, while in the second study only a marginal effect was found. The preceding studies hint on a positive effect of familiarity on interest, the effect size of which depends on other (unknown) influences.

Aside a direct effect of familiarity on interest, more support exists for an indirect effect via comprehensibility. Amongst others, Schraw et al. (1995) and Sadoski (2001) report substantial positive relations between familiarity and comprehensibility, and between comprehensibility and interest. These relations seem to operate separately from a direct influence of familiarity on interest. However, in some cases this indirect effect is non-existent (e.g., study 1 of Sadoski, 2001), due to an absence of an influence of familiarity on comprehensibility. This can be explained by indications that prior knowledge is only significant when the text is not explicatory (Wade et al., 1999; Schraw et al., 1995). Hence, texts which are informationally complete do not require the reader to be familiar with the topic in order to be comprehensible (Wade et al., 1999).

The described findings on the relation between familiarity and interest show different relations for aesthetic stimuli and epistemic stimuli. Although for aesthetic stimuli evidence exists for an inverted-U, for epistemic stimuli the effect is mostly positive, either direct or indirect (i.e., via comprehensibility). The positive effect common for epistemic stimuli seemingly contradicts the appraisal theory of interest (Silvia, 2006) as well as Berlyne (1970)'s collative variables, which state that novelty invokes interest. An explanation might be that novelty and familiarity are not necessarily contradictory of each other. For example, a high level of domain knowledge or familiarity does not exclude the possibility of a novel stimulus that extends this knowledge base (Schraw and Lehman, 2001). Even more so, the impact or effect of a novel stimulus might be higher when it connects to a larger domain knowledge (Kintsch, 1980). This suggests that novelty and familiarity can be understood as distinct, non-antagonistic phenomena which both contribute to interest.

4.2.5 Individual Differences

The appraisals that lead to interest can be influenced by individual differences. Although several different conceptualizations of individual traits related to curiosity exist (e.g., openness to experience, need for cognition, openness to ideas, typical intellectual engagement), their discriminant validity is questioned (Mussel, 2010). Instead, the Epistemic Curiosity Trait (ECT), which consists of a diversive and a specific component (Litman and Silvia, 2006), "*can be seen as a well defined construct, with a structure of highly correlated dimensions, for which validated measures have been developed*" (Mussel, 2010, p. 507).

Although the construct of ECT may be well-defined (Mussel, 2010), its relation to the emotion of interest is less clear (Silvia, 2001, 2008a). As Silvia (2008a) concludes: "*little is known about why curious people experience interest in response to specific situations*" (p. 94). Some evidence exists for the role of ECT as a corollary of the emotion of interest for poems and visual art (Silvia, 2008a). In these experiments, the effect of ECT on the emotion of interest was fully mediated by the appraisal of coping potential. Furthermore, Litman et al. (2005) showed that curiosity (as a state) could be explained by either high levels of uncertainty or by ECT, which suggests a distinction between the influence of the specific and diversive motivations on interest. The limited available evidence supports the influence of ECT on interest via the appraisal of coping potential, and indicates a difference between the influence of the diversive and specific components of the ECT.

Another salient individual difference is described by the long-term interests of an individual. The momentary emotion of interest differs from the long-term interests often implemented in IF&R systems as sole determinant. Interests do not necessarily lead to interest and interest does not directly lead to (but is a requirement for) the development of interests (Silvia, 2001). Although there are many theories on the development of long-term interests, the consensus is that the repeated experience of interest is needed for the development of interests. The contemporary model by Hidi and Renninger (2006) illustrates this consensus. They propose a four-stage model of the development of long-term interests based on a wide range of findings on the topic: 1) triggered "situational interest";

2) maintained "situational interest"; 3) emerging "individual interest"; and 4) well-developed "individual interest". Here, "situational interest" can be regarded similar to the emotion of interest and "individual interest" to long-term interests (Silvia, 2006). This four-stage model clearly shows the importance of interest for the development of interests, although it also indicates this is a lengthy and complicated process for which a repeated experience of interest and a prolonged motivation are required. Moreover, the model suggests that long-term interests and familiarity go hand in hand.

4.2.6 Interest and the Information eXperience

The importance of interest has long been noted for various cognitive processes. Following from this, the emotion of interest is key to the experience during interaction with information; that is, the IX. Interest is believed to be pivotal to each of the prototypical experiences (see Section 2.6): an overall positive experience, an engaging/flow experience, and a learning experience.

For an overall positive experience, the *"quality of experience seems to be an epiphenomenon of interest"* (Schiefele, 1996, p. 13). Although interest is a positive emotion, interest is distinctively different from enjoyment. Where interest generally occurs for novel and complex stimuli (see Section 4.2.2), enjoyment occurs for familiar and (somewhat) easy stimuli. This indicates an orthogonal relation between the appraisals of interest and enjoyment. This type of relation has been confirmed by a qualitative study on the emotions of adolescents during search, showing interest can lead to explorative behavior yet also cause frustration (Bowler, 2010). However, this is not a necessity. As discussed next, theories of user engagement and flow illustrate that complex stimuli which provoke an interest response can be followed by an enjoyable and motivated experience (O'Brien and Toms, 2008).

For an engaging experience, the emotion of interest has been associated with the onset of the experience (O'Brien and Toms, 2008). Furthermore, the determinants of a Flow experience are of a similar nature to the appraisals of interest; namely, complexity corresponds with challenge and comprehensibility can be understood as similar to skill(Csikszentmihalyi, 1991). In general, interest is a primary determinant of motivation. Motivation starts with interest by initiating attention and exploratory behavior, subsequently interest interacts with enjoyment and sustains persistence in an activity (Reeve, 1989). This shows that aside from a positive emotion, interest is key to an engaging/flow experience that, in turn, possibly leads to an enjoyable experience.

Finally, interest is also linked to intelligent tutoring systems and a learning experience, where interest motivates learning for its own sake (D'Mello et al., 2007; Silvia, 2008b). Subjects acquire a higher depth of comprehension, apply better learning strategies, and have an overall more enjoyable (learning) experience when texts evoke interest (Schiefele and Krapp, 1996; Schiefele, 1996). As Jonassen (2000) summarized it: *"Students think harder and process material more deeply when they are interested"* (p. 71).

The preceding outline indicates the importance of interest for a positive, fruitful IX. With that, the value for information systems to be able to foster an interest response is clear. To explore whether and how it is possible to foster an interest response, the following sections will report on an experiment that attempts to influence and explain the occurrence of interest for news articles.

4.3 Method

In a user experiment, 18 news articles were shown to 30 participants. Each of the articles differed on their textual complexity, as indicated by the model of textual complexity developed in Chapter 3, and by the topical familiarity of the participant with the article. The dependent variables were appraised complexity, appraised comprehensibility, and interest. Furthermore, ECT was included as a control variable.

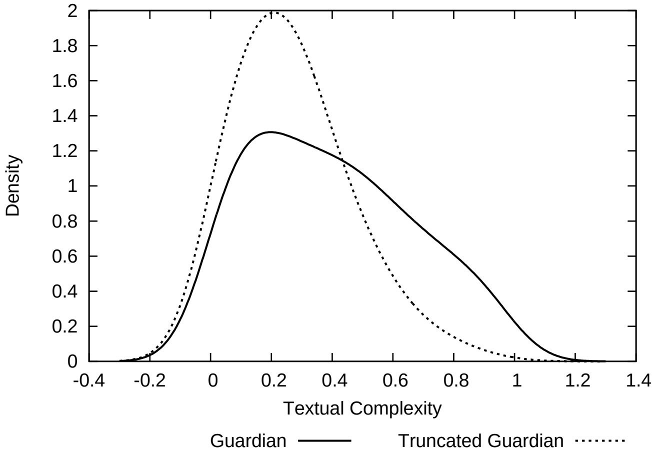
4.3.1 Participants

A total of 30 participants (22 male, 8 female) with an average age of 28.60 ($SD =$ 6.06) voluntarily joined in the experiment. None of the participants were native English speakers, but all graded their reading literacy as high ($M = 4.63$; $SD =$ .62; range 1 to 5, 5 highest). All participants were well-educated, they either had a university degree or were enrolled as a student at a university. Participants were recruited via e-mail, by distributing flyers, and by personally inviting local students and employees.

4.3.2 Data set

As data set a collection of 14 856 articles from The Guardian[1] was used. The Guardian is a widely known newspaper published in the UK. From the entire collection of 14 856 articles a final selection of 18 articles was made to be used in the experiment. The selection procedure is further explained in the next section. The final selection consisted of articles from the following news feeds: culture;

[1] www.guardian.co.uk



Note. The textual complexity scores range 0 (low complexity) to 1 (high complexity). Because the estimated density was derived using a Gaussian Kernel with a bandwidth of .1, the limits of the graph extend this range.

Figure 4.2: Estimated density of textual complexity scores for the Guardian data sets (Section 4.3.2).

environment; financial, market and economics; commentary; life and style; and science and technology. To reduce variation that originates from differences in article length, all articles were truncated after 1 200 characters. The cut-off point was placed before the end of the word at position 1 200 and three dots were added to indicate that the story would normally continue. Any layout was stripped from the articles, leaving only the title and textual content.

4.3.3 Filtering

A selection was performed by applying a version of the model of textual complexity from Chapter 3 on the data set of The Guardian. This version was specifically designed for this selection procedure and is concisely described in Appendix A. Furthermore, it is described in detail in Van der Sluis et al. (ip). The resulting distribution is given in Figure 4.2. The distribution was derived from the predictions of the preliminary model, using a kernel density estimation with a Gaussian kernel and a bandwidth of .1. Given that the figure shows a somewhat normal distribution of the predictions over the full range of possible values, the application of the classifier seems reliable. Nonetheless, this does not confirm its validity for this data set yet; its validity can be confirmed after comparison with subjective ratings of complexity. Furthermore, Figure 4.2 shows that, in comparison to the original texts, the truncated data set was evaluated as less complex and less variable.

The distribution of the truncated texts was used to filter articles in two steps. First, articles from the lower, middle, and upper part of the distribution of textual complexity were pre-selected. Then, a final selection consisting of 18 articles was performed based on suitability. For the final selection the following criteria were applied. Firstly, having a participant pool of international origin and being a news source of national origin, texts were to be of international orientation. Secondly, a comments section was included in the data source. Although at a higher level of textual complexity this contained elaborative background articles, at a lower level of textual complexity this included user-generated content submitted by children. Although belonging to the lower level of complexity, such articles were deemed unsuitable. Thirdly, selected news items differed in topic in order to ensure a variation in topical familiarity would exist.

To measure textual complexity, the remainder of this chapter uses the classifiers resulting from Chapter 3. Nine classifiers are available based on three types of classifiers (Support Vector Machine (SVM), Random Forest (RF), and Logistic Regression Model (LRM)) that were each trained on texts of three different sizes (articles, sections, and paragraphs). Each of these classifiers is trained on a binary problem (simple vs. complex). Accordingly, the resulting metrics give a value ranging from 0 to 1 indicating easy or complex. Considering that all features developed in Chapter 3 were directional and largely semantically independent, it is reasonable to use the resulting value as a continuous scale.

4.3.4 Instruments

Throughout the experiment, four measurement instruments were used. Besides a starting questionnaire, three questionnaires were applied after the reading of each article to measure interest, complexity, and comprehensibility.

The first instrument was a basic demographics questionnaire that queried the following items: gender, age, nationality, educational background, English reading proficiency, and visual acuity. These items were included to control for their potential intervening influence.

Topical Familiarity Questionnaire

The abilities or skills of the subject with regards to the topics that were discussed in the selected articles were measured using 7-point Likert scales, which asked subjects' familiarity with the main topics of each article. Familiarity ratings are commonly used to indicate the memory strength for a topic (Wixted and Stretch, 2004). Moreover, it was expected that this topic-based indication of familiarity would resemble what is commonly modeled in user models of long-term interests or knowledge (cf. Brusilovsky and Millán, 2007): namely, the strength of familiarity for a topic. To measure topic familiarity the keywords supplied by The Guardian were used as topics. These keywords represent the topics on a general level (e.g., viruses) as well as a concrete level (e.g., bird flu). More specific keywords can be expected to give a more specific indication of the relevant prior knowledge of the subject. Therefore, for every article only the three most specific keywords were presented to the participant. When articles had overlapping keywords, the keywords were only queried once. The resulting questionnaire is given in Appendix B.

The resulting scores were averaged per article and will hereafter be referred to as topical familiarity. Since the topical familiarity scale consisted of a different set of questions per article, Cronbach's alpha can only be computed per article. The mean alpha for the topical familiarity scales was .639 (SD = .176), based on N = 30 observations per scale. This alpha score can be considered reasonable.

Epistemic Curiosity Scale

To measure epistemic curiosity, the epistemic curiosity scale (Litman and Spielberger, 2003) was used. This scale consists of two subscales: one for the diversive and one for the specific component of trait curiosity. Both scales contained five items and were measured using 7-point Likert scales. The scales are further detailed in Appendix B.

For each subscale, the items were merged into one scale by taking their mean. Cronbach's alpha was .831 for the diversive subscale and .763 for the specific subscale (N = 30).

Interest-Appraisal Scales

Silvia (2008a) presents two scales, one for appraised complexity and one for appraised comprehensibility. Both are based on 7-point semantic-differential scales. The appraised complexity scale (Silvia, 2008a) consisted of just one differential: complex-simple. To improve scale reliability another item was added to measure appraised complexity: easy to read-difficult to read. Cronbach's alpha for this scale was at a good level of .824 (N = 540), confirming the value of the added item. Comprehensibility was measured by the appraised comprehensibility scale

(Silvia, 2008a), consisting of the following three differentials: comprehensible-incomprehensible, coherent-incoherent, and easy to understand-hard to understand. Cronbach's alpha for this scale was at a good level of .893 ($N = 540$).

To verify the topical familiarity questionnaire, one question was asked after participants read each article. Using a 7-point Likert scale, this question queried the actual familiarity with the article. It asked whether the participant thought of the content as "not familiar to me" or "very familiar to me". This question will hereafter be referred to as the article familiarity.

In accordance with related studies (e.g., Silvia, 2006, 2008a), interest was measured using two 7-point differentials: *interesting-uninteresting* and *boring-exciting* (Silvia, 2008a). Furthermore, a 7-point Likert scale was added to benefit from the shortened texts (see Section 4.3.2), asking the participant to agree with the statement *"I would be interested in reading more of this text"*. All three questions formed a reliable scale, confirmed by an excellent Cronbach's alpha of .921 ($N = 540$).

The complete interest-appraisal questionnaire is further detailed in Appendix B.

4.3.5 Design and Procedure

The experiment used a within-subjects design which showed all articles to each participant. The independent variables were topical familiarity and textual complexity. The dependent variables were appraised complexity, appraised comprehensibility, and interest. Furthermore, ECT was included as a control variable.

For each participant, the articles were grouped based on the average topical familiarity score per article. Three blocks were created based on this score, containing the six highest, six lowest, and six miscellaneous articles. The order of the blocks was counterbalanced, giving a total of 3! (= 6) conditions. The articles were not grouped according to their level of textual complexity. Within each block the order of the articles was according to their familiarity score, in descending order.

The experiment started with the demographics questionnaire and the diversive and specific ECT scales, followed by the topical familiarity questionnaire. Then, the participants proceeded through the three blocks. Every block was preluded by a short textual instruction which described the task as simply, "read the news". Per block, six articles were shown sequentially to the participants for self-paced reading. After each article followed the interest-appraisals scales. And, after each block a questionnaire about the past experience was given. The goal of the experiment was not made explicit and all participants were instructed to read thoroughly and on their own pace. The total experiment lasted on average 50 minutes. The participants indicated that the length of the experiment made it somewhat demanding, especially since not all texts were experienced as interesting.

4.3.6 Analyses

Although most of the statistical techniques used for the analyses are straightforward in their application and interpretation, a few require more detail. Namely, outlier detection, the interpretation of Pearson's correlation, smoothing techniques, and the creation of a Structural Equation Model (SEM). Furthermore, all statistical tests and illustrated confidence intervals are based on an α level of .95.

Outliers were explored for the dependent variables by t-tests. Based on the mean appraised complexity one article was indicated as an outlier and excluded from further analyses. By using a series of t-tests it was shown that the complexity for this article was appraised significantly different in comparison with the other articles ($t(17) = 2.10$, $p < .05$).

As guideline for interpretation of Pearson's correlation Cohen's rule of thumb for effect sizes will be used. This interpretation divides and interprets the strength of correlation (r) as follows: $.100 \leq r < .300$ small, $.300 \leq r < .500$ medium, $r \geq .500$ large (Cohen, 1992).

Furthermore, to illustrate the relation between the independent variables of topical familiarity and textual complexity and the dependent variables of article familiarity, appraised complexity, and interest, in Figure 4.3 (p. 130) and 4.5 (p. 134) a smoothing technique was used. This technique fits a polynomial surface determined by the independent variable, based on local fitting[2]. That is, the fit at a point x is made using those points that lay in the neighborhood of x, weighted by their distance from x. Standard parameters were used for both the size of the neighborhood (i.e., the nearest 75% of all points) and the weights (i.e., $(1 - \frac{\text{distance}}{\text{maximumdistance}}^3)^3$). Fitting was performed using a (weighted) least squares method. To indicate a goodness-of-fit, confidence intervals were included in the figures as well.

SEM is a multi-regression technique, solving multiple regression equations. It differs from normal regression in that dependent variables can be independent variables as well. Using SEM the latent and observed variables need to be identified. The scales presented in Section 4.3.4 were used as latent variables and their items as observed variables. Furthermore, two extra variables were included and one was excluded. First, the output of the model of textual complexity (see Section 4.3.3) was used as sole indicator for the latent variable of textual complexity. Second, the latent variable of familiarity was based on both topical (predicted)

[2] `http://stat.ethz.ch/R-manual/R-patched/library/stats/html/loess.html`

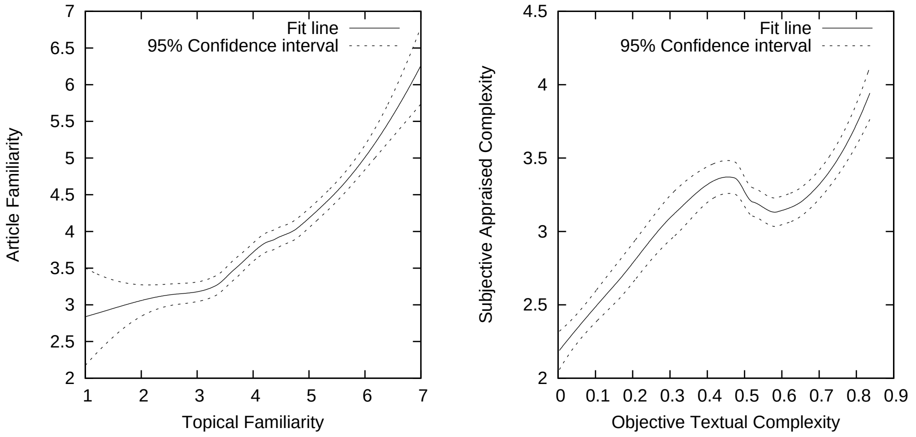
and article (actual) familiarity, given that no hypotheses exist for a difference in influence and the behavior of both items was found to be similar (see Figure 4.3a, p. 130). Finally, of the two subscales of ECT, only diversive ECT was included in the SEM. Both scales were found to behave distinctively. In line with the theoretical expectations and the hypotheses, of both ECT scales only diversive ECT formed a valid predictor in the SEM. The SEM was developed with the R package Lavaan (Rosseel, 2012) and was based on a covariance matrix. The resulting SEM is shown in Figure 4.6. It was modeled after standardization of the observed as well as the latent variables, causing the coefficients in Figure 4.6 to represent the strength of a relation in the amount of standard deviations.

A SEM is only valuable when it forms a good representation of the data, as indicated by a plethora of fit indices. However, most indices suffer from being very sensitive to sample size or number of parameters. Here, Iacobucci (2010) was followed in defining three more robust indices: relative X^2 (X^2_r), Standardized Root Mean Square Residual (SRMR), and Comparative Fit Index (CFI). Each will be introduced next. First, the X^2_r is the inferential X^2 statistic divided by the degrees of freedom. It is very sensitive to the sample size, where already a modest sample size (e.g., $N = 200$) gives a high (and, thus, significant) X^2. Its value is regarded good if below 5 (Schumacker and Lomax, 2010). Second, SRMR is a robust badness-of-fit index comparing the model against the actual data. It is also, yet to a lesser extent, influenced by sample size, where higher sample sizes give better results. The maximum allowed value for SRMR is .090 (Iacobucci, 2010). Third, the CFI is a goodness-of-fit index, which compares the hypothetical model to a simpler model (without any defined paths). It is the most robust metric of the three, adjusting for model parsimony and relatively invariant of sample size (in particular for $N \geq 200$). The consensus for this index is that it should be “close to” .950 (Iacobucci, 2010). Section 4.4.3 will report on the three fit indices for the developed SEM (see Figure 4.6).

4.4 Results

4.4.1 Familiarity

The extent to which the independent variables predict their dependent counterparts needs to be verified before further analysis. The extent to which topical familiarity (measured at the beginning of the experiment) predicts article familiarity (measured after the participants read an article) is illustrated in Figure 4.3a. The relation that is apparent in the figure is confirmed by a correlation of medium strength between topical familiarity and article familiarity ($r = .344$,



(a) Familiarity

(b) Complexity

Figure 4.3: The ability of (a) topical familiarity to predict article familiarity and (b) objective textual complexity to predict subjective appraised complexity.

$t(508) = 8.26$, $p < .001$). Figure 4.3a shows that the first half of topical familiarity has a limited influence on article familiarity, whereas the second half of topical familiarity differentiates especially well between different levels of article familiarity.

For each of the conditions of topical familiarity (i.e., low, medium, and high topical familiarity), Figure 4.4 shows the interest, its appraisals of complexity and comprehensibility, topical familiarity as asked beforehand, and article familiarity as asked after each article, including the confidence intervals. Figure 4.4 shows that each variable differs over the conditions and that each of the variables behaves in the direction as expected. An increase in topical familiarity leads to a decrease in appraised complexity and an increase in article familiarity, comprehensibility, and interest. In particular the positive relation to interest suggests that familiarity is, indeed, an important determinant of the emotion of interest.

Although Figure 4.4 indicates the manipulation by topical familiarity was overall successful, the difference between the low and medium conditions is small. The confidence intervals in Figure 4.4 indicate that the manipulation was successful for the high versus medium or low conditions, and not for the medium versus low conditions. This suggests a threshold for the influence of topical familiarity exists, which was also shown in Figure 4.3a.

A within-subjects (or repeated measures) Multivariate ANalysis Of VAriance (MANOVA) was conducted to test for an effect of the conditions of familiarity

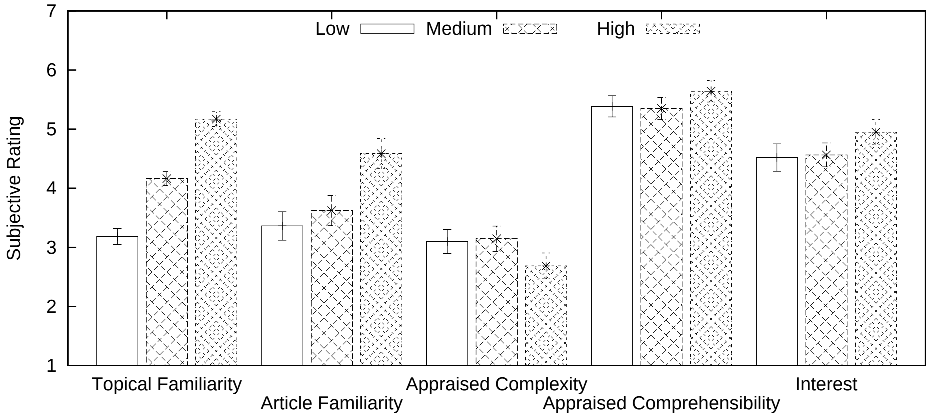


Figure 4.4: Means and confidence intervals for the familiarity, appraisals, and interest for each of three conditions (i.e., articles of low, medium, and high topical familiarity).

on article familiarity, appraised complexity, appraised comprehensibility, and interest. There was an overall significant effect of the conditions, Wilks' $\Lambda = .153$, $F(8, 22) = 15.232$, $p < .001$, indicating that the manipulation by familiarity was successful, with a very strong overall effect size ($\eta^2 = .847$). Follow-up MANOVAs were used to make post hoc comparisons between the conditions of familiarity. A within-subjects (or repeated measures) MANOVA confirms a significant effect of the manipulation by topical familiarity on the following variables: article familiarity ($\eta^2 = .777$, $p < .001$), appraised complexity ($\eta^2 = .249$, $p < .01$), and interest ($\eta^2 = .145$, $p < .05$) (see Table 4.1). A trend was found for the influence of topical familiarity on comprehensibility ($\eta^2 = .106$, $p < .1$). The order of presentation is significant for article familiarity, appraised complexity, and interest as well, either directly or through an interaction with the condition. The influence of the presentation order can be expected from the procedure (see Section 4.3.5), since within each condition the articles were presented in the order of their topical familiarity.

