When Complexity Becomes Interesting

Frans van der Sluis
Human Media Interaction Group, Faculty of Electrical Engineering, Mathematics, and Computer Science, University of Twente, P.O. Box 217, 7500AE, Enschede, The Netherlands. E-mail: f.vandersluis@acm.org;

Egon L. van den Broek
Human Media Interaction Group, Faculty of Electrical Engineering, Mathematics, and Computer Science, University of Twente, P.O. Box 217, 7500AE, Enschede, The Netherlands; and Department of Information and Computing Sciences, Utrecht University, Princetonplein 5, 3584CC, Utrecht, The Netherlands. E-mail: vandenbroek@acm.org

Richard J. Glassey
School of Computing, Robert Gordon University, Garthdee Road, AB10 7QJ, Aberdeen, United Kingdom. E-mail: r.j.glassey@rgu.ac.uk

Elisabeth M. A. G. van Dijk
Human Media Interaction Group, Faculty of Electrical Engineering, Mathematics, and Computer Science, University of Twente, P.O. Box 217, 7500AE, Enschede, The Netherlands. E-mail: e.m.a.g.vandijk@utwente.nl

Franciska M. G. de Jong
Human Media Interaction Group, Faculty of Electrical Engineering, Mathematics, and Computer Science, University of Twente, P.O. Box 217, 7500AE, Enschede, The Netherlands. E-mail: f.m.g.dejong@utwente.nl

How to provide users a positive experience during interaction with information (i.e., the “Information eXperience” (IX)) is still an open question. As a starting point, this work investigates how the emotion of interest can be influenced by modifying the complexity of the information presented to users. The appraisal theory of interest suggests a “sweet spot” where interest will be at its peak: information that is novel and complex yet still comprehensible. This “sweet spot” is approximated using two studies. Study One develops a computational model of textual complexity founded on psycholinguistic theory on processing difficulty. The model was trained and tested on 12,420 articles, achieving a classification performance of 90.87% on two classes of complexity. Study Two puts the model to its ultimate test: Its application to change the user’s IX. Using 18 news articles the influence of complexity on interest and its appraisals is unveiled. A structural equation model shows a positive influence of complexity on interest, yet a negative influence of comprehensibility, confirming a seemingly paradoxical relationship between complexity and interest. By showing when complexity becomes interesting, this paper shows how information systems can use the model of textual complexity to construct an interesting IX.

Received October 23, 2012; revised July 31, 2013; accepted July 31, 2013

© 2014 ASIS&T • Published online 13 March 2014 in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/asi.23095

Introduction

The following is among the biggest challenges of information systems but, in parallel, also the one that is ignored most: Provide users a positive experience during interaction with information (Belkin, 2008), that is, a positive “Information eXperience” (IX). This challenge exists across many domains, whether users are making critical professional decisions or seeking casual social awareness. Although systems can efficiently retrieve, aggregate, rank, filter, and recommend information, the IX of a user is not evaluated as part of a system’s usefulness. For example, information filtering and recommending (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 should lead to a positive IX (cf. Konstan & Riedl, 2012). However, it could be argued that “more of the same” may cause “diminishing returns,” generate filter bubbles with a limited degree of novelty (Ricci, Rokach, Shapira, & Kantor, 2009) and be detrimental to the IX. Similarly in information retrieval (IR) systems, Kuhlthau (2004), Arapakis, Jose, and Gray (2008), and Bowler (2010) have observed the occurrence of a range of negative emotions (e.g., irritation, anxiety, and despair) during retrieval tasks.

To evaluate users' IX, either positive or negative, one cannot rely on system behavior alone. Moreover, it is unclear what exactly constitutes a positive IX; nor is it clear which emotional experiences are desirable or "useful" during interaction or what their causes or effects are (Arapakis et al., 2008; Belkin, 2008; Bowler, 2010; Kuhlthau, 2004). There is a clear need to delineate what constitutes a better IX and correspondingly a more positive User eXperience (UX)—the complex fabric of thoughts, feelings, and actions experienced during user interaction (Hassenzahl, 2013). The IX is considered a subset of the UX. The latter describes the experience during interaction with all facets of a product, including its design and interface. On the contrary, the former focuses only on one aspect: the information. Although UX is a difficult concept to operationalize, it hints at the utility of incorporating emotion into the realm of IX. A more proactive system might actively attempt to affect the emotional state of the user, by targeting emotions closely related to information seeking, such as interest (Glassey & Azzopardi, 2011), certainty (Kuhlthau, 2004), and surprise (Arapakis et al., 2008).

As a starting point, this work investigates how the experience of the emotion of interest can be influenced by modifying the complexity of the information presented to users. Defining *interest* as an emotion allows one to characterize interest by its cognitive, subjective, and physiological and expressive response components (Silvia, 2008b), concretizing three aspects of the UX: the thoughts, feelings, and actions, respectively. Furthermore, this allows one to identify the causes of interest—that is, learn the relationships between stimuli and responses (see Interest section; Silvia, 2008b), in particular, the relationship between textual complexity and an interest response. The emotion of interest differs from the long-term interests usually modeled in IR and IF&R systems. Whereas the importance of long-term interests for explaining relevance decisions has been confirmed (Ruthven, Baillie, & Elsweiler, 2007), the role of interest during information interaction is yet to be confirmed. Interest is believed to be key to a positive IX: The "quality of experience seems to be an epiphenomenon of interest" (Schiefele, 1996, p. 13) and to be part of an engaging UX (O'Brien & Toms, 2008). Furthermore, the emotion of interest has the potential to influence all other relevance criteria users apply (i.e., affective relevance; Cosijn & Ingwersen, 2000). By taking interest as a primary goal for information systems, this article operationalizes the holistic concept of the IX as the amendable goal of predicting if and when a stimulus leads to an interest response.

Interest is regarded as 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 & Krapp, 1996). However, a particular text cannot easily be categorized as interesting or uninteresting by looking at objective features derived solely from its content. 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 later) (Schraw & Lehman, 2001). 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 most likely do 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. Although 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. The relation between textual complexity and interest has been studied extensively from an educational perspective, focusing mainly on one side of the complexity spectrum: for complex stimuli, comprehensibility enhances interest and learning. For example, Schraw et al. (1995) found 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. Similar results were obtained by numerous other studies (cf., Hidi, 1990; Schraw & 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.

Textual complexity is expected to be an important factor in predicting interest, aside from the topical familiarity of the user that is often implemented in IF&R systems (Van der Sluis, Glassey, & Van den Broek, 2012). It has also been identified as part of the relevance criteria users apply when using IR systems (Barry & Schamber, 1998; Xu & Chen, 2006), besides topicality, which is generally regarded a precondition to the importance of other types of relevance (Spink

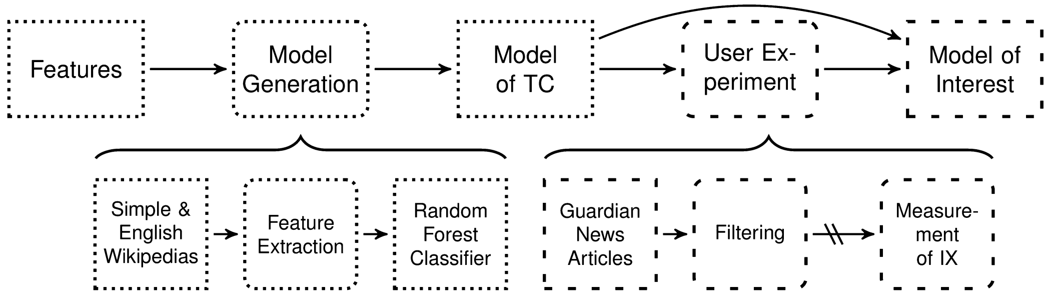

FIG. 1. Flow diagram of the design of and relation between Study 1 and Study 2. Processes are denoted by rounded rectangles and objects (e.g., results or data sets) by square rectangles. Study 1 and Study 2 are contrasted by the borders: respectively, a dotted and a dashed border. Each of the rectangles represents a subsection of this paper.

& Greisdorf, 2001). When included in its set of information metrics, a metric of textual complexity allows an information system to select information that is appraised within the sweet spot of interest. However, for a metric of textual complexity to actually influence the experience of interest, the metric needs to be generic, that is, applicable to a variety of data sets and reflective of subjective, experienced complexity. This ensures that the metric actually predicts experienced complexity for a multitude of (genres of) stimuli. Although some studies explore the feasibility of a generic metric, the actual genericity generally remains untested (Benjamin, 2012). Notable exceptions are provided by Collins-Thompson, Bennett, White, de la Chica, and Sontag (2011), who use a readability metric to predict the time a user of an IR system spends on a web page, and by Vor der Brück, Hartrumpf, and Helbig (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. Hence, although some steps have been made toward a generic model of textual complexity, the ability of such a metric to actually influence the IX is unclear.

The hypothesis guiding this article is that information systems can foster the experience of interest using a metric of textual complexity, selecting information within the sweet spot of interest—novel-complex while remaining comprehensible. Consequently, this sets the prediction that an information system can actively construct part of the IX. This article explores the hypothesis through two subsequent studies. Study 1 develops a model of textual complexity that can be applied in information systems. The model can be expected to be predictive of experienced complexity by integrating psycholinguistic findings on processing difficulty. A classifier is trained and evaluated using a large data set containing stimuli distinctive on textual complexity. Study 2 evaluates if and when textual complexity influences interest. The model of Study 1 is directly applied in Study 2 to filter complex and easy stimuli from a news corpus, making Study 2 an additional test case for the generic model from Study 1. The effect of textual complexity is evaluated for both consecutive appraisals, appraised complexity and appraised comprehensibility, and in relation to interest. Study 2 combines the model of Study 1 into an explanatory path model.

The resulting combined model allows one to reflect on whether and when complexity becomes interesting. The connections between the two studies are shown in Figure 1. The figure maps the development of the model of textual complexity, which, in turn, is applied to foster an interest response, to the resulting combined model of interest. The combination of both studies shows the feasibility for future systems to target the sweet spot of interest by adding a model of textual complexity to its set of features. This is a first step in better understanding the interplay of information interaction and the experience of interest, and, accordingly, at operationalizing the IX.

The remainder of this article is organized as follows: The Background section provides a theoretical overview of textual complexity, interest, and a review of the methods used by IF&R and IR systems to select relevant information. Study 1 is presented in the Study 1: Generic Model of Textual Complexity section. Details are given on how a model of textual complexity is built and trained and of a data-driven evaluation of the model. Study 2 is presented in the Study 2: Influencing Interest section. A user study is described in which an interest response is fostered using the developed model of textual complexity. Finally, the General Discussion section discusses the findings and implications of both studies.

Background

This section introduces the necessary background knowledge on interest (see Interest section) and textual complexity (see Textual Complexity section). Furthermore, the methods and evaluative approaches on related systems are briefly reviewed (see Information Systems section).

Interest

The study of interest and related epistemological emotions has a long history. At the top of his list of passions, Descartes (1649) listed the emotion of wonder, stating its role in motivating people toward certain actions, such as the

DOI: 10.1002/asi

desire to learn. A tamer version of wonder, interest, has received considerable attention: Its determinants, consequences, and components have been studied (Silvia, 2008b). Interest is characterized by a cognitive component (appraisal), a subjective component (feeling), and physiological and expressive components (movement of muscles in the forehead and eyes, faster speech rate, and greater frequency range) (Banse & Scherer, 1996; Hess & Polt, 1960). These are all typical features of emotions (Lazarus, 1991), supporting its status as an emotion. This section elaborates on both the consequences and the determinants of interest, highlighting the importance of interest and theories on the causes of interest.

The momentary emotion of interest differs from the long-term interests[1] often implemented in IR and IF&R systems (see Relevance subsection). Although interests are important to explain the relevance decisions that users make (Ruthven et al., 2007), they do not necessarily lead to interest and interest does not directly lead to (but is a requirement for) the development of interests (Silvia, 2001). In other words, interests can be seen as a determinant as well as a consequence of interest, the importance of which is highlighted in the following two subsections.

Consequences. The importance of interest has long been noted for various cognitive processes, for example, for learning (Schiefele & Krapp, 1996), problem solving (Bowler, 2010), and motivation in general (Reeve, 1989). Subjects acquire a higher depth of comprehension, apply better learning strategies, and have an overall more enjoyable (learning) experience when texts evoke interest (Schiefele, 1996; Schiefele & Krapp, 1996). As Jonassen (2000) summarized: "Students think harder and process material more deeply when they are interested" (p. 71). Three consequences are highlighted: long-term interests, motivation, and the overall IX.

Although many theories on the development of long-term interests exist, 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: (a) triggered "situational interest," (b) maintained "situational interest," (c) emerging "individual interest," and (d) well-developed "individual interest." In this article, "situational interest" can be regarded similar to the emotion of interest and "individual interest" to long-term interests. 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.

Interest is a primary determinant of motivation: Motivation starts with interest by arousing the initiation of attention and exploratory behavior and then interacts with enjoyment in sustaining a persistence in an activity (Reeve, 1989). Although interest is a positive emotion, interest is distinctively different from enjoyment. Where interest generally occurs for novel and complex things (see Determinants section), enjoyment occurs for familiar and (somewhat) easy things. This indicates an orthogonal relation between 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. Theories of user engagement illustrate this (O'Brien & Toms, 2008). For example, a flow experience, the feeling of complete and energized focus in an activity with a high level of enjoyment and fulfillment, shows that interest and enjoyment can be aligned (Csikszentmihalyi, 1991). For flow, a balance between challenge and skills is of prime importance, implying that complex things provoking an interest response can be followed by an enjoyable and motivated experience.

The distinction between interest and enjoyment makes the relation between interest and IX less than trivial, where interest does not always lead to an enjoyable experience (Bowler, 2010) but can coexist (e.g., during a flow experience) (O'Brien & Toms, 2008). Furthermore, there can be too much of a good a thing: Too many interesting objects could lead to a negative IX (Glassey & Azzopardi, 2011). For the evaluation of information systems, this posits a problem: Although the task of a recommender system is to invoke interest (see Performance and Evaluation subsection), a balance may be required to achieve a positive IX (Glassey & Azzopardi, 2011).

Determinants. Early theories on interest focus mainly on textual characteristics causing interest. Berlyne (1960) was among the first to comprise a list of collative variables, properties of textual stimuli associated with an interest response: novelty, surprisingness, incongruity, complexity, variation, and puzzlement. Interest, however, is particularly difficult to predict because of within- and between-subject variation. The former indicates that something that is interesting today may not be so tomorrow or the next month. The latter indicates that something interesting for one person may be uninteresting to another person (Silvia, 2008b).

Later theories, including a later position taken by Berlyne (1975), emphasize the role of subjective judgments and related personal states and traits, acknowledging the variation of interest. Proceeding with this position, Schraw and Lehman (2001) extended and categorized 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 is an interplay between text characteristics, task context, and the knowledge of the reader. With a focus solely on short-term interest, the appraisal theory of interest provides insight into the complex interaction between user, task, and

[1] The plural interests are used for stable, topical interests. The singular interest refers to the emotion of interest.

event. It explains using two subjective appraisals when the emotion of interest occurs and how to predict it.

The appraisal theory of interest (Silvia, 2008b) fits into the general appraisal theories of emotion (Ellsworth & Scherer, 2003). The contemporary appraisal theories of emotion state that the experience of an emotion, including interest, is dependent on two consecutive appraisals (Ellsworth & Scherer). The primary appraisal evaluates a situation's significance regarding a person's well-being. This appraisal is a highly automated evaluation of stimulus characteristics. At a primitive level this includes an evaluation of stimulus intensity. At a higher level this appraisal depends on the predictability or familiarity of a stimulus and on the valence or intrinsic pleasantness of the stimulus with regard to a person's needs, goals, and values. The secondary appraisal assesses our ability to deal with the situation, that is, the coping potential. The ability of a person to change the situation and its consequences determines the response (e.g., fight or flight; Ellsworth & Scherer, 2003). The appraisal is proactive, going beyond the current stimulus and evaluating possible outcomes.

For interest, 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. This appraisal is a subjective successor of the (objective) collative variables identified by Berlyne (1960). The "appraised comprehensibility" is the operationalization of the secondary appraisal for interest. It determines the coping potential related to prior knowledge, available resources, and so forth (Silvia, 2005). For example, if a stimulus is too complex, the coping abilities will probably not suffice, leading to a different emotion than interest. 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 it (Connelly, 2011), and in a study on learning from text, where interestingness was found to correlate with comprehensibility, familiarity, and concreteness (Sadoski, 2001).

Ample evidence exists of the importance of other determinants next to long-term interests in evoking an interest response. Textual complexity is one of them, as it is expected to influence both the primary and secondary appraisal (see the Introduction).

Textual Complexity

Many attempts have been made to predict the processing difficulty a user has with a text via the analysis of textual complexity—those features of a text that make it easier or harder to process. Processing difficulty is commonly described as the measurable effort required to process a new token of information (Jaeger & Tily, 2011). Processing difficulty, according to the verbal efficiency theory (Perfetti, 1988), propagates bottom-up: when lower-level interpretations of (the tokens of) a text fail, higher-level interpretations are also incorrect. Accordingly, three levels of interpretation can be discerned: word, sentence, and discourse. Different influences on processing difficulty can be identified at each level, namely, at word level, word decoding and vocabulary access; at sentence level, syntactical analyses and semantic interpretation; and, at discourse-level sentence integration and inference processing.

For comprehensibility, evidence for the propagation of processing difficulty is not unequivocal: less-than-optimal word processing may be sufficient for good discourse comprehension (Long, Wilson, Hurley, & Prat, 2006). This indicates a difference between processing difficulty and comprehension, where difficulties at the word-level may contribute to processing difficulty, but not per se hurt comprehensibility. A similar relation can be expected for interest, where processing difficulty may enhance the primary novelty-complexity appraisal without significantly influencing the secondary appraisal of coping potential and, consequently, increase the experience of interest.

To prevent a lengthy recap of models of reading, only the core effects known to influence processing difficulty are reviewed, divided into word, sentence, and discourse-level effects. For the interested reader, we refer to Rayner and Reichle (2010) for an overview of contemporary models of the reading process.

Word-level effects. Using a memory-based theory of word-level processing difficulty, two types of effects can be discerned: isolated and interword effects. In a memory-based account, each word has a certain level of activation, where a higher level of activation eases the access to a word. The baseline activation of a word is explained by isolated word effects (Jaeger & Tily, 2011). Besides the baseline activation, the level of activation can be increased by foregoing words. These interword effects give rise to (a) boosts in activation due to previous retrievals (i.e., repeated access) and (b) spreading activation to similar elements in memory (Jaeger & Tily, 2011).

Isolated word effects are of a lexical nature, implying that a reader has some kind of innate lexicon in which words are searched and meaning is found. Two types of variables can be differentiated: orthographic and semantic. Key orthographic variables of influence on word recognition are word length, word frequency, familiarity, and age of acquisition. Measures for word length are given in Equations 1 and 2 (see Traditional subsection) and for word frequency in Equations 3 (see Traditional subsection) and 5 (see Lexical Familiarity subsection). Key semantic variables are concreteness and meaningfulness (Balota, Yap, & Cortese, 2006). All of these isolated word variables are correlated. More frequent words tend to be shorter, more familiar, and acquired earlier on in life (Balota, Cortese, Sergent-Marshall, Spieler, & Yap, 2004).

Interword effects are those where the processing of a word is influenced by a foregoing word (i.e., priming effects). Again, two types of effects can be differentiated:

orthographic and semantic. With orthographic priming, word recognition is improved when a word shares (more) letters with its prime. With semantic priming, word recognition is improved when a word is associated with its prime (Balota et al., 2006). Examples are the combinations "couch-touch" for orthographic priming and "feel-touch" for semantic priming. Measures for orthographic priming are given by Equations 6 and 7 (see Priming subsection).

Sentence-level effects. Theories of sentence-level effects on processing difficulty fall into two broad categories—memory-based accounts, explaining difficulty because of some limited resource, and constraint-satisfaction accounts, explaining difficulty by the probability of a processed structure: Infrequent or unexpected words or structures are more difficult to process.

A contemporary memory-based theory is the dependency-locality theory (DLT). The DLT states that a reader, while reading, performs a moment-by-moment integration of new information sources. This implies an evolving structure is kept in mind, keeping track of what is 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 that 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 DLT is particularly capable of explaining the processing cost of normal sentences, contrary to syntactically ambiguous ones. This makes the DLT particularly suitable as a measure of processing difficulty. Equation 8 (see Dependency-Locality subsection) defines a measure of integration costs.

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. The bigger the decrease in probability mass, the higher the informativeness of a new piece of information was, which, in turn, leads to higher processing complexity and longer reading times. However, this account is mainly able to differentiate between common and rare (e.g., syntactically ambiguous) sentences (Jaeger & Tily, 2011), making it less suitable to indicate the processing difficulty of common sentences.

Discourse-level 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 & van Dijk, 1978). The integration of incoming information with the current situational model can be facilitated by the degree to which a reader can connect incoming information with prior information either in the stimulus (i.e., cohesion) or in memory (i.e., knowledge). The resulting mental model is more coherent when either the text is cohesive or the reader has enough knowledge to fill gaps of information (Benjamin, 2012).[2]

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 characterized as having dense lexical chains. Such chains can be identified on a semantic level and on a linguistic level. At 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. At a linguistic level, repeated mentions referring to the same (linguistic) entity provide cues on how to relate incoming information to the active mental representation of a text. These effects can be explained using a memory-based account, where spreading activation eases the connection of new information. Partly, cohesion is measured by the orthographic repetition measures defined in Equations 6 and 7 (see Priming subsection).

The effects outlined in this section form the theoretical environment used to operationalize the concept of textual complexity in Study 1 (see Study 1: Generic Model of Textual Complexity section). The use of well-known core psycholinguistic effects contributes to the validity of the manipulation of novelty-complexity and comprehensibility in Study 2 (see Study 2: Influencing Interest section).

Information Systems

This section consists of two parts. First, a short overview is given on the notion of relevance and its implementation in both IR and IF&R systems. Second, the performance of both types of systems is reviewed with a focus on methodology related to the complexity of defining and measuring relevance.

Relevance. As posited by Belkin and Croft (1992), IF&R systems and IR systems form two sides of the same coin. Both deliver information that is relevant to the user. Even when relevance is not explicitly implemented through algorithmic relevance, it operates as an "invisible hand" (Saracevic, 2007, p. 1916) underlying a wide range of information systems. This is described next for IR and IF&R systems while highlighting the importance of other factors often not covered by algorithmic relevance.

IR systems retrieve information that is relevant to a query, information need, and task (Saracevic, 2007). The contemporary implementation of relevance in IR systems compares

[2] *Coherence* refers to how propositions are connected in a reader's mental representation (a psychological construct), whereas *cohesion* refers to the connectedness of propositions to one another within in a text (a textual construct).

the topics expressed in a query with the topics expressed in documents to determine the degree of similarity, that is, the degree of topicality. This implementation is based on the assumption that topicality is objective or static and excludes other relevance criteria. Nonetheless, subjective relevance, which encompasses the relevance criteria that users apply during information interaction, includes relevance criteria other than topicality that also play a role during an information search session. For example, Barry and Schamber (1998) derived a set of criteria from two studies: depth or scope, validity, clarity, currency, tangibility, quality, accessibility, availability, verification, and affectiveness. Xu and Chen (2006) showed that numerous criteria can be reduced to a core set of five that indicate relevance: topicality, novelty, reliability, understandability, and scope. These studies indicate that aspects of novelty, complexity, and comprehensibility are part of the subjective relevance criteria users apply. For IR tasks, these criteria likely follow after topicality has been satisfied (Spink & Greisdorf, 2001) and are usually not reflected by algorithmic relevance. The value of novelty and complexity has been explored in the context of IR systems. For example, the Text RE-trieval Conference (TREC) novelty track defined the challenge of finding information that is both relevant and novel. However, here novelty refers to returning a diverse set of information with respect to a query instead of novel information with respect to a user (Carbonell & Goldstein, 1998; Soboroff & Harman, 2005). Furthermore, Collins-Thompson et al. (2011) showed the value of a readability metric in predicting the dwell time (indicative of relevance) of a user for a search result. The difference between snippet and page readability explained 69% of dwell times less than 120 seconds, suggesting the importance of complexity for IR.