The results confirm the expected importance of familiarity for the occurrence of interest within the context of a homogeneous set of novel news articles. Nevertheless, the relatively low effect size for interest indicates that more factors, including appraised complexity and appraised comprehensibility, play a role in predicting interest. And, the results indicate that these factors themselves are affected by

Table 4.1: MANOVAs of within-subjects contrasts for condition (sets of articles of high, medium, or low topical familiarity) and order of presentation with η^2 denoting their effect sizes.

		df			
Variable	F	model	error	p^*	η^2
Article Familiarity					
Condition	100.859	1	29	$< .001$	.777
Order	6.162	1	29	$< .05$	.175
Condition * Order	5.705	1	29	$< .05$	.164
Appraised Complexity					
Condition	9.605	1	29	$< .01$	.249
Order	6.333	1	29	$< .05$	.179
Condition * Order	6.903	1	29	$< .05$	.192
Appraised Comprehensibility					
Condition	3.434	1	29	$< .1$	.106
Order	2.470	1	29	n.s.	.078
Condition * Order	.351	1	29	n.s.	.012
Interest					
Condition	4.935	1	29	$< .05$	.145
Order	.034	1	29	n.s.	.001
Condition * Order	4.869	1	29	$< .05$	.144

*Note**. Non-significance is denoted by "n.s."

the user's familiarity with the articles.

4.4.2 Textual Complexity

The relation between objective textual complexity and its subjective counterpart, appraised complexity, is dependent on the used classifier. Given that in Chapter 3 three types of classifiers (SVM, RF, and LRM) were trained on texts of three different sizes (articles, sections, and paragraphs from the Wikipedia data sets), a total of nine classifiers can be evaluated in comparison to the mean appraised complexity of each of the stimuli. The correlations between the mean subjective complexity and objective complexity are shown in Table 4.2 for each of the classifiers. The scatter plots on which each of these correlations is founded are given in Appendix Da. With a correlation of $r = .704$ ($t(15) = 3.84$, $p < .01$) the

Table 4.2: Correlations between objective complexity and mean subjective complexity for each of the classifiers developed in Chapter 3.

Length	SVM	RF	LRM	Average
Articles	.355*	.632**	.197	.395
Sections	.675**	.616**	.704**	.665
Paragraphs	.699**	.558*	.633**	.630
Average	.576	.602	.511	.563

Note. Significance levels: * $p < .05$; and ** $p < .01$.

sections-based LRM showed the best performance in predicting mean appraised complexity and will, henceforth, be used for all further analyses. The very large effect size indicated by the correlation of $r = .704$ confirms the effectiveness of the model in influencing the appraised complexity.

The evaluations shown in Table 4.2 and Appendix D indicate that every classifier performed well and, therefore, testifies to the predictive validity of the model of textual complexity. The only exceptions were the paragraph-based LRM and, to a lesser extent the paragraph-based SVM. This can partly be explained by the length of the (training) texts, for which Table 4.2 shows that it was of influence on the performance. This suggests that the influence of text length needs to be acknowledged for achieving the predictive validity. In addition to Appendix Da, Appendix Db gives the scatter plots and correlations for the data without outlier correction (i.e., for all 18 stimuli). Although this gave a decrease in performance, the indicators of textual complexity generated by the best classifier (i.e., the paragraph-based SVM classifier) still correlated with the subjective complexity, with $r = .592$ ($t(16) = 2.94$, $p < .01$). Aside from being a large effect, this indicates that the model of textual complexity is robust as well.

The relation between objective and subjective complexity is shown in detail in Figure 4.3b for the sections-based LRM classifier. The fit line clearly shows the ability of the model of textual complexity to predict appraised complexity. The fit line also indicates that the two variables are not on a par with each other near the value of .5 for objective complexity. Essentially, a value of .5 means that the classifier is uncertain whether the classified text is simple or complex. An objective complexity of .5, thus, has the highest chance of being not in line with the subjective complexity.

The correlation between objective textual complexity and interest is small and non-significant ($r = .163$, $t(15) = 0.64$, $p > .1$). However, further analyses showed

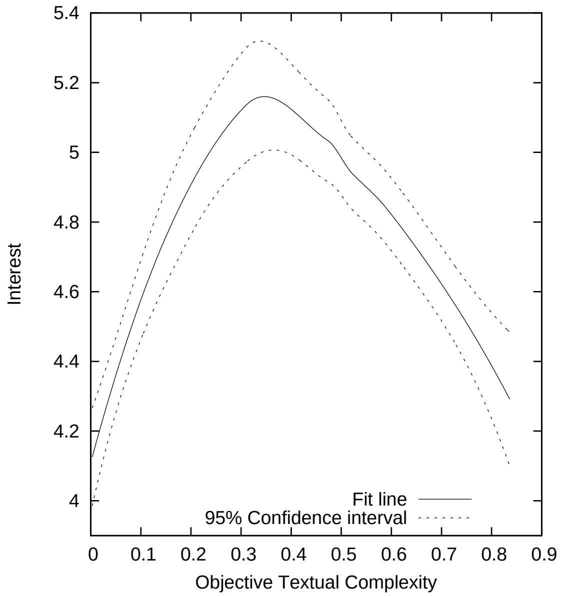
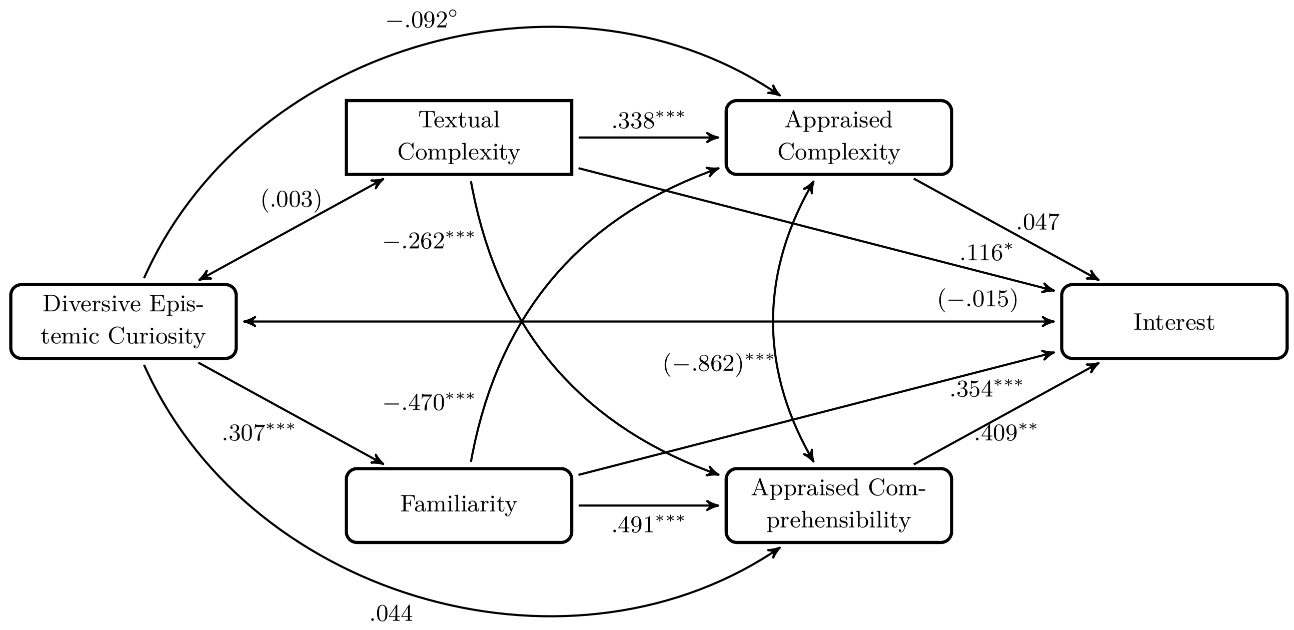


Figure 4.5: Inverted-U or Wundt-curve that describes the relation between objective textual complexity and subjective interest.

that while an increase in complexity is initially accompanied by an increase in interest, an additional increase in complexity actually shows a decrease in interest. To further explicate this effect, Figure 4.5 shows a fit line of the relation between objective textual complexity and interest. The figure clearly shows that after an initial increase in interest, objective complexity decreases interest. In other words, the relation between textual complexity and interest resembles an inverted-U or Wundt-curve (see Figure 4.1). However, the confidence intervals surrounding the fit line indicate that other factors mediate the relation between complexity and interest as well. To further explicate the effect of textual complexity on interest and verify its influence in relation to other variables in explaining interest, the next section presents a SEM.

4.4.3 Explaining Interest

To further explicate the complicated relation between familiarity, complexity, the appraisals, and interest, a SEM was developed (see Section 4.3.6). The resulting path model is shown in Figure 4.6, together with the regression coefficients and



Note. $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$.

Note. Subjective measures are denoted by rounded boxes, objective measures (i.e., textual complexity) by squared boxes.

Figure 4.6: Path diagram showing regression coefficients and covariance (between parentheses) of the objective variable (squared box) and subjective variables (rounded boxes) that together explain interest responses ($R^2 = .378$).

Table 4.3: Correlations between dependent and independent variables.

	Correlations					
Variable[†]	1	2[††]	3[‡]	4[‡]	5	6
1. Familiarity						
2. Tex. Com.[††]	.363					
3. Epi. Cur. (D)[‡]	.465**					
4. Epi. Cur. (S)[‡]	.155		.265			
5. App. Compl.	-.315***	.704**	-.481**	-.188		
6. App. Compr.	.345***	-.501*	.391*	.275	-.763***	
7. Interest	.387***	.163	.430*	.172	-.352***	.485***

Note. Significance levels: * $p < .05$; ** $p < .01$; and *** $p < .001$.

Note. [†] Full variable names: 1. Familiarity; 2. Textual Complexity; 3. Epistemic Curiosity Trait (Diversive); 4. Epistemic Curiosity Trait (Specific); 5. Appraised Complexity; 6. Appraised Comprehensibility; 7. Interest.

Note. [‡] Excluding within-subject variance.

Note. [††] Excluding within-stimulus variance.

covariances connecting each of the latent variables congruent with the hypotheses. Furthermore, Table 4.3 shows the correlations between each of the measured variables.

Before allowing further discussion, the fit of the developed SEM to the data was checked using the X_r^2, SRMR, and CFI indices. First, with $X_r^2 = 5.24$, the developed SEM scores just above the maximum allowed value of 5. However, this small deviation can be fully explained by the large sample size of 510 samples, which is confirmed by the other indices. Second, with a score of SRMR = .055, the SEM corresponds to the requirement of SRMR $\leq$ .090. Third, the developed SEM gave a CFI value of .925, also corresponding to the consensus of "close to" .950. Taken together, the three indices showed a good fit and justify further discussion of the SEM shown in Figure 4.6.

In total 37.80% of variance in interest responses was explained. Interest was explained by familiarity ($\beta = .354$, $r = .387$), textual complexity ($\beta = .116$, $r = .163$) and appraised comprehensibility ($\beta = .409$, $r = .485$). Appraised comprehensibility had the largest contribution to reported interest, as indicated by both its coefficient within the SEM ($\beta = .409$) and its correlation ($r = .485$), which makes it the most important predictor of interest. Surprisingly, appraised complexity was not a significant determinant for reported interest. Objective complexity together with appraised comprehensibility and familiarity seemingly captured the influential aspects of appraised complexity on interest.

The SEM of interest (see Figure 4.6) shows that textual complexity was a significant determinant for both the highest predictor of interest (i.e., the appraisal of comprehensibility) and the reported interest itself. Paradoxically, it had a positive direct effect ($\beta = .116$, $r = .163$) as well as an indirect negative effect (via appraised comprehensibility, $\beta = .262$, $r = -.501$) on interest. This paradoxical influence shows that textual complexity is co-dependent on comprehensibility to foster an interest response, which is also confirmed by a non-significant direct correlation (Table 4.3) yet a significant direct regression coefficient (Figure 4.6). The co-dependence explains the Wundt-curve between textual complexity and interest depicted in Figure 4.5.

In Table 4.3 and Figure 4.6, familiarity is shown as one construct. Given the lack of hypotheses that indicate a difference between the two measured types of familiarity in their effect on interest, and given that the behavior of both items was found to be similar (see Figure 4.3a), the added value of a separate treatment was limited. Familiarity was found to be an important predictor of interest and its appraisals. It had a significant direct contribution to interest ($\beta = .354$, $r = .387$) as well as a significant indirect effect via comprehensibility ($\beta = .491$, $r = .345$). This indicates that, given that news articles were used as stimuli, familiarity can be confirmed as a relevant contributor to interest.

The two measures of ECT were found to have good discriminant validity, measuring distinct aspects of the ECT with only a small correlation between the aspects. Of the two aspects, only diversive ECT was found to be of influence on interest (see Table 4.3). This is in line with the hypotheses. Diversive ECT was found to influence interest only indirectly; that is, via its appraisals. Although a direct correlation between diversive ECT and interest exists ($r = .430$), the SEM shows that this relation is mediated via familiarity ($\beta = .307$, $r = .465$). The medium correlation between diversive ECT and familiarity indicate that the diversive ECT is an important predictor of familiarity. Furthermore, diversive ECT also showed a trend to influence the appraised complexity of the articles. This is indicated by a significant direct correlation ($r = -.315$) which is partly mediated by familiarity in the SEM ($\beta = -.092$).

4.5 Discussion

In a study with 30 participants and 18 news articles, the model of textual complexity was put to the ultimate test and confirmed to predict mean subjective ratings of appraised complexity. Moreover, interest was manipulated by the (objective) metric of textual complexity and according to a measure of (subjective) familiarity. The manipulations were successful for interest, either directly, indirectly via its underlying appraisals, or both. As such, textual complexity and familiarity were shown to be important in approaching the "sweet spot" of interest; novel-complex yet comprehensible. A SEM showed that complexity (objective and subjective), comprehensibility, familiarity, and diversive ECT explained 37.80% of variance in interest responses and confirms the importance of familiarity and complexity for the construction of an IX.

4.5.1 Textual Complexity

The model of textual complexity developed in Chapter 3 has been compared to its subjective counterparts, in particular appraised complexity, completing the evaluation of the model of textual complexity. With a maximum correlation of .704, the predictive power of textual complexity on appraised complexity is very large. This confirms that the model of textual complexity indeed reflects appraised complexity and testifies to its predictive validity. Besides the maximum correlation, the average correlation of .563 for all types of classifiers and training data sets further testifies to the predictive validity of the model. Both the maximum and average correlations can be considered an excellent performance. In the first place because the objective model of complexity was used to predict subjectively appraised complexity, and secondly because the model was trained on a distinctly different data set (encyclopedia articles) from the data set it was applied to (news articles). Furthermore, given the precise and operable specification of the model in Chapter 3, this allows for its direct application in information systems. A further discussion of the relation between the different classifiers and predictive validity is postponed until Section 5.2.2.

A key finding of this study is the existence of a Wundt-curve or inverted-U for epistemic stimuli. Hitherto, no evidence existed that confirmed a Wundt-curve for high-level cognitive processes, including the experience of epistemic interest. Moreover, the existence of a Wundt-curve for other processes, such as aesthetic preference for random polygons, has even been called into question (Martindale et al., 1990). In general, attempts to find a Wundt-curve for interest have been abandoned, because, as Silvia (2006) notes, "*explaining (and demonstrating) the inverted-U has proved to be difficult*" (p. 32). In light of a history of almost 120 years (Wundt, 1896) in which the Wundt-curve has been embraced (Berlyne, 1960, 1970) and abandoned again (Martindale et al., 1990; Silvia, 2006), the long-sought Wundt-curve can be considered a remarkable finding, in particular for epistemic interest.

As Silvia (2006) indicates, explaining the occurrence of a Wundt-curve is similarly important to demonstrating it. The SEM partly explains the occurrence of the Wundt-curve. Namely, the SEM showed that textual complexity had a positive (direct) effect as well as a negative (via appraised comprehensibility) effect on interest. This paradoxical influence shows that textual complexity is co-dependent on comprehensibility to foster an interest response. This is in line with the expectations from the appraisal theory of interest, which predicts that the primary appraisals of novelty-complexity will increase preference until the secondary appraisals of coping potential become affected and preference will decrease again. Hence, the findings generally support the original Wundt-curve (Wundt, 1896), Berlyne (1970)'s collative variables that operationalize the curve, as well as the appraisal theory of interest (Silvia, 2006) that explains its shape. However, there is one exception: the non-significant influence of subjective complexity. As will be discussed below the non-significance can be explained from both a methodological perspective as well as from a theoretical perspective.

Methodologically, measuring the positive and negative components that constitute an emotion (Cacioppo and Berntson, 1994) is difficult. The components respond inversely when measured during short time spans, which are typical for emotions (Diener and Emmons, 1984). For the emotion of interest this difficulty is confirmed by the high correlation between appraised complexity and appraised comprehensibility ($r = -.763$). Its strong relation can be expected: less ability will change the perception of complexity. Whether or not an (in)dependence between the (positive and negative) components is found also depends on the scales used (Egloff, 1998). Future research is needed to formulate measurement instruments for the appraisals of complexity and comprehensibility that operate (more) independently. Nonetheless, given the short time span of an emotional response, the extent to which this is feasible remains questionable.

Historically, the importance of objective variables that generate intrinsic pleasantness seems ubiquitous. Aside from being theorized by Wundt (1896) as stimulus intensity, and by Berlyne (1960, 1971) via collative variables, the appraisals theorists also note the importance of objective variables for the appraisal of intrinsic pleasantness. Namely, as Ellsworth and Scherer (2003) conclude: "*it is important to note that the intrinsic pleasantness or unpleasantness detected is mostly a characteristic of the stimulus*" (p. 577). Still, they also acknowledge individual differences (e.g., learning, memory) to play a substantial role in the appraisal. Combined with the methodological difficulty in measuring the distinct attractive and aversive components of emotions, these arguments show why an objective indicator of textual complexity is particularly apt in predicting intrinsic pleasantness and explaining interest.

The objective model of textual complexity seems to uncover salient and otherwise difficult, if not impossible to detect aspects of textual complexity. Since this indicator is based on "small data" of common observations about processing difficulty during reading, it is fair to assume this indicator reflects those aspects of a text that generally increase processing difficulty. The excellent correlation between objective and mean subjective complexity further supports this claim. Also the choice for a rather simple, linear classifier (the LRM) and the training on, and application to texts of comparable size (sections) heighten the transparency of this claim. Notwithstanding, imperfections in the objective indicator of textual complexity could have influenced the results. Although such imperfections are not unique, for example when compared to the problems of a purely subjective approach driven by introspection, it shows that such an objective approach has, aside from clear advantages, also limitations.

4.5.2 Familiarity

The manipulation by familiarity was successful, as shown by the relation between predicted, topical familiarity and actual, article familiarity as well as by the influence of familiarity on interest and its appraisals. When we look at article familiarity as a function of topical familiarity, the former showed a delayed response to the latter. This suggests a difference between the scope of prior knowledge as measured. If we put the measured aspects of familiarity on a continuum from a broad to a specific realm of knowledge, the scope is more specific for article familiarity than for predicted, topical familiarity. Actual, article familiarity, then, requires a relatively high level of topical familiarity in order to occur.

Familiarity was found to be the independent variable with the largest influence on interest. This influence was either direct or indirect, via comprehensibility, and was mainly positive and linear. The linear, positive relation is in line with findings from other, related studies. And, for the indirect influence via comprehensibility, the relation is also in line with theory. However, based on the general assumption that familiarity operates antagonistically with novelty (cf. Berlyne, 1970), the primary appraisals or "*novelty check*" (Silvia, 2006, p. 57) of interest would predict a negative effect of familiarity on interest.

The positive influence of familiarity on interest can be explained by defining the notion of epistemic effect; that is, the extent to which a piece of information changes the knowledge structures of an individual. This cognitive perspective is reflected in a theory by Kintsch (1980) and the knowledge-schema theory by Yarlas and Gelman (1998), which both highlight the importance of how the current knowledge structures of an individual are changed by information. The epistemic effect increases through an interaction between novelty and familiarity. This explains a positive influence of familiarity on interest, in particular in the current study where recent and, likely, novel articles were used. Moreover, in line with the the appraisal-theory of interest, novelty coincides with comprehensibility to, hypothetically, show a Wundt-curve (see Figure 4.1). As Kintsch (1980) summarizes: "*if there is no relevant knowledge structure, or if too great a deviation from it is encountered, there is no way one can integrate the new knowledge*" (p. 93).

Aside from the theoretical implications, the findings support the assumption of "more of the same" that commonly underlies filtering techniques applied in IR and IF&R systems. This interpretation is further supported by the the way in which topical familiarity was measured. Namely, the measurement resembled the algorithmic approaches to modeling prior knowledge and interests. Whether or not topical familiarity actually reflected the predictions otherwise made by these user models is uncertain. This implicates that the extent to which such user models predict actual, article familiarity is an important, open question, to which an answer is needed in order to ascertain the importance of familiarity, as inferred from user models, in predicting interest.

4.5.3 Epistemic Curiosity Trait

The findings on the influence of ECT supply novel insights into how individual differences affect interest responses. Namely, the SEM shows that diversive ECT only influences interest via familiarity. A direct influence on interest, or a direct influence on comprehensibility, were not found. These results explain the conclusions of earlier studies (Silvia, 2008a). Namely, that interest was indeed influenced by diversive ECT via comprehensibility. However, these results also contradict earlier studies by showing that the influence of diversive ECT on comprehensibility was, in turn, fully mediated by familiarity. In other words, it is not comprehensibility but familiarity that mediates the effects of diversive ECT on interest. This suggests that individuals high on ECT estimate their knowledge higher, contrary to their coping potential.

The importance of familiarity does not explain how diversive ECT might have contributed to interest. Either individuals high on diversive ECT over-estimate their prior knowledge, or they actually have more prior knowledge. No indications exist for the former explanation. The latter explanation is more likely, as curious people are driven by a desire to know (Berlyne, 1954; Litman, 2005). However, this posits familiarity as both a consequence[3] and antecedent of interest and leads to a circular argumentation. A possible solution is in the finding of a trend of

[3] As explained in Section 4.2.5, the repeated experience of interest hypothetically leads to long-term interests and, accordingly, heightened familiarity.

diversive ECT to influence the appraised complexity, which shows next to the indirect influence via familiarity. This suggests that individuals high on diversive ECT perceive the world as less complex, instead of appraising their own coping potential higher. However, in light of the previous discussion about the difficulty of distinguishing between both positive and negative components of an emotion, this conclusion needs to be interpreted with caution.

The findings support the distinction between the diversive and specific components of ECT. As expected, diversive ECT had a substantial influence, whereas specific ECT had a marginal, non-significant influence on interest in an experimental situation that lacked a clear information need. Furthermore, the correlation between the diversive and specific components of ECT was of small strength. These findings confirm the distinction between the diversive and specific components of ECT, as well as their relation to interest.