Herlocker, Konstan, Terveen, and Riedl (2004) stated that the goal of an IF&R system is to “recommend items based on the likelihood that they will meet a specific user’s taste or interest” (p. 23). This suggests that the value of IF&R is to foster the emotion of interest. Yet, current IF&R systems do not directly focus on the experience but rather on the long-term interests of a user. Three types of filtering methods can be distinguished: content based, collaborative, and property based (Hanani, Shapira, & Shoval, 2001). These techniques can be summarized as selecting information based on an area of interest based on what similar users found interesting and based on a model of the knowledge or interests of the user. This can be interpreted as the topicality of information relative to the long-term interests of the user and is comparable with the interpretation of objective relevance for IR systems. The focus on topicality leads to the so-called serendipity problem, where IF&R systems “produce recommendations with a limited degree of novelty” (Ricci et al., 2009, p. 79). Methods that diversify the recommendations have been proposed to counter the limited degree of novelty (Konstan & Riedl, 2012). The serendipity problem signifies the difference between long-term interests and the emotion of interest, and shows the value of other aspects of relevance such as novelty in creating the IX of interest.

Performance and evaluation. Algorithmic relevance is generally evaluated via objective relevance, that is, an objective relation between a query (or user model) and a document. However, as we show next, a complicated relation exists between objective relevance, subjective relevance, and the IX, which creates a need for novel evaluative approaches.

For IR, the ground truth that is used to evaluate algorithmic relevance is commonly generated by information experts who assess whether a document is relevant to a query (i.e., the Cranfield paradigm; Voorhees, 2002). The validity of this ground truth is testified to by a positive relation between objective relevance and satisfaction (Gluck, 1996; Huffman & Hochster, 2007), although it is unclear exactly what indicators give an optimal prediction of user satisfaction (Su, 1994), something partly shown by the continuous search for new relevance indicators (Borlund & Ingwersen, 1998; Demartini & Mizzaro, 2006). Various different evaluative approaches have been suggested that include the user in the evaluative model. The interactive IR evaluation model (Borlund, 2003) aims for ecological validity by mimicking real tasks and, accordingly, creating real information needs in actual users. Inherent to the model is the use of alternative performance indicators that compare objective relevance with subjective relevance and, possibly, the user experience (Borlund & Ingwersen, 1998).

The classic approach to evaluating the performance of IF&R systems is by their accuracy. Consider a system containing (e.g., movie) ratings by a user, then the accuracy is the ability of the system to predict (withheld) ratings. A “magic barrier” in performance seems to have been reached where many algorithms do not cross a mean absolute error of 0.73, presumably because of natural variability (Herlocker et al., 2004). This idea has been supported by the finding that people provide inconsistent ratings for the same item when asked at different times (Hill, Stead, Rosenstein, & Furnas, 1995). The magic barrier of performance as measured in accuracy or error rate highlights the need for other evaluation methods and metrics, more predictive of user satisfaction (Konstan & Riedl, 2012). Novelty, as opposed to accuracy, and even serendipity have been proposed as important evaluative criteria for IF&R systems (Herlocker et al., 2004).

The relation between objective relevance, subjective relevance, and possibly the IX is unclear (Borlund, 2003). Similarly, the performance on the core task of an IF&R system, as measured by accuracy, provides little insight into the resulting IX (or UX) of a recommender system (Konstan & Riedl, 2012). Knijnenburg, Willemsen, Gantner, Soncu, and Newell (2012) presented a user-centric evaluation framework for IF&R systems aimed at their UX. It consists of three sets of variables: objective system aspects (i.e., algorithmic elements), subjective system aspects (i.e., appraisals of the objective system aspects), and experiential aspects (i.e., system, process, and outcome related). This framework overlaps with the perspective outlined by the interactive IR evaluation model (Borlund, 2003) in that it differentiates between objective relevance (i.e., objective system aspects) and subjective relevance (i.e., subjective system aspects). The approach taken in this article fits in the evaluative framework for IF&R systems (Knijnenburg et al.): Textual complexity is evaluated as an objective system variable in Study 1 (see Study 1: Generic Model of Textual Complexity section), as well as by its subjective counterparts appraised complexity and appraised comprehensibility in Study 2 (see Study 2: Influencing Interest section). As the remainder of this article shows, such a user-centric evaluation allows us to evaluate different aspects of relevance, such as textual complexity, and include (part of) the IX of the user.

Study 1: Generic Model of Textual Complexity

Introduction

The facets of processing difficulty described in the Textual Complexity section can be translated to metrics, allowing for an information system to include an indication of processing difficulty in its judgments. So-called readability metrics have been devised to make inferences at each level of processing difficulty: word, sentence, and discourse. As early as 1923, Lively and Pressey introduced the first objective readability formula. Such traditional formulas use basic word length as indication of word-level effects (e.g., decoding and lexical access) and sentence length as indicator of sentence-level processing difficulty (e.g., syntactic complexity). Examples are the Flesch Reading Ease Scale (Flesch, 1948), ranging from 0 to 100 and based on the words per sentence and the syllables per word; the Flesch-Kincaid Readability formula (Kincaid et al., 1975), indicating the reading grade level; and the New Dale-Chall Readability Formula, based on the words per sentence and number of unfamiliar words (Fry, 2002).

Modern approaches to readability analysis define more accurate indicators, using next to the traditional formulas two types of state-of-the-art techniques: language models and deep features. A language model gives the probability that a piece of text is written using a certain language (model). By its definition it is a model of language rather than a model of readability. A language model can be applied as a model of readability when, for a certain group of users, a representative data set is available, that is, representative of the language used at a certain level of complexity. Language models often achieve good results. For example, Collins-Thompson and Callan (2005) showed it is possible to distinguish 9 (of 12) grade levels with a maximum error of 2 grade levels. Applied to general text classification tasks, language models can achieve very high classification performance. For example, on the 20 Newsgroups data set, Peng and Schuurmans (2003) achieve 89.08% performance in distinguishing 20 categories using language models as input to a naive Bayes classifier.

Contrary to shallow features, which are simple proxies of textual complexity, deep features are reflective of psychological processes (Vor der Brück et al., 2008). Few systems apply deep features to detect (aspects of) textual complexity. Coh-Metrix is a well-known system, applying cognitively inspired indexes including deep features to indicate the cohesion of a text (Graesser, McNamara, Louwerse, & Cai, 2004). Combined with a few shallow (traditional) measures, these 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, Louwerse, McCarthy, & Graesser, 2010). Applied to textual complexity research, Crossley, Greenfield, and McNamara (2008) showed how a model based on three criteria of the Coh-Metrix system correlates highly ($r = .925$) with Cloze test results.[3] Another more recent system is DeLite (Vor der Brück et al., 2008). It uses deep features to heighten its (construct) validity. Compared with 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 accounted for only 28% of the variance among ratings.

The most successful approaches combine traditional methods and state-of-the-art techniques to predict readability. For example, Feng, Jansche, Huenerfauth, and Elhadad (2010), using the power of language models together with a wide range of criteria, achieved a classification accuracy of 74% against 37.8% baseline accuracy. However, as explained next, a lack of construct validity and possible overfitting make it hard to compare and interpret the reported performance by state-of-the-art systems. Moreover, although a pivotal requirement, the predictive validity of the systems generally remains untested (Benjamin, 2012).

Predictive validity refers to the capability of measures of textual complexity to actually predict subjectively experienced complexity (e.g., appraised complexity) (Cronbach & Meehl, 1955). This was already stressed for the evaluation of IF&R and IR systems (see Performance and Evaluation subsection): Objective system aspects should reflect the subjective perception of these aspects. To assure the potential for predictive validity, this study develops a generic model of textual complexity not susceptible to overfitting and ensuring construct validity. However, the actual test of the predictive validity of the model is done in Study 2 (see Study 2: Influencing Interest section), applying the model to a different data set and comparing its predictions with ratings of appraised complexity.

Overfitting is caused by either a lack of separation or a lack of difference between training and test sets, leading to unexpected results when a model is applied to different data sets. Two guidelines are adhered to in this study to prevent overfitting: independency of semantics and independency of text length. First, semantic independence allows the applicability to a broad range of data sets. Although a

[3] In a Cloze test, the x^{th} word is left out for test subjects to fill in.

dependency on a syntactic level remains, restricting the applicability to the English language, it can be achieved on a semantic level by excluding any features that (also) model the meaning of words (e.g., language models). Second, length independence is required for a generic model of processing difficulty. Although sometimes implemented in models of readability, the amount of information is different from processing difficulty. For example, an expository text will be longer yet easier to read than an encyclopedic text, yet the former likely contains more information, both seen as text length and the width of concepts. Accordingly, this study approaches processing difficulty as the moment-by-moment effort needed to read part of a text, irrespective of the total length of a text.

Construct validity is the degree to which a metric actually measures its associated construct (Cronbach & Meehl, 1955). For a measure of textual complexity, construct validity is concretized as the degree to which the methods are reflective of (subsets of) experienced processing difficulty, that is, deep features. A lack of construct validity raises doubts to what actually is classified. Hence readability metrics should reflect user-centered facets of processing difficulty. Accordingly, rather than aiming for classification performance, this study develops a generic model based on a user-centered notion of processing difficulty. The psycholinguistic facets of processing difficulty as described in the Textual Complexity section form the basis for this user-centered model.

Study 1 develops and evaluates a classifier system predictive of textual complexity, borrowing from the techniques used by similar systems, yet incorporating the theoretical background on textual complexity described in the Textual Complexity section. The model is largely semantically and length independent, making it a generic model of textual complexity. The Features section describes the core of the system: its features. The classifier is trained and evaluated on a large data set distinctive in textual complexity, described in the Method section. The resulting model is given in the Results section. The resulting classifier system is used in Study 2 (see Study 2: Influencing Interest section) to manipulate interest and its appraisals by filtering texts of different complexity, verifying its applicability to different data sets, and testing its predictive validity.

Features

Four approaches to compute textual complexity are briefly addressed: traditional, lexical familiarity, priming, and dependency-locality. These approaches are particularly useful for large-scale application because of their low computational complexity, as they either use simple word-based representational models or highly optimized parsers. Moreover, they cover word-level (both isolated and inter-word), sentence-level, and (indirectly) discourse-level effects. This section describes features for each of the approaches.

Because the goal is to give an indication of the textual complexity of a text, those features defined at a smaller granularity, such as at a sentence or word level, are aggregated by deriving their statistical mean.

Traditional. Many readability formulas have been devised, all relating a common set of variables to some practical, dependent variable such as a reading grade level (DuBay, 2007). Instead of reviewing all the different formulas, only the common parameters are added that are shared by most of these traditional formulas. The exact, usually linear, model connecting the parameters to the dependent variable are left over to the classification process described in the Classification section. Not all parameters used as traditional criteria can be captured using algorithms, and many parameters are subsets of a few basic ones. Hence only the basic parameters are reported.

First, an indication of word length is used as indication for the semantic difficulty of a word (Fry, 2002). Word length can be defined in characters and syllables.

Feature 1: word length in characters c per word w, $|c \in w|$.
Feature 2: word length in syllables s per word w, $|s \in w|$.

Second, an indication of the commonness of words is used, indicative of word frequency. A popular one is the Dale list of 3,000 common words (Chall & Dale, 1995).

Feature 3: frequency of words on the Dale list of common words D in a text T,

$$\frac{|\{w \in T | w \in D\}|}{|w \in T|}.$$

Third, the length of a sentence is related to syntactic difficulty (Fry, 2002). Length can also be defined in numerous units. Being the most common one, only words per sentence is defined.

Feature 4: sentence length in words w per sentence S, $|w \in S|$.

Lexical familiarity. Lexical familiarity indicates how familiar a reader is with a word. The most salient measure of lexical familiarity is printed word frequency. It influences a reader's fixation duration, where more frequent words take less initial processing time (Inhoff & Rayner, 1986). This effect is even observed when controlling for word length, number of syllables, and bigram and trigram frequency (i.e., the frequency of the sequence of a word and its, respectively, two or three neighboring words). Also, high-frequency words are more likely to be skipped than less frequent words (Reichle, Pollatsek, Fisher, & Rayner, 1998). Indicated in the Word-Level Effects section, word frequency is highly correlated with other key variables of influence on word recognition such as word length, familiarity, and age of acquisition, confirming the robust role of printed word frequency. Hence printed word frequency is defined as a proxy for how familiar a reader likely is with a word, that is, as a measure of lexical familiarity:

Feature 5: a logarithm of the word count c in a representative collection of writing, log c.

A representative corpus of writing is to be used for the frequency counts. In this study, the Google Books N-Gram corpus (Michel et al., 2011) is used (see Feature Extraction section). 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 metric that is common in IR, word frequency metrics are common in psycholinguistics and studied extensively in relation to processing difficulty. Hence the word frequency feature is preferred to ensure construct validity.

Priming. Numerous studies of priming have shown that a target string is better identified when it shares letters with the prime. This holds for identity priming (repeating the prime), as well as form priming (using a partly different string). Although more vulnerable when extrapolated to a sentential or discourse context, the lexical repetition effects remain. This is confirmed by eye-tracking studies, where, within a meaningful context, word repetition decreases early eye fixation measures indicative of lexical access (Ledoux, Camblin, Swaab, & Gordon, 2006).

From an information theoretic point of view, repetition creates a form of redundancy that can be measured in terms of entropy. The information rate of a channel is given by its entropy in bits per symbol. 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. Consider a variable χ with a probability distribution $p(x)$ for every value x in χ. Then the entropy is defined as (Shannon, 1948):

$$H(\chi) = -\sum_{x\in\chi} p(x)^2 \log p(x). \tag{1}$$

For longer sequences, entropy can be defined as well. If we define a range of variables $\chi_1 \ldots \chi_n$ containing the joint probabilities $p(x_1 \ldots x_n)$ of a sequence $x_1 \ldots x_n$, then the joint (or n-gram) entropy is given by (Cover & Thomas, 2006):

$$\begin{aligned} &H(\chi_1, \ldots, \chi_n) \\ &\quad = -\sum_{x_1\in\chi_1} \cdots \sum_{x_n\in\chi_n} p(x_1, \ldots, x_n)^2 \log p(x_1, \ldots, x_n). \end{aligned} \tag{2}$$

The variables $\chi_1, \ldots, \chi_n$ are derived directly from the occurrences of (joint) values in a text T. However, basing these variables on a whole text T gives a measure of the amount of information in T. Because the interest is in writing style rather than text size, a sliding window entropy (SWE) will be calculated with a window size w over a text length N:

$$\mathrm{H_w}(T) = \sum_{i=w}^{N} \frac{H(\chi(x_{i-w+1}, \ldots, x_i))}{N - w}. \tag{3}$$

Here, χ derives the joint distributions $\chi_1, \ldots, \chi_n$ from the sequence of symbols $x_{i-w}, \ldots, x_i$.

The SWE has several benefits over other size-corrected measures. First, when correcting for the size by calculating the entropy ratio, the ratio between $H(\chi)$ and the entropy of a uniform distribution, the influence of text size on the distribution is still profound: Longer samples have an inherently different distribution compared with shorter ones. Second, psycholinguistic effects of priming are vulnerable to distance: Farther away primes are less effective (Ledoux et al., 2006), making SWE a measure more in accordance with observations on psycholinguistic priming. Third, SWE can be interpreted as lexical chains indicative of cohesion (Morris & Hirst, 1991), where repeated mentions of discourse referents help readers integrate new information (Kintsch & van Dijk, 1978).

SWE gives a size invariant information rate measure, or in other words, an information density measure. Text with a higher repetition of symbols have a lower entropy rate. Using Equation 3, two features are defined using either characters or words as symbols:

Feature 6^n: character-n-gram SWE of a text.
Feature 7^n: word-n-gram SWE of a text.

Dependency-locality. The 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 is 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 that requires resources (i.e., integrations 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, Vasishth, & Dyke, 2006). The DLT is a particularly interesting theory because it explains processing cost of normal sentences, contrary to syntactically ambiguous ones, and its computation is fast and accurate, using state-of-the-art part-of-speech (POS) taggers and dependency resolvers (Cer, de Marneffe, Jurafsky, & Manning, 2010).

For normal sentences, the costs of the integrations 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 are not significant. When normal texts are used, such excessive storage requirements are rare. Hence the focus is on integration costs alone.

Integration costs are dependent on two factors: (a) the type of the element to be integrated, where new discourse elements require more resources than established ones, and (b) the distance between the to-be-integrated head and its referent, where distance is measured by the number of intervening discourse elements (Gibson, 2000). This is approximated by defining a (new) discourse referent as a noun, proper noun, or verb (phrase).

Consider a dependency d connecting nodes a and b. Let Y_d be the collection containing each POS tag y for the terminal nodes between and including nodes a and b, then the dependency length of dependency d is given by:

$$L_{\mathrm{DLT}}(d) = |\{y \in Y_d | y \in \{\text{noun, proper noun, verb}\}\}|. \quad (4)$$

For a whole sentence, this gives the following feature of sentence complexity:

Feature 8: integrations costs $I(D)$ of a sentence containing dependencies D,

$$I(D) = \sum_{d \in D} L_{\mathrm{DLT}}(d). \quad (5)$$

Method

To indicate the power of the proposed features, we compared two data sets that are overall similar but highly distinctive in their expected processing difficulty: Simple English Wikipedia and (normal) English Wikipedia.

Data set. To evaluate the proposed metrics, we needed a data set with a clear diversity in expected processing difficulty. One relatively large data set perfectly suited for this is the Wikipedia encyclopedia. Wikipedia is available in many languages, among which are 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, partly deliberately, less depth in which a topic is discussed.

The Wikipedia dump of August 3, 2011, was used. Only articles that were found in both languages were selected, allowing for a pairwise comparison. This gave a data set of 69,395 pairs of English and simple English articles, a total of 138,790 articles, 398,718 sections, and 1,459,370 paragraphs. Of this data set, only articles that were neither a stub (i.e., an incomplete article) nor a special, redirect, or disambiguation page were selected. Moreover, only the oldest 10,000 articles per language, a total of 20,000 articles, were used for classification purposes. The underlying assumption being that more matured articles better reflect the actual intention of both Wikipedia versions.

The following preprocessing steps were performed on the data set: First, the original data consisted of two dumps, containing all articles encoded as wikitext for each language. Both sets were imported into an 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.

Feature extraction. For all features, the Stanford CoreNLP word and sentence tokenizers were used (cf. Toutanova et al., 2003). To measure the number of syllables per word, we applied the Fathom toolkit (Ryan, 2012). As model representative for common English, the Google Books N-Gram corpus was used (Michel et al., 2011). For each word, the 1-gram frequencies were summed over the years starting from the year 2000. For dependency parsing, the Stanford Parser was used (Klein & Manning, 2003), a state-of-the-art dependency parser (Cer et al., 2010).

Entropy was based on n-grams of length $n = 1 \ldots 5$ and windows of size $w = 100$ (see the Priming subsection). The window size was based on a trade-off between, on the one hand, minimal required text length (in this case, 100 symbols) and 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, reducing each word to its root form. The stemming was used to reduce simple syntactical variance and, hence, give more significance to the semantic meaning of a word. Stemming was performed using the Snowball stemmer (Porter, 2001).

Classification. The classification pipeline consisted of three steps: preprocessing, classification, and validation. As regards preprocessing, first, variables containing more than 25% missing values were removed. Second, observations containing any missing value were removed. Because of the relatively few features, no further feature selection was performed during the preprocessing step.

As classifier a random forest was chosen, showing the best results in comparison with a support vector machine, neural network, and nearest neighbor classifier. This is in line with benchmark studies, showing support vector machines are among the best but are often outperformed by other techniques such as a random forest (e.g., Meyer, Leisch, & Hornik, 2003). A random forest is a bagging technique, building many decision trees based on random selections of features (Breiman, 2001). As such it is a feature selection technique as well, making separate feature selection unnecessary. 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. The ranges 1 to ${}^2\log k + 1$ (k the number of features) and 1 to 10 were used, respectively. The number of trees was set to 100, found to be an optimal amount (cf. Meyer et al., 2003).

The classifier was trained on 80% and tested on 20% of the data set to validate the classification performance. The data were balanced to assure it contained an equal number of

TABLE 1. Power, model significance, and correlations of the features.

			Interfeature correlations (*r*)														
Feature*	r_{pb}	MDA	1	2	3	4	5	6^1	6^2	6^3	6^4	6^5	7^1	7^2	7^3	7^4	7^5
1	.470	14.73															
2	.484	14.20	.913														
3	.271	15.71	.140	.106													
4	−.474	15.75	−.565	−.637	−.288												
5	−.191	14.31	−.305	−.285	−.213	.669											
6^1	.096	13.71	−.010	−.054	.201	−.361	−.470										
6^2	.272	14.04	.186	.177	.072	−.209	−.244	.609									
6^3	.265	13.74	.165	.183	.001	−.080	−.100	.319	.896								
6^4	.213	13.47	.069	.091	−.023	.018	−.018	.221	.800	.966							
6^5	.175	14.39	.013	.038	−.027	.062	.021	.179	.745	.927	.987						
7^1	.264	14.07	.326	.280	−.009	−.194	−.303	.296	.726	.755	.685	.642					
7^2	.175	14.56	.210	.197	−.062	−.024	−.093	.141	.625	.742	.750	.753	.817				
7^3	.084	14.97	.105	.112	−.062	.043	−.018	.067	.473	.605	.650	.685	.611	.899			
7^4	.039	14.65	.052	.068	−.059	.066	.024	.025	.355	.479	.533	.579	.460	.758	.942		
7^5	.021	15.53	.030	.044	−.034	.074	.046	.017	.281	.390	.445	.492	.355	.639	.847	.962	
8	.295	15.65	.263	.233	.254	−.349	−.373	.121	.080	.038	.009	−.004	.179	.061	.012	−.005	−.015

Note. *Features: 1: characters per word; 2: syllables per word; 3: words per sentence; 4: Dale frequency of common words; 5: logartihmic word frequency; 6^n: *n*-character-based SWE; 7^n: *n*-word-based SWE; and 8: dependency-locality.
MDA = mean decrease in accuracy (%); r_{pb} = point-biserial correlation.

simple and normal articles. The 80%-20% split was chosen to reduce computational load, and will likely not have affected the performance significantly because of the relatively large size of the data set.

All steps were implemented using R, a statistical package (Ihaka & 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 analyzing the results.

Analyses. The focus of the analyses is on the power of the individual features, their contribution to the model, and the performance of the model. The statistical techniques used are briefly described for each.

The point-biserial correlation, the correlation between a continuous and dichotomous variable, is an indicator of the effect size of individual features. As guideline for its interpretation, Cohen's rule of thumb for effect sizes is used. Translated to the point-biserial correlation, the interpretation is: $.100 \le r_{pb} < .243$ small, $.243 \le r_{pb} < .371$ medium, $r_{pb} \ge .371$ large (Rice & Harris, 2005).

The mean decrease in accuracy is reported to evaluate the contribution of each of the features to the resulting model of textual complexity. This indicator gives an indication of how much the performance (as measured in accuracy) decreases when a feature is removed from the model (Breiman, 2001).

A multitude of possibilities exists to evaluate the classification performance, of which four are 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 with 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-around measure, giving the probability that the classifier scores 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, when all measures give similar results, accuracy can be used in line with common practice.

Results

Study 1 develops a system differentiating on textual complexity. To this end, a random forest classifier was trained on 10,336 articles and tested on 2,584 articles, balanced to consist of an equal number of simple and normal articles. The resulting model consisted of 17 features and achieved a classification accuracy of 90.87%. Several tests confirm the classification accuracy: The AUC was .967, the F1-score was .908, and the Phi correlation or Matthew's correlation was .817. Note that the range of the Phi measures lies between −1 and 1.

Table 1 shows the features, their intercorrelations, their isolated effect size as calculated through the point-biserial correlation, and their importance within the model as indicated by the mean decrease in accuracy.

The effect size of the individual features, given in Table 1, indicates word length (Features 1 and 2) and the Dale list of frequent words (Feature 4) are strong indicators. All other features have either medium or small (word frequency, Feature 5) effect sizes. For the SWE features, there is a clear optimum in effect size for a bigram ($n = 2$) representation for characters (Feature 6^2) and a unigram ($n = 1$) representation for words (Feature 7^1; see Equation 2). Except for denoting the strength, the point-biserial correlation also gives the direction of the effect, confirming all correlations behaved as expected.

The mean decrease in accuracy, given in Table 1, shows all features had a unique and significant contribution to the model. In particular, 5-gram word SWE (Feature 7^5) and words per sentence (Feature 3) had a bigger contribution to the model than what could be expected from their individual effect sizes. Moreover, all SWE features (Features 6 and 7) had a significant contribution. Being all highly related to each other, the unique contribution as indicated by the decrease in mean accuracy gives an underestimation in comparison with their individual effect sizes.