4.5.4 Interest

Overall, the findings are supportive of the appraisal theory of interest, allowing us to cover 37.80% of variance in interest responses. In particular the paradoxical influence of complexity, which can both increase and decrease (i.e., mediated via comprehensibility) interest, gives very strong support to the appraisal model. And, in addition to complexity, the confirmation that the influence of ECT operates via the appraisals further supports the appraisal model. On the contrary, the findings on the influence of familiarity are not supportive of the currently posited set of appraisals of interest. This study as well as related studies suggest that the contribution of novelty to interest should be (re)defined non-antagonistically from familiarity, such as suggested through the notion of "epistemic effect".

Aside from confirming the appraisal theory of interest, the findings are also in line with Berlyne (1960, 1971)'s list of collative variables and the Wundt-curve (Wundt, 1896). In particular the importance of objective textual complexity was noted in explaining interest. This suggests that, contrary to Silvia (2006)'s statement that "*the effects of events on emotions do not stem from objective qualities of the events*" (p. 63), that it is not all subjective and the venerable Wundt-curve should *not* be laid to rest. Instead, given the right operationalization of the objective qualities, it is possible to demonstrate the Wundt-curve of interest and still explain it using the appraisal model. However, we do acknowledge that the relative homogeneity of the group of participants might have lessened the general importance of subjectivity. Nonetheless, the findings join both the appraisal theory and the Wundt-curve, which suggests that a midway between the subjective and objective approaches is valuable for explaining interest.

This study used 18 articles which probably differed on other textual characteristics as well, for example: quality, depth, scope, clarity, surprisingness, incongruity, variability, or puzzlement (Berlyne, 1960; Barry, 1994). Since textual complexity overarches many of these characteristics it is impossible to say, from the current studies, how other characteristics of the text influenced the IX. This problem is partly delineated by having articles from one source, assuring all texts adhere to one set of editorial criteria and raising the internal validity of the manipulation by textual complexity accordingly. Furthermore, this study lacked the notion of an information need. This simplified the experimental setup, yet leaves the influence of an information need on the experience of interest unexplored. Further research is needed to deepen the understanding of the relation between the underlying textual characteristics, (information) needs, and interest.

4.6 Conclusion

This chapter presented an extensive user study on the causes of interest. Several surprising theoretical contributions were made. Namely, familiarity cannot be regarded as opposite of novelty when relating it to interest. And, the ECT indeed influences interest via its appraisals. Yet, via different appraisals than earlier studies suggest. Finally, the remarkable finding of a Wundt-curve confirms early expectations by Wundt (1896) and Berlyne (1970). Overall, the presented study confirms the appraisal theory of interest, albeit with some notes. The explanatory ability of an objective measure of complexity especially is higher than expected from the appraisal theory of interest. This indicates the value of objective measures in relation to subjective experiences, which will further be discussed in Section 5.3.1. Hence, it seems that it is not all subjective.

The described findings are directly applicable in information systems. In particular the findings about familiarity and textual complexity have implications for information systems. Firstly, the importance of familiarity or “more of the same” has been confirmed empirically. This indicates that the assumptions underlying many state-of-the-art information filtering techniques can be confirmed experimentally. Secondly, the findings on textual complexity show how information systems can use textual complexity to foster interest. Both through the development of a generic model of textual complexity (see Chapter 3) and through validating the relationship between its predictions and interest. The robustness of this prediction was confirmed by a large correlation, even when including outliers. Together with the importance of familiarity, both the content (i.e., familiarity) and the format (i.e., complexity) of text can be used by information systems to predict the emotion of interest.

Taking interest as a primary goal allows for the operationalization of the holistic concept of the IX as the amendable goal of predicting if and when a stimulus leads to an interest response. The combination of textual complexity and familiarity shows the feasibility for future systems to target the “sweet spot” of interest by adding both these factors to its set of features. In essence, the model of textual complexity allows the selection of information based on *how* it is written, whereas a measure of familiarity or topicality can select information based on *what* is written (e.g., language models). This study is a first step in better understanding the interplay of information interaction and the experience of interest, and, accordingly, forms a proof-of-concept of the Information eXperience Framework (IXf).

5

General Discussion

Abstract

The studies in this dissertation covered a distinct set of topics, using a diverse set of methods. The upcoming chapter discusses in what way the whole of these studies is greater than the sum of its parts. In particular, the following inter-dependencies are highlighted: the function of the Information eXperience Framework (IXF) as blueprint; the unique, dual evaluation of the model of textual complexity on both a data-oriented task and a user-centered task; and the interest study as a proof of the feasibility of influencing the Information eXperience (IX). Moreover, the value of the somewhat uncommon approaches taken to the studies in this dissertation is discussed. The discussion is concluded by highlighting the value of the IX for the evaluation of information systems.

5.1 Introduction

This dissertation has outlined a novel perspective on the development and evaluation of information systems. Namely, the notion of Information eXperience (IX) was introduced as a salient target for information systems. This monograph has presented both conceptualizations that showed how, and studies that confirmed that information systems can influence the IX. Next, we will provide a concise overview of this monograph denoting the core issues discussed in each of the subsequent chapters.

Chapter 1 argued that information systems should not solely be evaluated on their ability to deliver topical documents, but also on the IX they provide. In particular information systems that support retrieval, filtering, and recommending information were highlighted as systems that can benefit from the IX as a more subjective evaluative approach. The IX can be very beneficial to the goal of finding, learning, and encountering. Moreover, the IX can even become the primary goal, such as with satisfaction, slow thinking, and engagement. Three topics were discussed: the Information eXperience Framework (IXF), textual complexity, and interest. Together, these three topics illustrate the possibility for information systems to foster a fruitful IX.

Chapter 2 explored the intricate relationship between relevance and IX. The chapter first introduced the key concepts of relevance and IX. The chapter continued to define the IXF. With the IXF, the chapter operationalized the potential for mutual reinforcement between relevance and IX. The IXF transformed the vague and ambiguous concepts that form an IX into a set of connected values, responses, and states. Moreover, throughout the chapter an attempt was made to concretize the values, responses, and states of the IX, such as by structuring the "forest of information related emotions" into a "feeling tree" and the "narrative of experience" into an "experience wheel". The chapter finished by showing how information systems can benefit from the synergistic potential between relevance and IX. Namely, how algorithmic relevance influences the IX and how the IX can inform algorithmic relevance.

Chapter 3 introduced a model of textual complexity. The goal of this model was to predict subjective, experienced complexity. It was a first step at implementing part of the IXF by allowing information systems to select texts differing in complexity and, accordingly, value. The construct of processing difficulty was introduced to operationalize appraised complexity. The challenge of predicting appraised complexity using an objective model has proven to be difficult (Benjamin, 2012). Given this difficulty, the validity of the model was of key importance. To attain validity, a set of features was introduced based on common observations about our processing difficulty. This was a novel approach in that it did not rely on either extensive theorizing, or on characteristics of the data. Subsequently, to assure large-scale applicability, these features were combined into a model and trained using two large data sets differing in their degree of textual complexity. In doing so, the contemporary data-driven approach was followed, which is often taken in text classification and related tasks. The results confirmed the success of combining the approaches to classifying texts of two different levels of complexity, with performances of up to 93.62%.

Chapter 4 presented a user study that explored the determinants of the emotion of interest. This study investigated the extent to which complexity, familiarity, and an individual's Epistemic Curiosity Trait (ECT) influenced interest. The model of textual complexity developed in Chapter 3 was used as indicator of complexity. This study served as the final step in the evaluation of this model. The objective model of textual complexity was clearly successful in predicting its subjective counterpart. The average subjective complexity could be predicted with very high confidence, as indicated by a correlation of $r = .704$. Objective textual complexity showed an inverted-U relation or Wundt-curve with subjective interest. To the best knowledge of the authors, hitherto, neither an inverted-U nor a positive effect of textual complexity on epistemic interest have been confirmed before. The finding of a Wundt-curve can thus be considered remarkable and highlights the salience of objective models to alter the subjective IX. As such, this chapter formed a proof-of-concept of the IXF and of the feasibility of changing the IX.

Chapter 5, which you are currently reading, will continue with a discussion of the limitations and implications of the presented contributions on interest, textual complexity, and the IXF. Given the inter-disciplinary nature of the work presented in this monograph, strong inter-dependencies exist throughout this dissertation. The discussion will take these inter-dependencies into account and mainly focus on the implications that arise from the combination of, and synergy between the preceding chapters. The discussion will continue with the novel position that this dissertation has taken within the following three bi-polar dimensions:

- objective and subjective,
- data and theory, and
- reason and feelings.

Moreover, the discussion will be concluded by highlighting the value of the IX for the evaluation of information systems. We will argue that a focus on the IX can be *the* next step in improving our interaction with information – both in terms of efficiency and experience.

5.2 Three Challenges

5.2.1 Interest

Challenge III was "*to identify the causes of the emotion of interest and to explore whether these causes can be influenced by an information system*" (p. 5). This challenge was tackled by executing a user study, which explored the influence of textual complexity, familiarity, and ECT on the experienced interest in news articles. Advanced statistical techniques were applied to show the inter-relations between the influences and a fitting technique to show the relation between independent and dependent variables. A Structural Equation Model (SEM), with an explained variance of 37.80%, confirms that interest can be influenced via complexity, familiarity, and ECT.

Textual complexity was shown to be able to partially predict interest through an inverted-U relation or Wundt-curve. This is a clear and remarkable finding that shows the possibility for information systems to change the experience of interest. Moreover, it confirms a historical relation (Wundt, 1896; Berlyne, 1970) that was yet to be proven for epistemic interest. Perhaps the most important aspect of the revealed Wundt-curve is its horizontal axis, which is based on a specific and rather lengthy processing pipeline. Each of the steps of this processing pipeline were clearly documented and all the decisions were underpinned. No tuning was performed in relation to interest, only in relation to objective textual complexity (i.e., training the classifiers; see Chapter 3) and in relation to subjective appraised complexity (i.e., choosing the classifier; see Chapter 4). This makes the finding of the Wundt-curve not only remarkable but robust as well.

Aside from the influence of complexity, familiarity has also found to be an important predictor of interest. Familiarity shows a linear relationship with interest, confirming that the so-called "filter bubbles" are indeed beneficial to the IX. The familiarity or prior knowledge of a user was measured subjectively. The measurement was set up to reflect contemporary, objective methods used in information science and Information Retrieval (IR). The participants were asked to indicate their familiarity with several keywords. Similarly, objective methods generally compare keywords or terms of a document with a personalized history of information use. Such methods are often applied within information systems to model the familiarity, prior knowledge, or long-term interests of a user (Brusilovsky and Millán, 2007). More generally, such methods are applied to improve the prediction of relevance. Subsequently, this suggests that objective models of relevance and familiarity might have value in explaining emotions as well, in particular the emotion of interest.

Lastly, ECT was shown to influence interest. Contrary to expectations, this influence was fully mediated by familiarity. Although not impossible, ECT was not measured with the intent to be applied in information systems. Its contribution was purely on a theoretical level, on which the findings generally support the interest-appraisal theory, yet raise several questions about why the influence of ECT on interest is almost fully mediated via familiarity.

Together, the influence of textual complexity, familiarity, and ECT show that interest can (partially) be explained and predicted. In particular textual complexity and familiarity can be measured by information systems, allowing them to influence the emotion of interest and, accordingly the IX. Consequently, the study on interest solves Challenge III.

5.2.2 Textual Complexity

Challenge II was "*to develop a model of textual complexity that is applicable to a variety of data sets and predictive of subjective appraisals of textual complexity. Moreover, a key challenge for this model is to be able to influence a user's* IX" (p. 5). To solve this challenge, Chapter 3 developed a set of features based on common observations about processing difficulty (i.e., "small data"). Subsequently, these features were combined into a model. The model was tuned using a large training set distinctive on complexity (i.e., "big data"). The resulting model of textual complexity was evaluated twice. First in Chapter 3 on the data-oriented task of classifying encyclopedic articles as either simple or normal. Second in Chapter 4 on the user-centered task of predicting the subjectively appraised complexity of news articles.

The applicability of the model was assured by developing the model independently of text length and semantics, assuring that article length and genre were of little influence on the classification outcome. This reduced the risk of overfitting and enabled the model to be applied to a profoundly different data set (i.e., news articles, as described in Chapter 4) than the data used to tune the model (i.e., encyclopedia articles, as described in Chapter 3). This is an indicator of the generic applicability of the model. Moreover, as shown in Chapter 3 the model was applied on a total of 138 790 articles. This shows that the model scales to large data sets as well, further confirming the applicability of the model.

The ability of the model to predict subjectively appraised complexity and to influence the IX was confirmed in Chapter 4. The user evaluation in Chapter 4 put the model to a final, thorough test and showed that the output of the model correlated highly with perceived complexity. Moreover, Chapter 4 confirmed the ability of the model to influence the IX. The user study showed that textual complexity influences the experienced interest in an inverted-U manner which is coherent with the appraisal theory of interest (see Section 5.2.1). The user evaluation gives a unique confirmation of the predictive validity of the model.

Testing the model of textual complexity on both a data-oriented and user-centered task indicates its effectiveness on multiple levels. This is illustrated by the mixed results on which type of classifier is optimal. Whereas the Logistic Regression Model (LRM) and Support Vector Machine (SVM) showed peak performances on respectively the user-centered and data-oriented tasks, the Random Forest (RF) showed the best overall performance. Given that the performance of the LRM classifier is often considered a baseline to which other classifiers are compared, it is remarkable that the LRM was responsible for the peak performance in predicting subjective complexity. An explanation might be found in the linearity of the LRM. Since all features were also devised to give a linear indication of an aspect of processing difficulty, it is not unlikely that a linear classifier achieves peak performance. Hence, the success of a linear classifier to predict subjective complexity further confirms the content validity of the features, as well as the predictive validity of the model.

The common approach to the evaluation of a model of textual complexity generally tests its ability to predict new and often objective ratings for the same data set. By testing the model on its ability to predict subjective appraisals of complexity for a new data set, this dissertation goes beyond this common approach. It assures the model actually predicts (part of) an IX. This type of application of a metric of textual complexity is an exception (for an exception, see Collins-Thompson et al., 2011)), in particular when used to manipulate the IX. Accordingly, the model forms a novel contribution that managed to solve Challenge II.

5.2.3 Information eXperience

Challenge I was "*to transform the fuzzy concept of the* IX *into an amendable target for information systems. This requires the identification and specification of the different aspects that constitute an* IX *and relate them to the notion of relevance that generally underlies information systems*" (p. 5). To achieve this goal, many aspects of the IX had to be identified and made operational. In particular a bridge needed to be built between relevance and experience.

Operationalizing the fuzzy concept of IX was done by identifying which aspects make up an experience. In doing so, the IX is defined as the combination of values, (emotional) responses, and (cognitive-affective-motivational) states. This definition is an abstraction of several models of User eXperience (UX) which already solve part of the fuzziness surrounding the topic. Furthermore, in following McCarthy and Wright (2004), these aspects of the IX were depicted as a narrative that evolved over time during an information interaction session. The multitude of aspects and their inter-connections were modeled into the IXF. For each of the values, responses, and states, a significant contribution was made. For values, the existence of non-instrumental relevances and values was proposed, and the relation between relevance(s) and value(s) was identified; for responses, a feeling tree was introduced by which the different (information) emotions were identified and their occurrence explained, and; finally, three key states were identified that can occur during information interaction. Considering that previous studies often used the triplet of thoughts, feelings, and actions to describe an experience (Hassenzahl, 2013; Kuhlthau, 2004), numerous extensions on the status quo were needed to transform the fuzzy concept of IX into a clear and operable IXF.

Transforming the fuzzy concept of IX into an amendable target for information systems is not achieved by only operationalizing it. To make the IX a workable target for information systems, the connections between (algorithmic) relevance and values, responses, and states of the user were identified. These connections were identified in two directions. On the one hand the possibility was illustrated that certain aspects of a text can give rise to certain values. These values can subsequently foster responses which, in turn, can change the state of the individual. Building on the IXF, exact routes of influence were identified, such as the role of complexity which was studied in this dissertation. On the other hand, the possibility was denoted for the responses and states to inform algorithmic relevance. This possibility was introduced by, amongst others, Arapakis (2010) and Wilson (2006), yet only marginally specified by them. In this dissertation detailed routes of information were defined, either as affective feedback (based on responses) and affective relevance (based on states).

As stated in the introduction (Chapter 1), it is unclear what exactly constitutes a fruitful IX: neither is it clear which emotional experiences are desirable or "useful" during interaction, nor what their causes or effects are (Kuhlthau, 2004; Arapakis et al., 2008; Belkin, 2008; Bowler, 2010). The IXF has been proposed to fill this gap. However, the IXF, like any other model, will never give a complete picture of reality. But, this is not necessary either since a model derives its value from being useful. This monograph, as a whole, forms a proof-of-concept of the phenomenon of IX and the usefulness of the IXF. The relation between relevance and IX was demonstrated in one direction: namely, algorithmic relevance can influence the values and responses. The model of textual complexity developed in Chapter 3 is an implementation of an algorithm of relevance. This was confirmed in Chapter 4, where the model influenced the appraisals (i.e., relevance criteria) of complexity and comprehensibility. Subsequently, the model was also shown to influence the emotion of interest. As such, this monograph posits and confirms the possibility that an algorithm can influence both value and response and, accordingly, shows the usefulness of the IXF.

5.3 Three Approaches

The approaches taken to the studies in this dissertation can be regarded somewhat unusual. The value of these approaches will be assessed next within the general context of their respective research fields.

5.3.1 Objective and Subjective

Objective properties have been a focus since Wundt (1896) introduced the relation depicted in the Wundt-curve. He compared this relation between stimulus intensity and emotion with the influence of brightness on the experience of color. However, attempts to replicate a Wundt-curve have been abandoned in favor of subjective approaches (Silvia, 2006; Martindale et al., 1990). This dissertation approached textual complexity as an objective property, and IX as a subjective phenomenon. As such, similarly to Wundt (1896) and Berlyne (1970), a mixed position was taken that included both objective and subjective variables. In the remainder of this section, the value of this mixed position for the study of emotion in general and interest in particular will be discussed.

The Wundt-curve, revealed in Chapter 4, is an excellent summary of the synergy that arises from a comparison between objective and subjective variables. The objective model of textual complexity manages to uncover unique and difficult, if not impossible, to detect aspects of textual complexity that have very clear effects on the experience of interest. This suggests that both Wundt (1896) and Berlyne (1970) were right in postulating the importance of objective variables for the experience of emotions in general and the importance of complexity for the experience of interest in particular. Notwithstanding, although the Wundt-curve seems quite robust, there is still potential for optimization. Namely, a more granular comparison between objective properties and subjective experience is possible. Such a comparison will nuance how certain stimulus properties influence the experience. That such a detailed analysis is now (also) possible with epistemic stimuli makes the Wundt-curve a novel finding and a promising direction for future research.

Moreover, this illustrates the value of the combination of objective and subjective variables.

It is remarkable to note that objective characteristics influence the IX, contrary to the subjective appraisals of these characteristics. This highlights that there are limitations to introspection with regards to the appraisals underlying emotions. On the one hand, some appraisals that are important for emotions are highly automatic and unconscious (Ellsworth and Scherer, 2003), making it difficult to measure them. This was already illustrated in Section 4.5.1 via the difficulty of measuring the distinct positive and negative parts of an emotional experience. On the other hand, perception is not always coherent with the properties of the perceived stimulus. For example, it is well-known that color perception is biased by expectations or top-down processing (Van den Broek et al., 2005, 2008).

Similarly, our perception of complexity and even of relevance might be colored. An example of this was given in Section 2.5.3, where hard-to-read fonts increase the perceived difficulty of written instructions (Song and Schwarz, 2008). In sum, the subjective appraisals are not always known consciously and are sometimes biased due to top-down expectations. Accordingly, this indicates why an objective approach can have value over a purely subjective approach.

Taking a rather philosophical stance, the extent of objectivity or what is referred to as objective can be questioned: because as Husserl (1931) argues it is the intersubjective experience that constitutes the objective spatio-temporal world. Hence, objectivity can be understood as the common denominator of subjective appraisals of a stimulus. This argument is confirmed by the model of textual complexity, which has been tuned to optimally predict (subjectively appraised) complexity. Nevertheless, it is objective in the sense that its sole input is a text and, thus, it only describes stimulus properties. A similar case can be made for implementations of relevance, where the Cranfield paradigm is called "objective relevance" (Saracevic, 2007), yet is based on expert ratings of relevance (Voorhees, 2002). Therefore, essentially objective can also be seen as the appraisal of a stimulus property defied of any inter-personal and intra-personal variance. Yet, as this excludes top-down biases and other momentary influences it still shares the aforementioned benefits of objectivity.

5.3.2 Data and Theory

This monograph presented an intermediate approach in between contemporary data-driven and theory-driven approaches (cf. Manhart, 1996). As such, it forms a nuanced approach to Artificial Intelligence (AI), adding some findings about human intelligence to the data-driven approach. This intermediate approach circumvents the problem of modeling (aspects of) intelligence not from a purely data-driven perspective, nor by modeling unifying theories of (aspects of) intelligence, but circumvents it by modeling basic findings about intelligence. Hence, the approach taken stepped back from comprehension to processing difficulty and from theory to “small data”. As was highlighted in Section 5.2.2 the resulting model performed very well on the classical task of text classification, and achieved an excellent prediction of subjectively appraised complexity. Furthermore, it attained both applicability and (predictive) validity. Beyond this practical value, the intermediate approach has theoretical value and limitations as well, which will be discussed next.

The intermediate approach shares some of the limitations of both the data-driven and theory-driven approaches. On the data side, the model is somewhat data-dependent, aside from sharing the applicability and validity. This dependency was reduced by deliberately excluding the influence of text length and semantics on the features. However, true data-independence is clearly infeasible given that the model was tuned using “big data”. This is also confirmed by the influence of the length of the texts used for training (i.e., “big data”) on the ability to predict subjectively appraised complexity. The higher the correspondence between the text lengths of training (i.e., the Wikipedia data sets) and test data (i.e., the Guardian data set), the better the predictive ability. On the theory side, the model shares some of the issues concerning content validity (Frijda, 1967). Namely, the extent to which the algorithms actually reflect the common observations they are devised to reflect. Although this problem is less pressing than when an actual theory is being modeled, it does not disappear. For example, numerous algorithms exist that measure cohesion. This suggests a difficulty in achieving content validity, at least for this particular feature.

Although an intermediate model largely excludes theoretical underpinnings, it is still feasible to derive theoretical implications from evaluating the model. Namely, by zooming in on the underlying features of the model of textual complexity, it is possible to evaluate each of the features on their ability to predict complexity. The features and model have known numerous iterations. Although not reported in this monograph, several other algorithms were explored before arriving at the current model. This process can be interpreted as iterative model building, in which a search for the most discriminative features lead to different versions of the model. All subsequent models were tested and compared on their explanatory capacity with respect to objective complexity. Moreover, the relation between the features and objective complexity was explored via the principal component analysis in Section 3.6.3, which showed the relations between the features and explanatory power of the principal components. It showed which features are most salient. Given that an an ecologically valid data set was used, the most discriminative features are probably also the most salient common observations. Hence, essentially the resulting model can already form a fundament for a unifying theory of processing difficulty.

Aside from comparing the features to complexity, it is also possible to compare components of the model to other phenomena. The choice for a simple, transparent LRM classifier in Chapter 4 highlights this possibility. For example, it is possible to differentiate between levels of processing difficulty and interest. It can be expected that particular aspects or levels of processing difficulty foster interest, whereas others have a neutral or maybe even negative contribution to interest. Given the opposite roles of comprehensibility and complexity for interest, it is expected that those aspects of processing difficulty that contribute to interest are partly different from those aspects of processing difficulty that contribute to comprehensibility. As such, the intermediate approach provides a (strongly needed; Long et al., 2006a) sophisticated model of common observations about processing difficulty. Moreover, it allows us to hypothesize and test (features of) the resulting model in relation to subjective consequences (e.g., comprehension, interest).