The interfeature correlations in Table 1 further confirm the uniqueness of the features. Particularly noteworthy are the following findings. First, the dependency-locality feature (Feature 8) correlates somewhat with sentence length (Feature 3, $r = .254$), indicating it indeed captures a unique aspect of syntactic complexity. Second, as expected (see the Textual Complexity section), word frequency (Feature 5) correlates reasonably with many other features. Third, all SWE features (Features 6 and 7) indeed correlate highly with each other, in particular when using the same representation (words or characters) and when using similar length n-grams.

The results indicate the features behaved according to theoretical expectations, confirming both the validity of the features and of the data set used to train the model.

Discussion

Gounded in psycholinguistic findings on processing difficulty, several features of textual complexity were introduced (see Features section). All features were of low computational complexity, using either word-based or fast and robust natural language processing representations. These features were tested on data sets that are distinctive in complexity, namely, Simple English Wikipedia and English Wikipedia. The evaluation showed a high accuracy performance, which was confirmed by numerous robust metrics of accuracy such as the F1-score and Matthew's correlation. The procedure and results are discussed in relation to three aspects: implementation, processing difficulty, and IX.

With using only 17 features, the prediction accuracy of 90.87% can be regarded an excellent performance (see the Introduction); in particular, when considering the 17 features were based on just eight unique features derived from only four underlying theories: priming, frequency, dependency-locality, and traditional. Undoubtedly, when more features are added, the prediction performance will rise. The Introduction already indicated a plethora of features gives optimal performance; here, we showed a carefully selected subset of deep features can already lead to a good performance.

Two particular characteristics of the model presented in this study are its length and semantic independence. All features were carefully devised to exclude text length as much as possible. This claim is supported in Study 2, where the model is applied to a different data set containing truncated articles (see the Data Set subsection). Semantic independence was achieved with two limitations: A list of word frequencies (see Lexical Familiarity subsection) and a POS and dependency parser were used. Because the scope of these dependencies is very broad, this merely limits the optimal level of performance to most of contemporary English language. The length and semantic independence ensure that the model is applicable to a broad range of texts, both of different sizes and meaning. In turn, this contributes to the validity of the model as a model of processing difficulty, rather than a model of language or length. These two characteristics ensure that the resulting model is a generic model of textual complexity not suffering from overfitting, allowing for its application in Study 2.

The evaluation showed that all features had a significant contribution to the resulting model of textual complexity. All metrics could indeed very clearly differentiate between different levels of complexity and, thus, formed a valid and reliable way to infer the processing difficulty of a text. Furthermore, the tests showed that the metrics measure different properties of complexity. This is a clear indication that part of the outlined facets of processing difficulty are indeed reflected in the features.

The features presented in the Features section cover all three psycholinguistic processing levels. Word-level effects are described by features indicative of lexical familiarity (isolated word effects) and priming (interword effects), covering the core effects at this level. In particular, word frequency (see Lexical familiarity subsection) is highly related to other isolated word effects. Moreover, although the features for priming effects are mainly orthographic, the addition of a stemming algorithm to the implementation gives more significance to the meaning of the (stemmed) word as well (see Feature Extraction section). At a sentence level, the dependency-locality metric clearly reflects processing difficulty, based on a contemporary theory particularly able to predict processing complexity of normal, common sentences.

Processing difficulty was covered less at the discourse level than at the word or sentence level because of the computational costs of a good metric for cohesion. However, often used metrics for discourse-level processing difficulty are based on a similar notion of orthographic repetition as the metrics devised for interword priming. Both share the idea of lexical chains connecting the discourse elements throughout a text (see Priming subsection). For example, the Coh-Metrix system defines stemmed (noun-)word overlap as a metric of coreference cohesion. This group of metrics was found to differentiate well between low- and high-cohesive texts (McNamara et al., 2010). Furthermore, Lapata and Barzilay (2005) define cohesion as sentence overlap, where overlap can be defined at an orthographic level as words or at a semantic level as concepts. Hence, although strictly speaking the interword priming effects are not measures of discourse-level processing difficulty, they are related to it, from a theoretical point of view (i.e., as lexical chains) and from a practical point of view (i.e., as common metrics of cohesion).

This study set out to create a generic model of textual complexity. By creating features based on key psycholinguistic findings, the objective system aspects are embedded in user-centered constructs. This coupling between objective and subjective constructs is requisite for the ability to affect the IX (see Performance and Evaluation subsection). The adherence to semantic and length independence allows to generalize to different data sets, differing from the encyclopedic style of Wikipedia. Furthermore, the use of features of low computational complexity allows for large-scale applications such as filtering or retrieval, where scalability is an issue. Hence the model is expected to generalize to the IX and different, large data sets.

The main goal of the model is not to achieve a high accuracy in classifying texts, it is to give an indication of the processing difficulty as part of the IX. Whether the model generalizes from classification performance to different data sets and whether the model actually influences the IX (i.e., its predictive validity) is tested in Study 2. In particular, Study 2 shows that the model can be used to select interesting information.

Study 2: Influencing Interest

Introduction

Interest is an important part of the IX, in particular for IF&R systems (see Performance and Evaluation subsection). As explained in the Interest subsection, interest is theorized to be dependent on two consecutive appraisals: novelty-complexity and comprehensibility. The model from Study 1 is applied to manipulate interest through its two consecutive subjective appraisals, as well as to explain why and when interest occurs. Study 2 completes the evaluation of the model created in Study 1. This study checks the model for overfitting and test its predictive validity by applying the model to a different data set and comparing its predictions with appraised complexity. In doing so, Study 2 combines an algorithmic approach to textual complexity with a user-centered approach to interest in explaining an emotional aspect of the IX. The resulting model of interest is referred to, accordingly, as the combined model of interest.

The relation between textual complexity and interest is expected to be positive as long as the secondary appraisal of comprehensibility stays positive. Numerous studies give support to a positive relation between comprehension and interest, for expository texts (Sadoski, 2001), news magazine articles (Schraw et al., 1995), fiction (Schraw, 1997), and science texts (Connelly, 2011). Less empirical support exists for a positive effect of complexity on interest. Schiefele (1996) is a notable exception, showing that encyclopedic texts that are below the reading grade level of a reader diminish interest. Indirect support for the relation between complexity and interest comes from studies on challenge and motivation. These studies show that a balance between skills and challenge leads to a feeling of competence and mastery important for the persistence in an activity (Csikszentmihalyi, 1991; Reeve, 1989).

Textual complexity is not the only variable of influence. The appraisal theory of interest points to two aspects in particular, which are highlighted next. First, key to the appraisal theory of emotion is that the appraisals are individual (Ellsworth & Scherer, 2003; Lazarus, 1991), pointing to one individual characteristic that is extra salient for the experience of interest: the comprehension ability. Second, next to complexity, novelty is part of the appraised complexity (Silvia, 2008b).

Much research on processing difficulty originated from studies on individual differences in comprehension ability. As indicated in the Textual Complexity section, different levels of processing put different requirements on the reader. At each level of processing, specific abilities can be identified that explain individual differences in processing difficulty: at word level, word-identification abilities and print exposure; at sentence level, working memory capacity; and at discourse level, background knowledge (Jaeger & Tily, 2011; Long et al., 2006). These individual differences are more profound for complex texts that put more requirements on the reader to be comprehended. The subjectivity of emotional appraisals is applicable to the influence of textual complexity as well: Although complexity is likely perceived as such, individual abilities interact with complexity in creating this perception.

For the appraised complexity, a second salient textual characteristic is the novelty or recency of information. This novelty is partly an appraisal by the reader and partly an intrinsic value of the information itself, better described as recency. Recency and novelty differ because recency reflects a publication date, whereas novelty is a subjective evaluation of the content. Although not a certainty, a recent document has a higher chance of being novel to the reader (Barry, 1994; Xu & Chen, 2006). Hence recent content is preferable to inspire an interest response.

Study 2 manipulates interest using the model of textual complexity resulting from Study 1. A popular news source is used as data source to have recent texts that are more likely to be appraised as novel; however, novelty is not the focus of this study. To oppose any effects of miscellaneous textual characteristics (e.g., surprisingness, incongruity, variability, and puzzlement; Berlyne, 1960), only one source is used. The relation between textual complexity, interest, and its appraisals is evaluated, resulting in a combined model of interest. The combined model allows us to reflect on the main hypothesis guiding this article, namely, that more complex stimuli are more interesting, if within the "sweet spot"—novel-complex yet comprehensible.

Method

Participants. A total of 30 participants (22 male, 8 female) with an average age of 28.60 ($SD = 6.06$) voluntarily took part in the experiment. None of the participants was a native English speaker, but all graded their reading literacy as high ($M = 4.63$, $SD = .62$, range 1–5, 5 highest). All participants were well-educated; they either had a university degree or were enrolled as a student at a university.

Data set. A collection of 14,856 articles from *The Guardian*[4] (a widely known newspaper published in the United Kingdom) was used as the data set. The data set consisted of articles from the following news feeds: culture; environment; financial, market and economics; commentary; life and style; and science and technology.

To reduce variation that originates from differences in article length, we truncated all articles after 1,200 characters. The cutoff point was placed before the end of the word at position 1,200, and three dots were added to indicate the story normally would continue. Any layout was stripped from the articles, leaving only the title and textual content.

Filtering. A selection was performed by applying the model of textual complexity from Study 1 on the data set of *The Guardian*. The resulting distribution is given in Figure 2. The distribution was derived from the predictions of the model using a kernel density estimation with a Gaussian kernel and a bandwidth of .1. The figure shows that the center of mass of the distribution for the truncated texts is at the lower end of the complexity spectrum. In comparison with 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 preselected. Then, a final selection consisting of 18 articles was performed based on suitability. For the final selection, the following criteria were applied. First, having a participant pool of international origin and being a news source of national origin, texts were to be of international orientation. Second, 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. Third, selected news items differed in topic to ensure a variation in topical familiarity would be existent. The textual complexity of the resulting selection of 18 articles, grouped by three levels of complexity (i.e., the experimental conditions), is depicted in Figure 3.

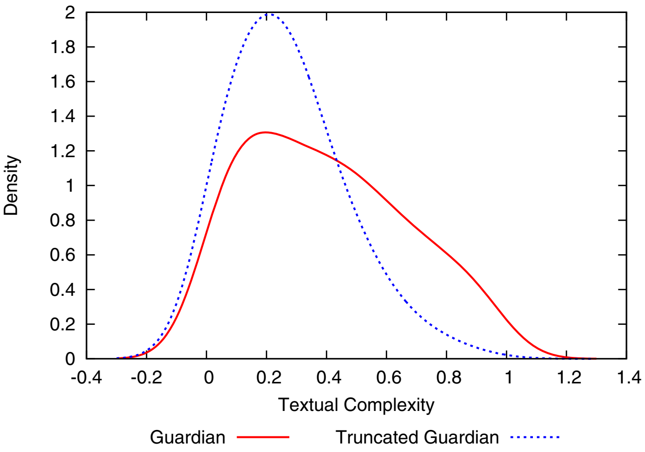

FIG. 2. Estimated density of textual complexity scores for *The Guardian* data sets (see Data Set subsection). The textual complexity scores range from 0 (low complexity) to 1 (high complexity). Because the estimated density was derived using a Gaussian Kernel with a bandwidth of .1, the limits of the graph extend this range. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Instruments. Four 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 and background questionnaire that addressed the following items: gender, age, nationality, educational background, prior knowledge, personality traits, English reading proficiency, and visual acuity. These items were included to control for their potential intervening influence.

Silvia (2008a) presents two scales, one for appraised complexity and one for appraised comprehensibility. Both are based on seven-point semantic-differential scales. The appraised complexity scale consisted of just one differential: complex-simple. To improve scale reliability, we added another item to measure appraised complexity: easy to read-difficult to read. Cronbach's alpha for this scale was at a good level of .82 ($N = 540$), 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 .89 ($N = 540$).

In accordance with related studies (e.g., Silvia, 2006, 2008a), interest was measured using two 7-point

[4] http://www.guardian.co.uk

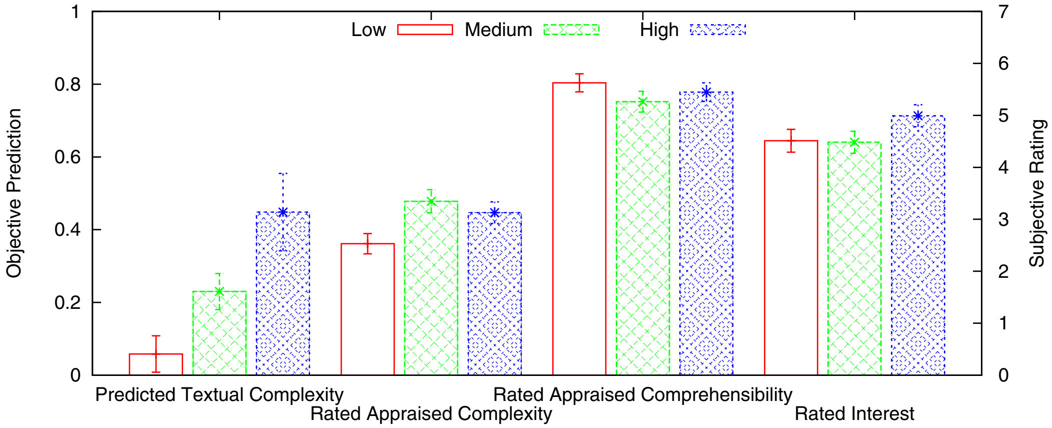

FIG. 3. Means and confidence intervals ($\alpha = .95$) for the predicted textual complexity and rated appraisals and interest for each of three conditions (i.e., articles of low, medium, and high complexity). [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

differentials: *interesting-uninteresting* and *boring-exciting* (Silvia, 2008a). Furthermore, a 7-point Likert scale was added to benefit from the shortened texts (see Data Set subsection), 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 .92 ($N = 540$).

Design and procedure. The experiment used a within-subject design with a randomized order of articles. The independent variable was textual complexity, grouped in three levels forming the experimental conditions. The dependent variables were appraised complexity, appraised comprehensibility, and interest.

The experiment consisted of two phases. The first started with instructions and a questionnaire on demographics and background. As instructions, the participants were told the experiment queried their interest in different news articles. Complexity was not mentioned as part of the experiment. The second phase of the experiment showed the articles, each followed by the novelty-complexity, comprehensibility, and interest scale. Before and between the phases a short explanatory text was shown. The reading of each of the articles was self-paced.

The full experiment had already lasted around 45 minutes. The participants indicated this required a lot of their concentration. Staying concentrated proved to be somewhat of a challenge, especially considering not all texts were experienced as interesting.

Analyses. The following techniques were used for the analyses: Pearson's correlation, multivariate analysis of variance (MANOVA), and structural equation modeling. The former two are straightforward in their application, the latter requires an elaboration on its parameters. Furthermore, based on the mean appraised complexity, one article was indicated as an outlier and excluded from further analysis. Using a series of *t* tests, we showed the complexity for this article was appraised significantly different as compared with the other articles ($t[17] = 2.10$, $p < .05$).

Structural equation modeling is a multiregression technique, solving multiple regression equations. It differs from normal regression in that dependent variables can be independent variables as well. For a structural equation model (SEM), the latent and observed variables need to be identified. The scales presented in the Instruments section were used as latent variables and their items as observed variables. Furthermore, the classifier output (see Filtering subsection) was used as an indicator for the latent variable of textual complexity. The SEM was developed with the R package Lavaan (Rosseel, 2012) and was based on a covariance matrix. The result is shown in Figure 4. It was modeled after standardization of the observed variables as well as the latent variables, causing the coefficients in Figure 4 to represent changes in SDs.

An SEM is valuable only 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 $X^2(X_r^2)$, standardized root mean square residual (SRMR), and comparative fit index (CFI). Each is described for the developed SEM (see Figure 4).

The X_r^2 is the inferential X^2 statistic divided by the degrees of freedom (*df*). 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 less than 5 (Schumacker & Lomax, 2010). With $X_r^2 = 9.42$, the developed SEM scores are above this maximum of 5. However, this can be fully explained by the sample size of 510 samples, which is confirmed by the other statistics.

SRMR is a robust badness-of-fit index comparing the model against the actual data. It is still influenced by sample size, where higher sample sizes give better results. The maximum value for SRMR is .090 (Iacobucci, 2010). With a

DOI: 10.1002/asi

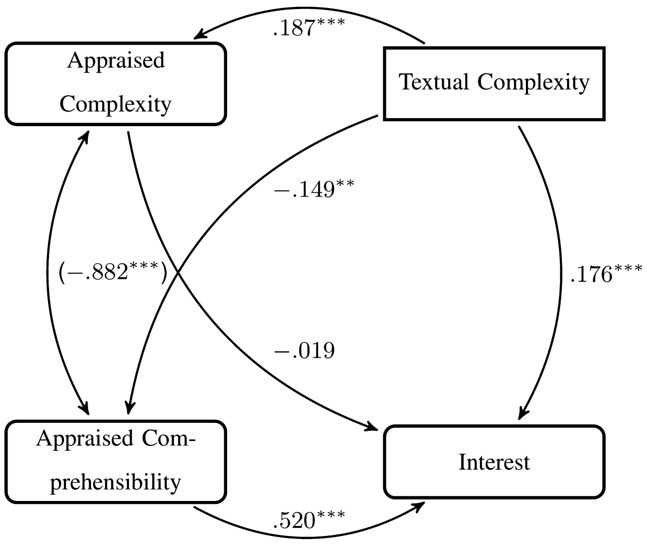

FIG. 4. Path diagram showing regression coefficients and covariances (between parentheses) of objective variables (squared boxes) and subjective variables (rounded boxes) that together explain interest responses ($R^2 = 29.10$), forming the combined model of interest. *$p < .05$, **$p < .01$, ***$p < .001$.

TABLE 2. MANOVAs of within-subject contrasts for condition (sets of articles of high, medium, or low complexity) and stimulus with η^2 denoting their effect sizes.

		df			
Variable	*F*	Model	Error	*p*	η^2
Appraised complexity					
Condition	23.590	1	29	<.001	.449
Stimulus	36.768	1	29	<.001	.559
Condition * stimulus	2.117	1	29	n.s.	.068
Appraised comprehensibility					
Condition	6.500	1	29	<.05	.183
Stimulus	9.184	1	29	<.01	.241
Condition * stimulus	7.559	1	29	<.05	.207
Interest					
Condition	15.912	1	29	<.001	.354
Stimulus	11.112	1	29	<.01	.277
Condition * stimulus	0.095	1	29	n.s.	.003

Note. n.s. = not significant.

score of SRMR = .061, the SEM corresponds to the requirement of SRMR ≤ .090.

The CFI is a goodness-of-fit index, comparing the hypothetical model with a simpler model (without any defined paths). It is the most robust metric of the three, adjusting for model parsimony and relatively invariance of sample size (in particular for $N \geq 200$). The consensus for this index is that it should be "close to" .950 (Iacobucci, 2010). The developed SEM gave CFI = .941, corresponding to the consensus.

Although the X_r^2 index does not conform to the guideline, this value is mainly inflated because of the number of observations ($N = 510$) and the modest *df* ($df = 24$). A smaller sample size easily leads to a X_r^2 well within range (Iacobucci, 2010). Taken together, two of three indices showed a good fit and one index indicated a bad fit, although this was highly influenced by the within-subject design. These indices confirm the SEM shown in Figure 4 and justify further discussion in the Results section.

Results

The experiment consisted of three conditions: a set of six articles of low textual complexity, a set of five articles (after exclusion of an outlier) of medium textual complexity, and a set of six articles of high textual complexity. The means and confidence intervals for each condition are shown in Figure 3, suggesting a profound influence of the condition, where (objective) textual complexity increased the appraised complexity, decreased the appraised comprehensibility, and had a positive influence on the reported interest. However, this holds only for the difference between the small and the medium or high conditions of textual complexity. Figure 3 shows that the difference between the medium and high levels of textual complexity is not reflected in the subjective appraisals of complexity.

A within-subject (or repeated-measures) MANOVA was conducted to test for an effect of the conditions of textual complexity and the stimuli (i.e., the 17 articles) on appraised complexity, appraised comprehensibility, and interest. There was an overall significant effect of the conditions, Wilks' $\Lambda = .132$, $F(6,24) = 26.33$, $p < .001$, indicating that the manipulation by textual complexity was successful, with a very strong overall effect size ($\eta^2 = .868$). Follow-up MANOVA were used to make post hoc comparisons between the levels of textual complexity. Table 2 shows that the influence of the conditions is confirmed for all three dependent variables: appraised complexity, appraised comprehensibility, and interest. As expected, the strongest effect was found for appraised complexity (see Table 2).

In addition to an influence of the conditions, the stimuli also had a separate influence on the reported interest and on the appraisals of complexity and comprehensibility. This suggests that some articles are more interesting, complex, or understandable than others unrelated to the condition to which they belonged. An interaction effect was found for appraised comprehensibility, indicating that, between conditions, there was an influence of the used articles. The nonsignificant interaction effect between the conditions of textual complexity and interest indicates that, although some articles were experienced with higher interest, on average, these articles were evenly distributed over the conditions.

The relation between objective complexity and the subjective variables was further analyzed using Pearson's correlation. Table 3 shows the correlation between each of the measured constructs. The correlation between appraised complexity and textual complexity ($r = .442$) confirms the effectiveness of the model in influencing the appraised complexity. However, as signified by a correlation of only $r = .195$, its (direct) influence on interest is small. This

TABLE 3. Correlations between latent variables.

	Correlations		
Variables	1*	2	3
1. Textual complexity*			
2. Appraised complexity	.442		
3. Appraised comprehensibility	−.358	−.763	
4. Interest	.195	−.352	.485

Note. *Excluding within-stimulus variance.

indicates that textual complexity is dependent on the appraisals in creating a significant contribution to interest.

To further explicate the complicated relation between complexity, the appraisals, and interest, a SEM was developed (see Analyses section). The resulting combined model of interest is shown in Figure 4, together with the regression coefficients and covariances connecting each of the four latent variables as expected from theory. Furthermore, Table 3 shows the correlations between each of the latent variables.

Interest was explained by textual complexity ($\beta = .176$, $r = .195$) and appraised comprehensibility ($\beta = .520$, $r = .485$). Surprisingly, appraised complexity was not a significant determinant for reported interest. Objective complexity together with appraised comprehensibility seemingly captured the influential aspects of appraised complexity on interest. Appraised comprehensibility had the largest contribution to reported interest, as indicated by its coefficient within the combined model ($\beta = .520$) and its correlation ($r = .485$), which makes it the highest correlated variable to interest. The combined model of interest (see Figure 4) shows that textual complexity was a significant determinant for each latent variable, both for the appraisals and the reported interest. It had a positive (direct) effect as well as a negative (via appraised comprehensibility) effect on interest. The total explained variance of interest with this model is $R^2 = .291$. Together with the corresponding results depicted in Table 2 and Figure 3, this shows the relation between textual complexity, appraised complexity, appraised comprehensibility, and the experience of interest.

Discussion

In a study with 30 participants, interest was manipulated by an (objective) metric of textual complexity. The manipulation was successful for interest and its underlying appraisals. As such, textual complexity was shown to be important in approaching the “sweet spot” of interest—novel-complex yet comprehensible. In total, 29.10% of variance in interest could be explained by a combined model of interest consisting of indicators of textual complexity, appraised complexity, and appraised comprehensibility. This section interprets these results; its implications are considered in the General Discussion section.

The model of textual complexity developed in Study 1 has been compared with its subjective counterparts, in particular, appraised complexity, completing the evaluation of the model of textual complexity. The model was shown to have manipulative power, as well as predictive value for appraised complexity. With a correlation of .442, the predictive power of textual complexity on appraised complexity is shown. This confirms that the model of textual complexity indeed reflects appraised complexity, testifying to its predictive validity.

As a model for textual complexity, a classifier trained on a binary problem (simple vs. complex) was used. The resulting metric is, accordingly, a value ranging from 0 to 1 indicating easy or complex. Considering all features developed in Study 1 were directional and largely semantically independent, it is reasonable to use the resulting value as a continuous scale.