5.3.3 Reason and Feelings

A central proposition guiding this dissertation is that reason and feelings are inherently inter-connected. In other words, cognition and emotion are intertwined (Van den Broek, 2011). As was introduced in Chapter 2, a contemporary and related distinction is noted between automated, fast Type 1 and controlled, slow Type 2 processes (Evans, 2008). Individuals generally apply Type 1 processes when dealing with problems, judgments, and preferences. When using Type 1 processes, individuals rely on existing knowledge (i.e., top-down processing) and on emotions (Schwarz and Clore, 2007). Although often flawed, it allows us to handle quickly and effortlessly. When necessary, deliberate Type 2 processes can modify or override Type 1 processes. As such, cognition is adaptively tuned to situational demands (i.e., the tuning hypothesis[1]). It appears that most studies that indicated fallacies in reasoning or Type 2 processes actually tested Type 1 processes (Kahneman, 2003; Schwarz and Clore, 2007). Seemingly, whereas most studies focus on reason, most individuals rely on feelings. As we will highlight next, this creates a gap between contemporary models and an opportunity for future, inter-disciplinary research.

The understanding that most individuals rely on feelings instead of reason clearly has implications for information systems and thinking about relevance.

[1] For a definition, see p. 47

The strong inter-relation between feelings and reason can partly explain the "unexplained natural variability" seen in the ratings used as ground truth for recommender systems (Hill et al., 1995; Herlocker et al., 2004). Furthermore it can also explain the low inter-rater agreement between experts in deciding on relevance with respect to a query (Voorhees, 2002). Following these insights, this monograph introduced the notion of non-instrumental relevance. Instead of solely looking at instrumental or task-related relevance, including non-instrumental relevance will contribute to a more adequate explanation of what makes information relevant. More generally, the notion that relevance is actually dependent on feelings should not be neglected. This monograph has outlined how relevance judgments depend on affect, such as via the affect-as-information mechanism. Contrary to the proposition that topicality is a pre-condition for relevance (Spink and Greisdorf, 2001), this suggests that what is relevant is not necessarily topical. In other words, non-topical information can also be relevant. A clear example of this is encountered during serendipitous information behavior when information is encountered unexpectedly (Foster and Ford, 2003).

When stating that feelings and reason are intertwined, this implies that feelings not only affect reason, but reason also affects feelings. This notion is well-established, for example via the cognitive-appraisal theories of emotion (Ellsworth and Scherer, 2003) and the interest-appraisal theory (Silvia, 2006). These theories state that an emotional experience is dependent on cognitive appraisals. In particular when determining the relevance and significance of a stimulus the influence of cognition is salient. The appraisals of these motivational bases depend on an individual's long-term goals, needs, and values (Ellsworth and Scherer, 2003). The motivational bases are often of a cognitive nature, such as with an information need. Relevance and pertinence[2] have been noted as important appraisals within the appraisal theories of emotion (Ellsworth and Scherer, 2003), while their significance is not clear yet. As Scherer et al. (2006) notes: "*It is not unreasonable, then, to assume that real affect is only induced when the stimulus is appraised as pertinent by the individual*" (p. 109). Hence, the influence of the motivational bases on the experience of emotions is an unresolved research endeavor (Scherer et al., 2006) to which (studies on) relevance can contribute.

The close relation between reason and feelings highlights the potential of an inter-disciplinary approach. The motivational bases underlying emotional responses illustrate this possibility. On the one hand, the influence of the motivational bases on the experience of emotions is an unresolved research endeavor. On the other hand, relevance and pertinence are primary subjects within information science

[2] Within the appraisal-theories, pertinence refers to the impact that a relevant stimulus has on an individual (Scherer et al., 2006).

and IR. Relevance has even been posited as "*a, if not even the, key notion in information science in general and information retrieval in particular*" (Saracevic, 2007, p. 1915). Moreover, the problems in modeling and predicting relevance further illustrate the potential inter-disciplinary synergy. Namely, current IR and Information Filtering and Recommending (IF&R) systems struggle with a dynamic, multi-dimensional, non-binary, and user-centered implementation of relevance. Yet, the IX in general and emotions in specific have been positioned in this monograph as a feasible and promising solution to this challenge. In sum, the relation between reason and feelings suggest a synergy between information science, information retrieval, and psychology.

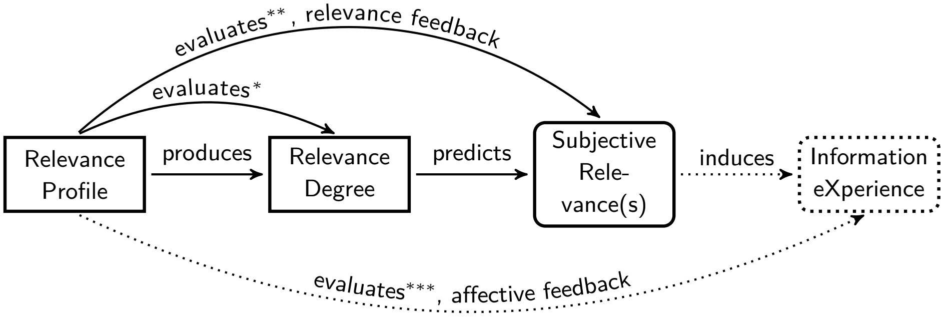
Aside from the possible inter-disciplinary synergy depicted for the relation between reason and feelings, the other two bipolar dimensions also indicate the potential of an inter-disciplinary approach. The relation between objective and subjective indicate the importance of AI to psychology. In a similar vain, the relation between data and theory stresses the importance of psychology (or, in the case of processing difficulty, psycholinguistics) to AI and, also, vice versa. This inter-disciplinary potential was opportunistically noted by Allport (1980): "*Artificial Intelligence will ultimately come to play the role vis-à-vis the psychological and social sciences that mathematics, from the seventeenth century on, has done for the physical sciences*" (p. 31; cited in Manhart, 1996). In agreement, although to a slightly lesser extent, we suggest that an inter-disciplinary opportunity exists that is as yet mostly unexplored.

5.4 Next Generation Information Systems

This section will discuss the implications of the results presented in this dissertation on interest, textual complexity, and the IXF for current information systems. Particular focus will be on the evaluation of IR and IF&R systems and the way the IX transforms this. Figure 5.1 gives a concise overview of current approaches to evaluate relevance and the possibilities that acknowledgment of the IX adds to this.

5.4.1 Relevance Profile

Current IR and IF&R systems struggle with the implementation and evaluation of multi-dimensional, dynamic, non-binary, and user-centered relevance (Saracevic, 2007; Borlund, 2003). To support the multi-dimensional nature of relevance, this dissertation introduced the notion of a relevance profile. This profile denotes an objective model containing features that describe the (relation between the)


evaluates**, relevance feedback
evaluates*
Relevance Profile
produces
Relevance Degree
predicts
Subjective Rele- vance(s)
induces
Information eXperience
evaluates***, affective feedback

Note. Evaluation based on: * Cranfield (Cleverdon et al., 1966); ** Interactive Information Retrieval (Borlund, 2003) and click-through data (Joachims, 2002); and *** Information eXperience framework

Note. Rounded boxes denote the user side, squared boxes the system side.

Figure 5.1: Concise overview of approaches to evaluate relevance. The possibilities that are added by the Information eXperience are indicated via dotted lines.

information, query, and (in this case) the user. It can be thought of as the objective counterpart of a subjective appraisal profile that underlies an emotional response (Scherer, 2004). Moreover, the idea of a relevance profile fits within the principle of polyrepresentation, which states that a combination of different, distinct cognitive representations improve the prediction of relevance (Ingwersen, 1996).

Key to a relevance profile is that it can be used to select and (re)rank search or recommender results. For this, a relevance profile is merged into a single value which can be compared to a relevance degree (See Figure 5.1). Moreover, this enables us to perform a distance measurement between relevance profiles and, accordingly, select the most relevant pieces of information. Although the addition of complexity to a relevance profile was thus far fairly unexplored, current implementations of a relevance profile often already exist in the form of a feature vector. As we will argue next, instead of the implementation of the relevance profile, the crux is in its evaluation.

The standard approach to evaluating relevance profiles is to use a singular relevance value or relevance degree. The predictions derived from the relevance profiles are often compared to binary relevance assessments. These assessments are, for example, obtained via the Cranfield-paradigm that allows information experts to rate the relevance of a document for a given query (Cleverdon et al., 1966). However, a singular, binary relevance degree is unable to fully capture the notion of relevance. This can be deduced from the finding of unexplained variability in the ground truth. Namely, generally a low inter-rater agreement is found between information experts when deciding whether a document is (topically) relevant given a query (Voorhees, 2002). And, the existence of a "magic barrier" for recommendation accuracy further highlights the existence of unexplained variability (Hill et al., 1995). This unexplained variance can partly be solved by acknowledgment of the non-binary, multi-dimensional nature of relevance (profiles).

The interactive IR evaluation model (Borlund, 2003) gives a clear framework on how to include non-binary relevance assessments in the evaluation of relevance (see Figure 5.1). By devising realistic information tasks, fully controlled laboratory experiments can be used to explain relevance decisions from multiple perspectives and in a non-binary fashion. As such, a larger proportion of the variability in relevance decisions can be explained. However, essentially this still evaluates relevance (profiles) as a singular degree; it does not give insight into the importance of the underlying features of a relevance profile. The IXF allows us to extend this evaluation and further the explanation of variability in multi-dimensional relevance.

In addition to evaluating a relevance degree, the IXF suggests that a relevance profile can be compared to a response (see Figure 5.1). Namely, the IXF indicates how different aspects of a text can lead to different responses, where the desired response is dependent on the goal (e.g., finding, encountering, learning). This implies that an activity can be associated with an optimal response (pattern), which in turn suggests that a relevance profile should be geared to a specific response as well. Aside from an optimal response, during the course of an information interaction a smorgasbord of responses can occur. As indicated by the IXF, these responses reflect on the underlying relevance profile. Tentatively, Chapter 2 concluded that the resulting information emotion is a reflection of the value and control that is experienced by the user. As such, a focus on responses allows us to evaluate not solely on a singular notion of relevance, but to evaluate on the salience of specific features of a relevance profile. This suggests the feasibility of evaluating a relevance profile on the responses it generates, as illustrated in Figure 5.1.

5.4.2 Interest Profile

An example of a relevance profile that supports the goal of encountering is one that is geared towards an interest response; that is, an interest profile. Key to this interest profile is, besides familiarity and ECT, the implementation of textual complexity.

It is easy to imagine a system that uses the model of complexity from Chapter 3 in its relevance profile to foster an interest response. Consider a service that stores an index of searchable documents that includes the analysis of textual complexity as presented in Chapter 3. This would already allow the addition of an indication of complexity to the relevance profile. Furthermore, a similar service can keep track of, and analyze the complexity of the texts that a user has read or written. This allows the personalization of the level of complexity such that it becomes optimal for a user and, subsequently, gives an indication of the comprehensibility of a text for that user. Together, these services allow the selection of information likely to be appraised in the "sweet spot" of interest; complex yet comprehensible. In essence, a service that provides analyses of complexity allows the selection of information based on *how* it is written, contrary to *what* is written (e.g., language models) or *how much* data (e.g., information theory) is required to store it.

It can be argued that methods currently applied for topicality and familiarity already indicate the complexity of a text as well. Current implementations of topicality in IR systems and familiarity in IF&R systems are often based on some form of similarity (Hanani et al., 2001; Saracevic, 2007), such as via language models (Goodman, 2001) that give the similarity between texts (e.g., documents or queries). With these similarity-based implementations, topicality and familiarity or "more of the same" will overlap with the level of complexity that a user seeks or normally reads. However, this is merely an overlap and not a full coverage of the effects otherwise indicated by a model of textual complexity. The contribution of topicality, familiarity, and complexity to relevance relate to each other, but are not the same. This can be considered a premise of the notion that relevance is multi-dimensional (Schamber, 1994). And, this can also be deduced from the finding that topicality is often merely a pre-condition for relevance (Spink and Greisdorf, 2001). Therefore, a distinction is needed between topicality, familiarity, and complexity to give a more precise prediction of, in general, relevance and, in particular, an interest response. To implement this distinction, one cannot solely rely on measures of similarity. Hence, to optimally predict an interest response, a relevance profile which includes a metric of complexity is a necessity.

The importance of evaluating a relevance profile on a response, such as interest, becomes apparent when comparing the relation between objective complexity, appraised complexity, and interest. On the one hand the classifier performance on objective complexity was largely, though limitedly, related to subjectively appraised complexity. For example, whereas in Chapter 3, the LRM classifier had the lowest performance out of three classifiers, in Chapter 4 the LRM showed the highest performance. On the other hand, the performance of the model of complexity debilitated with respect to its effect on the IX. Namely, a complex inverted-U relation was found between textual complexity and interest.

When using the IXF the difference between objective performance, subjective appraisals, and a user's responses can clearly be accounted for. The IXF highlights the way in which information metrics change the subjective perception of certain relevances. This perception in turn combines with other aspects of relevance and together leads to a response. Hence, it is of no surprise that a difference exists between objective performance, subjective appraisals, and a user's responses. Neither is it unlikely that this difference extrapolates to other indicators of relevance as well, including topicality. The limitations of the interpretation of algorithmic performance are salient and point to the importance of evaluating the effects on the IX through the lens of the IXF.

5.4.3 User-centered Evaluation

This monograph has shown the close connection that exists between (aspects of) relevance and the IX. It has shown that the two are thoroughly interweaved, which implies that the evaluation of information systems should reflect this and include both relevance and IX. As we will argue next, acknowledging the importance of the IX is (also) possible within large-scale evaluation of text retrieval methodologies. This will enable the evaluation of multi-dimensional, dynamic relevance as well as the evaluation of the IX itself as part of a system's usefulness.

Acknowledging the importance of the IX allows us to evaluate how (algorithmic) relevance profiles affect the user's perceived, subjective relevance(s) and experienced responses. This makes the IX an ideal next step in evaluating relevance. However, evaluating the IX is costly. Similar to the work presented in this dissertation, it requires both experimental rigor and algorithmic inventions and is therefore inherently inter-disciplinary and laborious. This creates a problem of scalability, which is particularly salient for IR systems given their large amount of searchable data. For the evaluation of relevance, the problem of scalability was solved by the Cranfield-paradigm in which a large ground truth was created consisting of relevant and non-relevant documents given a query (Cleverdon et al., 1966). Accordingly, Figure 5.1 can be thought of as a ruler, going from high scalability (left) and decreasing towards the right side when the user is included in the evaluation of information systems. A mid-way solution to including the user in the evaluation of IR systems is through relevance feedback and implicit measures of relevance, such as click-through data (see Figure 5.1). Essentially, this allows us to fine-tune any relevance profile and (partially) include the user in the evaluation.

The problem of scalability is one of "small data" and "big data", which has been discussed extensively in this monograph. Studies on the IX will inevitably result in “small data”, whereas algorithms need to be optimized for “big data”. Several solutions are foreseeable that solve this dichotomy. In the first place a two-step approach can be taken to relevance which differentiates between ranking and re-ranking. Namely, whereas the basic evaluation of algorithmic relevance is provided through “big data” (i.e., Cranfield; Cleverdon et al., 1966), the algorithms can subsequently be hand-tuned or machine learned (i.e., learning-to-rank; Liu, 2009) using “small data” resulting from studies on the IX. Secondly, a ground truth can be based on measurements of the IX of actual users. Using the resulting “small data”, a relevance profile can be created of the relation between objective variables (e.g., about the information, user, and query) and the resulting IX (cf. “simulated users”; Azzopardi et al., 2011). Basic statistical assumptions state that the relations found in properly derived “small data” extrapolate to a larger scale. The task of algorithmic optimalization can then focus on a less costly optimalization of the objective features underlying the relevance profile. Hence, similar to the approach common for large-scale evaluation of text retrieval methodologies (e.g., TREC; Voorhees, 2005), the algorithmic optimization can be separated from evaluating the IX.

An alternative to optimizing algorithms upfront is to change them dynamically. Numerous (implicit) features can be used to increase or decrease the weight of several features in the relevance profile. In the case of relevance feedback, measures such as click-through data are used to (de)emphasize certain query terms (see Figure 5.1). Similar to relevance feedback, the affective response of a user is particularly promising as feedback. Hypothetically, the IX can be measured using various proxies derived from interaction data (Agichtein et al., 2006), query analysis (Ruthven, 2012), eye-tracking, or psychophysiological measures (Arapakis, 2010). When applied as affective feedback, these measures can be used to give feedback to an information system about a user’s responses to information (see Figure 5.1). The method of affective feedback is similar to Arapakis (2010)’s affective relevance feedback, yet with adherence to the multi-dimensional nature of relevance. It is not feedback on a singular notion of relevance, but feedback on a relevance profile. Similar to the regular evaluative function of the IXF, real-time feedback about the IX can improve the prediction of relevance and the IX dynamically, in particular on the value and control that is experienced by the user.

Aside from evaluating (algorithmic) relevance profiles with the concept of IX, the IX itself can be evaluated in light of an activity and goal that an information system is developed to support. The extent to which the resulting IX is beneficial to a target goal can be regarded part of a system’s usefulness. For this, a target IX should be identified and related to task performance. This monograph has already proposed three prototypical experiences and related them to an information interaction outcome. Namely, slow thinking was proposed as beneficial for learning; engagement for encountering; and, satisfaction for searching and common ad-hoc retrieval. However, a positive influence of these experiences on these outcomes is yet to be proven. Nonetheless, the potential of the IX to increase a system's usefulness is clear. A focus on the IX has the potential to counter negative emotions, filter bubbles, and even fast thinking. The IXF forms a method to evaluate the IX and accordingly, gives clear guidelines on how to include the user in the evaluation of information technology.

5.5 Conclusion

In this monograph we took on three challenges:

I The elusive concept of experience was translated into an IXF. We thus established the IX as an amendable target for information systems, and advanced the theoretical insight into the complex interaction between information and its users or, in other words, relevance and the IX.

II An advancement was made on modeling textual complexity. A model was developed that has an excellent performance on the traditional task of text classification, and besides sets a new benchmark performance on the challenging task of predicting subjectively appraised complexity on a foreign data set. Moreover, the introduction of the notion of processing difficulty may yield a next generation of models of textual complexity.

III The results of the interest study led to an increased understanding of the interplay between the determinants of the emotion of interest. The influence of textual complexity in particular led to surprising findings. We confirmed the existence of the "sweet spot" of interest, novel-complex yet comprehensible, and we revealed the Wundt-curve for epistemic interest. Given its long history (Wundt, 1896), the long-sought Wundt-curve can be considered an eminent result.

The work on each of these challenges resulted in significant advances on the state of the art. The significance of the work is the result of the three approaches that were used:

- Reason and feelings were considered inherently inter-connected. This allows us to extend beyond an ingrained distinction between faculties of the mind and create a synergy between them. This synergy was implemented in the

IXF through modeling the inter-connection between relevance and the IX. Accordingly, the IX was transformed into an amendable target for information systems.

- An intermediate approach was proposed to AI. This intermediate approach circumvents the challenge of modeling (aspects of) intelligence by modeling basic findings about intelligence (i.e., "small data"). This nuanced approach shares the applicability of the contemporary data-driven approach as well as the (predictive) validity of the theory-driven approach and explains the excellent performance achieved on both classifying objective and predicting subjective textual complexity.

- An integrative approach to the often found duality between user and system was proposed by combining both an objective and a subjective operationalization of constructs. This approach allows us to lift limitations of subjective appraisals and an objective operationalization. Essentially, where objectivity stops, subjectivity enters and vice versa. By combining the objective and the subjective, unique and difficult to detect relations between user and system can be revealed, such as the Wundt-curve.

These approaches by themselves are unconventional. Their combination is unique in information sciences and related areas and, more importantly, powerful, as is illustrated by the results presented in this monograph.

The interdisciplinary stance taken in this monograph holds promise for numerous key problems, such as the optimization of the prediction of relevance, modeling textual complexity, and explaining epistemic interest. Accordingly, the work presented here can serve as a next step in information interaction, both in understanding and improving it. The proposed shift of focus to the IX has the potential to foster lifelong learning, create and sustain interest, and even turn an information overload into a feeling of ecstasy. As such, the notion of an IX creates new dreams; simultaneously, the IXF shifts frontiers as well.

Through building on the IXF, this monograph is a proof-of-concept of the ability to understand and amend the IX. As such, this work is a starting point for contemplating the implications of information systems that are not only emotionally aware, but make an attempt to actively manipulate the IX. By continuing this line of inquiry, it will become possible to envision information systems that create and maintain a fruitful IX for their users.

A

Filter Model

A model of textual complexity was specificially set up to filter articles from the Guardian data set (see Section 4.3.3). This model was an early version of the model presented in Chapter 3. This appendix will give a concise description of the model. An extensive description and evaluation can be found in Van der Sluis et al. (ip).

Features

All computational features that constituted the model will be described. Starting:

len1 $= |c \in w|$, word length in characters c per word w (see Section 3.3.1).

len2 $= |s \in w|$, word length in syllables s per word w (see Section 3.3.1).

wps $= |w \in X|$, the number of words w in a sentence X (see Section 3.3.3).

fam $= \log_{10}\mathrm{cnt}(w)$, the logarithm of the term count cnt per word w (see Section 3.3.3). For the term count function cnt the Google Books N-Gram corpus was used.

loc $= I(D)$, sentential integration costs where D is the collection of dependencies within a sentence (Equation 3.13, p. 87) (see Section 3.3.3).

The following two features use a sliding window f_w (Equation 3.4, p. 84) of entropy H_n (Equation 3.2, p. 83) and probability mass function (PMF) $p(x)$ (Equation 3.6, p. 85) (see Section 3.3.2):

cha$_n$ $= f_w(X)$ with $f(X) = H_n(X)$, a sliding window of n-gram entropy using PMF $p(x)$ where X is an ordered collection of characters x.

wor$_n$ $= f_w(X)$ with $f(X) = H_n(X)$, a sliding window of n-gram entropy using PMF $p(x)$ where X is an ordered collection of words x.

One feature was included in the model that is not described elsewhere in this monograph:

dal $= \frac{|\{w \in T | w \in D\}|}{|w \in T|}$, frequency of words on the Dale list of 3000 common words D (Chall and Dale, 1995) in a text T.

Method and Results

The model was trained and evaluated in a similar way as described in Section 3.5. The differences will be highlighted. The same data sets (Simple Wikipedia and

English Wikipedia) were used as described in Section 3.5.1 with two exceptions. First, only articles were used. Second, per data set only the oldest 10 000 articles were used, a total of 20 000 articles. Furthermore, the same feature extraction techniques were used, with one exception: for the sliding windows, a window size of $w = 100$ was used for words as well as for character entropy (features wor_n and cha_n).

To evaluate the model, three steps were performed: preprocessing, classification, and validation. As preprocessing, first variables containing more than 25% missing values were removed. Second, observations containing any missing value were removed. For classification a Random Forest (RF) was used. The method was tuned the same way as described in Section 3.5.3. For validation the classifier was trained on 80% and tested on 20% of the data set. The data was balanced to assure it contained an equal number of articles from the Simple Wikipedia set and the (normal) English Wikipedia set.

The model was trained on 10 336 articles and tested on 2 584 articles. The resulting model consisted of a total of 17 features and achieved a classification accuracy of 90.87%. Several tests (all described in Section 3.5.3) confirm the classification accuracy: the Area Under Curve (AUC) was .967; the F1-score was .908; the Phi correlation or Matthew's correlation was .817. Note that the range of the Phi measures lies between -1 and 1.

B

Questionnaires

Several questionnaires were used in the experiment described in Chapter 4. Each will be detailed next.