The correlation between textual complexity and appraised complexity ($r = .442$) is lower than often found in comparable studies measuring Cloze test results or subjective readability. However, as indicated in the Introduction, comparable studies often lack predictive validity, construct validity, or suffer from overfitting. An exception is the recent DeLite system, which achieved a prediction of $r = .53$ with difficulty ratings made by the participants (Vor der Brück et al., 2008). There still exists one important difference between this study and the DeLite system: Contrary to ratings of difficulty, this study queried appraised complexity, which has a clear notion of individuality (subjectiveness) and experience (time and situation dependency) in it.

Besides the obvious individuality of the appraisals contrasting the generic model of textual complexity, Figure 3 showed another methodologic factor influenced the effect size: The experiment consisted primarily of not highly complex articles. Combined with the high level of education of the participants, the complex articles were mainly experienced as relatively easy, reducing the effect size. Likely, this explains the lack of a difference in appraisals between the medium and high levels of complexity as well.

Interest was positively influenced by textual complexity. This was confirmed by the combined model, as well as by analyzing the influence of textual complexity as a discrete problem (comparing the conditions of low, medium, and high complexity). Moreover, as illustrated using the within-subject MANOVA shown in Table 2, there was close to no dependence between the conditions and the stimuli ($\eta^2 = .003$). This allows the effects of textual complexity on interest to be interpreted independently of a baseline “interestingness” of the articles, that is, articles leading to, on average, more intense interest responses. Interest cannot be explained by textual complexity alone, it is dependent on individual appraisals as well. This dependency is confirmed by the small direct correlation between textual complexity and interest, and the success of the combined model of interest in identifying the determinants of an interest response. The combined model of interest joined the objective indicator of textual complexity with the subjective appraisals to explain the experience of interest.

In the combined model of interest, the nonsignificant influence of appraised complexity on interest is a surprising result. Seemingly, (objective) textual complexity together with appraised comprehensibility cover the influence of appraised complexity. This is partly confirmed by the high correlation between appraised complexity and appraised comprehensibility ($r = -.763$). Its strong relation can be expected: Less ability heightens the perceived complexity. Furthermore, this is partly confirmed by the correlation ($r = .442$) between textual complexity and appraised complexity, which in itself was one of the aims of the generic model of textual complexity. Both strong relations to appraised complexity do not cover all its variance, indicating there is a part of appraised complexity that is not of influence on interest. Looking at the appraisal theory of emotion, this is not unlikely: Appraised complexity only contributes to interest as far as it contributes to the intrinsic pleasantness of the stimulus. Moreover, in its primary appraisal, it joins with goal relevance as well (Ellsworth & Scherer, 2003). This indicates appraised complexity only influences interest as far as it interacts with other aspects of the primary appraisal, leaving part of appraised complexity unrelated to explaining interest.

This study used 18 articles differing in textual complexity and, at the same time, in their semantic content. Partly, the variance in interest responses can be attributed to an interaction of individual differences with the semantic differences between the stimuli. For example, familiarity with an article and with an article's topics alone explains 15.90% of variance in interest responses (Van der Sluis et al., 2012). This illustrates that, given the current set of stimuli, variables delineating individual differences such as long-term interests and their influence on appraised novelty and appraised complexity can help demystify part of the baseline "interestingness." By adding more variables to the combined model of interest, a higher proportion of the variance in interest responses can be explained. Hence, although the combined model confirms the influence of textual complexity on interest, it also signifies the importance of other textual characteristics and related subjective appraisals for fully predicting the occurrence of an interest response.

General Discussion

The two described studies confirm the main hypothesis guiding this article. Namely, information systems can select information within the "sweet spot" of interest: complex yet comprehensible. Although this "sweet spot" consists of two subjective appraisals of a text, it can be approximated using an objective metric of textual complexity. Study 1 showed that textual complexity could be classified, relatively independent of semantics and text length, with a classification performance of 90.87%. Subsequently, Study 2 showed that this classifier could be used to predict and explain the experience of interest via the appraisal model of interest. It showed that a positive effect of complexity yet a negative effect of comprehensibility lead to an overall positive effect of textual complexity on interest. The combined model showed complexity (objective and subjective), and comprehensibility explained 29.10% of variance in interest responses. Given that a multitude of possible determinants of an interest response exist, the 29.10% of explained variance shows that complexity is an influential factor for the construction of an IX.

Textual Complexity

The model of textual complexity developed in Study 1 gave an excellent classification performance while using just a few features. Study 2 continued the evaluation of the model, showing a correlation of $r = .442$ between (objective) textual complexity and (subjective) appraised complexity. 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. Testing the model on its ability to predict subjective appraisals of complexity for a new data set goes beyond this approach. It ensures that 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.

Study 1, as well as Study 2, supplied evidence that the model of textual complexity is truly generic. First, in the development of the features of textual complexity there was a strong dependence on psycholinguistic theories of processing difficulty (see Features section). A clear reflection of core effects known to influence processing difficulty in the features assured a high construct validity. Second, the data-oriented evaluation in Study 1, where the model was trained and tested on a total of 12,920 articles, showed the model scales to large data sets. Third, Study 1 reduced the risk of overfitting by developing the model independently of text length and semantics, ensures that article length and genre are of little influence on the classification outcome. This allowed Study 2 to apply the model to a profoundly different data set (i.e., news articles) than the model generation data (i.e., encyclopedia articles) used in Study 1. Finally, the user evaluation in Study 2 showed a positive relation between textual complexity and perceived complexity, giving a unique indication of the predictive validity of the model, as well as a relation between textual complexity and experienced interest coherent with the appraisal theory of interest. These four findings confirm the genericity of the model of textual complexity: It generalizes to different sets of data, as well as to part of an IX.

Textual complexity is likely to relate to other characteristics of texts as well, for example, quality, depth, scope, clarity, surprisingness, incongruity, variability, or puzzlement (Barry, 1994; Berlyne, 1960). Because 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. Yet, further research is needed to deepen the understanding of the relation between the underlying textual characteristics, textual complexity, and interest.

Interest

Besides the key finding of an overall positive effect of textual complexity on interest, a closer look at the relation between textual complexity, the appraisals, and interest shows that complexity can both increase and decrease interest. This paradoxical finding confirms the main hypothesis: More complex stimuli are more interesting, if within the "sweet spot"—novel-complex yet comprehensible. It is in line with the appraisal theory of interest and with other theories on motivation as well. For example, the flow theory states that a balance between challenge (e.g., textual complexity) and skills (e.g., comprehensibility) leads to an optimal experience with a peak level of motivation (Csikszentmihalyi, 1991).

Hitherto, little empirical support for a positive effect of complexity on interest existed (see Introduction). Although shown for simple stimuli such as polygons (Silvia, 2005) and hinted on for textual stimuli (Schiefele, 1996), this effect was yet to be proved for textual complexity. This study confirms an important role for complexity in inducing interest with textual stimuli. Moreover, this was done using an objective model of textual complexity and in an ecologically valid context, as a set of real-world news articles was filtered by the model of textual complexity and presented as an information stream.

The overall positive effect of complexity on interest reported in this article is limited to the lower end of the complexity spectrum: None of the articles was evaluated as particularly complex. Given that complexity can both foster (via the primary appraisal) and diminish (via the secondary appraisal) an interest response, at the higher end of the spectrum an opposite effect of complexity can be expected. Furthermore, interest is dependent on more variables than manipulated and measured in Study 2; for example, novelty and goal relevance are known determinants (Ellsworth & Scherer, 2003; Silvia, 2006) whose influence can be imagined salient given that 18 articles were used as stimuli in Study 2. The former, novelty, has been noted as important for IF&R systems (Konstan & Riedl, 2012). The latter, goal relevance, is partly reflected by topicality in IR systems and is considered a precondition for any other relevances to become significant (Spink & Greisdorf, 2001). In essence, the experience of interest is individualized, indicating its predictability will inevitably rise with the addition of more variables reflective of individual differences. The effects of an information need, long-term interests, and knowledge were not included in this article and are, therefore, subjects for further investigation.

The combined model and the appraisal theory of interest highlight the difference between long-term interests, which are often implemented in IF&R systems, and the short-term emotion of interest. The former is captured by selecting familiar items, or "more of the same," possibly involving the diversification of the selection (e.g., the TREC novelty track; Soboroff & Harman, 2005). The latter is approached by selecting novel, complex, yet comprehensible items. The findings from the combined model suggest that, instead of being used to generate so-called filter bubbles, proxies of long-term interests or knowledge should be used together with the proposed indicator of textual complexity to predict the novelty and comprehensibility of information.

Information Systems

This article took an integrated approach to the design and evaluation of the experience of information systems: from a theory-supported algorithm, via related subjective constructs, to the specific experience of interest. It follows the evaluative framework proposed by Knijnenburg et al. (2012) in linking objective system aspects (i.e., textual complexity), via subjective system aspects (i.e., the appraisals), to experiential aspects (i.e., the emotion of interest). The combination of studies shows that information systems can be designed that use a generic model of textual complexity to select information likely laying in the "sweet spot" of interest.

The unique approach of testing a classifier not only on data but also on its effect on (part of) the indicates the effectiveness of the classifier on multiple levels. Yet, it also highlights the difference between the evaluative approaches. The existence of a magic barrier in classification performance for recommendation accuracy already hints at this difference (see Performance and Evaluation subsection). Because of the use of relatively few features, the magic barrier for the classification of textual complexity has likely not been reached by Study 1. Still, an important difference was found between textual complexity and appraised complexity, and a complex relation was found between textual complexity and interest. Within the context of interest, the classification performance is largely unrelated to the final IX. This indicates the profound difference between the two methods and the importance of a user-centered evaluation next to a data-oriented evaluation. In other words, algorithmic performance provides little insight into the IX.

Both the combined model and the model of textual complexity show how information systems can construct (part of) an IX. However, although interest is clearly part of a hypothetical optimal IX for IF&R systems in specific and information systems in general, no direct evidence is given for the effect of interest on the holistic IX. Most likely the relation between interest and the final IX is not linear: A hypothetical system with only interesting information may actually lead to a negative affect (Glassey & Azzopardi, 2011), for example, because of fatigue or desensitization. This highlights the problem of concretizing the complex concept of IX, already a subset of the UX, to an amendable goal. During the course of an information interaction, a smorgasbord of emotions concur and combine with goal-related (e.g., usefulness) aspects in forming the final, holistic IX. This article proposes that, through identifying key (positive) emotions during information interaction, information systems can be designed that target a specific well-defined emotion and, accordingly, can construct part of an IX. Notwithstanding, what constitutes the optimal IX, in particular in terms of emotions, remains an open question.

Conclusion

In this work, IX was introduced as a term to describe the positive or negative experience during interaction with information via an information system. Systems that provide the most relevant results or recommendations are not guaranteed to provide a positive IX in all cases. Instead, there is a larger set of factors to be considered, such as the emotional reactions of the users and the complexity of the information they encounter.

The emotion of interest was selected for its perceived relevance to the process of interacting with information, operationalizing the IX as an amendable goal. It was hypothesized that if items of news were not only novel, but also approaching optimal complexity (i.e., not too easy; not too difficult to comprehend), then they would generate a higher level of interest. A model of textual complexity was developed and integrated into a combined model of interest to confirm this hypothesis. The model of textual complexity was developed in line with core psycholinguistic findings on the causes of processing difficulty and validated on its classification accuracy and predictive validity for subjective complexity, resulting in a state-of-the-art objective indicator of textual complexity. The combined model of interest shows how the objective indicator of textual complexity is reflected in its subjective counterparts and, consequently, influences the experience of interest. It confirms the hypothesis by showing that information systems can use a well-devised objective metric of textual complexity to approximate the optimal level of complexity, that is, by selecting information likely appraised as complex yet comprehensible.

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 directly influencing the emotion of interest. By continuing this line of enquiry, it is possible to foresee information systems that create and maintain a positive IX for their users.

Acknowledgments

The authors thank the anonymous reviewers for their valuable comments and suggestions to improve the quality of this paper. Furthermore, the authors thank Anton Nijholt for his valuable input. 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.

References

Arapakis, I., Jose, J.M., & Gray, P.D. (2008). Affective feedback: An investigation into the role of emotions in the information seeking process. In Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR ′08, pp. 395–402). New York, NY: ACM.

Balota, D.A., Cortese, M.J., Sergent-Marshall, S.D., Spieler, D.H., & 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., & Cortese, M.J. (2006). Visual word recognition: The journey from features to meaning (a travel update). In Handbook of Psycholinguistics (pp. 285–376). San Diego, CA: Academic Press.