Topical Familiarity Questionnaire

The abilities or skills of the subject with regards to the topics that were discussed in the stimuli were measured using 7-point Likert scales. Participants answered "*please indicate how familiar you are with each of the followings topics*" for the following keywords:

1. nuclear weapons
2. nuclear Armageddon
3. climate change
4. history of sex culture
5. sexual freedom
6. christian society
7. bird flu
8. global terrorism
9. viruses
10. Stephen Hawking
11. physics
12. black holes
13. science
14. European monetary union
15. eurozone crisis
16. economics
17. neuroscience
18. psychology
19. ACTA
20. intellectual property
21. piracy
22. theater
23. Iran
24. Iraq
25. Greenpeace
26. environmentalism
27. rainbow warrior
28. capitalism
29. public services policy
30. journalism
31. stem cells
32. embryos
33. ethics
34. global development
35. environmental sustainability
36. united nations
37. Herman van Rompuy
38. positive psychology
39. Higgs boson

40. CERN
41. particle physics
42. relationships
43. cloning
44. medical research
45. bipolar disorder
46. depression.

Epistemic Curiosity Scale

The epistemic curiosity scale (Litman and Spielberger, 2003) consists of two subscales: one for the diversive and one for the specific component of trait curiosity. Both scales contained five items and were measured using 7-point Likert scales. Participants were asked to "*Please indicate how much you agree with the following statements*".

The following items formed the diversive subscale:

1. I enjoy learning about subjects which are unfamiliar;
2. I find it fascinating to learn new information;
3. I enjoy exploring new ideas;
4. I like to learn something new/like to find out more; and,
5. I enjoy discussing abstract concepts.

Furthermore, the following items formed the specific subscale:

1. When I see a complicated piece of machinery, I ask someone how it works;
2. I like a new kind of arithmetic problem/enjoy imagining solutions;
3. When I see an incomplete puzzle, I try and imagine the final solution;
4. I am interested in discovering how things work; and,
5. When I see a riddle, I am interested in trying to solve it.

Interest-Appraisal Scales

Three separate scales were used for: appraised complexity, appraised comprehensibility, and interest. The scales were asked after the following introduction: "*please answer the following questions about how you experienced the previous text*" and "*I thought the content was*". The scales were based on 7-point semantic-differential scales.

The following differentials were used to measure appraised complexity:

- Complex – simple
- Easy to read – difficult to read

For appraised comprehensibility:

- Comprehensible – incomprehensible
- Coherent – incoherent
- Easy to understand – hard to understand

To measure interest:

- Interesting – uninteresting
- Boring – exciting

Furthermore, one more item was added to the measurement of interest based on a 7-point Likert-scale:

- I would be interested in reading more of this text

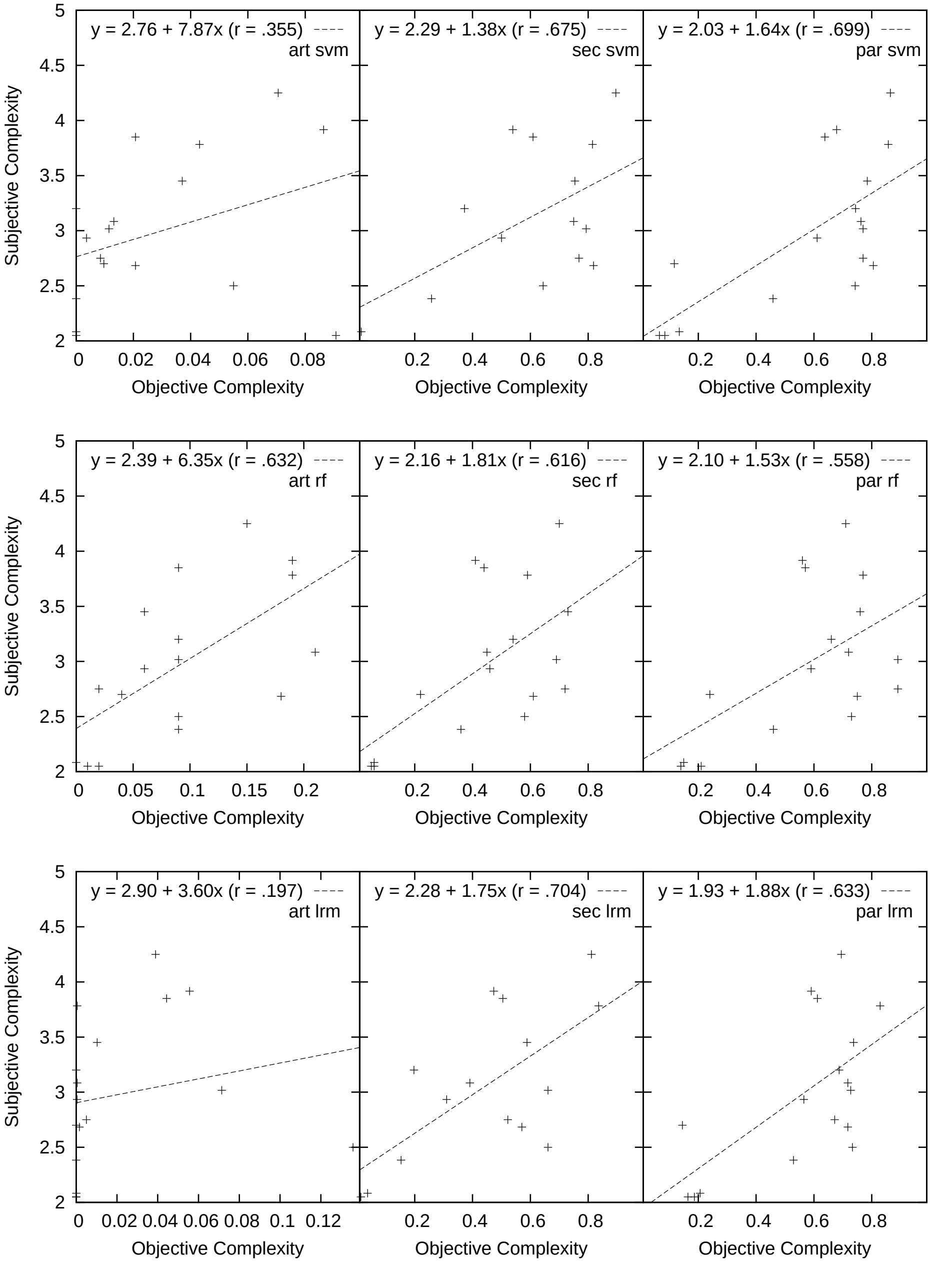
In addition to the preceding three scales, another semantic differential scale was asked to verify the familiarity of the participant with the article:

- Not familiar to me – very familiar to me

C

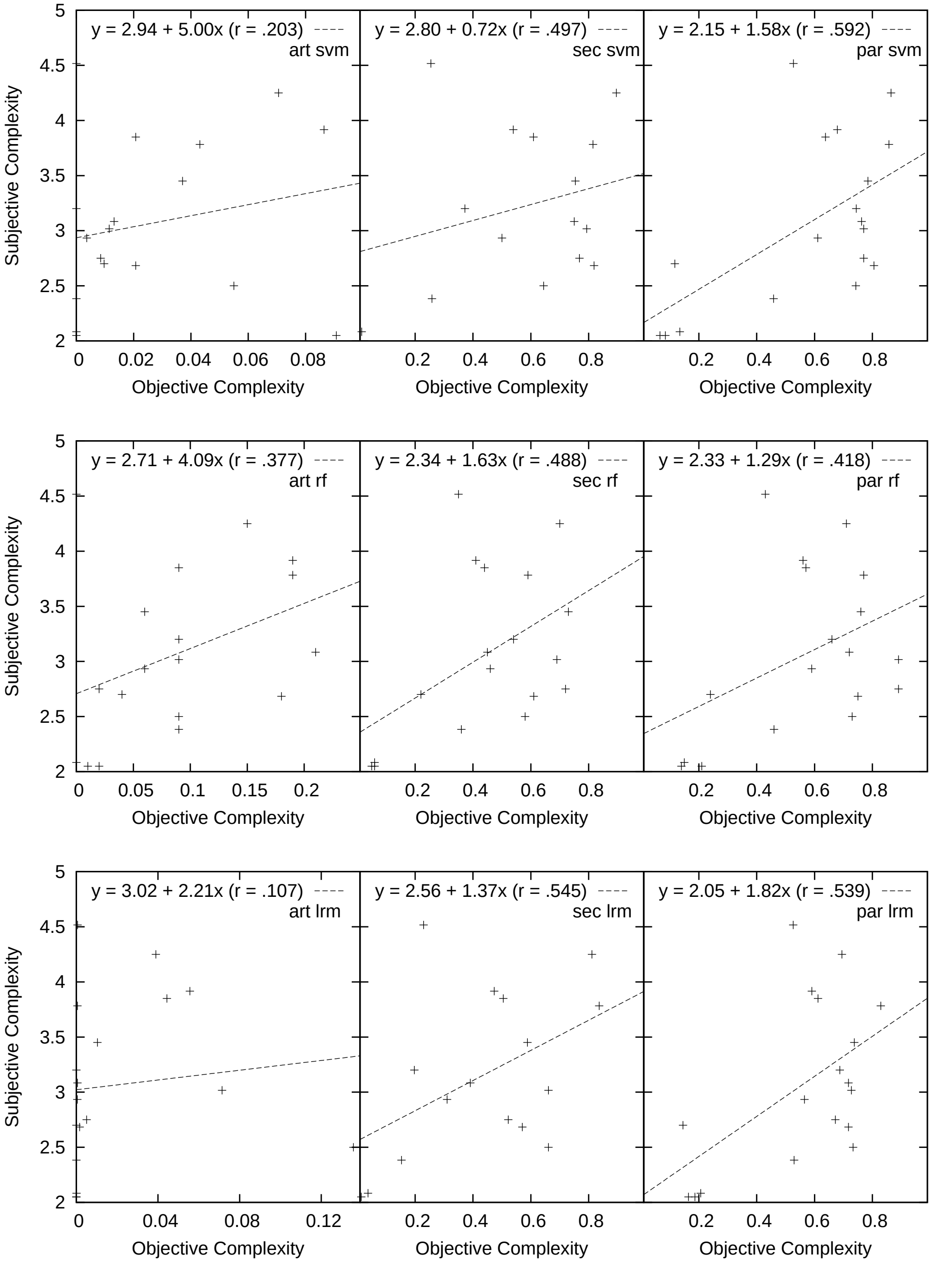
Predictions of Textual Complexity

Using two figures, a detailed description will be given of the relation between objective textual complexity and subjectively appraised complexity for the articles from the Guardian data set (see Chapter 4). In total 18 scatter plots show the congruence between objective complexity and mean subjective complexity. Figure C.1 shows the relations for the articles from the Guardian data set excluding one outlier (see Section 4.3.6), whereas Figure C.2 shows the relations for all the articles including the outlier.



Note. Scatter plots for articles (art), sections (sec), and paragraphs (par) using a support vector machine (svm), random forest (rf), or logistic regression model (lrm).

Figure C.1: Scatter plots and correlation r of the objective prediction of mean subjective complexity.



Note. Scatter plots for articles (art), sections (sec), and paragraphs (par) using a support vector machine (svm), random forest (rf), or logistic regression model (lrm).

Figure C.2: Scatter plots and correlation r of the objective prediction of mean subjective complexity, including outliers.

D

Publications

Following is a selection of publications and patent applications on topics directly or indirectly related to this thesis. Lists of published abstracts, presentations, technical reports, and deliverables are omitted for reasons of brevity.

Journal & Magazine papers

22. **Van der Sluis, F.**, Van den Broek, E.L., Glassey, R.J., Van Dijk, E.M.A.G., and De Jong, F.M.G. When complexity becomes interesting. *Journal of the American Society for Information Science and Technology. [in press]*

21. Van Dijk, E.M.A.G., **Van der Sluis, F.**, Perloy, L.M., and Nijholt, A. A user experience model for tangible interfaces for children. *International Journal of Arts and Technology. [in press]*

20. Van den Broek, E.L., **Van der Sluis, F.**, and Dijkstra, T. (2013). Cross validation of bi-modal health-related stress assessment. *Personal and Ubiquitous Computing, 17(2)*: 215–227. *[Open Access]*

19. Bergervoet, E.J., **Van der Sluis**, F., Van Dijk, E.M.A.G., and Nijholt, A. (2013). Bombs, fish, and coral reefs: the role of in-game explanations and explorative game behavior on comprehension. *The Visual Computer, 29(2)*: 99 – 110.

18. Van den Broek, E.L., **Van der Sluis, F.**, and Schouten, Th.E. (2010). User-centered digital preservation of multimedia. *European Research Consortium for Informatics and Mathematics (ERCIM) News, No. 80 (January)*, 45–47. *[Special issue: Digital Preservation]*

Book chapters

17. Van Dijk, E.M.A.G., **Van der Sluis, F.**, and Nijholt, A. (2012). Designing a museum multi-touch table for children. In Camurri, A. and Costa, C. (Eds.), *Intelligent Technologies for Interactive Entertainment*, Vol. 78 of *Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering*, p. 139 – 148. Springer Berlin / Heidelberg.

16. **Van der Sluis, F.**, Duarte Torres, S., Hiemstra, D., Van Dijk, E.M.A.G., and Kruisinga, F. (2011). Visual exploration of health information for children. In Clough, P., Foley, C., Gurrin, C., Jones, G., Kraaij, W., Lee, H., and

Mudoch, V. (Eds.), *Advances in Information Retrieval*, Vol. 6611 of *Lecture Notes in Computer Science*, p. 788 – 792. Springer Berlin / Heidelberg.

15. Van den Broek, E.L., **Van der Sluis, F.**, and Dijkstra, T. (2011). Telling the story and re-living the past: How speech analysis can reveal emotions in post-traumatic stress disorder (PTSD) patients. In J.H.D.M. Westerink, M. Krans, and M. Ouwerkerk (Eds.), *Sensing Emotions: The impact of context on experience measurements, Chapter 10*, p. 153–180. Series: Philips Research Book Series, Vol. 12. Dordrecht, The Netherlands: Springer Science+Business Media B.V. *[invited]*

14. **Van der Sluis, F.** and Van den Broek, E.L. (2010). Modeling user knowledge from queries: Introducing a metric for knowledge. In An, A., Lingras, P., Petty, S., and Huang, R. (Eds.), *Active Media Technology*, Vol. 6335 of *Lecture Notes in Computer Science*, p. 395–402. Springer Berlin / Heidelberg.

Full papers in proceedings (peer reviewed)

13. **Van der Sluis, F.**, Glassey, R.J., and Van den Broek, E.L. (2012). Making the news interesting: Understanding the relationship between familiarity and interest. In J. Kamps, W. Kraaij, and N. Fuhr (Eds.), *IIiX 2012: ACM Proceedings of the 4th symposium on Information Interaction in Context*, p. 314–317. August 21–24, Nijmegen, The Netherlands.

12. **Van der Sluis, F.**, Van Dijk, E.M.A.G., and Perloy, L.M. (2012). Measuring fun and enjoyment of children in a museum: Evaluating the smileyometer. In Spink, A. J., Grieco, F., Krips, O. E., Loijens, L. W. S., Noldus, L. P. J. J., and Zimmerman, P. H. (Eds.), *Proceedings of Measuring Behavior 2012: 8th International Conference on Methods and Techniques in Behavioral Research*, p. 86–89, August 28-31, Utrecht, The Netherlands.

11. **Van der Sluis, F.**, Dijkstra, T., and Van den Broek, E.L. (2012a). Computer aided diagnosis for mental health care: On the clinical validation of sensitive machines. In Conchon, E., Correia, C., Fred, A., and Gamboa, H. (Eds.), *HealthInf 2012: Proceedings of the 5th International Conference on Health Informatics*, p. 493–498, February 1–4, Vilamoura, Portugal.

10. **Van der Sluis, F.**, Van den Broek, E.L., and Dijkstra, T. (2011). Towards an artificial therapy assistant: Measuring excessive stress from speech. In

Traver, V., Fred, A., Filipe, J., and Gamboa, H. (Eds.), *HealthInf 2011: Proceedings of the 4th International Conference on Health Informatics*, p. 357 – 363, January 26–29, Rome, Italy.

9. Bergervoet, E., **Van der Sluis, F.**, Van Dijk, E., and Nijholt, A. (2011). Let the game do the talking: The influence of explicitness and game behavior on comprehension in an educational computer game. In Gavrilova, M.L. (Eds.), *Cyberworlds 2011: International Conference on*, p. 120 – 127, October 4–6, Banff, Canada.

8. **Van der Sluis, F.** and Van den Broek, E.L. (2010). Using complexity measures in information retrieval. In Belkin, N.J., Kelly, D. (Eds.) *IIiX 2010: ACM Proceedings of the 3th symposium on Information Interaction in Context*, p. 383–388, Augustus 18–22, New Brunswick, USA.

7. **Van der Sluis, F.**, Van den Broek, E.L., and Van Dijk, E.M.A.G. (2010). Information Retrieval eXperience (IRX): Towards a human-centered personalized model of relevance. In *WI-IAT 2010: IEEE Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, Vol. 3, p. 322 – 325, August 31 – September 3, Toronto, Canada.

6. **Van der Sluis, F.** and Van Dijk, E.M.A.G. (2010). A closer look at children's information retrieval usage: Towards child-centered relevance. In Serdyukov, P., Hiemstra, D., and Ruthven, I (Eds.), *Proceedings of the Workshop on Accessible Search Systems held at the 33st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, p. 3 – 10, July 23, Geneva, Switzerland.

5. **Van der Sluis, F.**, Van Dijk, E.M.A.G., and Van den Broek, E.L. (2010). Aiming for user experience in information retrieval: Towards User-Centered Relevance (UCR). In Chen, H.-H., Efthimiadis, E. N., Savoy, J., Crestani, F., and Marchand-Maillet, S. (Eds.), *SIGIR 2010: ACM Proceedings of the 33rd International Conference on Research and Development in Information Retrieval*, p. 924 – 924, July 19–23, Geneva, Switzerland.

4. Lingnau, A., Ruthven, I., Landoni, M., and **Van der Sluis, F.** (2010). Interactive search interfaces for young children – the PuppyIR approach. In Jemni, M., Sampson, D., Kinshuk, and Spector, J.M. (Eds.), *ICALT 2010: IEEE Proceedings of the 10th International Conference on Advanced Learning Technologies*, pages 389 – 390, July 5–7, Sousse, Tunisia.

3. **Van der Sluis, F.**, Van den Broek, E.L., and Dijkstra, T. (2010). Towards semi-automated assistance for the treatment of stress disorders. In Fred, A. and Filipe, J. and Gamboa, H. (Eds.), *HealthInf 2010: Proceedings of the 3th International Conference on Health Informatics*, p. 446 – 449, January 20–23, Valencia, Spain.

2. Van den Broek, E.L., **Van der Sluis, F.**, and Dijkstra, T. (2009). Therapy Progress Indicator (TPI): Combining speech parameters and the subjective unit of distress. In J. Cohn, A. Nijholt, M. Pantic (Eds.), *ACII 2009: IEEE Proceedings of the 3rd International Conference on Affective Computing and Intelligent Interaction*, Vol. 1, pages 381 – 386, September 10-12, Amsterdam, The Netherlands.

Patents

1. **Van der Sluis, F.** and Van den Broek, E.L. (2013). Method and computer server system for retrieving and presenting information to a user in a computer network. European Patent Application No. 13175793.2, filed on July 9, 2013.

Bibliography

Abdi, H. (2009). Centroids. *Wiley Interdisciplinary Reviews: Computational Statistics*, 1(2):259–260.

Abdi, H. and Williams, L. J. (2010). Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(4):433–459.

Agichtein, E., Brill, E., and Dumais, S. (2006). Improving web search ranking by incorporating user behavior information. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '06, pages 19–26, New York, NY, USA. ACM.

Al-Maskari, A. and Sanderson, M. (2010). A review of factors influencing user satisfaction in information retrieval. *Journal of the American Society for Information Science and Technology*, 61(5):859–868.

Alexander, P. A., Kulikowich, J. M., and Schulze, S. K. (1994). The influence of topic knowledge, domain knowledge, and interest on the comprehension of scientific exposition. *Learning and Individual Differences*, 6(4):379 – 397.

Alhakami, A. S. and Slovic, P. (1994). A psychological study of the inverse relationship between perceived risk and perceived benefit. *Risk Analysis*, 14(6):1085–1096.

Allport, D. A. (1980). Patterns and actions: Cognitive mechanisms are content-specific. *Cognitive psychology: New directions*, pages 26–64.

Alter, A. L. and Oppenheimer, D. M. (2009). Uniting the tribes of fluency to form a metacognitive nation. *Personality and Social Psychology Review*, 13(3):219–235.

Arapakis, I. (2010). *Affect-Based information retrieval.* PhD thesis, University of Glasgow.

Arapakis, I., Jose, J. M., and Gray, P. D. (2008). Affective feedback: An investigation into the role of emotions in the information seeking process. In *SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 395–402, New York, USA. ACM.

Arapakis, I., Konstas, I., and Jose, J. M. (2009). Using facial expressions and peripheral physiological signals as implicit indicators of topical relevance. In *Proceedings of the 17th ACM international conference on Multimedia*, MM '09, pages 461–470, New York, NY, USA. ACM.

Azzopardi, L., Järvelin, K., Kamps, J., and Smucker, M. D. (2011). Report on the sigir 2010 workshop on the simulation of interaction. *SIGIR Forum*, 44(2):35–47.

Babyak, M. A. (2004). What you see may not be what you get: A brief, nontechnical introduction to overfitting in regression-type models. *Psychosomatic Medicine*, 66(3):411–421.

Balota, D. A., Cortese, M. J., Sergent-Marshall, S. D., Spieler, D. H., and Yap, M. J. (2004). Visual word recognition of single-syllable words. *Journal of Experimental Psychology: General*, 133(2):283.

Balota, D. A., Yap, M. J., and Cortese, M. J. (2006). *Handbook of Psycholinguistics*, chapter Visual Word Recognition: The Journey from Features to Meaning (A Travel Update), pages 285–376. Academic Press, Inc.

Banse, R. and Scherer, K. R. (1996). Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology*, 70(3):614 – 636.

Barabási, A.-L. and Albert, R. (1999). Emergence of scaling in random networks. *Science*, 286(5439):509–512.

Barry, C. L. (1994). User-defined relevance criteria: An exploratory study. *Journal of the American Society for Information Science*, 45(3):149–159.

Barry, C. L. and Schamber, L. (1998). Users' criteria for relevance evaluation: A cross-situational comparison. *Information Processing & Management*, 34(2-3):219 – 236.

Beatty, J. (1982). Task-evoked pupillary responses, processing load, and the structure of processing resources. *Psychological bulletin*, 91(2):276.

Belkin, N. J. (2008). Some(what) grand challenges for information retrieval. *SIGIR Forum*, 42(1):47–54.

Benjamin, R. (2012). Reconstructing readability: Recent developments and recommendations in the analysis of text difficulty. *Educational Psychology Review*, 24(1):63–88.

Beresi, U., Kim, Y., Song, D., and Ruthven, I. (2010). Why did you pick that? visualising relevance criteria in exploratory search. *International Journal on Digital Libraries*, 11(2):59–74.

Berlyne, D. (1960). *Conflict, arousal and curiosity.* McGraw-Hill, New York.

Berlyne, D. (1970). Novelty, complexity, and hedonic value. *Perception & Psychophysics*, 8(5):279–286.

Berlyne, D. E. (1954). A theory of human curiosity. *British Journal of Psychology. General Section*, 45(3):180–191.

Berlyne, D. E. (1966). Curiosity and exploration. *Science*, 153(3731):25–33.

Berlyne, D. E. (1971). *Aesthetics and psychobiology.* Appleton-Century-Crofts, East Norwalk, CT, USA.

Berlyne, D. E. (1978). Curiosity and learning. *Motivation and Emotion*, 2:97–175.

Blei, D. M. and Lafferty, J. D. (2009). Text mining: Classification, clustering, and applications. In Srivastava, A. N. and Sahami, M., editors, *Text Mining: Classification, Clustering, and Applications*, chapter Topic Models, pages 71–94. CRC Press.

Bless, H., Clore, G. L., Schwarz, N., Golisano, V., Rabe, C., and Wölk, M. (1996). Mood and the use of scripts: Does a happy mood really lead to mindlessness? *Journal of personality and social psychology*, 71(4):665.

Blythe, M., Overbeeke, K., Monk, A., and Wright, P. (2004). *Funology: from usability to enjoyment*, volume 3. Springer, Berlin / Heidelberg.

Borlund, P. (2003). The concept of relevance in IR. *Journal of the American Society for Information Science and Technology*, 54(10):913–925.

Borlund, P. and Ingwersen, P. (1998). Measures of relative relevance and ranked half-life: performance indicators for interactive IR. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '98, pages 324–331, New York, NY, USA. ACM.

Bowler, L. (2010). The self-regulation of curiosity and interest during the information search process of adolescent students. *Journal of the American Society for Information Science and Technology*, 61(7):1332–1344.

Breiman, L. (2001a). Random forests. *Machine Learning*, 45(1):5–32.

Breiman, L. (2001b). Statistical modeling: The two cultures. *Statistical Science*, 16(3):199–215.

Broder, A. (2002). A taxonomy of web search. *SIGIR Forum*, 36(2):3–10.

Brusilovsky, P. and Millán, E. (2007). User models for adaptive hypermedia and adaptive educational systems. In Brusilovsky, P., Kobsa, A., and Nejdl, W., editors, *The Adaptive Web: Methods and Strategies of Web Personalization*, volume 4321 of *Lecture Notes in Computer Science*, pages 3–53. Springer, Berlin / Heidelberg.

Brysbaert, M., Drieghe, D., and Vitu, F. (2005). *Word skipping: Implications for theories of eye movement control in reading.*, chapter 6, pages 1–29. Oxford University Press.

Byström, K. and Hansen, P. (2005). Conceptual framework for tasks in information studies. *Journal of the American Society for Information Science and Technology*, 56(10):1050–1061.

Cacioppo, J. T. and Berntson, G. G. (1994). Relationship between attitudes and evaluative space: A critical review, with emphasis on the separability of positive and negative substrates. *Psychological Bulletin*, 115(3):401 – 423.

Cannon, W. B. (1927). The James-Lange theory of emotions: A critical examination and an alternative theory. *The American Journal of Psychology*, 39(1/4):106–124.

Chall, J. S. and Dale, E. (1995). *T1 - Readability Revisited: The New Dale-Chall Readability Formula.* PB - Brookline Books, Cambridge, Mass.

Chang, C.-C. and Lin, C.-J. (2011). LIBSVM: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.*, 2(3):27:1–27:27.

Chen, H. (2006). Flow on the net–detecting web users' positive affects and their flow states. *Computers in Human Behavior*, 22(2):221 – 233.

Chen, J. (2007). Flow in games (and everything else). *Commun. ACM*, 50(4):31–34.

Chomsky, N. (1956). Three models for the description of language. *Information Theory, IRE Transactions on*, 2(3):113 –124.

Cleverdon, C. W., Mills, J., and Keen, M. (1966). Factors determining the performance of indexing systems. Technical report, ASLIB Cranfield project, Cranfield.

Cohen, J. (1992). A Power Primer. *Psychological Bulletin*, 112(1):155–159.

Collins-Thompson, K., Bennett, P. N., White, R. W., de la Chica, S., and Sontag, D. (2011). Personalizing web search results by reading level. In Berendt, B., de Vries, A., Fan, W., Macdonald, C., Ounis, I., and Ruthven, I., editors, *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, CIKM '11, pages 403–412, New York, NY, USA. ACM.

Collins-Thompson, K. and Callan, J. (2005). Predicting reading difficulty with statistical language models. *Journal of the American Society for Information Science and Technology*, 56(13):1448–1462.

Coltheart, M., Rastle, K., Perry, C., Langdon, R., and Ziegler, J. (2001). DRC: a dual route cascaded model of visual word recognition and reading aloud. *Psychological review*, 108(1):204–256.

Connelly, D. A. (2011). Applying Silvia's model of interest to academic text: Is there a third appraisal? *Learning and Individual Differences*, 21(5):624 – 628.

Cosijn, E. and Ingwersen, P. (2000). Dimensions of relevance. *Information Processing & Management*, 36(4):533 – 550.

Cover, T. M. and Thomas, J. A. (2006). *Elements of Information Theory*, chapter Entropy, Relative Entropy, and Mutual Information, pages 13–56. John Wiley & Sons, Inc.

Crossley, S., Greenfield, J., and McNamara, D. (2008). Assessing text readability using cognitively based indices. *TESOL Quarterly*, 42(3):475–493.

Csikszentmihalyi, M. (1991). *The Psychology of Optimal Experience.* Harper Collins, New York.

Csikszentmihalyi, M. and LeFevre, J. (1989). Optimal experience in work and leisure. *Journal of personality and social psychology*, 56(5):815.

Davison, A. and Kantor, R. N. (1982). On the failure of readability formulas to define readable texts: A case study from adaptations. *Reading Research Quarterly*, 17(2):187–209.

Day, H. (1967). Evaluations of subjective complexity, pleasingness and interestingness for a series of random polygons varying in complexity. *Perception & Psychophysics*, 2(7):281–286.

De Sousa, R. (2008). Epistemic feelings. In Brun, G. and Kuenzle, D., editors, *Epistemology and emotions*, pages 185–204. Ashgate Publishing Company.

Degand, L., Lefèvre, N., and Bestgen, Y. (1999). The impact of connectives and anaphoric expressions on expository discourse comprehension. *Document Design*, 1(1):39–51.

DeLone, W. and McLean, E. (2002). Information systems success revisited. In *System Sciences, 2002. HICSS. Proceedings of the 35th Annual Hawaii International Conference on*, pages 2966 – 2976.

Demartini, G. and Mizzaro, S. (2006). A classification of IR effectiveness metrics. In Lalmas, M., MacFarlane, A., Rüger, S., Tombros, A., Tsikrika, T., and Yavlinsky, A., editors, *Advances in Information Retrieval*, volume 3936 of *Lecture Notes in Computer Science*, pages 488–491. Springer Berlin / Heidelberg.

Denning, P. J. (2006). Infoglut. *Communications of the ACM*, 49(7):15–19.

Descartes, R. (1989/1649). *The Passions of the Soul.* Hackett Publishing Company, Indianapolis, Indiana.

Diener, E. and Emmons, R. A. (1984). The independence of positive and negative affect. *Journal of Personality and Social Psychology*, 47(5):1105 – 1117.

D'Mello, S., Graesser, A., and Picard, R. (2007). Toward an affect-sensitive autotutor. *Intelligent Systems, IEEE*, 22(4):53 –61.

Dubay, W. H. (2004). The principles of readability. Technical report, Impact Information, Costa Mesa, CA, USA. http://www.impact-information.com/impactinfo/readability02.pdf. [Last accessed on August 10, 2012].

D'Mello, S. and Graesser, A. (2012). Dynamics of affective states during complex learning. *Learning and Instruction*, 22(2):145 – 157.

Eerola, T. (1997). The rise and fall of the experimental style of the beatles. the life span of stylistic periods in music. In Gabrielsson, A., editor, *Third Triennial European Society for the Cognitive Science of Music (ESCOM) Conference: Proceedings*, pages 377–381. Uppsala University, Uppsala, Sweden.

Egloff, B. (1998). The independence of positive and negative affect depends on the affect measure. *Personality and Individual Differences*, 25(6):1101 – 1109.

Eickhoff, C., Azzopardi, L., Hiemstra, D., De Jong, F., de Vries, A., Dowie, D., Duarte, S., Glassey, R., Gyllstrom, K., Kruisinga, F., Marshall, K., Moens, S., Polajnar, T., and Van der Sluis, F. (2012). EmSe: initial evaluation of a child-friendly medical search system. In *Proceedings of the 4th Information Interaction in Context Symposium*, IIIX '12, pages 282–285, New York, NY, USA. ACM.

Ellsworth, P. C. and Scherer, K. R. (2003). Appraisal processes in emotion. In Davidson, R. J., Scherer, K. R., and Goldsmith, H. H., editors, *Handbook of Affective Sciences*, chapter 29, pages 572–595. Oxford University Press.

Evans, J. S. B. (2008). Dual-processing accounts of reasoning, judgment, and social cognition. *Annu. Rev. Psychol.*, 59:255–278.

Feng, L., Jansche, M., Huenerfauth, M., and Elhadad, N. (2010). A comparison of features for automatic readability assessment. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 276–284, Stroudsburg, PA, USA. Association for Computational Linguistics.

Finnäs, L. (1989). How can musical preferences be modified? A research review. *Bulletin of the Council for Research in Music Education*, (102):1–58.

Finneran, C. and Zhang, P. (2005). Flow in computer-mediated environments: Promises and challenges. *Communications of the Association for Information Systems*, 15:82–101.

Fiore, A. M., Jin, H.-J., and Kim, J. (2005). For fun and profit: Hedonic value from image interactivity and responses toward an online store. *Psychology and Marketing*, 22(8):669–694.

Foster, A. and Ford, N. (2003). Serendipity and information seeking: an empirical study. *Journal of Documentation*, 59(3):321–340.

Fraser, B. (1999). What are discourse markers? *Journal of Pragmatics*, 31(7):931 – 952.

Fredrickson, B. L. and Kahneman, D. (1993). Duration neglect in retrospective evaluations of affective episodes. *Journal of personality and social psychology*, 65(1):45.

Frijda, N. H. (1967). Problems of computer simulation. *Behavioral Science*, 12(1):59–67.

Fulton, C. (2009). The pleasure principle: the power of positive affect in information seeking. In *Aslib Proceedings*, volume 61, pages 245–261. Emerald Group Publishing Limited.

Furner, J. (2002). On recommending. *Journal of the American Society for Information Science and Technology*, 53(9):747–763.

Gabrilovich, E. and Markovitch, S. (2009). Wikipedia-based semantic interpretation for natural language processing. *J. Artif. Int. Res.*, 34:443–498.

Gibson, E. (1998). Linguistic complexity: locality of syntactic dependencies. *Cognition*, 68(1):1 – 76.

Gibson, E. (2000). The dependency locality theory: A distance-based theory of linguistic complexity. In *Image, language, brain: Papers from the first mind articulation project symposium*, pages 95–126.

Giles, J. (2005). Internet encyclopaedias go head to head. *Nature*, 438(15):900–901.

Gluck, M. (1996). Exploring the relationship between user satisfaction and relevance in information systems. *Information Processing & Management*, 32(1):89 – 104.

Golder, S. A. and Huberman, B. A. (2006). Usage patterns of collaborative tagging systems. *Journal of Information Science*, 32(2):198–208.

Goodman, J. T. (2001). A bit of progress in language modeling. *Computer Speech & Language*, 15(4):403 – 434.

Graesser, A., McNamara, D., Louwerse, M., and Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods*, 36(2):193–202.

Granka, L., Feusner, M., and Lorigo, L. (2008). Eye monitoring in online search. In Hammoud, R. I., editor, *Passive Eye Monitoring*, Signals and Communication Technology, chapter 16, pages 346–371. Springer-Verlag.

Gwizdka, J. and Lopatovska, I. (2009). The role of subjective factors in the information search process. *Journal of the American Society for Information Science and Technology*, 60(12):2452–2464.

Hale, J. (2003). The information conveyed by words in sentences. *Journal of Psycholinguistic Research*, 32:101–123.

Hanani, U., Shapira, B., and Shoval, P. (2001). Information filtering: Overview of issues, research and systems. *User Modeling and User-Adapted Interaction*, 11(3):203–259.

Hargreaves, D. and North, A. (2010). Experimental aesthetics and liking for music. In Juslin, P. N. and Sloboda, J. A., editors, *Handbook of music and emotions: theory, research, applications*, pages 515–547. Oxford University Press, Oxford, UK.

Hargreaves, D. J. (1984). The effects of repetition on liking for music. *Journal of Research in Music Education*, 32(1):35–47.

Harter, S. P. (1992). Psychological relevance and information science. *Journal of the American Society for Information Science*, 43(9):602–615.

Hassenzahl, M. (2003). *The thing and I: Understanding the relationship between user and product*, chapter 3, pages 31–42. Kluwer Academic Publishers.

Hassenzahl, M. (2013). User experience and experience design. In Soegaard, M. and Dam, R. F., editors, *Encyclopedia of Human-Computer Interaction.* The Interaction-Design.org Foundation, Aarhus, Denmark, second edition. Available online at `http://www.interaction-design.org/encyclopedia/user_experience_and_experience_design.html` [Last accessed on July 23, 2013].

Hassenzahl, M., Diefenbach, S., and Göritz, A. (2010). Needs, affect, and interactive products – facets of user experience. *Interacting with Computers*, 22(5):353 – 362.

Hassenzahl, M. and Tractinsky, N. (2006). User experience - a research agenda. *The American Journal of Psychology*, 25(2):91–97.

Hatcher, E., Gospodnetic, O., and McCandless, M. (2010). *Lucene in Action.* Manning, second revised edition.

Hebb, D. O. (1955). Drives and the CNS (conceptual nervous system). *Psychological review*, 62(4):243.

Heinstrom, J. (2002). *Fast Surfers, Broad Scanners and Deep Divers - Personality and Information Seeking Behaviour.* PhD thesis, Abo (Turku): Abo Akademi University Press. Retrieved 10 March, 2009 from http://www.abo.fi/ jheinstr/thesis.htm.

Herlocker, J. L., Konstan, J. A., Terveen, L. G., and Riedl, J. T. (2004). Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems*, 22(1):5–53.

Hess, E. H. and Polt, J. M. (1960). Pupil size as related to interest value of visual stimuli. *Science*, 132(3423):349–350.

Hidi, S. (1990). Interest and its contribution as a mental resource for learning. *Review of Educational Research*, 60(4):549–571.

Hidi, S. and Renninger, K. A. (2006). The four-phase model of interest development. *Educational Psychologist*, 41(2):111–127.

Higgins, E. T. (2006). Value from hedonic experience and engagement. *Psychological review*, 113(3):439.

Hill, W., Stead, L., Rosenstein, M., and Furnas, G. (1995). Recommending and evaluating choices in a virtual community of use. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, CHI '95, pages 194–201, New York, NY, USA. ACM Press/Addison-Wesley Publishing Co.

Hoffman, D. L. and Novak, T. P. (2009). Flow online: Lessons learned and future prospects. *Journal of Interactive Marketing*, 23(1):23 – 34.

Huffman, S. B. and Hochster, M. (2007). How well does result relevance predict session satisfaction? In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '07, pages 567–574, New York, NY, USA. ACM.

Hull, C. L. (1943). *Principles of behavior: An introduction to behavior theory.* Appleton-Century.

Humphreys, G., Evett, L., and Taylor, D. (1982). Automatic phonological priming in visual word recognition. *Memory & Cognition*, 10:576–590.

Husserl, E. (1988,1931). *Cartesian Meditations.* Kluwer, Dordrecht, NL. Translated by D. Cairns.

Hutchison, K. (2003). Is semantic priming due to association strength or feature overlap? a microanalytic review. *Psychonomic Bulletin & Review*, 10:785–813.

Iacobucci, D. (2010). Structural equations modeling: Fit indices, sample size, and advanced topics. *Journal of Consumer Psychology*, 20(1):90–98.

Ihaka, R. and Gentleman, R. (1996). R: A Language for Data Analysis and Graphics. *Journal of Computational and Graphical Statistics*, 5(3):299–314.

Ingwersen, P. (1996). Cognitive perspectives of information retrieval interaction: Elements of a cognitive IR theory. *Journal of documentation*, 52(1):3–50.

Inhoff, A. and Rayner, K. (1986). Parafoveal word processing during eye fixations in reading: Effects of word frequency. *Attention, Perception, & Psychophysics*, 40(6):431–439.

Isen, A. M., Shalker, T. E., Clark, M., and Karp, L. (1978). Affect, accessibility of material in memory, and behavior: A cognitive loop? *Journal of personality and social psychology*, 36(1):1.

Jacobs, A. (2009). The pathologies of big data. *Commun. ACM*, 52(8):36–44.

Jaeger, T. F. and Tily, H. (2011). On language utility: processing complexity and communicative efficiency. *Wiley Interdisciplinary Reviews: Cognitive Science*, 2(3):323–335.

Jansen, B. J. and Rieh, S. Y. (2010). The seventeen theoretical constructs of information searching and information retrieval. *Journal of the American Society for Information Science and Technology*, 61(8):1517–1534.

Järvelin, K. and Kekäläinen, J. (2002). Cumulated gain-based evaluation of IR techniques. *ACM Trans. Inf. Syst.*, 20(4):422–446.

Joachims, T. (2002). Optimizing search engines using clickthrough data. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '02, pages 133–142, New York, NY, USA. ACM.

Jonassen, D. H. (2000). Toward a design theory of problem solving. *Educational Technology Research and Development*, 48(4):63–85.

Jordan, P. W. (1998). Human factors for pleasure in product use. *Applied Ergonomics*, 29(1):25 – 33.

Juhasz, B. J. (2005). Age-of-acquisition effects in word and picture identification. *Psychological Bulletin*, 131(5):684–712.

Just, M. and Carpenter, P. (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review*, 87:329–354.

Kahneman, D. (2003). *Well-being: The foundations of hedonic psychology*, chapter Objective Happiness, pages 3–25. Russell Sage Foundation Publications.

Kahneman, D., Fredrickson, B. L., Schreiber, C. A., and Redelmeier, D. A. (1993). When more pain is preferred to less: Adding a better end. *Psychological Science*, 4(6):401–405.

Kaplan, R. M. (1972). Augmented transition networks as psychological models of sentence comprehension. *Artificial Intelligence*, 3(0):77 – 100.

Karvonen, K. (2000). The beauty of simplicity. In *Proceedings on the 2000 conference on Universal Usability*, CUU '00, pages 85–90, New York, NY, USA. ACM.

Kate, R. J., Luo, X., Patwardhan, S., Franz, M., Florian, R., Mooney, R. J., Roukos, S., and Welty, C. (2010). Learning to predict readability using diverse linguistic features. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, pages 546–554, Stroudsburg, PA, USA. Association for Computational Linguistics.

Kincaid, J. P., Fishburne, Robert P., J., Rogers, R. L., and Chissom, B. S. (1975). Derivation of new readability formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for navy enlisted personnel. Technical report, National Technical Information Service, Springfield, Virginia.

King, M. (1996). Evaluating natural language processing systems. *Commun. ACM*, 39(1):73–79.

Kintsch, W. (1980). Learning from text, levels of comprehension, or: Why anyone would read a story anyway. *Poetics*, 9(1-3):87–98.

Kintsch, W. (1994). Text comprehension, memory, and learning. *American Psychologist*, 49(4):294 – 303.

Kintsch, W. and van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological Review*, 85(5):363 – 394.

Klein, D. and Manning, C. D. (2003). Accurate unlexicalized parsing. In *Proceedings of the 41st annual meeting on Association for computational linguistics*, volume 1 of *ACL '03*, pages 423–430, Stroudsburg, PA, USA. Association for Computational Linguistics.

Konstan, J. and Riedl, J. (2012). Recommender systems: from algorithms to user experience. *User Modeling and User-Adapted Interaction*, 22(1):101–123.

Kort, B., Reilly, R., and Picard, R. (2001). An affective model of interplay between emotions and learning: reengineering educational pedagogy-building a learning companion. In *Advanced Learning Technologies, 2001. Proceedings. IEEE International Conference on*, pages 43 –46.

Kreibig, S. D., Gendolla, G. H., and Scherer, K. R. (2010). Psychophysiological effects of emotional responding to goal attainment. *Biological Psychology*, 84(3):474 – 487.

Kreibig, S. D., Gendolla, G. H., and Scherer, K. R. (2012). Goal relevance and goal conduciveness appraisals lead to differential autonomic reactivity in emotional responding to performance feedback. *Biological Psychology*, 91(3):365 – 375.

Kuhlthau, C. C. (1993). A principle of uncertainty for information seeking. *Journal of Documentation*, 49(4):339–355.

Kuhlthau, C. C. (2004). *Seeking Meaning: A Process Approach to Library and Information Services.* Norwood, NJ: Ablex Pub. Corp.

LaBerge, D. and Samuels, S. J. (1974). Toward a theory of automatic information processing in reading. *Cognitive Psychology*, 6(2):293 – 323.

Lang, P. (1995). The emotion probe. Studies of motivation and attention. *The American Psychologist*, 50(5):372–385.

Lapata, M. and Barzilay, R. (2005). Automatic evaluation of text coherence: models and representations. In *IJCAI'05: Proceedings of the 19th international joint conference on Artificial intelligence*, pages 1085–1090, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Lazarus, R. S. (1991). Progress on a cognitive-motivational-relational theory of emotion. *The American psychologist*, 46(8):819–834.

Ledoux, K., Camblin, C. C., Swaab, T. Y., and Gordon, P. C. (2006). Reading words in discourse: The modulation of lexical priming effects by message-level context. *Behavioral and Cognitive Neuroscience Reviews*, 5(3):107–127.

Lee, H., Peirsman, Y., Chang, A., Chambers, N., Surdeanu, M., and Jurafsky, D. (2011). Stanford's multi-pass sieve coreference resolution system at the conll-2011 shared task. In Goldwater, S. and Manning, C., editors, *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, CONLL Shared Task '11, pages 28–34, Stroudsburg, PA, USA. Association for Computational Linguistics.

Lehmann, J., Lalmas, M., Yom-Tov, E., and Dupret, G. (2012). Models of user engagement. In Masthoff, J., Mobasher, B., Desmarais, M., and Nkambou, R., editors, *User Modeling, Adaptation, and Personalization*, volume 7379 of *Lecture Notes in Computer Science*, pages 164–175. Springer Berlin Heidelberg.

Lewis, R. L., Vasishth, S., and Dyke, J. A. V. (2006). Computational principles of working memory in sentence comprehension. *Trends in Cognitive Sciences*, 10(10):447 – 454.

Liaw, S.-S. and Huang, H.-M. (2006). Information retrieval from the world wide web: a user-focused approach based on individual experience with search engines. *Computers in Human Behavior*, 22(3):501 – 517.

Lindgaard, G. and Dudek, C. (2003). What is this evasive beast we call user satisfaction? *Interacting with Computers*, 15(3):429 – 452.

Lindgaard, G., Fernandes, G., Dudek, C., and Brown, J. (2006). Attention web designers: You have 50 milliseconds to make a good first impression! *Behaviour & Information Technology*, 25(2):115–126.

Lingnau, A., Ruthven, I., Landoni, M., and Van der Sluis, F. (2010). Interactive search interfaces for young children - the PuppyIR approach. In *Proceedings of the 2010 10th IEEE International Conference on Advanced Learning Technologies*, ICALT '10, pages 389–390, Washington, DC, USA. IEEE Computer Society.

Litman, J. (2005). Curiosity and the pleasures of learning: Wanting and liking new information. *Cognition & Emotion*, 19(6):793–814.

Litman, J., Hutchins, T., and Russon, R. (2005). Epistemic curiosity, feeling-of-knowing, and exploratory behaviour. *Cognition & Emotion*, 19(4):559–582.

Litman, J. A. and Silvia, P. J. (2006). The latent structure of trait curiosity: Evidence for interest and deprivation curiosity dimensions. *Journal of Personality Assessment*, 86(3):318–328.

Litman, J. A. and Spielberger, C. D. (2003). Measuring epistemic curiosity and its diversive and specific components. *Journal of Personality Assessment*, 80(1):75–86.

Liu, T.-Y. (2009). Learning to rank for information retrieval. *Found. Trends Inf. Retr.*, 3(3):225–331.

Lively, B. A. and Pressey, S. L. (1923). A method for measuring the "vocabulary burden" of textbooks. *Educational Administration and Supervision*, 9(7):389–398.

Loewenstein, G. F. and Prelec, D. (1993). Preferences for sequences of outcomes. *Psychological Review*, 100(1):91.

Long, D. L., Johns, C. L., and Morris, P. E. (2006a). *Handbook of Psycholinguistics*, chapter Comprehension Ability in Mature Readers, pages 801–833. Academic Press, Inc.

Long, D. L., Wilson, J., Hurley, R., and Prat, C. S. (2006b). Assessing text representations with recognition: The interaction of domain knowledge and text coherence. *Journal of experimental psychology. Learning memory and cognition*, 32(4):816–827.