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

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

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

Belkin, N.J., & Croft, W.B. (1992). Information filtering and information retrieval: Two sides of the same coin? Communications of the ACM, 35(12), 29–38.

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

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

Berlyne, D.E. (1975). Behaviourism? Cognitive theory? Humanistic psychology? To Hull with them all. Canadian Psychological Review, 16(2), 69–80.

Borlund, P. (2003). The IIR evaluation model: A framework for evaluation of interactive information retrieval systems. Information Research, 8(3), paper number 152. Retrieved from http://informationr.net/ir/8-3/paper152.html.

Borlund, P., & 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, pp. 324–331). New York, NY: 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. (2001). Random forests. Machine Learning, 45(1), 5–32.

Carbonell, J., & Goldstein, J. (1998). The use of MMR, diversity-based reranking for reordering documents and producing summaries. In Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR ′98, pp. 335–336). New York, NY: ACM.

Cer, D., de Marneffe, M.-C., Jurafsky, D., & Manning, C. (2010). Parsing to Stanford dependencies: Trade-offs between speed and accuracy. In Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC '10, Valletta, Malta). Paris, France: European Language Resources Association (ELRA).

Chall, J.S., & Dale, E. (1995). T1-readability revisited: The new Dale-Chall readability formula. Cambridge, MA: PB-Brookline Books.

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

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

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., & Ingwersen, P. (2000). Dimensions of relevance. Information Processing & Management, 36(4), 533–550.

Cover, T.M., & Thomas, J.A. (2006). Entropy, relative entropy, and mutual information. In Elements of Information Theory (pp. 13–56). Hoboken, NJ: John Wiley & Sons.

Cronbach, L.J., & Meehl, P.E. (1955). Construct validity in psychological tests. Psychological Bulletin, 52(4), 281–302.

Crossley, S., Greenfield, J., & McNamara, D. (2008). Assessing text readability using cognitively based indices. TESOL Quarterly, 42(3), 475–493.

Csikszentmihalyi, M. (1991). The Psychology of Optimal Experience. New York, NY: Harper Collins.

Demartini, G., & Mizzaro, S. (2006). A classification of IR effectiveness metrics. In M. Lalmas, A. MacFarlane, S. Rger, A. Tombros, T. Tsikrika, & A. Yavlinsky (Eds.), Advances in Information Retrieval: Lecture Notes in Computer Science (Vol. 3936, pp. 488–491). Berlin/Heidelberg, Germany: Springer.

Descartes, R. (1989/1649). The passions of the soul. Indianapolis, IN: Hackett Publishing Company.

DuBay, W. (2007). The classic readability studies. Technical report. Costa Mesa, CA: Impact Information.

Ellsworth, P.C., & Scherer, K.R. (2003). Appraisal processes in emotion. In R.J. Davidson, K.R. Scherer, & H.H. Goldsmith (Eds.), Handbook of affective sciences (pp. 572–595). New York, NY: Oxford University Press.

Feng, L., Jansche, M., Huenerfauth, M., & Elhadad, N. (2010). A comparison of features for automatic readability assessment. In Proceedings of the 23rd International Conference on Computational Linguistics: Posters (pp. 276–284). Stroudsburg, PA: Association for Computational Linguistics.

Flesch, R. (1948). A new readability yardstick. Journal of Applied Psychology, 32(3), 221–233.

Fry, E. (2002). Readability versus leveling. Reading Teacher, 56(3), 286.

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 (pp. 95–126). Cambridge, MA: The MIT Press.

Glassey, R., & Azzopardi, L. (2011). Finding interest in the stream. Proceedings of the American Society for Information Science and Technology, 48(1), 1–4.

Gluck, M. (1996). Exploring the relationship between user satisfaction and relevance in information systems. Information Processing & Management, 32(1), 89–104.

Graesser, A.C., McNamara, D.S., Louwerse, M.M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. Behavior Research Methods, Instruments, & Computers, 36(2), 193–202.

Hanani, U., Shapira, B., & Shoval, P. (2001). Information filtering: Overview of issues, research and systems. User Modeling and User-Adapted Interaction, 11(3), 203–259.

Hassenzahl, M. (2013). User experience and experience design. In M. Soegaard & R.F. Dam (Eds.), Encyclopedia of Human-Computer Interaction (2nd ed.). Aarhus, Denmark: The Interaction-Design.org Foundation. Retrieved from http://www.interaction-design.org/encyclopedia/user_experience_and_experience_design.html

Herlocker, J.L., Konstan, J.A., Terveen, L.G., & Riedl, J.T. (2004). Evaluating collaborative filtering recommender systems. ACM Transactions on Information Systems, 22(1), 5–53.

Hess, E.H., & 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., & Renninger, K.A. (2006). The four-phase model of interest development. Educational Psychologist, 41(2), 111–127.

Hill, W., Stead, L., Rosenstein, M., & 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, pp. 194–201). New York, NY: ACM Press/Addison-Wesley Publishing.

Huffman, S.B., & 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, pp. 567–574). New York, NY: ACM.

Iacobucci, D. (2010). Structural equations modeling: Fit indices, sample size, and advanced topics. Journal of Consumer Psychology, 20(1), 90–98.

Ihaka, R., & Gentleman, R. (1996). R: A language for data analysis and graphics. Journal of Computational and Graphical Statistics, 5(3), 299–314.

Inhoff, A., & Rayner, K. (1986). Parafoveal word processing during eye fixations in reading: Effects of word frequency. Attention, Perception, & Psychophysics, 40(6), 431–439.

Jaeger, T.F., & Tily, H. (2011). On language utility: Processing complexity and communicative efficiency. Wiley Interdisciplinary Reviews: Cognitive Science, 2(3), 323–335.

Jonassen, D.H. (2000). Toward a design theory of problem solving. Educational Technology Research and Development, 48(4), 63–85.

Kincaid, J.P., Fishburne, Robert P., J., Rogers, R.L., & 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. Springfield, VA: National Technical Information Service.

Kintsch, W., & van Dijk, T.A. (1978). Toward a model of text comprehension and production. Psychological Review, 85(5), 363–394.

Klein, D., & Manning, C.D. (2003). Accurate unlexicalized parsing. In Proceedings of the 41st Annual Meeting on Association for Computational Linguistics (Vol. 1, ACL ′03, pp. 423–430). Stroudsburg, PA: Association for Computational Linguistics.

Knijnenburg, B., Willemsen, M., Gantner, Z., Soncu, H., & Newell, C. (2012). Explaining the user experience of recommender systems. User Modeling and User-Adapted Interaction, 22(4), 441–504.

Konstan, J., & Riedl, J. (2012). Recommender systems: From algorithms to user experience. User Modeling and User-Adapted Interaction, 22(1), 101–123.

Kuhlthau, C.C. (2004). Seeking meaning: A process approach to library and information services. Norwood, NJ: Ablex Publishing Corp.

Lapata, M., & Barzilay, R. (2005). Automatic evaluation of text coherence: Models and representations. In Proceedings of the 19th International Joint Conference on Artificial Intelligence (IJCAI ′05, pp. 1085–1090). San Francisco, CA: Morgan Kaufmann Publishers.

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., & 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.

Lewis, R.L., Vasishth, S., & Dyke, J.A.V. (2006). Computational principles of working memory in sentence comprehension. Trends in Cognitive Sciences, 10(10), 447–454.

Lively, B.A., & Pressey, S.L. (1923). A method for measuring the “vocabulary burden” of textbooks. Educational Administration and Supervision, 9(7), 389–398.

Long, D.L., Johns, C.L., & Morris, P.E. (2006). Comprehension ability in mature readers. In Handbook of Psycholinguistics (pp. 801–833). San Diego, CA: Academic Press.

Long, D.L., Wilson, J., Hurley, R., & Prat, C.S. (2006). 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.

McNamara, D.S., Louwerse, M.M., McCarthy, P.M., & Graesser, A.C. (2010). Coh-Metrix: Capturing linguistic features of cohesion. Discourse Processes, 47(4), 292–330.

Meyer, D., Leisch, F., & Hornik, K. (2003). The support vector machine under test. Neurocomputing, 55(1–2), 169–186.

DOI: 10.1002/asi

Michel, J.-B., Shen, Y.K., Aiden, A.P., Veres, A., Gray, M.K., Pickett, J.P., . . . Aiden, E.L. (2011). Quantitative analysis of culture using millions of digitized books. Science, 331(6014), 176–182.

Morris, J., & Hirst, G. (1991). Lexical cohesion computed by thesaural relations as an indicator of the structure of text. Computational Linguistics, 17(1), 21–48.

O'Brien, H.L., & 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.

Peng, F., & Schuurmans, D. (2003). Combining naive bayes and n-gram language models for text classification. In F. Sebastiani (Ed.), Advances in information retrieval: Lecture notes in computer science (Vol. 2633, pp. 547–547). Berlin/Heidelberg, Germany: Springer.

Perfetti, C.A. (1988). Verbal efficiency in reading ability. In Reading research advances in theory and practice (Vol. 6, pp. 109–143). San Diego, CA: Academic Press.

Porter, M.F. (2001). Snowball: A language for stemming algorithms. Retrieved from http://snowball.tartarus.org/texts/introduction.html

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., & Reichle, E.D. (2010). Models of the reading process. Wiley Interdisciplinary Reviews: Cognitive Science, 1(6), 787–799.

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., & Rayner, K. (1998). Toward a model of eye movement control in reading. Psychological Review, 105(1), 125–157.

Ricci, F., Rokach, L., Shapira, B., & Kantor, P.B. (Eds.). (2009). Recommender systems handbook. Berlin/Heidelberg, Germany: Springer.

Rice, M., & Harris, G. (2005). Comparing effect sizes in follow-up studies: ROC area, Cohen's d, and r. Law and Human Behavior, 29(5), 615–620.

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. Journal of Statistical Software, 48(2), 1–36.

Ruthven, I., Baillie, M., & 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. Retrieved from http://search.cpan.org/%20kimryan/Lingua-EN-Fathom-1.15/lib/Lingua/EN/Fathom.pm

Sadoski, M. (2001). Resolving the effects of concreteness on interest, comprehension, and learning important ideas from text. Educational Psychology Review, 13(3), 263–281.

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.

Schiefele, U. (1996). Topic interest, text representation, and quality of experience. Contemporary Educational Psychology, 21(1), 3–18.

Schiefele, U., & Krapp, A. (1996). Topic interest and free recall of expository text. Learning and Individual Differences, 8(2), 141–160.

Schraw, G. (1997). Situational interest in literary text. Contemporary Educational Psychology, 22(4), 436–456.

Schraw, G., Dunkle, M.E., & Bendixen, L.D. (1995). Cognitive processes in well-defined and ill-defined problem solving. Applied Cognitive Psychology, 9(6), 523–538.

Schraw, G., & Lehman, S. (2001). Situational interest: A review of the literature and directions for future research. Educational Psychology Review, 13(1), 23–52.

Schumacker, R., & Lomax, R. (2010). A beginner's guide to structural equation modeling (3rd ed.). London: Routledge Academic.

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. New York, NY: Oxford University Press.

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.

Soboroff, I., & Harman, D. (2005). Novelty detection: The TREC experience. In Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT ′05, pp. 105–112). Stroudsburg, PA: Association for Computational Linguistics.

Spink, A., & 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.

Su, L.T. (1994). The relevance of recall and precision in user evaluation. Journal of the American Society for Information Science, 45(3), 207–217.

Toutanova, K., Klein, D., Manning, C.D., & Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (NAACL ′03, pp. 173–180). Morristown, NJ: Association for Computational Linguistics.

Van der Sluis, F., Glassey, R.J., & Van den Broek, E.L. (2012). Making the news interesting: Understanding the relationship between familiarity and interest. In ACM Proceedings of the 4th Symposium on Information Interaction in Context (IIiX, pp. 314–317). New York, NY: ACM.

Voorhees, E. (2002). The philosophy of information retrieval evaluation. In Evaluation of cross-language information retrieval systems (pp. 143–170). Berlin/Heidelberg, Germany: Springer.

Vor der Brück, T., Hartrumpf, S., & Helbig, H. (2008). A readability checker with supervised learning using deep indicators. Informatica, 32(4), 429–435.

Xu, Y.C., & 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.

Zipf, G.K. (1935). The psycho-biology of language: An introduction to dynamic philology. Boston, MA: Houghton Mifflin.