Maglaughlin, K. L. and Sonnenwald, D. H. (2002). User perspectives on relevance criteria: A comparison among relevant, partially relevant, and not-relevant judgments. *Journal of the American Society for Information Science and Technology*, 53(5):327–342.

Manhart, K. (1996). Artificial intelligence modelling: Data driven and theory driven approaches. In Troitzsch, K., Müller, U., Nigel, G., and Doran, J., editors, *Social Science Micro Simulation*, chapter 19, pages 416–431. Springer Verlag.

Manning, C. D., Raghavan, P., and Schütze, H. (2009). *Introduction to information retrieval*, volume 1. Cambridge University Press Cambridge, draft edition. Retrieved on 31-11-2012 from `http://nlp.stanford.edu/IR-book/pdf/irbookonlinereading.pdf`.

Marcus, M. P., Marcinkiewicz, M. A., and Santorini, B. (1993). Building a large annotated corpus of english: the penn treebank. *Comput. Linguist.*, 19(2):313–330.

Martindale, C., Moore, K., and Borkum, J. (1990). Aesthetic preference: Anomalous findings for Berlyne's psychobiological theory. *The American Journal of Psychology*, 103(1):53–80.

Maslow, A. H. (1943). A theory of human motivation. *Psychological review*, 50(4):370.

McCarthy, J. and Wright, P. (2004). Technology as experience. *interactions*, 11(5):42–43.

McDonald, S. A. and Shillcock, R. C. (2003). Low-level predictive inference in reading: the influence of transitional probabilities on eye movements. *Vision Research*, 43(16):1735 – 1751.

McGinnies, E., Comer, P., and Lacey, O. (1952). Visual-recognition thresholds as a function of word length and word frequency. *Journal of Experimental Psychology*, 44(2):65–69.

McNamara, D. S., Kintsch, E., Songer, N. B., and Kintsch, W. (1996). Are good texts always better? Interactions of text coherence, background knowledge, and levels of understanding in learning from text. *Cognition and Instruction*, 14(1):1–43.

McNamara, D. S. and Kintsch, W. (1996). Learning from texts: Effects of prior knowledge and text coherence. *Discourse Processes*, 22(3):247–288.

McNamara, D. S., Louwerse, M. M., McCarthy, P. M., and Graesser, A. C. (2010). Coh-Metrix: Capturing linguistic features of cohesion. *Discourse Processes*, 47(4):292–330.

Meyer, D., Leisch, F., and Hornik, K. (2003). The support vector machine under test. *Neurocomputing*, 55(1-2):169–186.

Meyer, D. E. and Schvaneveldt, R. W. (1971). Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. *Journal of Experimental Psychology*, 90(2):227–234.

Michailidou, E., Harper, S., and Bechhofer, S. (2008). Visual complexity and aesthetic perception of web pages. In *Proceedings of the 26th annual ACM international conference on Design of communication*, SIGDOC '08, pages 215–224, New York, NY, USA. ACM.

Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Pickett, J. P., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., and et al. (2011). Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014):176–182.

Miller, G. A. (1995). WordNet: A lexical database for english. *Commun. ACM*, 38(11):39–41.

Mizzaro, S. (1998). How many relevances in information retrieval? *Interacting with Computers*, 10(3):303 – 320.

Moon, J.-W. and Kim, Y.-G. (2001). Extending the TAM for a world-wide-web context. *Information & Management*, 38(4):217 – 230.

Mooney, C., Scully, M., Jones, G., and Smeaton, A. (2006). Investigating biometric response for information retrieval applications. In Lalmas, M., MacFarlane, A., Rüger, S., Tombros, A., Tsikrika, T., and Yavlinsky, A., editors, *Advances in Information Retrieval*, volume 3936 of *Lecture Notes in Computer Science*, pages 570–574. Springer Berlin / Heidelberg.

Morris, J. and Hirst, G. (1991). Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics*, 17(1):21–48.

Morton, J. (1969). Interaction of information in word recognition. *Psychological review*, 76(2):165.

Moshfeghi, Y. (2012). *Role of emotion in information retrieval.* PhD thesis, University of Glasgow.

Mummalaneni, V. (2005). An empirical investigation of web site characteristics, consumer emotional states and on-line shopping behaviors. *Journal of Business Research*, 58(4):526 – 532.

Munsinger, H. and Kessen, W. (1964). Uncertainty, structure, and preference. *Psychological Monographs: General and Applied*, 78(9):1–24.

Mussel, P. (2010). Epistemic curiosity and related constructs: Lacking evidence of discriminant validity. *Personality and Individual Differences*, 49(5):506 – 510.

Nagelkerke, N. (1991). A note on a general definition of the coefficient of determination. *Biometrika*, 78(3):691–692.

Nahl, D. (2004). Measuring the affective information environment of web searchers. *Proceedings of the American Society for Information Science and Technology*, 41(1):191–197.

Nahl, D. (2005). Affective load. In Fisher, K., Erdelez, S., and McKechnie, L., editors, *Theories of Information Behavior*, ASIST Monographs, pages 39–43. Information Today.

New, B., Ferrand, L., Pallier, C., and Brysbaert, M. (2006). Reexamining the word length effect in visual word recognition: New evidence from the english lexicon project. *Psychonomic bulletin review*, 13(1):45–52.

Niedenthal, P., Krauth-Gruber, S., and Ric, F. (2007). *Psychology of Emotion.* Psychology Press, New York, USA.

O'Brien, H. L. and Toms, E. G. (2008). What is user engagement? A conceptual framework for defining user engagement with technology. *Journal of the American Society for Information Science and Technology*, 59(6):938–955.

Oliver, R. L. (1993). Cognitive, affective, and attribute bases of the satisfaction response. *Journal of Consumer Research*, 20(3):418–430.

Ortony, A. and Turner, T. J. (1990). What's basic about basic emotions? *Psychological Review*, 97(3):315–31.

Ozuru, Y., Dempsey, K., and McNamara, D. S. (2009). Prior knowledge, reading skill, and text cohesion in the comprehension of science texts. *Learning and Instruction*, 19(3):228 – 242.

Page, L., Brin, S., Motwani, R., and Winograd, T. (1999). The PageRank citation ranking: Bringing order to the web. Technical report.

Pekrun, R. (2006). The control-value theory of achievement emotions: Assumptions, corollaries, and implications for educational research and practice. *Educational Psychology Review*, 18:315–341.

Pekrun, R., Goetz, T., Daniels, L., Stupnisky, R., and Perry, R. (2010). Boredom in achievement settings: Exploring control–value antecedents and performance outcomes of a neglected emotion. *Journal of Educational Psychology*, 102(3):531.

Pekrun, R. and Linnenbrink-Garcia, L. (2012). Academic emotions and student engagement. In Christenson, S. L., Reschly, A. L., and Wylie, C., editors, *Handbook of Research on Student Engagement*, pages 259–282. Springer, New York, NY, USA.

Perfetti, C. A. (1988). *Reading research Advances in theory and practice*, volume 6, chapter Verbal Efficiency in Reading Ability, pages 109–143. Academic Press, Inc.

Peters, E. (2006). The functions of affect in the construction of preferences. In Lichtenstein, S. and Slovic, P., editors, *The Construction of Preference*, pages 454–463. Cambridge University Press.

Petrov, S., Chang, P.-C., Ringgaard, M., and Alshawi, H. (2010). Uptraining for accurate deterministic question parsing. In Li, H. and Marquès, L., editors, *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 705–713, Stroudsburg, PA, USA. Association for Computational Linguistics.

Pfister, H. and Böhm, G. (2008). The multiplicity of emotions: A framework of emotional functions in decision making. *Judgment and Decision Making*, 3(1):5–17.

Porter, M. F. (2001). Snowball: A language for stemming algorithms. Available online at `http://snowball.tartarus.org/texts/introduction.html` [Last accessed on July 23, 2013].

Powers, D. M. W. (2011). Evaluation: From precision, recall and F-factor to ROC, informedness, markedness & correlation. *Journal of Machine Learning Technology*, 2(1):37–63.

Rayner, K. and Reichle, E. D. (2010). Models of the reading process. *Wiley Interdisciplinary Reviews: Cognitive Science*, 1(6):787–799.

Reber, R., Schwarz, N., and Winkielman, P. (2004). Processing fluency and aesthetic pleasure: Is beauty in the perceiver's processing experience? *Personality and Social Psychology Review*, 8(4):364–382.

Reeve, J. (1989). The interest-enjoyment distinction in intrinsic motivation. *Motivation and Emotion*, 13(2):83–103.

Reichle, E. D., Pollatsek, A., Fisher, D. L., and Rayner, K. (1998). Toward a model of eye movement control in reading. *Psychological Review*, 105(1):125–157.

Ricci, F., Rokach, L., Shapira, B., and Kantor, P. B., editors (2009). *Recommender Systems Handbook*. Springer, Berlin / Heidelberg.

Roark, B., Bachrach, A., Cardenas, C., and Pallier, C. (2009). Deriving lexical and syntactic expectation-based measures for psycholinguistic modeling via incremental top-down parsing. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, volume 1 of *EMNLP '09*, pages 324–333, Stroudsburg, PA, USA. Association for Computational Linguistics.

Rosenberg, D. (2003). Early modern information overload. *Journal of the History of Ideas*, 64(1):1–9.

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, 48(2):1–36.

Rotgans, J. I. and Schmidt, H. G. (2011). Situational interest and academic achievement in the active-learning classroom. *Learning and Instruction*, 21(1):58 – 67.

Russell, J. A. (2003). Core affect and the psychological construction of emotion. *Psychological review*, 110(1):145.

Russell, P. A. (1986). Experimental aesthetics of popular music recordings: Pleasingness, familiarity and chart performance. *Psychology of Music*, 14(1):33–43.

Ruthven, I. (2005). Integration approaches to relevance. In Spink, A. and Cole, C., editors, *New Directions in Cognitive Information Retrieval*, pages 61–80. Springer Netherlands.

Ruthven, I. (2012). Grieving online: the use of search engines in times of grief and bereavement. In *Proceedings of the 4th Information Interaction in Context Symposium*, IIIX '12, pages 120–128, New York, NY, USA. ACM.

Ruthven, I., Baillie, M., and Elsweiler, D. (2007). The relative effects of knowledge, interest and confidence in assessing relevance. *Journal of Documentation*, 63(4):482–504.

Ryan, K. (2012). Fathom - measure readability of english text. Available online at `http://search.cpan.org/~kimryan/Lingua-EN-Fathom-1.15/lib/Lingua/EN/Fathom.pm` [Last accessed on July 23, 2013].

Sadoski, M. (2001). Resolving the effects of concreteness on interest, comprehension, and learning important ideas from text. *Educational Psychology Review*, 13(3):263–281.

Salton, G. and Buckley, C. (1990). Improving retrieval performance by relevance feedback. *Journal of the American Society for Information Science*, 41(4):288–297.

Sanders, T. J. M. and Noordman, L. G. M. (2000). The role of coherence relations and their linguistic markers in text processing. *Discourse Processes*, 29(1):37–60.

Saracevic, T. (1975). 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.

Saracevic, T. (1996). Relevance reconsidered. In *Information science: Integration in perspectives. Proceedings of the Second Conference on Conceptions of Library and Information Science (CoLIS 2).*, pages 201–218, Copenhagen (Denmark), 14-17 Oct.

Saracevic, T. (2007). 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.

Schaik, P. v. and Ling, J. (2012). An experimental analysis of experiential and cognitive variables in web navigation. *Human–Computer Interaction*, 27(3):199–234.

Schamber, L. (1994). Relevance and information behavior. *Annual Review of Information Science and Technology (ARIST)*, 29:3–48.

Schamber, L., Eisenberg, M. B., and Nilan, M. S. (1990). A re-examination of relevance: toward a dynamic, situational definition. *Information Processing &amp; Management*, 26(6):755 – 776.

Schenkman, B. N. and Jönsson, F. U. (2000). Aesthetics and preferences of web pages. *Behaviour & Information Technology*, 19(5):367–377.

Scherer, K., Dan, E., and Flykt, A. (2006). What determines a feeling's position in affective space? A case for appraisal. *Cognition & Emotion*, 20(1):92 – 113.

Scherer, K. R. (1995). Plato's legacy: Relationships between cognition, emotion, and motivation. *Geneva Studies in Emotion and Communication*, 9(1):1–7.

Scherer, K. R. (2004). Feelings integrate the central representation of appraisal-driven response organization in emotion. In Manstead, A. S. R., Frijda, N. H., and Fischer, A., editors, *Feelings and emotions: the Amsterdam symposium*, pages 136–157. Cambridge University Press.

Schiefele, U. (1996). Topic interest, text representation, and quality of experience. *Contemporary Educational Psychology*, 21(1):3 – 18.

Schiefele, U. and Krapp, A. (1996). Topic interest and free recall of expository text. *Learning and Individual Differences*, 8(2):141 – 160.

Schraw, G., Dunkle, M. E., and Bendixen, L. D. (1995). Cognitive processes in well-defined and ill-defined problem solving. *Applied Cognitive Psychology*, 9(6):523 – 538.

Schraw, G. and Lehman, S. (2001). Situational interest: A review of the literature and directions for future research. *Educational Psychology Review*, 13(1):23–52.

Schumacker, R. and Lomax, R. (2010). *A Beginner's Guide to Structural Equation Modeling.* Routledge Academic, London, England, third edition.

Schwarm, S. E. and Ostendorf, M. (2005). Reading level assessment using support vector machines and statistical language models. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 523–530, Stroudsburg, PA, USA. Association for Computational Linguistics.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464.

Schwarz, N. (2000). Emotion, cognition, and decision making. *Cognition & Emotion*, 14(4):433–440.

Schwarz, N. (2010). Meaning in context: Metacognitive experiences. In Mesquita, B., Barrett, L. F., and Smith, E. R., editors, *The mind in context*, pages 105–125. The Guilford Press, New York, NY, USA.

Schwarz, N. and Clore, G. L. (2007). Feelings and phenomenal experiences. In Kruglanski, A. and Higgins, E. T., editors, *Social psychology. Handbook of basic principles.*, pages 385–407. Guilford, New York, USA, second edition.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Comput. Surv.*, 34(1):1–47.

Seidenberg, M. S. and McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psychological Review*, 96(4):523 – 568.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27(7, 10):379–423, 625–656.

Silvia, P. J. (2001). Interest and interests: The psychology of constructive capriciousness. *Review of General Psychology*, 5(3):270–290.

Silvia, P. J. (2005). What is interesting? Exploring the appraisal structure of interest. *Emotion*, 5(1):89 – 102.

Silvia, P. J. (2006). *Exploring the psychology of interest.* Oxford University Press, New York.

Silvia, P. J. (2008a). Appraisal components and emotion traits: Examining the appraisal basis of trait curiosity. *Cognition & Emotion*, 22(1):94–113.

Silvia, P. J. (2008b). Interest – the curious emotion. *Current Directions in Psychological Science*, 17(1):57–60.

Silvia, P. J. (2009). Looking past pleasure: Anger, confusion, disgust, pride, surprise, and other unusual aesthetic emotions. *Psychology of Aesthetics, Creativity, and the Arts*, 3(1):48.

Song, H. and Schwarz, N. (2008). If it's hard to read, it's hard to do: Processing fluency affects effort prediction and motivation. *Psychological Science*, 19(10):986–988.

Sparrow, B., Liu, J., and Wegner, D. M. (2011). Google effects on memory: Cognitive consequences of having information at our fingertips. *Science*, 333(6043):776–778.

Sperber, D. and Wilson, D. (1996). *Relevance: Communication and Cognition.* Wiley.

Spink, A. and Greisdorf, H. (2001). Regions and levels: Measuring and mapping users' relevance judgments. *Journal of the American Society for Information Science and Technology*, 52(2):161–173.

Steyvers, M. and Tenenbaum, J. B. (2005). The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. *Cognitive Science*, 29(1):41–78.

Taylor, R. S. (1962). The process of asking questions. *American Documentation*, 13(4):391–396.

Thorndike, E. L. (1921). *The teacher's word book.* Teachers College, Columbia University, New York, NY, USA.

Tiedens, L. Z. and Linton, S. (2001). Judgment under emotional certainty and uncertainty: the effects of specific emotions on information processing. *Journal of personality and social psychology*, 81(6):973.

Toutanova, K., Klein, D., Manning, C. D., and Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In *NAACL '03: Proceedings of the 2003 conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 173–180, Morristown, NJ, USA. Association for Computational Linguistics.

Tractinsky, N., Katz, A., and Ikar, D. (2000). What is beautiful is usable. *Interacting with Computers*, 13(2):127 – 145.

Tversky, A. and Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology*, 5(2):207 – 232.

Uitdenbogerd, A. and van Schnydel, R. (2002). A Review of Factors Affecting Music Recommender Success. In *Proceedings of 3rd International Conference on Music Information Retrieval*, Paris, France.

Van den Broek, E. L. (2011). *Affective Signal Processing (ASP): Unraveling the mystery of emotions.* PhD thesis, Human Media Interaction (HMI), Faculty of Electrical Engineering, Mathematics, and Computer Science, University of Twente, Enschede, The Netherlands.

Van den Broek, E. L., Kisters, P. M., and Vuurpijl, L. G. (2005). Content-based image retrieval benchmarking: Utilizing color categories and color distributions. *Journal of Imaging Science and Technology*, 49(3):293–301.

Van den Broek, E. L., Schouten, T. E., and Kisters, P. M. F. (2008). Modeling human color categorization. *Pattern Recognition Letters*, 29(8):1136–1144.

Van der Heijden, H. (2003). Factors influencing the usage of websites: the case of a generic portal in the netherlands. *Information & Management*, 40(6):541 – 549.

Van der Sluis, F., Duarte Torres, S., Hiemstra, D., Van Dijk, E. M. A. G., and Kruisinga, F. (2011). Visual exploration of health information for children. In Clough, P., Foley, C., Gurrin, C., Jones, G., Kraaij, W., Lee, H., and Mudoch, V., editors, *Advances in Information Retrieval*, volume 6611 of *Lecture Notes in Computer Science*, pages 788–792. Springer Berlin Heidelberg.

Van der Sluis, F., Glassey, R. J., and Van den Broek, E. L. (2012). Making the news interesting: Understanding the relationship between familiarity and interest. In Kamps, J., Kraaij, W., and Fuhr, N., editors, *IIiX 2012: ACM Proceedings of the 4th symposium on Information Interaction in Context*, pages 314–317, New York, NY, USA. ACM.

Van der Sluis, F. and Van den Broek, E. L. (2010). Using complexity measures in information retrieval. In Belkin, N. J. and Kelly, D., editors, *IIiX 2010: ACM Proceedings of the 3th symposium on Information Interaction in Context*, New York, USA. ACM.

Van der Sluis, F., Van den Broek, E. L., Glassey, R. J., van Dijk, E. M. A. G., and de Jong, F. M. G. (i.p.). When complexity becomes interesting. *Journal of the American Society for Information Science and Technology.* [in press].

Van der Sluis, F., Van den Broek, E. L., and Van Dijk, E. M. A. G. (2010). Information Retrieval eXperience (IRX): Towards a human-centered personalized model of relevance. In *Proceedings of the Workshop on Web Information Retrieval Support Systems at the IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*, volume 3, pages 322–325.

Van der Sluis, F. and Van Dijk, E. M. A. G. (2010). A closer look at children's information retrieval usage: Towards child-centered relevance. In Serdyukov, P., Hiemstra, D., and Ruthven, I., editors, *Proceedings of the Workshop on Accessible Search Systems held at the 33st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3–10, New York, USA. ACM.

Van der Sluis, F., Van Dijk, E. M. A. G., and Van den Broek, E. L. (2010). Aiming for user experience in information retrieval: Towards User-Centered Relevance (UCR). In Chen, H.-H., Efthimiadis, E. N., Savoy, J., Crestani, F., and Marchand-Maillet, S., editors, *SIGIR 2010: ACM Proceedings of the 33rd International Conference on Research and Development in Information Retrieval*, pages 924–924, New York, USA. ACM.

Verveer, E. M., Barry, H., J., and Bousfield, W. A. (1933). Change in affectivity with repetition. *The American Journal of Psychology*, 45(1):130–134.

Voorhees, E. (2002). The philosophy of information retrieval evaluation. In *Evaluation of Cross-Language Information Retrieval Systems*, pages 143–170. Springer Berlin / Heidelberg.

Voorhees, E. M. (2005). *TREC: Experiment and Evaluation in Information Retrieval.* Digital Libraries and Electronic Publishing. MIT Press.

Vor der Brück, T., Hartrumpf, S., and Helbig, H. (2008). A readability checker with supervised learning using deep indicators. *Informatica*, 32(4):429–435.

Wade, S. E., Buxton, W. M., and Kelly, M. (1999). Using think-alouds to examine reader-text interest. *Reading Research Quarterly*, 34(2):194–216.

Walker, E. L. (1981). The quest for the inverted u. In Day, H., editor, *Advances in intrinsic motivation and aesthetics*, chapter 3, pages 39–70. New York: Plenum Press.

Wang, P. and Soergel, D. (1998). A cognitive model of document use during a research project. Study I. Document selection. *Journal of the American Society for Information Science*, 49(2):115–133.

Watson, D. (2000). *Mood and temperament.* Guilford Press, New York, USA.

White, H. D. (2007a). Combining bibliometrics, information retrieval, and relevance theory, part 1: First examples of a synthesis. *Journal of the American Society for Information Science and Technology*, 58(4):536–559.

White, H. D. (2007b). Combining bibliometrics, information retrieval, and relevance theory, part 2: Some implications for information science. *Journal of the American Society for Information Science and Technology*, 58(4):583–605.

Wilks, Y. (1999). Evaluating natural language processing systems: An analysis and review. *Artificial Intelligence*, 107(1):165 – 170.

Wilson, T. (1981/2006). On user studies and information needs. *Journal of documentation*, 62(6):658–670.

Winkielman, P., Schwarz, N., Fazendeiro, T., and Reber, R. (2003). The hedonic marking of processing fluency: Implications for evaluative judgment. In Musch, J. and Klauer, K. C., editors, *Affective processes in cognition and emotion*, pages 189–217. Lawrence Erlbaum Associates, Mahwah, NJ, US.

Wixted, J. and Stretch, V. (2004). In defense of the signal detection interpretation of remember/know judgments. *Psychonomic Bulletin & Review*, 11:616–641.

Wundt, W. (1896). *Grundriss der Psychologie.* Wilhelm Engelmann.

Xu, Y. (2007). Relevance judgment in epistemic and hedonic information searches. *Journal of the American Society for Information Science and Technology*, 58(2):179–189.

Xu, Y. C. and Chen, Z. (2006). Relevance judgment: What do information users consider beyond topicality? *Journal of the American Society for Information Science and Technology*, 57(7):961–973.

Yarlas, A. S. and Gelman, R. (1998). Learning as a predictor of situational interest. In *Paper presented at the Annual Meeting of the American Educational Research Association (San Diego, California, April 1998).*

Yi, M. Y. and Hwang, Y. (2003). Predicting the use of web-based information systems: Self-efficacy, enjoyment, learning goal orientation, and the technology acceptance model. *International Journal of Human-Computer Studies*, 59(4):431 – 449.

Yu, L.-C., Wu, C.-H., and Jang, F.-L. (2009). Psychiatric document retrieval using a discourse-aware model. *Artificial Intelligence*, 173(7–8):817 – 829.

Zipf, G. K. (1935). *The psycho-biology of language: An introduction to dynamic philology*. Houghton Mifflin, Boston.

Zyngier, S., Van Peer, W., and Hakemulder, J. (2007). Complexity and foregrounding: In the eye of the beholder? *Poetics Today*, 28(4):653–682.

SIKS Dissertation Series

Since 1998, all dissertations written by Ph.D.-students who have conducted their research under auspices of a senior research fellow of the SIKS research school are published in the SIKS Dissertation Series. This monograph is the **426**th in the series.

2013-28 Frans van der Sluis (UT), *When Complexity becomes Interesting: An Inquiry into the Information eXperience*

2013-27 Mohammad Huq (UT), *Inference-based Framework Managing Data Provenance*

2013-26 Alireza Zarghami (UT), *Architectural Support for Dynamic Homecare Service Provisioning*

2013-25 Agnieszka Anna Latoszek-Berendsen (UM), *Intention-based Decision Support. A new way of representing and implementing clinical guidelines in a Decision Support System*

2013-24 Haitham Bou Ammar (UM), *Automated Transfer in Reinforcement Learning*

2013-23 Patricio de Alencar Silva(UvT), *Value Activity Monitoring*

2013-22 Tom Claassen (RUN), *Causal Discovery and Logic*

2013-21 Sander Wubben (UvT), *Text-to-text generation by monolingual machine translation*

2013-20 Katja Hofmann (UvA), *Fast and Reliable Online Learning to Rank for Information Retrieval*

2013-19 Renze Steenhuizen (TUD), *Coordinated Multi-Agent Planning and Scheduling*

2013-18 Jeroen Janssens (UvT), *Outlier Selection and One-Class Classification*

2013-17 Koen Kok (VU), *The PowerMatcher: Smart Coordination for the Smart Electricity Grid*

2013-16 Eric Kok (UU), *Exploring the practical benefits of argumentation in multi-agent deliberation*

2013-15 Daniel Hennes (UM), *Multiagent Learning - Dynamic Games and Applications*

2013-14 Jafar Tanha (UVA), *Ensemble Approaches to Semi-Supervised Learning Learning*

2013-13 Mohammad Safiri(UT), *Service Tailoring: User-centric creation of integrated IT-based homecare services to support independent living of elderly*

2013-12 Marian Razavian(VU), *Knowledge-driven Migration to Services*

2013-11 Evangelos Pournaras(TUD), *Multi-level Reconfigurable Self-organization in Overlay Services*

2013-10 Jeewanie Jayasinghe Arachchige(UvT), *A Unified Modeling Framework for Service Design.*

2013-09 Fabio Gori (RUN), *Metagenomic Data Analysis: Computational Methods and Applications*

2013-08 Robbert-Jan Merk(VU), *Making enemies: cognitive modeling for opponent agents in fighter pilot simulators*

2013-07 Giel van Lankveld (UT), *Quantifying Individual Player Differences*

2013-06 Romulo Goncalves(CWI), *The Data Cyclotron: Juggling Data and Queries for a Data Warehouse Audience*

2013-05 Dulce Pumareja (UT), *Groupware Requirements Evolutions Patterns*

2013-04 Chetan Yadati(TUD), *Coordinating autonomous planning and scheduling*

2013-03 Szymon Klarman (VU), *Reasoning with Contexts in Description Logics*

2013-02 Erietta Liarou (CWI), *MonetDB/DataCell: Leveraging the Column-*

store Database Technology for Efficient and Scalable Stream Processing

2013-01 Viorel Milea (EUR), *News Analytics for Financial Decision Support*

2012-51 Jeroen de Jong (TUD), *Heuristics in Dynamic Sceduling; a practical framework with a case study in elevator dispatching*

2012-50 Steven van Kervel (TUD), *Ontologogy driven Enterprise Information Systems Engineering*

2012-49 Michael Kaisers (UM), *Learning against Learning - Evolutionary dynamics of reinforcement learning algorithms in strategic interactions*

2012-48 Jorn Bakker (TUE), *Handling Abrupt Changes in Evolving Time-series Data*

2012-47 Manos Tsagkias (UVA), *Mining Social Media: Tracking Content and Predicting Behavior*

2012-46 Simon Carter (UVA), *Exploration and Exploitation of Multilingual Data for Statistical Machine Translation*

2012-45 Benedikt Kratz (UvT), *A Model and Language for Business-aware Transactions*

2012-44 Anna Tordai (VU), *On Combining Alignment Techniques*

2012-43 Withdrawn,

2012-42 Dominique Verpoorten (OU), *Reflection Amplifiers in self-regulated Learning*

2012-41 Sebastian Kelle (OU), *Game Design Patterns for Learning*

2012-40 Agus Gunawan (UvT), *Information Access for SMEs in Indonesia*

2012-39 Hassan Fatemi (UT), *Risk-aware design of value and coordination networks*

2012-38 Selmar Smit (VU), *Parameter Tuning and Scientific Testing in Evolutionary Algorithms*

2012-37 Agnes Nakakawa (RUN), *A Collaboration Process for Enterprise Architecture Creation*

2012-36 Denis Ssebugwawo (RUN), *Analysis and Evaluation of Collaborative Modeling Processes*

2012-35 Evert Haasdijk (VU), *Never Too Old To Learn – On-line Evolution of Controllers in Swarm- and Modular Robotics*

2012-34 Pavol Jancura (RUN), *Evolutionary analysis in PPI networks and applications*

2012-33 Rory Sie (OUN), *Coalitions in Cooperation Networks (COCOON)*

2012-32 Wietske Visser (TUD), *Qualitative multi-criteria preference representation and reasoning*

2012-31 Emily Bagarukayo (RUN), *A Learning by Construction Approach for Higher Order Cognitive Skills Improvement, Building Capacity and Infrastructure*

2012-30 Alina Pommeranz (TUD), *Designing Human-Centered Systems for Reflective Decision Making*

2012-29 Almer Tigelaar (UT), *Peer-to-Peer Information Retrieval*

2012-28 Nancy Pascall (UvT), *Engendering Technology Empowering Women*

2012-27 Hayrettin Gurkok (UT), *Mind the Sheep! User Experience Evaluation & Brain-Computer Interface Games*

2012-26 Emile de Maat (UVA), *Making Sense of Legal Text*

2012-25 Silja Eckartz (UT), *Managing the Business Case Development in Inter-Organizational IT Projects: A Methodology and its Application*

2012-24 Laurens van der Werff (UT), *Evaluation of Noisy Transcripts for Spoken Document Retrieval*

2012-23 Christian Muehl (UT), *Toward Affective Brain-Computer Interfaces: Exploring the Neurophysiology of Affect during Human Media Interaction*

2012-22 Thijs Vis (UvT), *Intelligence, politie en veiligheidsdienst: verenigbare grootheden?*

2012-21 Roberto Cornacchia (TUD), *Querying Sparse Matrices for Information Retrieval*

2012-20 Ali Bahramisharif (RUN), *Covert Visual Spatial Attention, a Robust Paradigm for Brain-Computer Interfacing*

2012-19 Helen Schonenberg (TUE), *What's Next? Operational Support for Business Process Execution*

2012-18 Eltjo Poort (VU), *Improving Solution Architecting Practices*

2012-17 Amal Elgammal (UvT), *Towards a Comprehensive Framework for Business Process Compliance*

2012-16 Fiemke Both (VU), *Helping people by understanding them - Ambient Agents supporting task execution and depression treatment*

2012-15 Natalie van der Wal (VU), *Social Agents. Agent-Based Modelling of Integrated Internal and Social Dynamics of Cognitive and Affective Processes.*

2012-14 Evgeny Knutov(TUE), *Generic Adaptation Framework for Unifying Adaptive Web-based Systems*

2012-13 Suleman Shahid (UvT), *Fun and Face: Exploring non-verbal expressions of emotion during playful interactions*

2012-12 Kees van der Sluijs (TUE), *Model Driven Design and Data Integration in Semantic Web Information Systems*

2012-11 J.C.B. Rantham Prabhakara (TUE), *Process Mining in the Large: Preprocessing, Discovery, and Diagnostics*

2012-10 David Smits (TUE), *Towards a Generic Distributed Adaptive Hypermedia Environment*

2012-09 Ricardo Neisse (UT), *Trust and Privacy Management Support for Context-Aware Service Platforms*

2012-08 Gerben de Vries (UVA), *Kernel Methods for Vessel Trajectories*

2012-07 Rianne van Lambalgen (VU), *When the Going Gets Tough: Exploring Agent-based Models of Human Performance under Demanding Conditions*

2012-06 Wolfgang Reinhardt (OU), *Awareness Support for Knowledge Workers in Research Networks*

2012-05 Marijn Plomp (UU), *Maturing Interorganisational Information Systems*

2012-04 Jurriaan Souer (UU), *Development of Content Management System-based Web Applications*

2012-03 Adam Vanya (VU), *Supporting Architecture Evolution by Mining Software Repositories*

2012-02 Muhammad Umair(VU), *Adaptivity, emotion, and Rationality in Human and Ambient Agent Models*

2012-01 Terry Kakeeto (UvT), *Relationship Marketing for SMEs in Uganda*

2011-49 Andreea Niculescu (UT), *Conversational interfaces for task-oriented spoken dialogues: design aspects influencing interaction quality*

2011-48 Mark Ter Maat (UT), *Response Selection and Turn-taking for a Sensitive Artificial Listening Agent*

2011-47 Azizi Bin Ab Aziz(VU), *Exploring Computational Models for Intelligent Support of Persons with Depression*

2011-46 Beibei Hu (TUD), *Towards Contextualized Information Delivery: A Rule-based Architecture for the Domain of Mobile Police Work*

2011-45 Herman Stehouwer (UvT), *Statistical Language Models for Alternative Sequence Selection*

2011-44 Boris Reuderink (UT), *Robust Brain-Computer Interfaces*

2011-43 Henk van der Schuur (UU), *Process Improvement through Software Operation Knowledge*

2011-42 Michal Sindlar (UU), *Explaining Behavior through Mental State Attribution*

2011-41 Luan Ibraimi (UT), *Cryptographically Enforced Distributed Data Access Control*

2011-40 Viktor Clerc (VU), *Architectural Knowledge Management in Global Software Development*

2011-39 Joost Westra (UU), *Organizing Adaptation using Agents in Serious Games*

2011-38 Nyree Lemmens (UM), *Bee-inspired Distributed Optimization*

2011-37 Adriana Burlutiu (RUN), *Machine Learning for Pairwise Data, Applications for Preference Learning and Supervised Network Inference*

2011-36 Erik van der Spek (UU), *Experiments in serious game design: a cognitive approach*

2011-35 Maaike Harbers (UU), *Explaining Agent Behavior in Virtual Training*

2011-34 Paolo Turrini (UU), *Strategic Reasoning in Interdependence: Logical and Game-theoretical Investigations*

2011-33 Tom van der Weide (UU), *Arguing to Motivate Decisions*

2011-32 Nees-Jan van Eck (EUR), *Methodological Advances in Bibliometric Mapping of Science*

2011-31 Ludo Waltman (EUR), *Computational and Game-Theoretic Approaches for Modeling Bounded Rationality*

2011-30 Egon L. van den Broek (UT), *Affective Signal Processing (ASP): Unraveling the mystery of emotions*

2011-29 Faisal Kamiran (TUE), *Discrimination-aware Classification*

2011-28 Rianne Kaptein(UVA), *Effective Focused Retrieval by Exploiting Query Context and Document Structure*

2011-27 Aniel Bhulai (VU), *Dynamic website optimization through autonomous management of design patterns*

2011-26 Matthijs Aart Pontier (VU), *Virtual Agents for Human Communication - Emotion Regulation and Involvement-Distance Trade-Offs in Embodied Conversational Agents and Robots*

2011-25 Syed Waqar ul Qounain Jaffry (VU)), *Analysis and Validation of Models for Trust Dynamics*

2011-24 Herwin van Welbergen (UT), *Behavior Generation for Interpersonal Coordination with Virtual Humans On Specifying, Scheduling and Realizing Multimodal Virtual Human Behavior*

2011-23 Wouter Weerkamp (UVA), *Finding People and their Utterances in Social Media*

2011-22 Junte Zhang (UVA), *System Evaluation of Archival Description and Access*

2011-21 Linda Terlouw (TUD), *Modularization and Specification of Service-Oriented Systems*

2011-20 Qing Gu (VU), *Guiding service-oriented software engineering - A view-based approach*

2011-19 Ellen Rusman (OU), *The Mind ' s Eye on Personal Profiles*

2011-18 Mark Ponsen (UM), *Strategic Decision-Making in complex games*

2011-17 Jiyin He (UVA), *Exploring Topic Structure: Coherence, Diversity and Relatedness*

2011-16 Maarten Schadd (UM), *Selective Search in Games of Different Complexity*

2011-15 Marijn Koolen (UvA), *The Meaning of Structure: the Value of Link Evidence for Information Retrieval*

2011-14 Milan Lovric (EUR), *Behavioral Finance and Agent-Based Artificial Markets*

2011-13 Xiaoyu Mao (UvT), *Airport under Control. Multiagent Scheduling for Airport Ground Handling*

2011-12 Carmen Bratosin (TUE), *Grid Architecture for Distributed Process Mining*

2011-11 Dhaval Vyas (UT), *Designing for Awareness: An Experience-focused HCI Perspective*

2011-10 Bart Bogaert (UvT), *Cloud Content Contention*

2011-09 Tim de Jong (OU), *Contextualised Mobile Media for Learning*

2011-08 Nieske Vergunst (UU), *BDI-based Generation of Robust Task-Oriented Dialogues*

2011-07 Yujia Cao (UT), *Multimodal Information Presentation for High Load Human Computer Interaction*

2011-06 Yiwen Wang (TUE), *Semantically-Enhanced Recommendations in Cultural Heritage*

2011-05 Base van der Raadt (VU), *Enterprise Architecture Coming of Age - Increasing the Performance of an Emerging Discipline.*

2011-04 Hado van Hasselt (UU), *Insights in Reinforcement Learning; Formal analysis and empirical evaluation of temporal-difference*

2011-03 Jan Martijn van der Werf (TUE), *Compositional Design and Verification of Component-Based Information Systems*

2011-02 Nick Tinnemeier(UU), *Organizing Agent Organizations. Syntax and Operational Semantics of an Organization-Oriented Programming Language*

2011-01 Botond Cseke (RUN), *Vari-*

ational Algorithms for Bayesian Inference in Latent Gaussian Models

2010-53 Edgar Meij (UVA), *Combining Concepts and Language Models for Information Access*

2010-52 Peter-Paul van Maanen (VU), *Adaptive Support for Human-Computer Teams: Exploring the Use of Cognitive Models of Trust and Attention*

2010-51 Alia Khairia Amin (CWI), *Understanding and supporting information seeking tasks in multiple sources*

2010-50 Bouke Huurnink (UVA), *Search in Audiovisual Broadcast Archives*

2010-49 Jahn-Takeshi Saito (UM), *Solving difficult game positions*

2010-48 Withdrawn,

2010-47 Chen Li (UT), *Mining Process Model Variants: Challenges, Techniques, Examples*

2010-46 Vincent Pijpers (VU), *e3alignment: Exploring Inter-Organizational Business-ICT Alignment*

2010-45 Vasilios Andrikopoulos (UvT), *A theory and model for the evolution of software services*

2010-44 Pieter Bellekens (TUE), *An Approach towards Context-sensitive and User-adapted Access to Heterogeneous Data Sources, Illustrated in the Television Domain*

2010-43 Peter van Kranenburg (UU), *A Computational Approach to Content-Based Retrieval of Folk Song Melodies*

2010-42 Sybren de Kinderen (VU), *Needs-driven service bundling in a multi-supplier setting - the computational e3-service approach*

2010-41 Guillaume Chaslot (UM), *Monte-Carlo Tree Search*

2010-40 Mark van Assem (VU), *Converting and Integrating Vocabularies for the Semantic Web*

2010-39 Ghazanfar Farooq Siddiqui (VU), *Integrative modeling of emotions in virtual agents*

2010-38 Dirk Fahland (TUE), *From Scenarios to components*

2010-37 Niels Lohmann (TUE), *Correctness of services and their composition*

2010-36 Jose Janssen (OU), *Paving the Way for Lifelong Learning; Facilitating competence development through a learning path specification*

2010-35 Dolf Trieschnigg (UT), *Proof of Concept: Concept-based Biomedical Information Retrieval*

2010-34 Teduh Dirgahayu (UT), *Interaction Design in Service Compositions*

2010-33 Robin Aly (UT), *Modeling Representation Uncertainty in Concept-Based Multimedia Retrieval*

2010-32 Marcel Hiel (UvT), *An Adaptive Service Oriented Architecture: Automatically solving Interoperability Problems*

2010-31 Victor de Boer (UVA), *Ontology Enrichment from Heterogeneous Sources on the Web*

2010-30 Marieke van Erp (UvT), *Accessing Natural History - Discoveries in data cleaning, structuring, and retrieval*

2010-29 Stratos Idreos(CWI), *Database Cracking: Towards Auto-tuning Database Kernels*

2010-28 Arne Koopman (UU), *Characteristic Relational Patterns*

2010-27 Marten Voulon (UL), *Automatisch contracteren*

2010-26 Ying Zhang (CWI), *XRPC: Efficient Distributed Query Processing on Heterogeneous XQuery Engines*

2010-25 Zulfiqar Ali Memon (VU), *Modelling Human-Awareness for Ambient Agents: A Human Mindreading Perspective*

2010-24 Dmytro Tykhonov, *Designing Generic and Efficient Negotiation Strategies*

2010-23 Bas Steunebrink (UU), *The Logical Structure of Emotions*

2010-22 Michiel Hildebrand (CWI), *End-user Support for Access to Heterogeneous Linked Data*

2010-21 Harold van Heerde (UT), *Privacy-aware data management by means of data degradation*

2010-20 Ivo Swartjes (UT), *Whose Story Is It Anyway? How Improv Informs Agency and Authorship of Emergent Narrative*

2010-19 Henriette Cramer (UvA),

People's Responses to Autonomous and Adaptive Systems

2010-18 Charlotte Gerritsen (VU), *Caught in the Act: Investigating Crime by Agent-Based Simulation*

2010-17 Spyros Kotoulas (VU), *Scalable Discovery of Networked Resources: Algorithms, Infrastructure, Applications*

2010-16 Sicco Verwer (TUD), *Efficient Identification of Timed Automata, theory and practice*

2010-15 Lianne Bodenstaff (UT), *Managing Dependency Relations in Inter-Organizational Models*

2010-14 Sander van Splunter (VU), *Automated Web Service Reconfiguration*

2010-13 Gianluigi Folino (RUN), *High Performance Data Mining using Bio-inspired techniques*

2010-12 Susan van den Braak (UU), *Sensemaking software for crime analysis*

2010-11 Adriaan Ter Mors (TUD), *The world according to MARP: Multi-Agent Route Planning*

2010-10 Rebecca Ong (UL), *Mobile Communication and Protection of Children*

2010-09 Hugo Kielman (UL), *A Politiele gegevensverwerking en Privacy, Naar een effectieve waarborging*

2010-08 Krzysztof Siewicz (UL), *Towards an Improved Regulatory Framework of Free Software. Protecting user freedoms in a world of software communities and eGovernments*

2010-07 Wim Fikkert (UT), *Gesture interaction at a Distance*

2010-06 Sander Bakkes (UvT), *Rapid Adaptation of Video Game AI*

2010-05 Claudia Hauff (UT), *Predicting the Effectiveness of Queries and Retrieval Systems*

2010-04 Olga Kulyk (UT), *Do You Know What I Know? Situational Awareness of Co-located Teams in Multidisplay Environments*

2010-03 Joost Geurts (CWI), *A Document Engineering Model and Processing Framework for Multimedia documents*

2010-02 Ingo Wassink (UT), *Work flows in Life Science*

2010-01 Matthijs van Leeuwen (UU), *Patterns that Matter*

2009-46 Loredana Afanasiev (UvA), *Querying XML: Benchmarks and Recursion*

2009-45 Jilles Vreeken (UU), *Making Pattern Mining Useful*

2009-44 Roberto Santana Tapia (UT), *Assessing Business-IT Alignment in Networked Organizations*

2009-43 Virginia Nunes Leal Franqueira (UT), *Finding Multi-step Attacks in Computer Networks using Heuristic Search and Mobile Ambients*

2009-42 Toine Bogers (UvT), *Recommender Systems for Social Bookmarking*

2009-41 Igor Berezhnyy (UvT), *Digital Analysis of Paintings*

2009-40 Stephan Raaijmakers (UvT), *Multinomial Language Learning: Investigations into the Geometry of Language*

2009-39 Christian Stahl (TUE, Humboldt-Universitaet zu Berlin), *Service Substitution – A Behavioral Approach Based on Petri Nets*

2009-38 Riina Vuorikari (OU), *Tags and self-organisation: a metadata ecology for learning resources in a multilingual context*

2009-37 Hendrik Drachsler (OUN), *Navigation Support for Learners in Informal Learning Networks*

2009-36 Marco Kalz (OUN), *Placement Support for Learners in Learning Networks*

2009-35 Wouter Koelewijn (UL), *Privacy en Politiegegevens; Over geautomatiseerde normatieve informatie-uitwisseling*

2009-34 Inge van de Weerd (UU), *Advancing in Software Product Management: An Incremental Method Engineering Approach*

2009-33 Khiet Truong (UT), *How Does Real Affect Affect Affect Recognition In Speech?*

2009-32 Rik Farenhorst (VU) and Remco de Boer (VU), *Architectural Knowledge Management: Supporting Architects and Auditors*

2009-31 Sofiya Katrenko (UVA), *A Closer Look at Learning Relations from Text*

2009-30 Marcin Zukowski (CWI), *Balancing vectorized query execution with bandwidth-optimized storage*

2009-29 Stanislav Pokraev (UT), *Model-Driven Semantic Integration of Service-Oriented Applications*

2009-28 Sander Evers (UT), *Sensor Data Management with Probabilistic Models*

2009-27 Christian Glahn (OU), *Contextual Support of social Engagement and Reflection on the Web*

2009-26 Fernando Koch (UU), *An Agent-Based Model for the Development of Intelligent Mobile Services*

2009-25 Alex van Ballegooij (CWI), *"RAM: Array Database Management through Relational Mapping"*

2009-24 Annerieke Heuvelink (VUA), *Cognitive Models for Training Simulations*

2009-23 Peter Hofgesang (VU), *Modelling Web Usage in a Changing Environment*

2009-22 Pavel Serdyukov (UT), *Search For Expertise: Going beyond direct evidence*

2009-21 Stijn Vanderlooy (UM), *Ranking and Reliable Classification*

2009-20 Bob van der Vecht (UU), *Adjustable Autonomy: Controling Influences on Decision Making*

2009-19 Valentin Robu (CWI), *Modeling Preferences, Strategic Reasoning and Collaboration in Agent-Mediated Electronic Markets*

2009-18 Fabian Groffen (CWI), *Armada, An Evolving Database System*

2009-17 Laurens van der Maaten (UvT), *Feature Extraction from Visual Data*

2009-16 Fritz Reul (UvT), *New Architectures in Computer Chess*

2009-15 Rinke Hoekstra (UVA), *Ontology Representation - Design Patterns and Ontologies that Make Sense*

2009-14 Maksym Korotkiy (VU), *From ontology-enabled services to service-enabled ontologies (making ontologies work in e-science with ONTO-SOA)*

2009-13 Steven de Jong (UM), *Fairness in Multi-Agent Systems*

2009-12 Peter Massuthe (TUE, Humboldt-Universitaet zu Berlin), *Operating Guidelines for Services*

2009-11 Alexander Boer (UVA), *Legal Theory, Sources of Law & the Semantic Web*

2009-10 Jan Wielemaker (UVA), *Logic programming for knowledge-intensive interactive applications*

2009-09 Benjamin Kanagwa (RUN), *Design, Discovery and Construction of Service-oriented Systems*

2009-08 Volker Nannen (VU), *Evolutionary Agent-Based Policy Analysis in Dynamic Environments*

2009-07 Ronald Poppe (UT), *Discriminative Vision-Based Recovery and Recognition of Human Motion*

2009-06 Muhammad Subianto (UU), *Understanding Classification*

2009-05 Sietse Overbeek (RUN), *Bridging Supply and Demand for Knowledge Intensive Tasks - Based on Knowledge, Cognition, and Quality*

2009-04 Josephine Nabukenya (RUN), *Improving the Quality of Organisational Policy Making using Collaboration Engineering*

2009-03 Hans Stol (UvT), *A Framework for Evidence-based Policy Making Using IT*

2009-02 Willem Robert van Hage (VU), *Evaluating Ontology-Alignment Techniques*

2009-01 Rasa Jurgelenaite (RUN), *Symmetric Causal Independence Models*